SIAM J. CONTROL OPTIM.
Vol. 35, No. 1, pp. 1–28, January 1997

LIMITS OF SINGULARLY PERTURBED CONTROL PROBLEMS WITH STATISTICAL DYNAMICS OF FAST MOTIONS*

ALEXANDER VIGODNER†

Abstract. We describe the limit behavior of admissible trajectories in a singularly perturbed control system as the small parameter tends to zero. A general case is considered where, in the limit, the fast motion may infinitely rapidly oscillate in time. Invariant measures of the parameterized fast flow are employed to describe the limit behavior and construct the limit control problem. The notion of relatively slow controls is introduced. Approximating properties of the limit problem within the families of relatively slow controls are verified. The results are illustrated by examples.

Key words. singular perturbations, invariant measure, statistical convergence, relaxed control, relatively slow control

AMS subject classifications. 49J15, 34D15, 34C35

PII. S0363012994264207

1. Introduction.

In this paper we consider a singularly perturbed control system which consists of two differential equations,

$$\begin{aligned} \dot{x} &= f(u,x,y), \\ \varepsilon\dot{y} &= g(u,x,y), \end{aligned} \tag{1.1}$$

together with initial conditions

$$x(0) = x^0,\ y(0) = y^0. \tag{1.2}$$

Here $x \in \mathbb{R}^m, y \in \mathbb{R}^n$, and ε is a positive small parameter. A dot denotes a derivative with respect to time t. The controls $u(t)$ are functions satisfying the inclusion $u(t) \in U$, where U is a prescribed compact metric space. As is customary, we refer to the first equation in (1.1) as generating the slow x-trajectory, while the second one generates the fast y-trajectory. This terminology may be justified by orders of magnitude of the derivatives $\dot{x}, \dot{y}$ which are $|f|$ and $|g|/\varepsilon$, respectively.

The control problem associated with these equations consists of minimizing the integral $J_\varepsilon(u)$ given by

$$J_\varepsilon(u) = \int_0^1 Q(u(t), x_\varepsilon(t), y_\varepsilon(t))dt, \tag{1.3}$$

where $(x_\varepsilon(t), y_\varepsilon(t))$ is the assumed unique solution of (1.1)–(1.2) for $t \in [0,1]$, induced by $u(\cdot)$.

The problems that we are interested in are as follows:

— to study the limit behavior of the trajectories of this control system and the cost as ε tends to zero,

— to construct a limit system which could in some sense approximate the behavior of the original ε-system for ε small.

*Received by the editors March 2, 1994; accepted for publication (in revised form) September 19, 1995.

http://www.siam.org/journals/sicon/35-1/26420.html

†Department of Theoretical Mathematics, Weizmann Institute of Science, 76100 Rehovot, Israel. Present address: Intel Israel, M.T.M., Haifa 31015, Israel (vigod@wisdom.weizmann.ac.il or vigodner@iil.intel.com).

Such problems were examined by many authors. See for instance the works of Bensoussan [Be1, Be2]; O'Malley [O'M]; Kokotovic [Ko1]; Kokotovic, O'Malley, and Sannuti [KOS]; Bensoussan and Blankenship [BB]; and Donchev and Veliov [DV1, DV2]. These papers consider the case where it is possible to identify a limit system for $\varepsilon \to 0$ in the form of a reduced system which is obtained by setting $\varepsilon = 0$ in (1.1), namely,

$$\begin{aligned} &\dot{x} = f(u, x, y),\ x(0) = x^0; \\ &0 = g(u, x, y). \end{aligned} \tag{1.4}$$

This reduced order scheme was suggested by Tichonov [Ti] for uncontrolled singularly perturbed differential equations and has been developed, among many others, by Vasil'eva [Va], Vasil'eva and Butuzov [VB], O'Malley [O'M2], and Campbell [Ca].

Under appropriate conditions the algebraic equation can be inverted with respect to y, resulting in a manifold $y = q(u, x)$. This manifold is then inserted into the dynamic part of (1.4) as follows:

$$\begin{aligned} &\dot{x} = f(u, x, q(u, x)),\ x(0) = x^0; \\ &y = q(u, x). \end{aligned} \tag{1.5}$$

The limit control problem is then to minimize the cost

$$J_0(u) = \int_0^1 Q(u(t), x_0(t), y_0(t))\, dt,$$

where $(x_0(t), y_0(t))$ is the trajectory of (1.5) induced by the control $u(t)$. Under appropriate conditions $x_\varepsilon(t) \to x_0(t)$ uniformly for $t \in [0, 1]$, $y_\varepsilon(t) \to y_0(t)$ uniformly for $t \in [\delta, 1]$, where $\delta > 0$ is arbitrary, and $J_\varepsilon(u) \to J_0(u)$. Moreover under additional conditions the value of the original problem converges to the value of the reduced problem. Namely $\inf_u J_\varepsilon(u) \to \inf_u J_0(u)$. A main condition for this scheme to be valid is that the point $q(u, x)$ (u, x are fixed) be a (locally) asymptotically stable equilibrium point for the (u, x)-parameterized time-invariant differential equation

$$\frac{dy}{d\tau} = g(u, x, y). \tag{1.6}$$

The problem of Mayer's type, that is, $J_\varepsilon(u) = Q(y_\varepsilon(1))$, was considered in [DV1]. Some related recent results are given in the works of Kabanov and Pergamenshchikov [KaP] and Tuan [Tu]. In the latter works the well-posedness of Mayer's problems is considered. The well-posedness is connected with the convergence as $\varepsilon \to 0$ of the reachable set of the original system to the reachable set of the limit problem. One can easily extend the results of this paper for Mayer's problems too.

In many works (see, e.g., Gaitsgory [G1, G2, G3]) it was noted that the reduced model approximates the original problem (1.1)–(1.3) only if the optimal control $u_\varepsilon(t)$ does not change rapidly as $\varepsilon \to 0$. If, for instance, in the optimal regime the control u and the fast variable y oscillate rapidly and the oscillation rate grows to infinity as ε tends to zero, then the reduced model cannot describe the limit behavior of the optimal trajectories of the original system. To define the limit control system in this case, Gaitsgory [G1, G2, G3] suggests another approach which is an extension of the averaging method described for uncontrolled motions in Volosov [Vo]. The limit problem in [G1, G2, G3] has the following form of a differential inclusion for the slow variable x:

$$\dot{x} \in V(x), \tag{1.7}$$

where $V(x)$ is a convex compact set-valued function, constructed by the special scheme of averaging of the function $f(u,x,y)$ over solutions of (1.6) on the infinite time interval. For a more recent result concerning topological dynamics and other properties of the differential inclusion in (1.7) see Grammel [Gr].

In this paper we consider a more general case where, even for constant in time controls $u \in U$, the reduced order system (1.4) may not be an appropriate nominal limit for (1.1)–(1.3) as $\varepsilon \to 0$. More specifically, the solutions of (1.6) for x, u fixed may not converge to an equilibrium. We establish conditions on the asymptotic topological dynamics of these trajectories and build a limit control system for (1.1)–(1.2). We develop here an approach suggested in [AV] for uncontrolled singularly perturbed dynamical systems.

We use *invariant measures* of (1.6) for x, u fixed instead of considering a root $y = q(u,x)$ of $0 = g(u,x,y)$. If there is a (locally) unique invariant measure of (1.6) (depending on parameters u, x), say $\nu(u,x)$, we suggest the following limit control system instead of (1.4):

$$\begin{aligned} \dot{x} &= \int_{\mathbb{R}^m} f(u,x,y)\,\nu(u,x)(dy),\ x(0) = x^0; \\ \mu(t) &= \nu(u(t),x(t)). \end{aligned} \tag{1.8}$$

The limit control problem is then to minimize over u the integral

$$J_0(u) = \int_0^1 Q(u(t),x_0(t),y)\,\mu_0(t)(dy), \tag{1.9}$$

where $(x_0(t),\mu_0(t))$ is the trajectory of (1.8) generated by a control $u(t)$. $\mu_0(t) = \nu(u(t),x_0(t))$ is a function which takes values in the space of probability measures on $\mathbb{R}^n$. The limit control problem (1.8)–(1.9) is similar to the chattering control problem as developed in [A1, A2, A3, A4]. The only complication is that in our model the measure-valued function $\mu_0(t)$ is a function of $x(t)$ and $u(t)$ rather than a parametric function. We then verify that this limit control problem approximates the original (1.1)–(1.3) problem on the special families of *relatively slow controls*. By relatively slow control we mean a function $u_\varepsilon(t)$ which may be "fast" in time t as ε tends to zero, but in time $\tau = \varepsilon^{-1}t$ the control $\tilde{u}_\varepsilon(\tau) = u_\varepsilon(\varepsilon\tau)$ is "slow." The families of relatively slow controls are deeply connected with the so-called *ergodic families of controls* with respect to the fast system (1.6) defined in [G1]. Note also that the average technique suggested in [G1, G2, G3] is applicable in our case too. However, our approach has two advantages. First, in the average technique, attention is paid only to the asymptotic behavior of the slow variable $x_\varepsilon(t)$. Asymptotic behavior of the fast variable $y_\varepsilon(t)$ is studied only in the reduced order case. Our approach employs the theory of measure-valued functions and determines completely the statistical asymptotic behavior of the fast variable $y_\varepsilon(t)$ in the general case. And second, we have in the limit, an explicit form of the control problem rather than the form of the differential inclusion (1.7).

The paper is organized as follows. Section 2 is devoted to the properties of invariant measures of dynamical system (flows) and their invariant measures depending on parameters. The results of this section are not new but presented in a convenient form. In section 3 we give a formal description of the singularly perturbed control problem and the technical conditions. We construct the limit control problem (1.8)–(1.9) and verify when this problem is well posed. In section 4 we employ *relaxed controls* and construct the limit relaxed control problem which guarantees existence

of the optimal solutions. We use the standard technique originally introduced by Warga [Wa]. In section 5 we define the notion of relatively slow controls and give two preliminary estimates on convergence for the trajectories of the original system. The main result is presented in section 6. It consists of four convergence theorems. The first three theorems state the same convergence result of the trajectories and the cost of the original system when, respectively, continuous, piecewise continuous, and measurable controls are applied. The difference between these theorems is only in appropriate assumptions. The fourth theorem determines the convergence result for the value of the original control problem to the value of the limit problem on the families of relatively slow controls. The proof of Theorems I–III is given in section 7. In the closing section we comment on the results obtained.

2. Dynamical systems, invariant measures, and convergence of motions. In this section we recall some properties of dynamical systems and invariant measures which depend on parameters. We establish the properties of asymptotic convergence of motions to invariant measure and prescribed sets. Consult with Nemytskii and Stepanov [NS] and Sell [Se] as a general reference and with Billingsley [Bi] as our reference on probability measures.

Consider a dynamical system (semiflow) $\phi(\tau, y)$ which is a continuous mapping

$$\phi(\tau, y) : \mathbb{R}_+ \times \mathbb{R}^n \to \mathbb{R}^n; \quad \phi(0, y) = y; \quad \phi(\tau_1, \phi(\tau_2, y)) = \phi(\tau_1 + \tau_2, y) \quad \forall \tau_{1,2} \in \mathbb{R}_+.$$

In what follows we shall often use the term flow instead of semiflow if no difference between properties of flows and semiflows appears.

The ω-limit set $\Omega(y)$ of a point y consists of all the points $z = \lim \phi(\tau_k, y)$ for some sequence $\tau_k \to +\infty$. A set A is *positively invariant* with respect to ϕ if $y \in A$ implies $\phi(\tau, y) \in A$ for $\tau \geq 0$.

We shall need also in the sequel the notion of *a prolongation of a set*; see, e.g., Bhatia and Szegö [BS]. Let K be a subset of R^n. The prolongation of K with respect to ϕ consists of all limit points $z = \lim \phi(\tau_i, y_i)$ with $r(y_i, K) \to 0$ as $i \to \infty$ and $\tau_i \geq 0$ for all i. Here r is a distance on $\mathbb{R}^n$.

Let Γ be a complete separable metric space, endowed with its Borel structure. We need some properties of measures defined on Γ. (We let Γ be an abstract space rather than, say $\mathbb{R}^n$, since in this paper we shall use measures on different spaces.)

Let Σ be a Borel σ-algebra on Γ. We denote the space of Borel probability measures on (Γ, Σ) by $\mathcal{P}(\Gamma)$. The space $\mathcal{P}(\Gamma)$ is endowed with the weak convergence of measures defined as follows. The sequence μ_k converges to μ if for every continuous bounded function $h(y) : \Gamma \to \mathbb{R}$

$$\int_{\leqq} h(y)\, \mu_k(dy) \to \int_{\Gamma} h(y)\, \mu(dy). \tag{2.1}$$

$\mathcal{P}(\Gamma)$ is metrizable. For definiteness we choose the Prohorov metric ρ defined as follows (see [Bi, p. 238]):

$$\rho(\mu_1, \mu_2) = \inf\{\eta : \mu_1(A) \leq \mu_2(A^\eta) + \eta,\ \mu_2(A) \leq \mu_1(A^\eta) + \eta\ \forall A \in \Sigma\}, \tag{2.2}$$

where A^η denotes the η-neighborhood of A in Γ.

A measure ν in $\mathcal{P}(\mathbb{R}^n)$ is said to be an *invariant measure* with respect to ϕ if for every $\tau \in \mathbb{R}_+$ and for every bounded and continuous function $h(\cdot)$

$$\int_{\mathbb{R}^n} h(\phi(\tau, y))\, \nu(dy) = \int_{\mathbb{R}^n} h(y)\, \nu(dy). \tag{2.3}$$

(2.3) is thus equivalent to the relation $\nu(A) = \nu(\phi(-\tau, A))$ for any Borel A in Γ.

In what follows we shall denote the Lebesgue measure on $\mathbb{R}$ by λ.

DEFINITION 2.1. *Let $w(\tau) : \mathbb{R}_+ \to \mathbb{R}^n$ be continuous, and let ν be a probability measure on R^n. We say that $w(\cdot)$ converges asymptotically in distribution to ν if the probability measures μ_s, given by*

$$\mu_s(A) = \frac{1}{s}\lambda(\{\tau : \tau \in [0, s],\ w(\tau) \in A\}),\ A \in \Sigma,$$

converge weakly to ν as $s \to \infty$.

Parameterized flows. Consider now the following differential equation depending on the parameter $p \in P$ where P is a metric space:

$$\frac{dy}{d\tau} = g(p, y) \tag{2.4}$$

with $y \in \mathbb{R}^n$. We assume that $g(\cdot, \cdot)$ is continuous on $P \times \mathbb{R}^n$ and that with any initial condition $y(0) = y$ the equation has a unique solution $\phi(\tau, p, y)$ for $\tau \in \mathbb{R}_+$ such that $\phi(0, p, y) = y$. Thus the mapping ϕ is continuous on $\mathbb{R}_+ \times P \times \mathbb{R}^n$, and for p fixed, ϕ is a continuous semiflow. For $\phi(\cdot, p, \cdot)$ we shall often use the notations ϕ_p or $\phi_p(\tau, y)$.

We state now an assumption under which we work throughout.

Assumption 2.2. For each p let $G(p)$ be an open subset of R^n and such that

(i) $G(p)$ is positively invariant with respect to ϕ_p;

(ii) for open $P' \subset P$, the set $\{(p, y) : p \in P',\ y \in G(p)\}$ is an open subset of $P \times \mathbb{R}^n$;

(iii) a compact $K(p) \subset G(p)$ exists such that the ω-limit set of $\phi_p(\tau, y)$ for p fixed, is in $K(p)$ for all $y \in G(p)$, and for compact $P' \subset P$ the set $\{(p, y) : p \in P',\ y \in K(p)\}$ is compact in $P \times R^n$.

Let $L(p)$ be the prolongation of the set $K(p)$ with respect to ϕ_p for p fixed. It is clear that $L(p) \supset K(p)$.

PROPOSITION 2.3. *Under Assumption* 2.2 *the set-valued function $L(p)$ has compact values included in $G(p)$ and $L(\cdot)$ is upper semicontinuous.*

Proof. See [AV, Proposition 8.2]. □

Note that by the Krilov–Bogolubov theorem (see, e.g., [NS, p. 493] the parameterized flow $\phi(\tau, p, y)$ for each p has at least one invariant measure supported in $G(p)$ if $G(p)$ is bounded. In our case $G(p)$ may not be bounded (actually $G(p)$ can coincide with $\mathbb{R}^n$). Nevertheless the existence of the compact subset $K(p)$ with properties (iii) of the latter assumption implies (see [AV, Proposition 3.2]) that the set of all ϕ_p-invariant measures in $G(p)$ is not empty, compact, and convex. Moreover, all these measures are supported on $K(p)$.

PROPOSITION 2.4. *Suppose that Assumption* 2.2 *is fulfilled. Then for $(p, y) \in$ graph G*

(i) *the average integral*

$$\frac{1}{s}\int_0^s d((p, \phi(\tau, p, y)), \mathit{graph}\, K)\, d\tau$$

tends to zero as $s \to \infty$. (Here d is a distance on $P \times \mathbb{R}^n$.)

(ii) *the pair $(p, \phi(\tau, p, y))$ converges to graph L as $\tau \to \infty$. Namely, $d((p, \phi(\tau, p, y)),$ graph $L) \to 0$ as $\tau \to \infty$.*

(iii) *the convergences in* (i) *and* (ii) *are uniform on compact subsets of pairs (p, y) in graph G.*

If additionally $\nu(p)$ for each p is the unique invariant measure of (2.4) *supported on $G(p)$, then*

(iv) *$\nu(p)$ is supported on $K(p)$ and continuous on P (in topology on $\mathcal{P}(\mathbb{R}^n)$).*

(v) *the parameterized flow $\phi(\tau, p, y)$ converges asymptotically in distribution to $\nu(p)$ uniformly on compact subsets of pairs (p, y) in graph G.*

Proof. (i) follows from [AV, Lemma 3.1 and Proposition 3.2] and properties of the weak convergence of measures. (ii) follows from (i) and the definition of a prolongation of a set. Uniformity of these convergences follows from [AV, Proposition 3.5]. (iv) and (v) follow also from [AV, Proposition 3.7]. The full proof of the results can be found in [Vi, Proposition 2.4.9]. □

Remark 2.5. The openness of *graph* G is essential for further considerations. However it can be easily proved that all latter results hold true if the graph $\{(p, y) : p \in P', y \in G(p)\}$ is closed for closed P'. Then, in particular, $G(p)$ can be compact and $K(p)$ may coincide with $G(p)$. For the details see [Vi, Chapter II].

It may seem that the stronger continuity property (for instance the Lipschitz continuity) of function g with respect to p implies the Lipschitz continuity of the invariant measure $\nu(p)$ in the sense of the Prohorov metric (2.2). But in general this is not true. As an example consider the one-dimensional system $\dot{y} = y(p - r(y))$, $p \in [1/2, 2]$; $r(y) = \sin(y)$ if $y \le \pi/2$ and $r(y) = (y - \pi/2)^2 + 1$ if $y > \pi/2$. This system is C^1 in p, y. For any p there exists then the equilibrium point $y(p) = \arcsin(p)$ if $p \le 1$ and $y(p) = \sqrt{(p-1)} + \pi/2$ if $p > 1$ which is locally asymptotically stable. The set is $G(p) = (0, \pi)$ and $K(p) = \{y(p)\}$. The invariant measure is concentrated on the point $y(p)$. It is clear that for $p^* = 1$ we have $\rho(\nu(p^*), \nu(p))/|p - p^*| \to \infty$ as $p \to 1$. Then $\nu(p)$ is not Lipschitz continuous in p on the interval $[1/2, 2]$.

The next result is concerned with time-varying perturbation of the variable p.

PROPOSITION 2.6. *Suppose that Assumption* 2.2 *is fulfilled. Let E be a compact subset of graph G. Then for any $\eta > 0$, there exist an $s_0 > 0$ and a $\theta > 0$ such that whenever $|p(\tau) - p_0| < \theta$ for $\tau \in [0, 2s_0]$ and $w(\tau)$ is a solution of $\dot{y} = g(p(\tau), y)$, with $(p_0, w(0)) \in E$, then for $s \in [s_0, 2s_0]$ the following inequalities hold:*

$$\frac{1}{s}\lambda(\{\tau : \tau \in [0, s], d((p(\tau), w(\tau)), graph\, K) > \eta\}) < \eta,$$
$$d((p(s), w(s)), graph\, L) < \eta.$$

Proof. It follows from Proposition 2.4 and the continuous dependence results in ordinary differential equations. □

The next proposition states a correspondence of the Prohorov metric between two measures and the difference between the corresponding integrals. This is a general property of the Prohorov metric without any connection to dynamical systems and invariant measures.

PROPOSITION 2.7. *Let $h : \Gamma \to \mathbb{R}$ be a bounded uniformly continuous function; namely, there exist constants a, b and a nondecreasing function $\omega(s) \to 0$ as $s \to 0$ such that*

$$a \le h(x) \le b;\ |h(x) - h(y)| \le \omega(|x - y|)\ \forall x, y \in \Gamma.$$

Assume that $\mu_1, \mu_2 \in \mathcal{P}(\Gamma)$. Let $\delta = \rho(\mu_1, \mu_2)$.

Then

$$\left| \int_\Gamma h(s)\mu_2(ds) - \int_\Gamma h(s)\mu_1(ds) \right| \le \omega(\delta) + (b - a)\delta. \tag{2.5}$$

Proof. We can write the following representation of the integrals for $i = 1, 2$:

$$\begin{aligned}\int_\Gamma h(s)\mu_i(ds) &= \int_{\{x:h(x)\geq a\}} h(s)\mu_i(ds)\\ &= a + \int_a^\infty \mu_i\{x : h(x) \geq s\}ds = a + \int_a^b \mu_i\{x : h(x) \geq s\}ds\end{aligned} \tag{2.6}$$

(see [Bi, p. 222]). We denote the set $A_s = \{x : h(x) \geq s\}$. Take the sequence $\delta_k > \delta$, $\delta_k \to \delta$. From the continuous property of h it follows that $A_s^{\delta_k} \subset A'_s$ where $A'_s = \{x : h(x) > s - w(\delta_k)\}$. Then $\mu_i(A'_s) \geq \mu_i(A_s^{\delta_k})$. Therefore from (2.6) we can make the following estimate:

$$\begin{aligned}&\int_\Gamma h(s)\mu_2(ds) - \int_\Gamma h(s)\mu_1(ds) = \int_a^b \mu_2(A_s) - \mu_1(A_s)ds\\ &= \int_a^b \mu_2(A_s) - \mu_1(A'_s)ds + \int_{a-\omega(\delta_k)}^a \mu_1(A_s)ds - \int_{b-\omega(\delta_k)}^b \mu_1(A_s)ds\\ &\leq \int_a^b \mu_2(A_s) - \mu_1(A_s^{\delta_k})ds + \omega(\delta_k) \leq \omega(\delta_k) + (b-a)\delta_k.\end{aligned} \tag{2.7}$$

Passing to the limit as $k \to \infty$ and according to the symmetry in (2.7), we obtain estimate (2.5). This completes the proof. □

In what follows we shall denote the metric on the abstract metric space, say Γ, by d_Γ.

3. The setting. In this section we give a formal description of the singularly perturbed control problem (we refer to this problem as the ε-problem) and introduce the conditions under which we work.

The control problem (1.1–1.3) is identified by the following data functions:

$$\begin{aligned}f(u, x, y) &: U \times \mathbb{R}^m \times \mathbb{R}^n \to \mathbb{R}^m,\\ g(u, x, y) &: U \times \mathbb{R}^m \times \mathbb{R}^n \to \mathbb{R}^n,\\ Q(u, x, y) &: U \times \mathbb{R}^m \times \mathbb{R}^n \to \mathbb{R}.\end{aligned}$$

The trajectories $x_\varepsilon(t), y_\varepsilon(t)$ of the system (1.1)–(1.2) generated by some admissible control function $u(t)$ induce the cost $J_\varepsilon(u)$ given by (1.3). We define the value of the ε-problem as follows:

$$\mathrm{val}_\varepsilon(\mathcal{U}) = \inf_{u(\cdot)\in\mathcal{U}} J_\varepsilon(u).$$

Here $\mathcal{U}$ is the family of admissible control functions $u(\cdot)$:

$$u(t) : [0, 1] \to U.$$

In this paper we consider three classes of admissible controls:

(1) $\mathcal{U}^c$ is a family of all continuous functions.

(2) $\mathcal{U}^p$ is a family of all piecewise continuous functions. By piecewise continuous functions we mean the functions which are continuous on $[0, 1]$ except possibly at a finite number of points and these points are discontinuities of the first type.

(3) $\mathcal{U}^m$ is a family of all measurable functions.

Let us introduce now the technical conditions assumed throughout the paper. We construct these conditions in order to study just a phenomenon of singular perturbations and to avoid other pathologies.

Remark. In the various estimates we shall often use the value of the function $u(t+z)$ where $t \in [0,1]$ and z is some number. Since $t+z$ may not be included in $[0,1]$ we shall suppose that $u(t)$ is extended for $t \in \mathbb{R}$ such that $u(t) = u(0)$ for $t < 0$ and $u(t) = u(1)$ for $t > 1$. No confusion should arise.

Assumption 3.1 (continuity of the data).

(i) The functions f, g, Q are continuous on $U \times \mathbb{R}^m \times \mathbb{R}^n$.

(ii) For any (u, x, y^0) fixed, the system $\dot{y} = g(u, x, y), \quad y(0) = y^0$ has a unique solution $\phi(\tau, u, x, y^0)$ defined for all $\tau \geq 0$.

Note that the pair $(u, x) \in U \times \mathbb{R}^m$ can be denoted as the parameter $p \in P$ from the previous section. Here $P = U \times \mathbb{R}^m$. With this notation the parameterized flow $\phi(\tau, p, y)$ (or $\phi(\tau, u, x, y)$) is a solution of (1.6). The following assumption copies the assumptions under which the results of section 2 were obtained.

Assumption 3.2. For any $p = (u, x)$ there exist sets $G(u, x)$ and $K(u, x)$ such that the corresponding set-valued maps $G(p)$ and $K(p)$ satisfy Assumption 2.2.

The following growth condition is there to ensure that for ε small, the trajectories $x_\varepsilon(t)$ stay bounded over $[0, 1]$.

Assumption 3.3 (growth condition). For some $c > 0$ the function $f(u, x, y)$ satisfies

$$\sup_{y \in G(u,x)} |f(u, x, y)| \leq c(1 + |x|) \quad \forall u \in U.$$

Assumption 3.4 (unique ergodicity). For each (u, x), the (u, x)-parameterized flow $\phi(\tau, u, x, y)$ has a unique invariant measure in $G(u, x)$; we denote this by $\nu(u, x)$.

Note that by Proposition 2.4 the mapping $\nu(u, x) : U \times \mathbb{R}^m \to \mathcal{P}(\mathbb{R}^n)$ is continuous on $U \times \mathbb{R}^m$.

Assumption 3.5 (Lipschitz continuity).

(i) The function $\nu(\cdot, \cdot)$ is Lipschitz continuous in x on every compact subset of $\mathbb{R}^m$ in the sense of the Prohorov metric and uniformly with respect to $u \in U$. Namely, for any compact D from $\mathbb{R}^m$ there exists a number $L > 0$ such that for any $u \in U$ and $x_1, x_2 \in D$

$$\rho(\nu(u, x_1), \nu(u, x_2)) \leq L|x_1 - x_2|.$$

(ii) On each compact subset of $\mathbb{R}^m \times \mathbb{R}^n$ the function f is Lipschitz continuous in x, y uniformly with respect to $u \in U$.

The limit control problem. Under Assumptions 3.1–3.4 we can formally build the limit control problem for $\varepsilon = 0$ in the following form:

$$\begin{aligned} &\underset{u(\cdot) \in \mathcal{U}}{\text{minimize}} \int_0^1 \int_{\mathbb{R}^n} Q(u(t), x(t), y)\, \mu(t)(dy)\, dt \\ &\text{subject to } \dot{x} = \int_{\mathbb{R}^n} f(u(t), x, y)\, \mu(t)(dy), \ x(0) = x^0, \\ &\mu(t) = \ \nu(u(t), x(t)). \end{aligned} \tag{3.1}$$

Here $\mu(t) \in \mathcal{P}(\mathbb{R}^n)$ for each t.

We do not specify here the family of admissible controls $\mathcal{U}$. For $u(\cdot) \in \mathcal{U}$ fixed we denote by $(x_0(t), \mu_0(t))$ $(\mu_0(t) = \nu(u(t), x_0(t)))$ an admissible trajectory of (3.1).

We define also the value of the limit problem

$$\mathrm{val}_0(\mathcal{U}) = \inf_{u(\cdot)\in\mathcal{U}} J_0(u).$$

In the next sections we shall show that this limit problem actually approximates in some sense the ε-problem (1.1)–(1.3). But in order to verify the approximating properties of the limit problem we need that the equations in (3.1) be uniquely solvable for $t \in [0,1]$ with respect to $x(t), \mu(t)$ for any given control $u(\cdot) \in \mathcal{U}$. To do that it is enough to assume that for the right-hand side of the differential equation

$$\dot{x} = \int_{\mathbb{R}^n} f(u,x,y)\,\nu(u,x)(dy) \tag{3.2}$$

is Lipschitz continuous in x on any compact subset of $\mathbb{R}^m$ uniformly with respect to $u \in U$. Assumptions 3.3–3.5 imply this property.

LEMMA 3.6. *Under Assumptions* 3.3–3.5 *the right-hand side of* (3.2) *is Lipschitz continuous in x on any compact subset of $\mathbb{R}^m$ uniformly with respect to $u \in U$.*

Proof. It follows directly from Proposition 2.7 and that by Assumption 3.2 for any compact D from $\mathbb{R}^m$ all measures $\{\nu(u,x) : u \in U, x \in D\}$ are supported in the compact set $\{y : u \in U, x \in D, y \in K(u,x)\}$. □

Therefore by Lemma 3.6 and under all assumptions in this section the system in (3.1) is uniquely solvable on the interval $t \in [0,1]$ for any given $u(\cdot) \in \mathcal{U}$.

Comments on the assumptions. Assumption 3.1 is standard. Without this assumption we could get various deviations which are not connected with singular perturbations.

Assumption 3.2 is crucial. Openness of the set $G(u,x)$ is essential for the proof that the triple $(u_\varepsilon(t), x_\varepsilon(t), y_\varepsilon(t))$ is defined for all $t \in [0,1]$, and it is included in *graph* G for ε small enough. On the other hand, without any additional requirements we can define (1.1) on the metric space $X \times Y$ where X is homeomorphic to $\mathbb{R}^m$ and Y is a closed subset of $\mathbb{R}^n$. Then the triple $(u_\varepsilon(t), x_\varepsilon(t), y_\varepsilon(t))$ is automatically included in the set $U \times X \times Y$. Hence G can be taken as Y and by Remark 2.5 the main results of this paper can be extended to the case where G is not open.

Assumption 3.3 implies that if the fast trajectory $y_\varepsilon(t)$ induced by a control $u_\varepsilon(t)$ exists uniquely on the time interval $[0,1]$ and $y_\varepsilon(t) \in G(u_\varepsilon(t), x_\varepsilon(t))$ (we shall prove this), then the slow trajectory $x_\varepsilon(t)$ is also defined for $t \in [0,1]$. This assumption can be omitted if it is known a priori that $x_\varepsilon(t)$ can be defined for all $t \in [0,1]$. We could assume that the function g satisfies the same growth condition too. But then we would rule out interesting examples.

Assumption 3.5 is needed only to prove the uniqueness of the solutions of (3.2). Thus instead of these assumptions we could assume the uniqueness directly. The similar condition is introduced in the steady state approach (Assumption (E) in Wasow [Was, p. 253]), but certainly in the steady state approach this condition can be checked easier.

Assumption 3.4 is crucial. Without this assumption the limit differential equation in (3.2) makes no sense, since the measure $\nu(u,x)$ is not uniquely defined. Instead, in this equation a limit differential inclusion must be defined; see [AV, Theorem I]. On the other hand, the local (in G) uniqueness of the invariant measure is equivalent to that for all continuous f of the following convergence:

$$\frac{1}{S}\int_0^S f(u,x,\phi(\tau,u,x,y))\,d\tau \to \int_{\mathbb{R}^n} f(u,x,y)\,\nu(u,x)(dy)$$

holds for $S \to \infty$ uniformly with respect to y, x, u (Proposition 2.4 (v)). This property reflects the deep connection between our technique and the averaging approach; see [Vo] and [G1, G2, G3]. In some sense the unique ergodicity is the weakest possible condition which guarantees applicability of the averaging technique for an arbitrary f. The simplest cases of the unique (local) ergodicity are where $\dot{y} = g(u, x, y)$ has an asymptotic stable equilibrium $q(u, x)$ or it has a stable limit cycle. In the first case $\nu(u, x)$ is the Dirac measure concentrated at the point $q(u, x)$ and in the second case $\nu(u, x)$ is distributed on the corresponding periodic curve. In a more general case the verification of the unique ergodicity is more difficult.

Measure-valued functions. Note that the trajectory $\mu_0(t) = \nu(u(t), x_0(t))$ of (3.1) is a measure-valued function. We want to display how this function approximates the fast trajectory $y_\varepsilon(t)$ of the ε-system. But $y_\varepsilon(t)$ and $\mu_0(t)$ take values in different spaces. Thus we need to define the type of convergence.

Let us describe briefly the properties of measure-valued functions. We use here the description of *chattering parametric functions* given in [A3]. By measure-valued functions we mean the family $\mathcal{M}$ of measurable mapping

$$\mu(t) : [0, 1] \to \mathcal{P}(\Gamma).$$

Here Γ is a separable complete metric space. A measure-valued function is assumed measurable if for every measurable $A \subset \Gamma$ the real-valued map $\mu(t)(A) : [0, 1] \to [0, 1]$ is measurable (see [A3, section 3]). Here $\mu(t)(A)$ is the weight the measure assigns to A. Any measurable function $r(t)$ with values in Γ can be presented as an element of $\mathcal{M}$ which for each $t \in [0, 1]$ is the measure concentrated on the point $\{r(t)\}$. We denote this measure-valued function by $\delta_r(t)$. Each measure-valued function $\mu(\cdot)$ can be identified with the measure $\boldsymbol{\mu}$ on the product $[0, 1] \times \Gamma$ which is the direct integral

$$\boldsymbol{\mu} = (D) \int_0^1 \mu(t)\, dt \tag{3.3}$$

defined as follows. On a Borel set $E \subset [0, 1] \times \Gamma$ it is given by

$$\boldsymbol{\mu}(E) = \int_0^1 \mu(t)(E_t)\, dt,$$

where $E_t = \{y : (t, y) \in E\}$ is a t-section of the set E.

Convergence on $\mathcal{M}$ is taken to be the weak convergence defined in (2.1)–(2.2) of the corresponding probability measures on $[0, 1] \times \Gamma$. (See also [A3, section 4].)

Note that since Γ, hence $[0, 1] \times \Gamma$, is complete and separable, it follows (see, e.g., Artstein [A3]) that $\mathcal{M}$ itself is a complete metric space. If Γ is compact, then $\mathcal{M}$ is also compact.

The distance on $\mathcal{M}$ between $\mu_1(\cdot)$ and $\mu_2(\cdot)$ is denoted by $\boldsymbol{\rho}(\mu_1, \mu_2)$ and equal to the Prohorov distance between the corresponding measures $\boldsymbol{\mu}_1$ and $\boldsymbol{\mu}_2$, namely $\boldsymbol{\rho}(\mu_1, \mu_2) = \rho(\boldsymbol{\mu}_1, \boldsymbol{\mu}_2)$. No confusion should arise with the distance $\boldsymbol{\rho}$ on $\mathcal{M}$ between two functions $\mu_1(\cdot)$ and $\mu_2(\cdot)$ and the distance on $\mathcal{P}(M)$ between two measures $\mu_1(t)$ and $\mu_2(t)$ for the t fixed.

Remark 3.7. For the sake of convenience we consider the problem on the unit interval of time $t \in [0, 1]$. Hence all measures $\boldsymbol{\mu}$ defined by (3.3) are probability measures namely $\boldsymbol{\mu}([0, 1] \times \Gamma) = 1$. If we take an interval $[0, T]$ instead of $[0, 1]$ then $\boldsymbol{\mu}([0, T] \times \Gamma) = T$ and T is common for all $\boldsymbol{\mu}$. Therefore the weak convergence of

measures on $\tilde{\Gamma} = [0,T] \times \Gamma$ is also defined by (2.1). The distance $\rho_T(\boldsymbol{\mu}_1, \boldsymbol{\mu}_2)$ can be taken as $T\rho(\boldsymbol{\mu}_1/T, \boldsymbol{\mu}_2/T)$. Thus all further results hold true for the general finite time interval case.

DEFINITION 3.8. *Let $\mu(t)$ be an element of $\mathcal{M}$. Let $y_k(t)$ be a sequence of measurable functions defined on* [0,1] *with values in* Γ. *We say that* $y_k(\cdot)$ statistically *converges to $\mu(\cdot)$ if the corresponding measure-valued functions $\delta_{y_k}(\cdot)$ converge to $\mu(\cdot)$ on $\mathcal{M}$.*

4. Relaxed controls. Optimal solutions of the limit control problem. We want the limit problem (3.1) to be an approximation of the ε-problem in some sense. In particular we need the optimal trajectories of the limit problem to be ε-close to some admissible trajectories of the ε-system. But it is well known that the solution of the optimal control problem may fail to exist; see Warga [Wa]. In this section we recall the notion of relaxed controls to provide (generalized) solutions to the limit optimal control problem (3.1). We apply the standard technique originally introduced by Warga [Wa] and developed by Artstein [A1, A2, A3, A4].

Following the ideas of Warga we introduce the relaxed controls as follows. We allow the control function to assign to each t a probability measure on U. Namely, the relaxed control $v(\cdot)$ is a measure-valued function:

$$v(t) : [0,1] \to \mathcal{P}(U). \tag{4.1}$$

Denote the family of all measurable relaxed controls (4.1) by $\mathcal{V}$. Convergence on $\mathcal{V}$ is taken to be the convergence of measure-valued functions.

Relaxed limit control problem. A relaxed control $v(\cdot)$ affects (3.1) by integration with respect to the measure $v(t)$, for each t, the effects of the points $u \in U$. To employ another integral notation in (3.1) could complicate the formulas. Thus we use here the standard notation. If $h(u)$ is a function of the variable u on U and if v is a measure on U, then $\bar{h}(v))$ is a function of v on $\mathcal{P}(U)$, which is the average $\int_U h(u)v(du)$. Define now the new data functions $\boldsymbol{Q}$ and $\boldsymbol{f}$:

$$\begin{aligned} \boldsymbol{Q}(u,x) &= \int_{\mathbb{R}^n} Q(u,x,y)\, \nu(u,x)(dy), \\ \boldsymbol{f}(u,x) &= \int_{\mathbb{R}^n} f(u,x,y)\, \nu(u,x)(dy). \end{aligned} \tag{4.2}$$

With this notation, the limit optimal control problem (3.1) with the availability of relaxed controls has the following form:

$$\begin{aligned} &\underset{v(\cdot)\in V}{\text{minimize}} \int_0^1 \bar{\boldsymbol{Q}}(v(t),x(t))\, dt \\ &\text{subject to } \dot{x} = \bar{\boldsymbol{f}}(v(t),x); \quad x(0) = x^0, \\ &\mu(t) = \bar{\nu}(v(t),x(t)). \end{aligned} \tag{4.3}$$

Here $\mu(t) \in \mathcal{P}(\mathbb{R}^n)$ and by $\bar{\nu}(v,x)$ we understand the following integration for every measurable set $A \subset \mathbb{R}^n$:

$$\bar{\nu}(v,x)(A) = \int_U \nu(u,x)(A)\, v(du);$$

namely, $\mu(t)$ for each t is an average of the measures $\nu(u, x(t))$ on U with the measure $v(t)$ on U.

We shall denote the cost of the relaxed problem by $\bar{J}_0(v)$, namely,

$$\bar{J}_0(v) = \int_0^1 \bar{\boldsymbol{Q}}(v(t), x_0(t))\, dt.$$

It is clear that if $v(t)$ for each t is a Dirac measure concentrated on the point $u(t)$, then $\bar{J}_0(v) = J_0(u)$.

The limit optimal control problem (3.1) with ordinary controls can be rewritten as follows:

$$\begin{gathered} \underset{u(\cdot)\in\mathcal{U}}{\text{minimize}} \int_0^1 \boldsymbol{Q}(u(t), x(t))\, dt \\ \text{subject to } \dot{x} = \boldsymbol{f}(u(t), x); \quad x(0) = x^0, \\ \mu(t) = \nu(u(t), x(t)). \end{gathered} \tag{4.4}$$

Approximation of relaxed controls.

PROPOSITION 4.1. *$\mathcal{V}$ is compact. The family of ordinary measurable controls $\mathcal{U}^m$ is dense in $\mathcal{V}$.*

Proof. The space of measures supported on a common compact set with the weak convergence of measures is compact. Since the convergence on $\mathcal{V}$ is taken to be the weak convergence of measures on $[0, 1] \times U$, the first claim follows. The approximation by ordinary controls follows from [A3, Proposition 4.5]. □

Suppose now that U can be included in a separable Banach space—say, Z. Then the Lebesgue integral $\int_0^t u(t)\, dt$ is well defined. If U is convex, then any measurable function $u(\cdot)$ can be approximated by a continuous function $u_\delta(t) \in U$ such that $u_\delta(t) \to u(t)$ as $\delta \to 0$ for almost every $t \in [0, 1]$ and defined as follows:

$$u_\delta(t) = \frac{1}{2\delta} \int_{t-\delta}^{t+\delta} u(s)\, ds. \tag{4.5}$$

COROLLARY 4.2. *If U is a convex subset of a separable Banach space, then the family $\mathcal{U}^c$ is dense in $\mathcal{V}$. Moreover, for any $u^0 \in U$ and $v_0(\cdot) \in \mathcal{V}$ there exists a sequence $u_k(\cdot) \in \mathcal{U}^c$ statistically converging to $v_0(\cdot)$ and such that $u_k(0) = u^0$ for any k.*

Proof. The first statement directly follows from Proposition 4.1 and (4.5). Suppose now that $\tilde{u}_k(\cdot)$ statistically converges to $v_0(\cdot)$, and take a $u^0 \in U$. Construct a new sequence $u_k(\cdot)$ such that $u_k(0) = u_0$, $u_k(t) = \tilde{u}_k(t)$ for $t \geq 1/k$, and affine otherwise. Then clearly $u_k(\cdot)$ statistically converges to $v_0(\cdot)$. This completes the proof. □

COROLLARY 4.3. *The family of piecewise constant functions $u(t) : [0, 1] \to U$ is dense in $\mathcal{V}$.*

Proof. By Proposition 4.1 it is enough to show that piecewise constant functions are dense in $\mathcal{U}^m$ in some strong topology, say L^1. Since U is compact, we can just show that for any *simple* function $u^0(t) : [0, 1] \to U$ with a finite number of different values u^i, $i = 1, k + 1$, there exists a sequence u_j of piecewise constant functions converging *in measure* to u. The proof is standard. It follows from the property that for any measurable subset $T \subset [0, 1]$ and any $\delta > 0$ there exists a subset $E \subset [0, 1]$ which is a union of a finite number of disjoint intervals of $[0, 1]$ and such that $\lambda(T \Delta E) < \delta$; see Halmos [Ha, p. 56]. □

Existence of optimal solutions to (4.3) and their approximation follow from the general theory of Warga [Wa, Chapter 4]. We formulate these results in the following lemma.

LEMMA 4.4. *Under Assumptions* 3.1–3.5 *the optimization problem* (4.3) *has an optimal relaxed control—say,* $v(\cdot)$. *There is a sequence of ordinary measurable controls* $u_k(\cdot) \in \mathcal{U}^m$ *which statistically converges to* $v(\cdot)$. *In particular the cost* $J_0(u_k)$ *converges to the value* $\mathrm{val}_0(\mathcal{U}^m) = \mathrm{val}_0(\mathcal{V})$.

5. Relatively slow controls. Preliminary convergence estimates. In this section we present the results on the limit behavior of the admissible trajectories $(x_\varepsilon(t), y_\varepsilon(t))$, with respect to prescribed sets. We establish also a preliminary result on the statistical convergence of the fast variable $y_\varepsilon(t)$.

The trajectories $(x_\varepsilon(t), y_\varepsilon(t))$ are induced by special families of *relatively slow controls*.

Relatively slow controls. We introduce now the special families of continuous, piecewise continuous, and measurable controls, depending additionally on the parameter ε. We call these families *relatively slow controls* since the control $u_\varepsilon(t)$ may change rapidly in time $t \in [0,1]$ as $\varepsilon \to \infty$, but for $\tau = \varepsilon^{-1}t$ the function $\tilde{u}_\varepsilon(\tau) = u_\varepsilon(\varepsilon\tau)$ converges to constant as $\varepsilon \to 0$ while τ is fixed. For instance if $u_\varepsilon(t) = \cos(t/\sqrt{\varepsilon})$, then $\tilde{u}_\varepsilon(\tau) = \cos(\sqrt{\varepsilon}\tau) \to 1$ as $\varepsilon \to 0$.

Let $F > 0$ be the diameter of the set U. Define two special positive real-valued functions $M(z)$ and Δ_ε of variables $z \in [0,1]$, $\varepsilon > 0$ such that for some $L > 0$

$$\begin{gathered} Lz \le M(z) \le F, \\ M(z) \to 0, \text{ as } z \to 0 \text{ and } \Delta_\varepsilon \to 0, \quad \varepsilon\Delta_\varepsilon^{-1} \to 0, \text{ as } \varepsilon \to 0. \end{gathered} \tag{5.1}$$

DEFINITION 5.1. *Let* $z \in [0, \Delta_\varepsilon]$. *A control function* $u_\varepsilon(\cdot)$ *is said to be a* relatively slow continuous control *with respect to the functions* $M(\cdot)$ *and* Δ_ε *if for any* $t \in [0,1]$

$$d_U(u_\varepsilon(t+z), u_\varepsilon(t)) \le M(\Delta_\varepsilon^{-1}z). \tag{5.2}$$

A control function $u_\varepsilon(\cdot)$ *is said to be a* relatively slow piecewise continuous control *if*

(i) *the minimal length of interval of continuity is not less than* Δ_ε,

(ii) *for* $t, t+z$ *from every common interval of continuity,* (5.2) *is satisfied.*

A control function $u_\varepsilon(\cdot)$ *is said to be a* relatively slow measurable control *if*

$$\int_0^1 d_U(u_\varepsilon(t+z), u_\varepsilon(t))\, dt \le M(\Delta_\varepsilon^{-1}z).$$

We denote by $\mathcal{U}_\varepsilon^c$,$\mathcal{U}_\varepsilon^p$, and $\mathcal{U}_\varepsilon^m$, respectively, the families of all relatively slow continuous, piecewise continuous, and measurable controls with respect to the fixed functions $M(\cdot)$ and Δ_ε.

In what follows we assume that the functions $M(z)$ and Δ_ε are given, and therefore the families $\mathcal{U}_\varepsilon^c$, $\mathcal{U}_\varepsilon^p$, $\mathcal{U}_\varepsilon^m$ are completely determined.

Remark 5.2. Let $M(\cdot)$ and Δ_ε be fixed functions. Note then that if $u(\cdot)$ is *piecewise constant* (with a finite number of discontinuities), then for small ε, $u(\cdot) \in \mathcal{U}_\varepsilon^p$ and $u(\cdot) \in \mathcal{U}_\varepsilon^m$. If $u(\cdot)$ is continuous and *piecewise linear* (if U is certainly included in a vector space), then for sufficiently small ε, $u(\cdot) \in \mathcal{U}_\varepsilon^c$. These inclusions obviously follow from Definition 5.1 and the properties of the latter functions. On the other

hand, by the standard approximation arguments, any continuous function can be uniformly approximated by a continuous piecewise linear function. Besides, by the proof of Corollary 4.3, any measurable function with values in a compact space can be approximated in L^1-metric by a piecewise constant function.

Therefore, for $\varepsilon \to 0$

(i) for an arbitrary measurable function $u(\cdot) \in \mathcal{U}^m$ there exists a function $u_\varepsilon(\cdot) \in \mathcal{U}^m_\varepsilon$ such that

$$\int_0^1 d_U(u(t), u_\varepsilon(t))\, dt \to 0 \text{ as } \varepsilon \to 0, \tag{5.3}$$

(ii) for an arbitrary piecewise continuous function $u(\cdot) \in \mathcal{U}^p$ there exists a function $u_\varepsilon(\cdot) \in \mathcal{U}^p_\varepsilon$ satisfying (5.3),

(iii) for an arbitrary continuous function $u(\cdot) \in \mathcal{U}^c$ (if U is convex) there exists a function $u_\varepsilon(\cdot) \in \mathcal{U}^c_\varepsilon$ satisfying

$$\sup_{t\in[0,1]} d_U(u(t), u_\varepsilon(t)) \to 0 \text{ as } \varepsilon \to 0.$$

In this sense the families $\mathcal{U}^c_\varepsilon$, $\mathcal{U}^p_\varepsilon$, $\mathcal{U}^m_\varepsilon$ converge as $\varepsilon \to 0$ to the basic families $\mathcal{U}^c$, $\mathcal{U}^p$, $\mathcal{U}^m$, respectively.

Preliminary estimates. We now establish two important lemmas which will be useful in what follows. The first lemma states convergence of the trajectory $(x_\varepsilon(t), y_\varepsilon(t))$ to prescribed sets. The second lemma determines statistical convergence of $y_\varepsilon(t)$ to the corresponding measure-valued function.

Suppose that $u_\varepsilon(\cdot)$ is a relatively slow control, and consider the fast equation from the system (1.1) separately:

$$\varepsilon \dot{y} = g(u_\varepsilon(t), x_\varepsilon(t), y). \tag{5.4}$$

LEMMA 5.3. *Suppose that Assumptions* 3.1–3.3 *are satisfied. Let D be a compact subset of $\operatorname{graph} G$.*

Then for any $\eta_0 > 0$ there exists an $\varepsilon_0 > 0$ such that if $\varepsilon < \varepsilon_0$, then for $(u_\varepsilon(0), x_\varepsilon(0), y_\varepsilon(0)) \in D$ and $u_\varepsilon(\cdot) \in \mathcal{U}^c_\varepsilon$

(i) *the pair $(x_\varepsilon(t), y_\varepsilon(t))$ can be extended to $t \in [0,1]$, and $(u_\varepsilon(t), x_\varepsilon(t), y_\varepsilon(t)) \in \operatorname{graph} G$. Moreover there exists a compact $X \subset \mathbb{R}^m$ such that $x_\varepsilon(t) \in X$;*

(ii) $\lambda\{t : t \in [0,1], d((u_\varepsilon(t), x_\varepsilon(t), y_\varepsilon(t)), \boldsymbol{K}) > \eta_0\} < \eta_0$;

(iii) $d((u_\varepsilon(t), x_\varepsilon(t), y_\varepsilon(t)), \boldsymbol{L}) < \eta_0$ *for* $t \in [\Delta_\varepsilon, 1]$.

Here $\boldsymbol{K}$, $\boldsymbol{L}$ are compact subsets of $\operatorname{graph} G$ defined as follows:

$$\begin{aligned} \boldsymbol{K} &= \{(u,x,y) : u \in U, x \in X, y \in K(u,x)\}, \\ \boldsymbol{L} &= \{(u,x,y) : u \in U, x \in X, y \in L(u,x)\}. \end{aligned} \tag{5.5}$$

Proof. (i). The result follows directly from [AV, Lemma 4.7]. (ii) and (iii). Denote $p = (u,x)$ and $p_\varepsilon(t) = (u_\varepsilon(t), x_\varepsilon(t))$. Take $\eta < \eta_0$ such that $\boldsymbol{L}^\eta \subset \operatorname{graph} G$. Such an η exists by Assumption 3.2 and Proposition 2.3. Denote $C = \operatorname{cl} L^\eta$. Choose $s_0 > 0$ and $\theta > 0$ provided by Proposition 2.6 with $E = C \cup D$. Take $\varepsilon_0 > 0$ such that for $\varepsilon < \varepsilon_0$ we have $\varepsilon s_0 < \eta$, $2\varepsilon s_0 < \Delta_\varepsilon$ and the following inequality holds for $s \in [s_0, 2s_0]$ and $t \in [0, 1 - \varepsilon s]$:

$$M(\varepsilon \Delta_\varepsilon^{-1} s) + |x_\varepsilon(t + \varepsilon s) - x_\varepsilon(t)| < \theta.$$

Such an ε_0 exists by property (5.1) of the functions $M(z)$, Δ_ε. Then since $u_\varepsilon(\cdot)$ satisfies (5.2) we obtain that for $s \in [s_0, 2s_0]$ and $t \in [0, 1-\varepsilon s]$:

$$d_P(p_\varepsilon(t+\varepsilon s), p_\varepsilon(t)) < \theta.$$

Define a subset of $[0,1]$:

$$R_\varepsilon = \{t : (p_\varepsilon(t), y_\varepsilon(t)) \in E\}.$$

Then Proposition 2.6 implies that for $t' \in R_\varepsilon$ and $s \in [s_0, 2s_0]$

$$\begin{gathered} \frac{1}{\varepsilon s}\lambda(\{t : t \in [t', t'+\varepsilon s], d((p_\varepsilon(t), y_\varepsilon(t)), \boldsymbol{K}) > \eta\}) < \eta, \\ d((p_\varepsilon(t'+\varepsilon s), y_\varepsilon(t'+\varepsilon s)), \boldsymbol{L}) < \eta. \end{gathered} \tag{5.6}$$

Construct a sequence $t_{i+1} = t_i + \varepsilon s_0$, $i = 0, k-1$, $t_0 = 0, t_{k+1} = 1$ such that $1 - t_k < \varepsilon s_0$. Then since $t_0 \in R_\varepsilon$ we obtain from (5.6) that $t_i \in R_\varepsilon$ for $i = 1, k-1$. Then from (5.6) it follows that for $i = 0, k$,

$$\begin{gathered} \lambda(\{t : t \in [t_i, t_{i+1}], d((p_\varepsilon(t), y_\varepsilon(t)), \boldsymbol{K}) > \eta\}) < \eta(t_{i+1} - t_i), \\ d((p_\varepsilon(t), y_\varepsilon(t)), \boldsymbol{L}) < \eta \quad \forall t \in [t_{i+1}, t_{i+2}]. \end{gathered} \tag{5.7}$$

By an extension of (5.7) to the interval $[0,1]$ and as $\varepsilon s_0 < \Delta_\varepsilon$ we obtain the desired result. This completes the proof. □

LEMMA 5.4. *Suppose that Assumptions* 3.1–3.5 *are fulfilled and the function $u_\varepsilon(\cdot)$ is a relatively slow measurable control. Furthermore, assume that $(x_\varepsilon(t), y_\varepsilon(t))$ for small ε can be extended to $t \in [0,1]$ and there exists a compact set C, $C \subset \operatorname{graph} G$, such that*

$$\lambda\{t : (u_\varepsilon(t), x_\varepsilon(t), y_\varepsilon(t)) \notin C\} \to 0 \text{ as } \varepsilon \to 0. \tag{5.8}$$

Then the distance $\boldsymbol{\rho}(\delta_{y_\varepsilon}, \mu_\varepsilon)$ tends to 0 as $\varepsilon \to 0$. Here $\mu_\varepsilon(t) = \nu(u_\varepsilon(t), x_\varepsilon(t))$ is invariant for the flow $\phi(\tau, u_\varepsilon(t), x_\varepsilon(t), y)$ for any t fixed.

Proof. Denote $p_\varepsilon(t) = (u_\varepsilon(t), x_\varepsilon(t))$. Take any bounded uniformly continuous function $h(t,y) : [0,1] \times \mathbb{R}^n$ and an arbitrary $\eta > 0$. Define the set

$$I_\varepsilon = \{t : (p_\varepsilon(t), y_\varepsilon(t)) \notin C\}.$$

From Propositions 2.4 and 2.7 it follows that there exists an $s_0 > 0$ such that for $t \in [0,1] \setminus I_\varepsilon$

$$\left| \int_{\mathbb{R}^n} h(t,y)\nu(p_\varepsilon(t))(dy) - \frac{1}{s_0}\int_0^{s_0} h(t, \phi(\tau, p_\varepsilon(t), y_\varepsilon(t)))\, d\tau \right| < \frac{\eta}{4}. \tag{5.9}$$

Note that

$$\int_0^1 h(t, y_\varepsilon(t))\, dt = \int_{-\varepsilon\tau}^{1-\varepsilon\tau} h(t+\varepsilon\tau, y_\varepsilon(t+\varepsilon\tau))\, d\tau.$$

Since h is bounded and uniformly continuous, we can find a positive $\varepsilon_1 > 0$ such that for $\varepsilon < \varepsilon_1$ and $\tau \in [0, s_0]$

$$\left| \int_0^1 h(t, y_\varepsilon(t))\, dt - \int_0^{1-\varepsilon_1 s} h(t, y_\varepsilon(t+\varepsilon\tau))\, dt \right| < \frac{\eta}{4}. \tag{5.10}$$

A change of time scale t to $\tau = (t-t')/\varepsilon$ yields that $y_\varepsilon(t'+\varepsilon\tau)$ solves the equation

$$\frac{dy}{d\tau} = g(p_\varepsilon(t'+\varepsilon\tau), y), \qquad y(0) = y_\varepsilon(t'). \tag{5.11}$$

The flow $\phi(\tau, p_\varepsilon(t'), y_\varepsilon(t'))$ for t' fixed solves the following equation:

$$\frac{dy}{d\tau} = g(p_\varepsilon(t'), y), \qquad y(0) = y_\varepsilon(t'). \tag{5.12}$$

Denote

$$n(\varepsilon, t') = \frac{1}{s_0}\int_0^{s_0} d_P(p_\varepsilon(t'+\varepsilon\tau), p_\varepsilon(t'))\, d\tau.$$

Take $\theta > 0$ such that if $n(\varepsilon, t') < \theta$, then for $\tau \in [0, s_0]$ the solutions of (5.11) and (5.12) satisfy

$$w(|y_\varepsilon(t'+\varepsilon\tau) - \phi(\tau, p_\varepsilon(t'), y_\varepsilon(t'))|) < \frac{\eta}{4}. \tag{5.13}$$

Here the function $w(s)$ is the modulus of continuity of the function $h(t, y)$. Such a θ exists by standard continuous dependence arguments.

Define the set T_ε:

$$T_\varepsilon = \{t' : n(\varepsilon, t') \geq \theta\}.$$

By property $x_\varepsilon(t)$ and $u_\varepsilon(t)$ (Definition 5.1) $\lambda(T_\varepsilon) \to 0$ as $\varepsilon \to 0$. This implies that there exists an $\varepsilon_2 > 0$ (and $\varepsilon_2 \leq \varepsilon_1$) such that for $\varepsilon < \varepsilon_2$ we get

$$\int_{I_\varepsilon} h(t, y_\varepsilon(t))\, dt + \int_{T_\varepsilon} h(t, y_\varepsilon(t+\varepsilon\tau))\, dt < \frac{\eta}{4}. \tag{5.14}$$

Then from (5.9), (5.10), (5.13), and (5.14) we obtain that for $\varepsilon < \varepsilon_2$

$$\left| \int_0^1 h(t, y_\varepsilon(t))\, dt - \int_0^1 \left(\int_{\mathbb{R}^n} h(t, y)\nu(p_\varepsilon(t))(dy) \right) dt \right| < \eta.$$

This implies that $\boldsymbol{\rho}(\delta_{y_\varepsilon}(t), \nu(p_\varepsilon(t))) \to 0$, which completes the proof. □

6. Approximating properties of the limit problem: The main results. In this section we present the main approximating properties of the limit control problem (3.1) and the relaxed limit problem (4.3).

Approximation of relaxed controls by relatively slow controls.

Claim 6.1. Let $v_0(\cdot) \in \mathcal{V}$ be an arbitrary relaxed control. Then

(i) there exists a relatively slow *measurable* control $u_\varepsilon(\cdot)$ statistically converging to $v_0(\cdot)$ as $\varepsilon \to 0$;

(ii) there exists a relatively slow *piecewise continuous* control $u_\varepsilon(\cdot)$ statistically converging to $v_0(\cdot)$ as $\varepsilon \to 0$;

(iii) if U is a convex subset of a separable Banach space, there exists a relatively slow continuous control $u_\varepsilon(\cdot)$ statistically converging to $v_0(\cdot)$ as $\varepsilon \to 0$. Moreover, for any $u^0 \in U$ there exists a relatively slow continuous control $u_\varepsilon(\cdot)$ statistically converging to $v_0(\cdot)$ and such that $u_\varepsilon(0) = u_0$ for any ε.

Proof. The result directly follows from Proposition 4.1, Corollaries 4.2 and 4.3, and Remark 5.2. □

Convergence result with continuous control. We establish now the first convergence result. We refer to the original ε-problem, the limit problem, and the limit relaxed problem. Let $v_0(\cdot)$ be an arbitrary admissible relaxed control defined in (4.1). Let the pair $(x_0(t), \mu_0(t))$ be the trajectory of (4.3) induced by $v_0(t)$. Let the pair $(x_\varepsilon(t), y_\varepsilon(t))$ be the trajectory of (1.1)–(1.2) induced by the control $u_\varepsilon(t)$ and an initial data $x^0_\varepsilon \to x^0$ and $y^0_\varepsilon \to y^0$. Note that if $u_\varepsilon(\cdot)$ is a relatively slow continuous control, then by Lemma 5.3 this trajectory indeed exists on $[0,1]$ and the triple $(u_\varepsilon(t), x_\varepsilon(t), y_\varepsilon(t))$ is included in $graph\, G$ if $(u_\varepsilon(0), x_\varepsilon(0), y_\varepsilon(0))$ is in a compact set E of $graph\, G$ for all $\varepsilon > 0$ small enough.

THEOREM I. *Under Assumptions* 3.1–3.5*, suppose that there exists a relatively slow continuous control $u_\varepsilon(\cdot)$ converging to $v_0(\cdot)$, and a compact set $E \subset graph\, G$ such that $(u_\varepsilon(0), x_\varepsilon(0), y_\varepsilon(0)) \in E$ for all $\varepsilon > 0$ small enough. Then*

(i) $\sup_{t\in[0,1]} |x_\varepsilon(t) - x_0(t)| \to 0$ *as* $\varepsilon \to 0$*;*

(ii) $y_\varepsilon(t)$ *statistically converges to the measure-valued function* $\mu_0(t)$*, where* $\mu_0(t) = \bar\nu(v_0(t), x_0(t))$*;*

(iii) $J_\varepsilon(u_\varepsilon) \to \bar J_0(v_0)$ *as* $\varepsilon \to 0$*.*

COROLLARY 6.2. *Suppose that $v_0(t) = \delta_{u_0}(t)$ where $u_0(\cdot) \in \mathcal{U}^c$. Then under Assumptions* 3.1–3.5 *if $y^0 \in G(u_0(0), x^0)$, then the trajectory $(x_\varepsilon(t), y_\varepsilon(t))$ and the cost $J_\varepsilon(u)$ of the ε-system induced by the control $u_0(t)$ converge, respectively, to the trajectory $(x_0(t), \mu_0(t))$ and the cost $J_0(u_0)$ of the limit system* (3.1)*. This convergence is uniform for (x^0, y^0) in compact sets of the graph $G(u_0(0), \cdot)$.*

Proof. It follows immediately from Theorem I. □

Remark 6.3. Corollary 6.2 coincides with [AV, Theorem II], where the result has been proved for uncontrolled systems.

Remark 6.4. For some $u^0 \in U$ take a compact set $H \subset graph\, G(u^0, \cdot)$. If U is a convex subset of a separable Banach space, then Claim 6.1 implies that there exists a relatively slow continuous control $u_\varepsilon(\cdot)$ converging to $v_0(\cdot)$ and such that $u_\varepsilon(0) = u^0$. In this case if $(x^0, y^0) \in H$, then the ε-system can always approximate any trajectories and the cost (in particular, the value) of the limit relaxed problem. But if U is not convex, then, in general, this approximation does not work.

Example 6.5. We provide now an example illustrating the convergence result in Theorem I. (Compare with [AV, Example 7.4].)

Consider the system with $U = [2,4]$, $x \in \mathbb{R}$, and $y = (y_1, y_2) \in \mathbb{R}^2$:

$$\begin{aligned} \dot x &= 2\frac{y_1^2}{|y|} - 3x, \\ \varepsilon \dot y_1 &= y_1\left(1 - \frac{|y|}{xu}\right) - y_2, \\ \varepsilon \dot y_2 &= y_2\left(1 - \frac{|y|}{xu}\right) + y_1, \\ x(0) &= x^0, \quad y(0) = y^0. \end{aligned}$$

Here by $|y|$ we denote the Euclidean norm of the vector y. Define

$$Q(u,x,y) = (x-1)^2 + (u-2)^2(u-4)^2 + y_1.$$

Then

$$J_\varepsilon(u) = \int_0^1 (x_\varepsilon(t) - 1)^2 + (u(t) - 2)^2(u(t) - 4)^2 + y_{1,\varepsilon}(t)\, dt.$$

The fast system for $u > 0, x > 0$ fixed, converges to the limit cycle $|y| = ux$ if $|y(0)| \neq 0$. The sets $G(u,x)$ and $K(u,x)$ can be chosen as follows: $G(u,x) = \{y : 3ux > |y| > 0\}$ and $K(u,x) = \{(y : |y| = xu\}$. The prolongation $L(u,x)$ of $K(u,x)$ coincides with $K(u,x)$. It is easily checked that all conditions of Theorem I are satisfied. The invariant measure $\nu(u,x)$ is equally distributed on the circle $|y| = xu$, namely,

$$\nu(u,x)(A) = \frac{1}{2\pi}\int_0^{2\pi} \chi_A(ux\cos\theta, ux\sin\theta)\,d\theta,$$

where $A \subset \mathbb{R}^2$ and χ_A is the indicator function of the set A. Then it is easy to see that the limit problem has the following form:

$$\begin{aligned} &\text{minimize} \int_0^1 (x(t)-1)^2 + (u(t)-4)^2(u(t)-2)^2\,dt \\ &\text{subject to } \dot{x} = x(u(t)-3); \quad x(0) = x^0, \\ &\mu(t) = \nu(u(t), x(t)). \end{aligned} \tag{6.1}$$

Thus the slow solution $x_\varepsilon(t)$ tends to $x_0(t) = x^0\exp(\int_0^t (u(t)-3)\,dt)$, while the fast solution tends to the measure-valued function $\nu(u(t), x_0(t))$. Suppose that $x^0 = 1$. It is easy to see that the optimization problem (6.1) does not have an optimal solution. The infimal cost in this problem is zero, but clearly it cannot be achieved with an ordinary control. But we can construct a sequence of controls u_k such that $J_0(u_k) \to 0$. To do this we define the relaxed limit problem in the following form:

$$\begin{aligned} &\text{minimize} \int_0^1 (x(t)-1)^2 + \int_U (u-4)^2(u-2)^2\,v(t)(du)dt \\ &\text{subject to } \dot{x} = x\int_U (u-3)\,v(t)(du), \\ &\mu(t) = \int_U \nu(u, x(t)\,v(t)(du). \end{aligned}$$

The optimal relaxed control $v_0(\cdot)$ assigns to each t the values 2 and 4 with equal probability. Then $\dot{x}_0(t) = 0$ and $x_0(t) = x^0 = 1$, and $\mu(t) = 1/2\nu(2,1) + 1/2\nu(4,1)$. The relaxed control can certainly be approximated by the sequence of continuous controls $\tilde{u}_k$ which can be defined as follows. The interval $[0,1]$ can be divided on $2k$ intervals of length $1/k - 1/k^2$ and $1/k^2$ where the small and large intervals alternate. Then $\tilde{u}_k(t)$ is chosen equal to 2 or 4 on the large alternate intervals and affine on the small intervals.

We now have to find a relatively slow continuous control $u_\varepsilon(t)$ statistically converging to $v_0(t)$ and such that $y^0 \in G(u_\varepsilon(0), 1)$. Suppose that $y^0 = (1,0)$. Then $(1,0) \in G(u,1)$ for $u \in [2,4]$ and we can take any initial values of $u_\varepsilon(0)$. To find u_ε we should define the functions $M(z)$ and Δ_ε. Take $\Delta_\varepsilon = \sqrt{\varepsilon}$ and $M(z) = z$. Define an integer-valued function $k(\varepsilon)$ such that $k(\varepsilon) < \varepsilon^{-1/4}$. Then it is easy to verify that $u_\varepsilon(t) = \tilde{u}_{k(\varepsilon)}$ is a relatively slow continuous control with respect to the given $M(z)$ and Δ_ε.

Suppose now that the set U is not convex and consists only of two points 2 and 4. Namely, $U = \{2,4\}$. Then we have only two admissible continuous functions $u(t) \equiv 2$

or 4 which certainly cannot approximate the relaxed control v_0. We should take a piecewise continuous control to approximate the optimal solution of the limit relaxed problem.

However, if $u(\cdot)$ is a piecewise continuous function, then the assumptions of Theorem I are not enough to provide the convergence result even in the case when the invariant measure is a Dirac measure; namely, the reduced order system (1.4) can be defined.

Example 6.6. Consider only the fast equation with $u(t) \in U = [1/2, 1]$,

$$\varepsilon \dot{y} = y\left(1 - \frac{y}{2u}\right)\left(1 - \frac{y}{u}\right);\ y(0) = y^0 > 0.$$

Then we can define the open set $G(u) = (0, 2u)$ and the corresponding compact $K(u) = \{u\}$. All conditions of Theorem I are satisfied. Suppose now that $u(t)$ is the piecewise constant function $u(t) = 1$ if $t \in [0, 1/2]$ and $u(t) = 1/2$ if $t \in [1/2, 1]$. Then if $y^0 > 1$, at time $t = 1/2$, set $G(u(t)) = G(1)$ switches to $G(1/2) = (0, 1)$ and the fast solution $y_\varepsilon(1/2) > 1$, and therefore $y_\varepsilon(t) \notin G(1/2)$. Thus $y_\varepsilon(t)$ goes to infinity as $\varepsilon \to 0$ and the desired convergence does not hold.

Convergence result with piecewise continuous controls.

Assumption 6.7. There exists a set $G(u, x)$ such that for any $u \in U, u' \in U$, and x, the following inclusion is valid: $G(u, x) \supset L(u', x)$.

Note that in Example 6.6 there is no $G(u, x)$ satisfying Assumption 6.7.

Remark 6.8. Assumption 6.7 is satisfied automatically if $G(u, x)$ can be found not depending on u.

The following theorem extends Theorem I to the case where $u_\varepsilon(\cdot)$ is a relatively slow piecewise continuous control. Notice that if Assumption 6.7 is satisfied then there exists a compact set H such that $H \subset G(u, \cdot)$ for any $u \in U$.

THEOREM II. *Suppose that Assumptions* 3.1–3.5*, Lemma* 3.6*, Remark* 3.7*, and Assumption* 6.7 *are fulfilled. Let $u_\varepsilon(\cdot)$ be a relatively slow piecewise continuous control statistically converging to $v_0(\cdot)$. Let H be a compact subset of graph $G(u_\varepsilon(0), \cdot)$ for any ε and $(x^0_\varepsilon, y^0_\varepsilon) \in H$. Then all statements of Theorem* I *hold.*

COROLLARY 6.9. *Suppose that $v_0(t) = \delta_{u_0}(t)$, where $u_0(\cdot) \in \mathcal{U}^p$. Then under Assumptions* 3.1–3.5 *and* 6.7 *and Lemma* 3.6 *if $(x^0, y^0) \in H$, the trajectory $(x_\varepsilon(t), y_\varepsilon(t))$ and the cost $J_\varepsilon(u)$ of the ε-system induced by the control $u_0(t)$ converge, respectively, to the trajectory $(x_0(t), \mu_0(t))$ and the cost $J_0(u_0)$ of the limit system* (3.1)*. This convergence is uniform for $(x^0, y^0) \in H$.*

Proof. It follows directly from Theorem II. □

Remark 6.10. Take a compact set $H \subset graph\, G(w, \cdot)$ for all $w \in U$. By Assumption 6.7 such an H exists. Therefore if $(x^0, y^0) \in H$, then the ε-system can always approximate any trajectories and the cost (in particular the value) of the limit relaxed problem if the family $\mathcal{U}^p_\varepsilon$ is applied.

Example 6.11. Consider the system from Example 6.5 and suppose that $U = \{2, 4\}$. Note that the prolongation of the set $K(u, x)$ coincides in this example with $K(u, x)$, namely $L(u, x) = K(u, x) = \{y : |y| = xu\}$. Recall that we take $G(u, x) = \{y : 3ux > |y| > 0\}$. Since $6x > 4x$ we see that Assumption 6.7 is fulfilled. Then if we define $\Delta_\varepsilon = \sqrt{\varepsilon}$, we can take a relatively slow piecewise continuous control which is equal to 2 or to 4 on alternative intervals with the same length equal to $\delta(\varepsilon)$ converging to zero and such that $\delta(\varepsilon) \geq \sqrt(\varepsilon)$.

Remark 6.12. We emphasize that piecewise continuous functions considered by us have only discontinuities of the first type. The next example illustrates the case

when discontinuity of a control $u_\varepsilon(\cdot)$, even at a single point (say, $t=0$), may falsify the consequences of Theorem II and even Corollary 6.9 if the right limit $u(0^+)$ does not exist.

Example 6.13. This example employs the same idea of resonance as in [AV, Example 9.2]. We could use the example from [AV], but let us show that the same phenomenon arises in control systems which are linear with respect to control. We shall show that even the convergence result in Corollary 6.9 does not hold in general if we deal with measurable controls. We consider the case when $u_0(t)$ has only the discontinuity of the second type in 0.

Define a nonlinear two-dimensional controlled oscillator with $U=[-1,1]$ as follows:

$$\begin{aligned} \frac{dy_1}{d\tau} &= y_2, \\ \frac{dy_2}{d\tau} &= -k\left(1-\frac{|y|}{4}\right)y_2 - y_1 + \beta(y)u. \end{aligned} \tag{6.2}$$

Here $k>0$ is small and

$$\beta(y)=\begin{cases} 1, & |y|<4, \\ 2(1-\frac{|y|}{8}), & 8>|y|\geq 4, \\ 0, & |y|\geq 8. \end{cases}$$

The corresponding fast system for $t\in[0,1]$ has the following form:

$$\varepsilon\dot{y}=g(u,y); \qquad g(u,y)=\begin{pmatrix} y_2 \\ -k(1-\frac{|y|}{4})y_2 - y_1 + \beta(y)u \end{pmatrix}. \tag{6.3}$$

For u fixed, the point $(u,0)$ is asymptotically stable for (6.2) if $|y(0)|<4$. For any u fixed, the sets $K(u)$ and $G(u)$ can be defined as follows: $K(u)=\{(u,0)\}$ and $G(u): G(u)=\{(y_1,y_2):(y_1-u)^2+y_2^2<9\}$. It is clear that for any $u\in[-1,1]$

$$\{y:|y|<2\}\subset G(u)\subset\{y:|y|<4\}.$$

The prolongation $L(u)$ of $K(u)$ coincides with $K(u)$ in this example. Namely $L(u)=K(u)$ and then $L(u)\subset G(v)$ for any $u,v\in[-1,1]$. All assumptions of Theorem II are satisfied, and we can guarantee the statistical convergence of the solution $y_\varepsilon(t)$ of (6.3) to the measure $\delta_{(u(t),0)}$ if any piecewise continuous control $u(\cdot)$ is applied and if $|y_\varepsilon(0)|<4$. In this example the statistical convergence of $y_\varepsilon(t)$ is reduced to the L^1 convergence to the function $(u(t),0)$ (see [A3, Remark 4.4]). Besides for small ε: $y_\varepsilon(t)\in G(u(t))$ and therefore $|y_\varepsilon(t)|<4$.

Suppose now that in (6.2) $u=\sin(\tau)$. Then it can be proved that for small k because of the resonance, the solution tends to infinity for any initial condition. This convergence is uniform with respect to $y(0)$ in compact sets of the set $\{y:|y|<8\}$. Since for $|y(0)|\geq 8$ the solution $y(\tau)$ never comes back in the ball $\{y:|y|<8\}$ we conclude that there exists a number $T>0$ such that for any initial data $y(0)$ we have $y(T)>8$.

Now we are ready to construct a control $u_0(t)$ which is continuous on $(0,1]$ and such that for any $y^0\in G(u(0))$, $y_\varepsilon(t)$ does not converge to the function $(u_0(t),0)$. As in [AV, Example 9.2] we identify a sequence of disjoint intervals $[a_j,b_j]$ with $a_j>0$ and $a_j,b_j\to 0$ as $j\to\infty$. Next we denote $\varepsilon_j=(b_j-a_j)T^{-1}$.

Define $u_0(t)$ as follows:
(1) $u_0(0) = 0$.
(2) For $t \in [a_j, b_j]$ let $u_0(t) = \sin(\varepsilon_j^{-1}(t - a_j))$.
(3) For $t \in (b_j, a_{j-1})$ define $u_0(t)$ such that $u_0(t)$ becomes continuous on $[b_j, a_{j-1}]$.
Note that $u_0(0^+)$ does not exist.

Consider now the solution $y_\varepsilon(t)$ with $|y_\varepsilon(0)| < 2$ (namely, $y_\varepsilon(0) \in G(u)$ for all $u \in [-1, 1]$) and take a sequence $\varepsilon_j \to 0$. Then by the properties of the sequence a_j, b_j for any j large, we have $|y_{\varepsilon_j}(b_j)| > 8$ and therefore $|y_{\varepsilon_j}(t)| > 8$ for all $t \in [b_j, 1]$. So $y_{\varepsilon_j}(t)$ stays away from the set $\{(y_1, y_2) : y_1 \in [-1, 1], y_2 = 0\}$ for $t \in [b_j, 1]$ and cannot converge to the function $(u(t), 0)$. The counterexample is complete.

Convergence result with measurable controls. We present now a similar result as in Theorems I and II but in the general case when the relatively slow measurable control is applied. An additional condition of the following theorem may not seem constructive since it is difficult to check it in the general case. However we shall display an assumption and the corresponding example when these conditions are fulfilled automatically.

THEOREM III. *Suppose that Assumptions* 3.1–3.5 *are satisfied. Let $v_0(\cdot)$ be a relaxed control and $u_\varepsilon(t)$ be a relatively slow measurable control statistically converging to $v_0(t)$. Let H be a compact subset of $\mathbb{R}^m \times \mathbb{R}^n$, $(x_\varepsilon^0, y_\varepsilon^0) \in H$, and $(x_\varepsilon^0, y_\varepsilon^0) \to (x^0, y^0)$. Assume, additionally, that there exists a compact set $\boldsymbol{C} \subset \mathit{graph}\, G$ such that*

$$\lambda(\{t : (u_\varepsilon(t), x_\varepsilon(t), y_\varepsilon(t)) \notin \boldsymbol{C}\}) \to 0 \text{ as } \varepsilon \to 0. \tag{6.4}$$

Then all convergence statements of Theorem I *hold.*

Remark 6.14. It is easy to show that condition (6.4) (which is identical to (5.8) in Lemma 5.4) actually is a necessary condition for the desired convergence in Theorems I and II. We omit the details. Notice just that in Example 6.13 the fast motion $y_\varepsilon(t)$ goes to infinity as $\varepsilon \to 0$ for $t \in [\delta, 1]$, where $\delta > 0$ is an arbitrary fixed positive number. Thus the desired convergence cannot hold.

The following assumption implies condition (6.4) of Theorem III. Recall that by Remark 2.5, the set $G(u, x)$ may be taken closed and upper semicontinuous in (u, x) if the existence of the triple $(u_\varepsilon(t), x_\varepsilon(t), y_\varepsilon(t)) \in \mathit{graph}\, G$ for $t \in [0, 1]$ is known a priori.

Assumption 6.15. Suppose that G is compact and does not depend on u, x. Assume additionally that for any measurable function $u(\tau) : \mathbb{R}_+ \to U$, x fixed and $y^0 \in G$, the solution of the system

$$\frac{dy}{d\tau} = g(u(\tau), y, x), \quad y(0) = y^0, \quad x\text{-fixed}, \tag{6.5}$$

is unique and is included in the set G for any $\tau \geq 0$. We denote this by $\pi(\tau, u, x, y)$.

LEMMA 6.16. *Assume that $y^0 \in G$ and $x^0 \in X_0$, where X_0 is a compact subset of $\mathbb{R}^m$. Then under Assumptions* 3.1, 3.3–3.5, *and* 6.15 *there exists a compact $X \subset \mathbb{R}^m$ such that for any $\varepsilon > 0$ and any measurable $u_\varepsilon(t)$, the solution of the ε-system exists uniquely for $t \in [0, 1]$ and $x_\varepsilon(t) \in X$, $y_\varepsilon(t) \in G$.*

Proof. From Assumption 6.15 we immediately see that the solution of (6.5) exists uniquely for any $\tau \geq 0$ if x is a piecewise constant function. Take an arbitrary $\varepsilon > 0$ and let $\tau' = \varepsilon^{-1}$. Then since any continuous function on $[0, \tau']$ can be approximated uniformly by a sequence of piecewise constant functions, we see by the continuous dependence and compactness arguments that for $\tau \in [0, \tau']$ the solution of (6.5) exists uniquely and is included in G for any continuous function $x(\tau)$. After changing

the time scale from τ to $t = \varepsilon\tau$ and from Assumption 3.3, we obtain the desired result. □

Lemma 6.16 immediately implies that under Assumption 6.15 the conditions of Theorem III are fulfilled. Indeed the set C can be defined then as the product $C = U \times X \times G$. Moreover $(u_\varepsilon(t), x_\varepsilon(t), y_\varepsilon(t)) \in C$ for every $t \in [0,1]$. In the following example we consider a case in which Assumption 6.15 is satisfied.

Example 6.17. This example displays the convergence result if any measurable function $u(\cdot)$ is applied. The fast subsystem is defined on $\mathbb{R}^2$ and is presented in polar coordinates (r, θ):

$$\begin{aligned} \dot{x} &= f(x, r, \theta, u); x(0) = x^0, \\ \varepsilon\dot{r} &= r(1-r);\ r(0) = r^0, \\ \varepsilon\dot{\theta} &= \sin^2(\theta/2) + ux^2;\ \theta(0) = \theta^0. \end{aligned}$$

Here $u(t) \in [0,1]$. The associated system has the following form:

$$\begin{aligned} \frac{dr}{d\tau} &= r(1-r), \\ \frac{d\theta}{d\tau} &= \sin^2\left(\frac{\theta}{2}\right) + ux^2. \end{aligned}$$

It is clear that the sets $G = \{(r,\theta)) : a \le |r| \le b\}$ are positively invariant for $0 < a < 1 < b$ if any measurable $u(\tau)$ is used. We can also define $K = L = G$. Hence all conditions of Lemma 6.16, and thus of Theorem III, are satisfied.

For $ux^2 > 0$ the system has a stable limit cycle and the unique invariant measure $\nu(u,x)$ which is supported on the circle $\{y : |y| = 1\}$ with density $\beta(u,x,\theta)$ for $\theta \in [0, 2\pi]$ given by

$$\beta(u,x,\theta) = \left((\sin^2\left(\frac{\theta}{2}\right) + ux^2) \int_0^{2\pi} \frac{ds}{\sin^2(\frac{s}{2}) + ux^2} \right)^{-1}.$$

If $ux^2 = 0$ then the system has an attracting point $r = 1, \theta = 0$ and the unique invariant measure is a Dirac measure concentrated at the point $(1,0)$.

Thus for all measurable $u(\cdot)$ the solutions of the original system converge to the solution of the limit system:

$$\begin{aligned} \dot{x} &= \int_0^{2\pi} f(x,1,\theta)\nu(u,x)(d\theta);\ x(0) = x^0, \\ \nu(u,x)(d\theta) &= \begin{cases} \beta(u,x,\theta)d\theta, & ux^2 > 0, \\ \delta(\theta)d\theta, & ux^2 = 0. \end{cases} \end{aligned}$$

(Here $\delta(\cdot)$ is the delta-function.)

If we take a relaxed control $v(t)$, we can find a relatively slow measurable control $u_\varepsilon(t)$ statistically converging to $v(t)$ and such that the corresponding trajectory $(x_\varepsilon(t), y_\varepsilon(t))$ converges to the trajectory of the relaxed limit problem

$$\begin{aligned} \dot{x} &= \int_U \left(\int_0^{2\pi} f(x,1,\theta)\nu(u,x)(d\theta) \right) v(\tau)(du);\ x(0) = x^0, \\ \bar{\nu}(v(\tau),x) &= \int_U \nu(u,x)v(\tau)(du). \end{aligned}$$

Continuity of the value. The last result of this paper deals with continuity of the value of the ε-problem with respect to ε on the families of relatively slow controls, namely, $\mathrm{val}_\varepsilon(\mathcal{U}_\varepsilon) \to \mathrm{val}_0(\mathcal{U})$. Note that in general the convergence $\mathrm{val}_\varepsilon(\mathcal{U})$ to $\mathrm{val}_0(\mathcal{U})$ does not hold. Sometimes we can find a control $u_\varepsilon(t)$ such that the cost $J_\varepsilon(u_\varepsilon)$ may be much less than $\mathrm{val}_0(\mathcal{U})$. Usually a phenomenon of resonance and nonconvexity of the cost function Q is employed. We refer to Gaitsgory [G1, G2] for various examples.

Example 6.18. Consider the fast system of Example 6.13. Define $Q(u,y) = u^2 - |y|^2$. The limit system is defined only by the algebraic equation $y_1 = u, y_2 = 0$. Then $Q(u,y)$ is equal to zero on admissible trajectories and thus $\mathrm{val}_0(\mathcal{U}) = 0$. On the other hand if we apply a control $u = \sin(t/\varepsilon)$ to the fast system we obtain $|y_\varepsilon(t)| \to \infty$ as $\varepsilon \to 0$ for $t \in (0,1]$. Therefore $\mathrm{val}_\varepsilon(\mathcal{U}) \to -\infty$ as $\varepsilon \to 0$ and the limit control problem cannot approximate optimal or near optimal trajectories of the ε-system.

Remark 6.19. In the classic case where the invariant measure $\nu(u,x)$ is concentrated on the point $q(u,x)$ (the reduced system (1.4),(1.5) is well defined) the conditions under which $\mathrm{val}_\varepsilon(\mathcal{U}) \to \mathrm{val}_0(\mathcal{U})$ can be found for instance in Bensoussan [Be] and Gaitsgory [G1, G2]. However, for the general model, to the best of the author's knowledge, the conditions for the validity of this approximation still have not been provided.

THEOREM IV. *Suppose that Assumptions* 3.1–3.5 *and Lemma* 3.6 *are satisfied. Let $v_0(\cdot)$ be the optimal relaxed control for the limit problem. Then*

(i) *if U is convex, then* $\mathrm{val}_\varepsilon(\mathcal{U}^c_\varepsilon) \to \mathrm{val}_0(\mathcal{U}^c) = \mathrm{val}_0(\mathcal{U}^m)$*;*

(ii) *if Assumption* 6.7 *is fulfilled, then* $\mathrm{val}_\varepsilon(\mathcal{U}^p_\varepsilon) \to \mathrm{val}_0(\mathcal{U}^p) = \mathrm{val}_0(\mathcal{U}^m)$*;*

(iii) *if there exists a compact $C \subset \mathit{graph}\, G$ such that* (5.8) *is satisfied for any $u_\varepsilon(\cdot) \in \mathcal{U}^m_\varepsilon$, then* $\mathrm{val}_\varepsilon(\mathcal{U}^m_\varepsilon) \to \mathrm{val}_0(\mathcal{U}^m)$*;*

(iv) *if in* (i)–(iii) *$u_\varepsilon(\cdot)$ statistically converges to $v_0(\cdot)$, then $|J_\varepsilon(u_\varepsilon) - \mathrm{val}_\varepsilon(\mathcal{U}_\varepsilon)| \to 0$ for the corresponding family of relatively slow controls.*

Proof. Theorems I–III with Claim 6.1 imply upper semicontinuity of $\mathrm{val}_\varepsilon(\mathcal{U}_\varepsilon)$ where $\mathcal{U}_\varepsilon$ is a corresponding family of relatively slow controls. Indeed, for $u_\varepsilon(\cdot) \to v_0(\cdot)$ we have that

$$\mathrm{val}_\varepsilon(\mathcal{U}_\varepsilon) \le J_\varepsilon(u_\varepsilon) \quad \text{and} \quad J_\varepsilon(u_\varepsilon) \to \mathrm{val}_0(\mathcal{U}^m).$$

This implies that $\limsup \mathrm{val}_\varepsilon(\mathcal{U}_\varepsilon) \le \mathrm{val}_0(\mathcal{U}^m)$.

Lower semicontinuity follows from the compactness of the family $\mathcal{V}$ (Proposition 4.1). The pair $(x_\varepsilon(t), y_\varepsilon(t))$ for each ε takes value in a compact set. Then by continuity properties of the function Q, the value $\mathrm{val}_\varepsilon(\mathcal{U}_\varepsilon)$ is finite for each ε. Take $u_\varepsilon(\cdot) \in \mathcal{U}_\varepsilon$ such that $J_\varepsilon(u_\varepsilon) - \mathrm{val}_\varepsilon(\mathcal{U}_\varepsilon)$ tends to zero. Take any sequence $\varepsilon_k \to 0$ as $k \to \infty$ and find a subsequence $\varepsilon_{k(j)}$ such that $u_{\varepsilon_{k(j)}}$ tends to some $v(\cdot) \in \mathcal{V}$. Then from Theorems I–III we have that $J_{\varepsilon_{k(j)}}(u_{\varepsilon_{k(j)}}) \to J_0(v) \ge \mathrm{val}_0(\mathcal{U}^m)$ and therefore as the sequence ε_j is arbitrary, we have

$$\liminf \mathrm{val}_\varepsilon(\mathcal{U}_\varepsilon) \ge \mathrm{val}_0(\mathcal{U}^m).$$

Then $\lim \mathrm{val}_\varepsilon(\mathcal{U}_\varepsilon) = \mathrm{val}_0(\mathcal{U}^m)$, and this completes the proof of (i)–(iii). The proof of (iv) follows immediately from Theorems I–III and (i)–(iii). □

7. Proof of Theorems I–III. Let $E \subset \mathit{graph}\, G$ be a compact set. The first part of the proof verifies that the trajectories $x_\varepsilon(t)$ of the solution $(x_\varepsilon(t), y_\varepsilon(t))$ of (1.1), induced by a control function $u_\varepsilon(\cdot)$ and initial conditions $(u_\varepsilon(0), x^0_\varepsilon, y^0_\varepsilon)$ in E, indeed converge to the x-trajectory $x_0(t)$ of (3.1).

Note that $x_\varepsilon(t)$ solves the differential equation

$$\dot{x} = f(u_\varepsilon(t), x, y_\varepsilon(t)), \quad x(0) = x^0_\varepsilon, \tag{7.1}$$

which is a time-varying equation. Our method now is to modify (7.1) and construct a differential equation

$$\dot{x} = f_\varepsilon(t,x) \tag{7.2}$$

which, on the one hand has $x_\varepsilon(t)$ as a solution and on the other hand will converge to the equation (3.2) as $\varepsilon \to 0$ in a sense that guarantees the continuous dependence of solutions. The function $f_\varepsilon(t,x)$ is constructed as follows:

$$\begin{aligned} f_\varepsilon(t,x) = \alpha(\varepsilon,x,t)\int_{\mathbb{R}^n} f(u_\varepsilon(t),x,y)\nu(u_\varepsilon(t),x)(dy) \\ +(1-\alpha(\varepsilon,x,t))f(u_\varepsilon(t),x,y_\varepsilon(t)) \end{aligned} \tag{7.3}$$

with $\alpha(\varepsilon,x,t)$ given by

$$\alpha(\varepsilon,x,t) = \begin{cases} 0, & |x-x_\varepsilon(t)| \le \varepsilon, \\ |x-x_\varepsilon(t)|/\varepsilon - 1, & \varepsilon < |x-x_\varepsilon(t)| \le 2\varepsilon, \\ 1, & |x-x_\varepsilon(t)| > 2\varepsilon. \end{cases} \tag{7.4}$$

Note that $\alpha(\varepsilon,x_\varepsilon(t),t) \equiv 0$ and hence $x_\varepsilon(t)$ solves (7.2) as well.

The function $f_0(t,x)$ is defined as follows:

$$f_0(t,x) = \bar{\boldsymbol{f}}(v_0(t),x), \tag{7.5}$$

where we recall that $\bar{\boldsymbol{f}}$ is defined by (4.2) as

$$\bar{\boldsymbol{f}}(v_0,x) = \int_U \left(\int_{\mathbb{R}^n} f(u,x,y)\,\nu(u,x)(dy)\right) v_0(du).$$

We now explain in what sense we claim that $f_\varepsilon(t,x)$ converges to $f_0(t,x)$.

LEMMA 7.1. *Suppose that $\zeta_\varepsilon(t) : [0,1] \to R^n$ are continuous functions, uniformly converging to $\zeta_0(t)$. Then for every $s \in [0,1]$*

$$\int_0^s f_\varepsilon(t,\zeta_\varepsilon(t))dt \to \int_0^s f_0(t,\zeta_0(t))dt, \tag{7.6}$$

if $\boldsymbol{\rho}(\delta_{y_\varepsilon},\mu_\varepsilon) \to 0$ as $\varepsilon \to 0$. Here $\mu_\varepsilon(t) = \nu(u_\varepsilon(t),x_\varepsilon(t))$.

Proof. We interpret the right-hand side of (7.6) as an integral,

$$\int_0^s \bar{\boldsymbol{f}}(v_0(t),\zeta_0(t))\,dt,$$

which by the definition of $\bar{\boldsymbol{f}}$ is given by

$$\int_0^s \int_U \left(\int_{\mathbb{R}^n} f(u,\zeta_0(t),y)\,\nu(u,\zeta_0(t))(dy)\right) v_0(t)(du)\,dt. \tag{7.7}$$

We interpret the left-hand side of (7.6) as an integral,

$$\int_0^s \left(\int_{\mathbb{R}^n} f(u_\varepsilon(t),\zeta_\varepsilon(t),y)\,\beta_\varepsilon(t)(dy)\right) dt, \tag{7.8}$$

with $\beta_\varepsilon(t) : [0,1] \to \mathcal{P}(\mathbb{R}^n)$ a measure-valued function given by

$$\beta_\varepsilon(t) = \alpha(\varepsilon,\zeta_\varepsilon(t),t)\nu_\varepsilon(t) + (1-\alpha(\varepsilon,\zeta_\varepsilon(t),t))\delta_{y_\varepsilon}(t),$$

where

$$\nu_\varepsilon(t) = \nu(u_\varepsilon(t), \zeta_\varepsilon(t)). \tag{7.9}$$

We also define another function $\tilde{\beta}_\varepsilon(t)$ given by

$$\tilde{\beta}_\varepsilon(t) = \alpha(\varepsilon, \zeta_\varepsilon(t), t)\nu_\varepsilon(t) + (1 - \alpha(\varepsilon, \zeta_\varepsilon(t), t))\mu_\varepsilon(t).$$

By definition (7.4) of the function $\alpha(\varepsilon, x, t)$ and continuity of the function $\nu(u, x)$ it follows that $\boldsymbol{\rho}(\tilde{\beta}_\varepsilon, \nu_\varepsilon) \to 0$. By an assumption of the lemma, $\boldsymbol{\rho}(\delta_{y_\varepsilon}, \mu_\varepsilon) \to 0$. Therefore $\boldsymbol{\rho}(\tilde{\beta}_\varepsilon, \beta_\varepsilon) \to 0$ and then $\boldsymbol{\rho}(\beta_\varepsilon, \nu_\varepsilon) \to 0$. Thus from (7.8) we have that for every $s \in [0, 1]$

$$\left| \int_0^s f_\varepsilon(t, \zeta_\varepsilon(t))\, dt - \int_0^s \left(\int_{\mathbb{R}^n} f(u_\varepsilon(t), \zeta_\varepsilon(t), y)\, \nu_\varepsilon(t)(dy) \right) dt \right| \to 0,$$

and hence we have just to prove that

$$\int_0^s \left(\int_{\mathbb{R}^n} f(u_\varepsilon(t), \zeta_\varepsilon(t), y)\, \nu_\varepsilon(t)(dy) \right) dt \to \int_0^s \bar{\boldsymbol{f}}_0(v_0(t), \zeta_0(t))\, dt, \tag{7.10}$$

where the right-hand side is defined in (7.7) But from (7.9) we can interpret the left-hand side of (7.10) as an integral

$$\int_0^s \bar{\boldsymbol{f}}(v_\varepsilon(t), \zeta_\varepsilon(t))\, dt,$$

where $v_\varepsilon(t) = \delta_{u_\varepsilon}(t)$ is a measure-valued interpretation of the ordinary function $u_\varepsilon(t)$. Since $v_\varepsilon(\cdot)$ statistically converges to $v_0(\cdot)$ and $\zeta_\varepsilon(\cdot)$ uniformly converges to $\zeta_0(\cdot)$ and from the uniform continuity of $\bar{\boldsymbol{f}}(v, x)$ on compact sets we conclude that

$$\int_0^s \bar{\boldsymbol{f}}(v_\varepsilon(t), \zeta_\varepsilon(t))\, dt \to \int_0^s \bar{\boldsymbol{f}}(v_0(t), \zeta_0(t))\, dt.$$

This completes the proof. □

LEMMA 7.2. *Under the assumptions of Theorems* I–III, $\boldsymbol{\rho}(\delta_{y_\varepsilon}, \mu_\varepsilon) \to 0$, *where* $\mu_\varepsilon(t) = \nu(u_\varepsilon(t), x_\varepsilon(t))$.

Proof. For Theorem III the result immediately follows from Lemma 5.4. Now define set C as follows:

$$C = \lambda(\{t : d((u_\varepsilon(t), x_\varepsilon(t), y_\varepsilon(t)), \boldsymbol{K}) \le \eta\}), \tag{7.11}$$

where K is defined in (5.5) and η is chosen such that $\boldsymbol{C} \subset \mathit{graph}\, G$. Then under assumptions of Lemma 5.3 (and hence of Theorem I) conditions of Lemma 5.4 are satisfied. Therefore the result follows for Theorem I also. Thus we have to prove the claim of the lemma only for Theorem II, namely, for the case where relatively slow piecewise continuous controls are employed.

We just have to show that statements of Lemma 5.3 hold true. Indeed, let t_ε^i, $i = 1, k(\varepsilon)$ be the points of discontinuity of the function $u_\varepsilon(\cdot)$. Then $t_\varepsilon^{i+1} - t_\varepsilon^i \ge \Delta_\varepsilon$. On each interval $[t_\varepsilon^i, t_\varepsilon^{i+1})$ the function $u_\varepsilon(\cdot)$ is relatively slow continuous with the common modulus of continuity $M(\Delta_\varepsilon^{-1} z)$, and then on each interval of continuity we can use Lemma 5.3 if we prove that the pairs $(x_\varepsilon(t_\varepsilon^i), y_\varepsilon(t_\varepsilon^i))$ stay in a common compact set and the triples $(u_\varepsilon(t), x_\varepsilon(t), y_\varepsilon(t))$ are included in $\mathit{graph}\, G$. Let $E = U \times H$ be a compact subset of $\mathit{graph}\, G$ and $(x_\varepsilon(0), y_\varepsilon(0)) \in H$. Let X be a compact subset

of $\mathbb{R}^m$ such that if $(u_\varepsilon(t), x_\varepsilon(t), y_\varepsilon(t)) \in \mathit{graph}\, G$ for $t \in [0,s]$ then $x_\varepsilon(t) \in X$ for $t \in [0,s]$. Such a set exists by Assumption 3.3. Define the set $\tilde{L}(x) = \cup_{u\in U} L(u,x)$. By the upper semicontinuity of L the mapping $\tilde{L}(\cdot)$ is compact for each x and upper semicontinuous. Thus by Assumption 6.7 $\mathit{graph}\, G$ contains the compact set

$$\tilde{\boldsymbol{L}} = \{(u,x,y) : u \in U, x \in X, y \in \tilde{L}(x)\}$$

(with, say, an η-neighborhood small enough).

We denote now a new compact set

$$\tilde{E} = E \cup \{(u,x,y) : u \in U, x \in X, d((u,x,y), \tilde{\boldsymbol{L}}) \le \eta\}.$$

Since by the definition the set $\tilde{\boldsymbol{L}}$ contains the set $\mathbf{L}$ defined in (5.5), we can employ estimates (5.6) and (5.7) of Lemma 5.3 to conclude that the triples $(u_\varepsilon(t), x_\varepsilon(t), y_\varepsilon(t))$ are included in $\mathit{graph}\, G$ for $t \in [0,1]$. Moreover for $(u_\varepsilon(t^i_\varepsilon), x_\varepsilon(t^i_\varepsilon), y_\varepsilon(t^i_\varepsilon)) \in \tilde{E}$ the triples $(u_\varepsilon(t^{i+1}_\varepsilon), x_\varepsilon(t^{i+1}_\varepsilon), y_\varepsilon(t^{i+1}_\varepsilon))$ are also included in $\tilde{E}$ for ε small enough. Hence the first inequality in (5.7) can be extended for $t \in [0,1]$, and we obtain the convergence in (5.8) (or (6.4)) if the set C is defined by (7.11). Therefore by Lemma 5.4 the result follows from the conditions of Theorem II too. □

Now we can proceed with the proof of Theorems I–III.

From Lemmas 7.1 and 7.2 it follows by the standard continuous dependence arguments (see, e.g., [AV, Lemma 4.1]) that $x_\varepsilon(t) \to x_0(t)$. Since by Lemma 7.2 $\boldsymbol{\rho}(\delta_{y_\varepsilon}, \mu_\varepsilon) \to 0$ where $\mu_\varepsilon(t) = \nu(u_\varepsilon(t), x_\varepsilon(t))$ and $\nu(u,x)$ is continuous in (u,x), we obtain $\boldsymbol{\rho}(\delta_{y_\varepsilon}, \nu_\varepsilon) \to 0$, where $\nu_\varepsilon(t) = \nu(u_\varepsilon(t), x_0(t))$. By an assumption $u_\varepsilon(\cdot)$ statistically converges to $v_0(\cdot)$, and then for any continuous function $h(t,u)$

$$\int_0^1 h(t, u_\varepsilon(t))\, dt \to \int_0^1 \int_U h(t,y)\, v_0(t)(du)\, dt. \tag{7.12}$$

For an arbitrary continuous function $\tilde{h}(t,y)$ define

$$h(t,u) = \int_{\mathbb{R}^n} \tilde{h}(t,y)\, \nu(u, x_0(t))(dy). \tag{7.13}$$

Then from (7.12) and (7.13) it follows that for any continuous $\tilde{h}(t,y)$

$$\int_0^1 \left(\int_{\mathbb{R}^n} \tilde{h}(t,y) \nu(u_\varepsilon(t), x_0(t))(dy) \right) dt \to \int_0^1 \left(\int_{\mathbb{R}^n} \tilde{h}(t,y)\, \bar{\nu}(v_0(t), x_0(t))(dy) \right) dt.$$

Therefore $\nu_\varepsilon(t)$ statistically converges to $\bar{\nu}(v_0(t), x_0(t))$, and since $\boldsymbol{\rho}(\delta_{y_\varepsilon}, \nu_\varepsilon) \to 0$ we get that $y_\varepsilon(t)$ statistically converges to $\bar{\nu}(v_0(t), x_0(t))$. Since the function $Q(u,x,y)$ is continuous, the convergence of the cost $J_\varepsilon \to J_0$ follows. This completes the proof of Theorems I–III.

8. Concluding remarks. Note that in our consideration U is an abstract metric compact space. This means that U itself may be a space of probability measures—say, on a compact set W of $\mathbb{R}^k$. Then the controls $u(t) \in U$ are themselves relaxed controls. In this case the relaxed controls $v(t) \in \mathcal{P}(U)$ are relaxed with respect to the controls $u(t)$.

Our convergence results may have another interpretation. Suppose we have an original control problem in the limit form (3.1). This type of problem arises for

instance in systems with uncertainty when we make appropriate averaging of the data functions. Then if we can find a function g such that the flow induced by the corresponding differential equation has an invariant measure ν from the model (3.1), we can construct the ε-system (1.1)–(1.3) which approximates (3.1).

We emphasize that our limit model completely approximates only the ε-problem on the families of relatively slow controls. We do not consider the fast controls $u(\tau) = u(t/\varepsilon)$. If, for instance, $u(\tau)$ is a periodic function, then the limit of $\mathrm{val}_\varepsilon(\mathcal{U})$ may be much less than $\mathrm{val}_0(\mathcal{U})$, as was mentioned in Example 6.18. (Consult also Gaitsgory [G1, G2, G3].) However, periodic controls induce the *discrete time* dynamical systems which are Poincaré maps. This gives an opportunity to apply the dynamical system approach to construct an extension of the limit control problem (3.1) when the fast periodic control is allowed.

Acknowledgment. This work is a part of a Ph.D. dissertation carried out under the supervision of Professor Zvi Artstein. The author would like to express his gratitude to Prof. Artstein for very useful discussions of these results.

REFERENCES

[A1] Z. Artstein, *A variational convergence that yields chattering systems*, Ann. Inst. H. Poincaré Anal. Non Linéaire, supp., 6 (1989), pp. 49–71.

[A2] Z. Artstein, *Chattering linear systems: A model of rapidly oscillating coefficients*, Math. Control Signals Systems, 2 (1989), pp. 49–71.

[A3] Z. Artstein, *Chattering variational limits of control systems*, Forum Math., 5 (1993), pp. 369–403.

[A4] Z. Artstein, *Rapid oscillations, chattering systems, and relaxed controls*, SIAM J. Control Optim., 27 (1989), pp. 940–948.

[AV] Z. Artstein and A. Vigodner, *Singularly perturbed ordinary differential equations with dynamic limits*, Proc. Roy. Soc. Edinburgh Sect. A, (1996), pp. 541–569.

[Be1] A. Bensoussan, *Singular perturbations for deterministic control problems*, in Singular Perturbations and Asymptotic Analysis in Control Systems, Lecture Notes in Control and Information Sciences, Springer-Verlag, Berlin, 1987, pp. 9–170

[Be2] A. Bensoussan, *Perturbation Methods in Optimal Control Problems*, John Wiley, New York, 1989.

[BB] A. Bensoussan and G. Blankenship, *Singular perturbations in stochastic control*, in Singular Perturbations and Asymptotic Analysis in Control Systems, Lecture Notes in Control and Inform. Sci., Springer-Verlag, Berlin, 1987, pp. 171–287.

[BS] N. P. Bhatia and G. P. Szegö, *Stability Theory of Dynamical Systems*, Springer-Verlag, Berlin, 1970.

[Bi] P. Billingsley, *Convergence of Probability Measures*, John Wiley, New York, 1968.

[Ca] S. L. Campbell, *Singular Systems of Differential Equations*, Pitman, San Francisco, 1980.

[DV1] A. L. Dontchev and V. M. Veliov, *Singular perturbations in Mayer's problem for linear systems*, SIAM J. Control Optim., 21 (1983), pp. 566–581.

[DV2] A. L. Dontchev and V. M. Veliov, *Singular perturbations in linear control systems with weakly coupled stable and unstable fast subsystems*, J. Math. Anal. Appl., 110 (1985), pp. 1–30.

[G1] V. Gaitsgory, *Systems with Divided Motions: Control and Suboptimization*, Nauka, Moscow, 1991 (in Russian).

[G2] V. Gaitsgory, *Suboptimization of singularly perturbed control systems*, SIAM J. Control Optim., 30 (1992), pp. 1228–1249.

[G3] V. Gaitsgory, *Suboptimal control of singularly perturbed systems and periodic optimization*, IEEE Trans. Automat. Control, 38 (1993), pp. 888–903.

[Gr] G. Grammel, *Averaging of singularly perturbed systems*, Nonlinear Anal., to appear.

[Ha] P. R. Halmos, *Measure Theory*, Van Nostrand Reinhold, New York, 1950.

[KaP] Yu. Kabanov and S. Pergamenshchikov, *Optimal control of singularly perturbed stochastic linear systems*, Stochastics Stochastics Rep., 36 (1991), pp. 109–135.

[Ko1] P. V. Kokotovic, *Applications of singular perturbations techniques to control problems*, SIAM Rev., 26 (1984), pp. 501–550.

[KOS] P. V. Kokotovic, R. E. O'Malley, and P. Sannuti, *Singular perturbations and order reduction in control theory*, Automatica, 12 (1976), pp. 123–132.

[NS] V. V. Nemytskii and V. V. Stepanov, *Qualitative Theory of Differential Equations*, Princeton University Press, Princeton, NJ, 1960.

[O'M] R. E. O'Malley, *Boundary layer methods for certain non-linear singularly perturbed optimal control problems*, J. Math. Anal. Appl., 45 (1974), pp. 468–484.

[O'M2] R. E. O'Malley, *Introduction to Singular Perturbations*, Academic Press, New York, 1974.

[Se] G. R. Sell, *Topological Dynamics and Ordinary Differential Equations*, Van Nostrand Reinhold, London, 1971.

[Ti] A. N. Tichonov, *Systems of differential equations containing small parameters near derivatives*, Mat. Sb., 31 (1952), pp. 575–586.

[Tu] H. D. Tuan, *On reachable set of singularly perturbed differential inclusions and optimal control problems*, Optimization, 26 (1992), pp. 325–338.

[Va] A. B. Vasil'eva, *Asymptotic behavior of solutions to certain problems involving nonlinear differential equations containing a small parameter multiplying the highest derivative*, Russian Math. Surveys, 18 (1963), pp. 15–86.

[VB] A. B. Vasil'eva and V. F. Butuzov, *Asymptotic expansions of solutions to singularly perturbed equations*, Nauka, Moscow, 1973 (in Russian).

[Vi] A. Vigodner, *Limits of singularly perturbed control problems: Dynamical systems approach*, Ph.D. thesis, Weizmann Institute of Sciences, Rehovot, Israel, 1995.

[Vo] V. M. Volosov, *Averaging in systems of differential and functional equations*, Uspekhi Mat. Nauk, 17 (1962), pp. 3–126.

[Wa] J. Warga, *Optimal Control of Differential and Functional Equations*, Academic Press, New York, 1972.

[Was] W. Wasow, *Asymptotic Expansions for Ordinary Differential Equations*, Wiley-Interscience, New York, 1965.

SIAM J. CONTROL OPTIM.
Vol. 35, No. 1, pp. 29–35, January 1997

002

ON POSITIVE ORTHANT CONTROLLABILITY OF BILINEAR SYSTEMS IN SMALL CODIMENSIONS*

YURI L. SACHKOV†

Abstract. For the bilinear control system of the form $\dot{x} = (A + \sum_{i=1}^{m} u_i B_i)x$, $x \in \mathbf{R}^n$, $u_i \in \mathbf{R}$, with A essentially nonnegative and B_i constant diagonal matrices, the following global controllability question is studied: when can any two points in $\mathbf{R}^n$ with positive coordinates be connected by a trajectory of this system? The answers for $m = n - 1$ and $m = n - 2$ for any $n > 2$ are given; some necessary conditions for other cases are proven.

Key words. bilinear system, global controllability, positive orthant

AMS subject classifications. 93B05, 93B27

PII. S0363012994270898

1. Introduction. The bilinear systems proved to be an interesting class of nonlinear control systems, both in theory and for applications. This class is one of the rare ones for which global controllability questions can be answered rather completely. This paper is devoted to one of such questions proposed by Boothby in [1].

Consider the bilinear control system

$$\dot{x} = \left(A + \sum_{i=1}^{m} u_i B_i \right) x, \tag{1}$$

where $x \in \mathbf{R}^n$; $u_1, \dots, u_m$ are the piecewise continuous scalar unbounded inputs; and A, $B_1, \dots, B_m$ are the constant $n \times n$ matrices.

The attainability set for system (1) from a point $x \in \mathbf{R}^n$ for all nonnegative times will be denoted by $\mathbf{A}(x)$.

We will denote *the open positive orthant* $\{x \in \mathbf{R}^n : x > 0\}$ by $\overset{\circ}{\mathbf{R}}{}^n_+$.

System (1) is called *controllable in* $\overset{\circ}{\mathbf{R}}{}^n_+$ if for any $x \in \overset{\circ}{\mathbf{R}}{}^n_+$ we have $\mathbf{A}(x) = \overset{\circ}{\mathbf{R}}{}^n_+$.

In what follows we will suppose that all trajectories of system (1) starting in $\overset{\circ}{\mathbf{R}}{}^n_+$ do not leave $\overset{\circ}{\mathbf{R}}{}^n_+$, i.e.,

(1) the matrix A is essentially nonnegative: $A = (a_{ij})$, $a_{ij} \geq 0$ for all $i \neq j$;

(2) the matrices $B_1, \dots, B_m$ are diagonal: $B_i = \operatorname{diag}(b_i)$, $b_i \in \mathbf{R}^n$, $i = 1, \dots, m$.

We will also suppose that the matrices B_i (or, equivalently, the vectors b_i) are linearly independent: for the linear hull $l = \operatorname{span}(b_1, \dots, b_m)$ we have $\dim l = m$. This can be achieved, if necessary, by eliminating some B_i and decreasing m.

The problem of controllability of system (1) in $\overset{\circ}{\mathbf{R}}{}^n_+$ under the conditions (1) and (2) was studied first by Boothby in [1]; he obtained some results for $m = 1$ and showed that for $m = n$ system (1) is controllable in $\overset{\circ}{\mathbf{R}}{}^n_+$. A complete solution for $m = 1$, $n = 2$ was obtained by Bacciotti in [2]. In [3] it was proven that for $m = 1$, $n > 2$, system (1) is, generically, noncontrollable in $\overset{\circ}{\mathbf{R}}{}^n_+$. So the problem was solved for the extreme codimensions 0 and $n - 1$.

In this paper we propose a solution of the problem for the systems of codimension 1 and 2 (i.e., for $m = n - 1$ and $m = n - 2$) and give some conditions sufficient for noncontrollability for $m \leq n - 2$.

The main idea is natural: if a simply connected state space is stratified into the integral manifolds of codimension one of the fields $B_i x$, then system (1) is globally

*Received by the editors July 11, 1994; accepted for publication (in revised form) September 29, 1995.

http://www.siam.org/journals/sicon/35-1/27089.html

†Program Systems Institute, Pereslavl-Zalessky 152140, Russia (sachkov@sys.botik.ru).

controllable iff all these manifolds are intersected by the field Ax in both directions. The corresponding general result was obtained by Bacciotti and Stefani in [4].

The structure of this paper is as follows. In section 2 we introduce a function determining the direction of intersection of the integral manifolds of the fields $B_i x$ by the field Ax and obtain conditions of change of sign of this function. In section 3 these conditions are applied to obtain the controllability conditions for $m = n-1$. In section 4 we give a test of controllability in $\overset{\circ}{\mathbf{R}}{}^n_+$ in terms of the notion of directional controllability. Finally, in sections 5 and 6 we apply the above results and give the conditions of controllability for other cases.

2. Conditions of change of sign. For every vector $h = (h_1, \dots, h_n) \in \mathbf{R}^n$ we will consider the corresponding function H defined on $\overset{\circ}{\mathbf{R}}{}^n_+$ by the equality

$$H(x) = x_1^{h_1} x_2^{h_2} \cdots x_n^{h_n}.$$

LEMMA 2.1. *Let $B = diag(b)$ be an $n \times n$ diagonal matrix and let $h \in \mathbf{R}^n$. The corresponding function H is an integral of the field Bx iff vectors h and b are orthogonal.*

Here and below we use the scalar product in $\mathbf{R}^n$: $\langle x, y \rangle = \sum_{i=1}^n x_i y_i$.

Proof. $\langle \operatorname{grad} H(x), Bx \rangle = \langle h, b \rangle H(x)$. □

Consider the function

$$\phi(x) = \langle \operatorname{grad} H(x), Ax \rangle / H(x),$$

which determines the direction of intersection of the level surfaces $\{H(x) = C\}$ by the field Ax. In the further part of this section we obtain, in terms of the vector h and matrix A, conditions that

$$\forall C > 0 \quad \text{the function } \phi(x)|_{\{H(x)=C\}} \text{ changes its sign.} \tag{2}$$

These conditions (Theorems 2.2 and 2.3) will be applied in the following sections to obtain the conditions of controllability of system (1) in $\overset{\circ}{\mathbf{R}}{}^n_+$.

Let us recall that an $n \times n$ matrix A is called *permutations irreducible* if the corresponding linear operator A has no k-dimensional invariant coordinate subspaces in $\mathbf{R}^n$ with $0 < k < n$; $A = (a_{ij})$ is called *essentially positive* if $a_{ij} > 0$ for all $i \neq j$.

THEOREM 2.2. *Let $h = (h_1, \dots, h_n) \in \mathbf{R}^n$ and let $\sum_{i=1}^n h_i \neq 0$.*

(1) *If the matrix A is permutations irreducible and the vector h has a pair of components with the mutually opposite signs, then condition* (2) *is satisfied.*

(2) *If $h_i \geq 0$ for all $i = 1, \dots, n$ and $\sum_{i=1}^n h_i a_{ii} \geq 0$, then condition* (2) *is not satisfied.*

Proof. Statement (1). Without loss of generality we can assume that $h_1 > 0$ and $h_n < 0$. H is a homogeneous function of order $\sum_{i=1}^n h_i \neq 0$, and every ray through x with the vertex in the origin intersects all hypersurfaces $\{H(x) = C\}$. So condition (2) holds iff $\phi(x)$ changes its sign in $\overset{\circ}{\mathbf{R}}{}^n_+$.

For $x_2 = x_3 = \cdots = x_n = 1$ and small $x_1 > 0$ the sign of the function $\phi(x) = \sum_{i,j=1}^n a_{ij} h_i x_j / x_i$ is determined by the large positive term $h_1 \sum_{j=2}^n a_{1j} x_j / x_1$. For $x_1 = x_2 = \cdots = x_{n-1} = 1$ and small $x_n > 0$ the function $\phi(x)$ is negative because of the term $h_n \sum_{j=1}^{n-1} a_{nj} x_j / x_n$. That is why $\phi(x)$ changes its sign in $\overset{\circ}{\mathbf{R}}{}^n_+$. Statement (1) is proven.

Statement (2) follows directly from the expansion

$$\phi(x) = \sum_{i \neq j} a_{ij} h_i x_j / x_i + \sum_{i=1}^n a_{ii} h_i. \qquad \square$$

THEOREM 2.3. *Let the matrix A be essentially positive, $h = (h_1, \dots, h_n) \in \mathbf{R}^n$, and $\sum_{i=1}^n h_i = 0$. Then condition (2) holds iff vector h has at least two positive and at least two negative components.*

This theorem follows from Lemmas 2.4–2.6.

LEMMA 2.4. *Let $n = 4$ and let the matrix A be essentially positive; suppose that $\sum_{i=1}^4 h_i = 0$; $h_1 > 0$, $h_2 > 0$, $h_3 < 0$, $h_4 < 0$. Then condition (2) is satisfied.*

Proof. Suppose that there exists $C > 0$ such that $\phi(x)$ does not change its sign on $\{H(x) = C\}$. We can assume that $\phi(x) \geq 0$ on $\{H(x) = C\}$.

In the homogeneous coordinates $u_i = x_i/x_4$, $i = 1, 2, 3$, $u_4 \equiv 1$, we have $H(u) = u_1^{h_1} u_2^{h_2} u_3^{h_3}$, $\phi(u) = \sum_{i,j=1}^4 h_i a_{ij} u_j / u_i$, and $\phi(u) \geq 0$ on $\{H(u) = C\}$.

On the surface $\{H(u) = C\}$ we have $u_3 = C^{1/h_3} u_1^{p_1} u_2^{p_2}$, where

$$p_1 = -h_1/h_3 > 0, \quad p_2 = -h_2/h_3 > 0.$$

(A) First we show that $p_2 \leq 1$.

Introduce the following family of curves parametrized by α:

$$\gamma_\alpha(s) = (s, s^\alpha, C^{1/h_3} s^{p_1 + \alpha p_2}). \tag{3}$$

Note that for any $\alpha \in \mathbf{R}$ and $s > 0$ we have $H(\gamma_\alpha(s)) \equiv C$.

Let $\alpha < 0$. On the curve $\gamma_\alpha(s)$ for $s \to +\infty$ we have $u_1 \to \infty$, $u_2 \to 0$. Then the largest positive terms in $\phi(u)$ are $h_2 a_{21} u_1/u_2$ and $h_2 a_{23} u_3/u_2$. Let us show that if $p_2 > 1$, then the parameter α can be chosen such that these positive terms have absolute value less than the negative term $h_3 a_{31} u_1/u_3 = h_3 a_{31} C^{-1/h_3} s^{1 - p_1 - \alpha p_2}$. Actually, on the curve $\gamma_\alpha(s)$ we have $u_1/u_2 = s^{1-\alpha}$, $u_3/u_2 = C^{1/h_3} s^{p_1 + \alpha p_2 - \alpha}$, and existence of the indicated α follows from compatibility of the system of inequalities

$$\begin{cases} 1 - p_1 - \alpha p_2 > 1 - \alpha \\ 1 - p_1 - \alpha p_2 > p_1 + \alpha p_2 - \alpha \\ \alpha < 0 \end{cases} \quad \underset{(\text{if } p_2 > 1)}{\Longleftrightarrow} \quad \begin{cases} \alpha < p_1/(1 - p_2) \\ \alpha < (2p_1 - 1)/(1 - 2p_2) \\ \alpha < 0. \end{cases}$$

So the indicated α exists, and the contradiction with $(\phi(u)|_{H(u)=C}) \geq 0$ shows that statement (A) is proven.

(B) Then we show, in the same way, that $p_1 \leq 1$.

(C) Then we show that $p_1 + p_2 \leq 1$. We consider family (3) for $\alpha > 0$ and $s \to 0$ and show, analogously to (A), that $p_1 + p_2 > 1$, $p_1 \leq 1$, $p_2 \leq 1$ imply that $\phi(\gamma_\alpha(s)) < 0$ for these α and s.

But inequality $p_1 + p_2 \leq 1$ is equivalent to $h_1 + h_2 + h_3 \leq 0$, which contradicts the conditions $h_1 + h_2 + h_3 + h_4 = 0$ and $h_4 < 0$. This contradiction shows that $\phi(x)$ changes its sign on the surface $\{H(x) = C\}$. □

Then we generalize Lemma 2.4 for any $n > 4$.

LEMMA 2.5. *Let $n > 4$ and let the matrix A be essentially positive; $h = (h_1, \dots, h_n)$ $\in \mathbf{R}^n$, $\sum_{i=1}^n h_i = 0$; $h_1 > 0$, $h_2 > 0$, $h_3 < 0$, $h_4 < 0$. Then condition (2) is satisfied.*

Proof. The general case is reduced to the case $n = 4$ by "freezing" the superfluous coordinates.

In the homogeneous coordinates $u_i = x_i/x_4$, $i \neq 4$, $u_4 \equiv 1$, we consider for any $K > 0$ the plane

$$\Pi_K = \{u = (u_1, \dots, u_n) : u_5 = \dots = u_n = K\}.$$

$H(u)|_{\Pi_K} = u_1^{h_1} u_2^{h_2} u_3^{h_3} K^{h_5 + \dots + h_n}$, and the plane Π_K intersects with the surface $\{H(u) = C\}$ for any $C > 0$, $K > 0$. So to prove this lemma it is sufficient to show that

$$\forall C > 0 \quad \exists K > 0 \text{ such that } \phi(u)|_{\{H(u)=C\} \cap \Pi_K} \text{ changes its sign.} \tag{4}$$

The direct calculations show that

$$\phi(u)|_{\Pi_K} = \sum_{i,j=1}^{4} h_i \tilde{a}_{ij}(K) u_j/u_i,$$

where

$$\tilde{a}_{ij}(K) \equiv a_{ij} \quad \text{for } i,j = 1,2,3,$$
$$\tilde{a}_{i4}(K) = a_{i4} + K\sum_{j=5}^{n} a_{ij} \quad \text{for } i = 1,2,3,$$
$$\tilde{a}_{4j}(K) = a_{4j} + 1/K\sum_{i=5}^{n} h_i/h_4 a_{ij} \quad \text{for } j = 1,2,3,$$
$$\tilde{a}_{44}(K) = a_{44} + K\sum_{j=5}^{n} a_{4j} + 1/K\sum_{i=5}^{n} h_i/h_4 a_{i4} + \sum_{i,j=5}^{n} h_i/h_4 a_{ij}.$$

So the function $\phi(u)|_{\Pi_K}$ coincides with the function of change of sign $\tilde{\phi}_K(x) = \langle \operatorname{grad} \tilde{H}(x), \tilde{A}(K)x\rangle/\tilde{H}(x)$ with $x \in \mathbf{R}^4$, $\tilde{H}(x) = x_1^{h_1}x_2^{h_2}x_3^{h_3}x_4^{h_4}$, $\tilde{A}(K) = (\tilde{a}_{ij}(K))$, $i,j = 1,\dots,4$ (after $\tilde{\phi}_K(x)$ is expressed in the usual homogeneous coordinates u_i). For a sufficiently large K the matrix $\tilde{A}(K)$ is essentially positive, so we can use conditions of change of sign for $n = 4$ obtained before. If $h_1 + h_2 + h_3 + h_4 \neq 0$, then it follows from statement (1) of Theorem 2.2 that $\tilde{\phi}_K(x)|_{H(x)=C}$ changes sign for any $C > 0$. And if $h_1 + h_2 + h_3 + h_4 = 0$, the same result follows from Lemma 2.4. So statement (4) is proven, as is Lemma 2.5. □

Now we will consider the case when h has only one component with the sign opposite to the signs of all other components. (There always exists at least one such component under the assumption $\sum_{i=1}^{n} h_i = 0$, $h \neq 0$.)

LEMMA 2.6. *Let the matrix A be essentially positive, $h \in \mathbf{R}^n$, $\sum_{i=1}^{n} h_i = 0$; $h_1 \geq 0$, $h_2 \geq 0, \dots, h_{n-1} \geq 0$, $h_n < 0$. Then condition (2) is not satisfied.*

Proof. In the coordinates $u_i = x_i/x_n$, $i = 1,\dots,n-1$, $u_n \equiv 1$, we have $H(u) = u_1^{h_1}u_2^{h_2}\dots u_{n-1}^{h_{n-1}}$. It follows from the inequalities $h_1 \geq 0$, $h_2 \geq 0$, …, $h_{n-1} \geq 0$ that

$$\lim_{C\to 0} \sup_{H(u)=C} \min_{h_i>0} \{u_i\} = 0;$$

i.e., for small $C > 0$ at least one of those components u_i of vector u, for which $h_i > 0$, becomes small on the surface $\{H(u) = C\}$. Choose sufficiently small $C > 0$, and let u_k be the small component in some neighborhood of u, $H(u) = C$, with $h_k > 0$.

In expansion $\phi(u) = \sum_{i,j=1}^{n} h_i a_{ij} u_j/u_i$ the negative terms are $h_n \sum_{j=1}^{n-1} a_{nj}u_j$ and, maybe, $\sum_{i=1}^{n} h_i a_{ii}$. But for small u_k the absolute values of all these terms are less than the large positive term $h_k \sum_{j\neq k} a_{kj}u_j/u_k$. Actually, we decompose the negative terms in the form

$$h_n\left(\sum_{j=1}^{k-1} a_{nj}u_j + a_{nk}u_k + \sum_{j=k+1}^{n-1} a_{nj}u_j\right) + \sum_{i=1}^{n} h_i a_{ii}$$

and the positive terms in the form

$$h_k\left(\sum_{j=1}^{k-1} a_{kj}u_j/u_k + \sum_{j=k+1}^{n-1} a_{kj}u_j/u_k + a_{kn}/u_k\right).$$

For sufficiently small u_k we have

$$h_k \sum_{j=1}^{k-1} a_{kj}u_j/u_k > \left| h_n \sum_{j=1}^{k-1} a_{nj}u_j \right|,$$

$$h_k \sum_{j=k+1}^{n-1} a_{kj}u_j/u_k > \left| h_n \sum_{j=k+1}^{n-1} a_{nj}u_j \right|,$$

$$h_k a_{kn}/u_k > |h_n a_{nk} u_k| + \left| \sum_{i=1}^{n} h_i a_{ii} \right|.$$

So the positive terms dominate all negative terms, and ϕ is positive in the neighborhood of the chosen u. Since u is arbitrary, $\phi(u)|_{H(u)=C} > 0$ for small $C > 0$. □

3. Systems of codimension one. In this section we suppose that $m = n - 1$ and obtain conditions of controllability of system (1) in $\overset{\circ}{\mathbf{R}}{}^n_+$ for this case.

There exists a unique (up to a scalar factor) nonzero vector $h \in \mathbf{R}^n$ orthogonal to the hyperplane $l = \operatorname{span}(b_1, \ldots, b_{n-1})$. We fix such vector $h = (h_1, \ldots, h_n)$ and the corresponding function $H(x) = x_1^{h_1} \ldots x_n^{h_n}$.

LEMMA 3.1. *System* (1) *is controllable in* $\overset{\circ}{\mathbf{R}}{}^n_+$ *iff the field* Ax *intersects any level surface* $\{H(x) = C\}$ *in both directions.*

Proof. We use the theory of global controllability of systems with $n - 1$ inputs on a manifold of dimension n, developed by Bacciotti and Stefani in [4]. It follows from Theorem 5.1 in [4] that system (1) is controllable in $\overset{\circ}{\mathbf{R}}{}^n_+$ iff for any $x \in \overset{\circ}{\mathbf{R}}{}^n_+$ the maximal integral manifold of the fields B_1x, …, $B_{n-1}x$ through the point x is intersected by the field Ax in both directions. But the family of fields B_1x, …, $B_{n-1}x$ is involutive (as all Lie brackets $[B_ix, B_jx] = [B_i, B_j]x$ vanish), so this integral manifold has dimension $n - 1$. On the other hand, the function H is an integral of the fields B_1x, …, $B_{n-1}x$ (see Lemma 2.1). The level surfaces of H are connected, so they coincide with the maximal integral manifolds of these fields. □

The direction of intersection of a level surface of H by the field Ax is determined by the sign of the function $\phi(x) = \langle \operatorname{grad} H(x), Ax \rangle / H(x)$, so Lemma 3.1 gives the following theorem.

THEOREM 3.2. *Let* $m = n - 1$, $h \perp l$, $h \neq 0$. *System* (1) *is controllable in* $\overset{\circ}{\mathbf{R}}{}^n_+$ *iff for any* $C > 0$ *the function* $\phi(x)|_{H(x)=C}$ *changes its sign.*

Now we apply the conditions of change of sign of ϕ obtained before (Theorems 2.2 and 2.3) and get the controllability conditions in the following form.

THEOREM 3.3. *Let* $m = n - 1$, $h \perp l$, $\sum_{i=1}^n h_i \neq 0$.

(1) *If the matrix* A *is permutations irreducible and the vector* h *has a pair of components with the mutually opposite signs, then system* (1) *is controllable in* $\overset{\circ}{\mathbf{R}}{}^n_+$.

(2) *If* $h_i \geq 0$ *for all* $i = 1, \ldots, n$ *and* $\sum_{i=1}^n h_i a_{ii} \geq 0$, *then system* (1) *is not controllable in* $\overset{\circ}{\mathbf{R}}{}^n_+$.

THEOREM 3.4. *Let* $m = n - 1$, $h \perp l$, $h \neq 0$, $\sum_{i=1}^n h_i = 0$. *Suppose that the matrix* A *is essentially positive. System* (1) *is controllable in* $\overset{\circ}{\mathbf{R}}{}^n_+$ *iff the vector* h *has at least two positive and two negative components.*

4. Directional controllability. In this section we apply the notion of directional controllability of system (1) and obtain controllability test in $\overset{\circ}{\mathbf{R}}{}^n_+$.

System (1) is *directionally controllable in* $\overset{\circ}{\mathbf{R}}{}^n_+$ if for any $x, y \in \overset{\circ}{\mathbf{R}}{}^n_+$ there exists $r \in \mathbf{R}_+$ such that $ry \in \mathbf{A}(x)$.

We will say that a point $x \in \mathbf{R}^n \setminus \{0\}$ determines *the direction entering (leaving) origin* for system (1) if $rx \in \mathbf{A}(x)$ for all $r \in (0;1)$ (respectively, $rx \in \mathbf{A}(x)$ for all $r \in (1;+\infty)$).

THEOREM 4.1. *System* (1) *is controllable in* $\overset{\circ}{\mathbf{R}}{}^n_+$ *iff*

(1) *it is directionally controllable in* $\overset{\circ}{\mathbf{R}}{}^n_+$*;*

(2) *there exist the vectors in* $\overset{\circ}{\mathbf{R}}{}^n_+$ *determining a direction entering the origin and a direction leaving the origin.*

Proof. Necessity is obvious.

Sufficiency. Let x determine an entering direction and y determine a leaving direction; $x, y \in \overset{\circ}{\mathbf{R}}{}^n_+$. Then we can move along the interval $\{rx : r \in (0;1)\}$ toward origin and along the ray $\{ry : r \in (1;+\infty)\}$ away from origin. But directional controllability of system (1) in $\overset{\circ}{\mathbf{R}}{}^n_+$ means that $\mathbf{A}(z)$ meets every ray of the form $\{rw : r \in \mathbf{R}_+\}$ in $\overset{\circ}{\mathbf{R}}{}^n_+$. So for any $z \in \overset{\circ}{\mathbf{R}}{}^n_+$ the attainable set $\mathbf{A}(z)$ is invariant under homotheties with respect to the origin, and that is why $\mathbf{A}(z) = \overset{\circ}{\mathbf{R}}{}^n_+$. □

5. Systems of codimension two. In this section we consider the case $m = n-2$. We use the controllability test from the previous section and reduce this case to the case of codimension one.

Let the $(n-2)$-dimensional plane $l = \operatorname{span}(b_1, \ldots, b_{n-2})$ not contain the vector $e = (1, 1, \ldots, 1)$. Then we fix a nonzero vector $h = (h_1, \ldots, h_n) \in \mathbf{R}^n$, orthogonal to the hyperplane $\operatorname{span}(e, l)$, and the corresponding functions $H(x) = x_1^{h_1} x_2^{h_2} \cdots x_n^{h_n}$ and $\phi(x) = \langle \operatorname{grad} H(x), Ax \rangle / H(x)$.

Note that $\sum_{i=1}^n h_i = 0$ for the chosen vector h.

LEMMA 5.1. *Let the matrix A be essentially positive. System* (1) *is directionally controllable in* $\overset{\circ}{\mathbf{R}}{}^n_+$ *iff the vector h has at least two positive and at least two negative components.*

Proof. Consider the auxiliary system

$$\dot{x} = \left(A + \sum_{i=1}^{m} u_i B_i + u_{m+1} E \right) x, \tag{5}$$

where A, B_i, u_i, $i = 1, \ldots, m$, are the same as in system (1), E is the identity $n \times n$ matrix, and u_{m+1} is an unbounded scalar input. It may easily be seen that system (1) is directionally controllable in $\overset{\circ}{\mathbf{R}}{}^n_+$ iff system (5) is controllable in $\overset{\circ}{\mathbf{R}}{}^n_+$. But system (5) has codimension 1, and we can apply Theorem 3.4 to obtain the conditions of controllability of system (5) $\overset{\circ}{\mathbf{R}}{}^n_+$. □

LEMMA 5.2. *Let the matrix A be permutations irreducible. If there exists a vector $b \in l \cap \overset{\circ}{\mathbf{R}}{}^n_+$ with pairwise distinct components, then system* (1) *has a direction entering origin and a direction leaving origin in* $\overset{\circ}{\mathbf{R}}{}^n_+$.

Proof. We apply Proposition 4.3 of [1] to the matrix $B = \operatorname{diag}(b)$ and obtain that for sufficiently large u (respectively, sufficiently negative u) all eigenvalues of the matrix $A + uB$ are positive (respectively, all negative). But the matrix $A + uB$ is essentially nonnegative and permutations irreducible. So we apply Frobenius's theorem on spectral properties of the nonnegative irreducible matrices [5] and the argument of Boothby from section 4 of [1] and obtain that the corresponding eigenvectors x with positive eigenvalue and y with negative eigenvalue belong to $\overset{\circ}{\mathbf{R}}{}^n_+$. But then x and y determine the direction leaving the origin and the direction entering the origin in $\overset{\circ}{\mathbf{R}}{}^n_+$, respectively. □

Now we apply Theorem 4.1 and Lemmas 5.1 and 5.2 and obtain the following controllability conditions for the systems of codimension 2.

THEOREM 5.3. *Let $m = n-2$, $e \notin l$, $h \perp span\,(e,l)$. Let the following conditions additionally hold:*

(1) *the vector h has at least two positive and at least two negative components;*
(2) *the matrix A is essentially positive;*
(3) *there exists a vector $b \in l \cap \overset{\circ}{\mathbf{R}}{}^n_+$ with pairwise distinct components.*

Then system (1) *is controllable in $\overset{\circ}{\mathbf{R}}{}^n_+$.*

THEOREM 5.4. *Let $m = n-2$, the matrix A be essentially positive, $e \notin l$, $h \perp span\,(e,l)$. Suppose that for some $i = 1,\dots,n$ we have $h_i > 0$ and $h_j \le 0$ for all $j \ne i$. Then system* (1) *is not controllable in $\overset{\circ}{\mathbf{R}}{}^n_+$.*

6. Systems of arbitrary codimensions. For the systems not covered by controllability conditions of sections 3 and 5 we can give some conditions sufficient for noncontrollability (i.e., in fact, necessary for controllability) due to the following simple consideration: if system (1) can be complemented to a system

$$\dot{x} = \left(A + \sum_{i=1}^{m} u_i B_i + \sum_{i=m+1}^{m+k} u_i B_i \right) x$$

for some $k > 0$ in such a way that the above system is noncontrollable in $\overset{\circ}{\mathbf{R}}{}^n_+$, then the initial system (1) is noncontrollable in $\overset{\circ}{\mathbf{R}}{}^n_+$ too.

THEOREM 6.1. *Let $m < n-1$, and let there exist a vector $h \in \mathbf{R}^n$, $h \perp l$, such that $h_i \ge 0$ for all $i = 1,\dots,n$ and $\sum_{i=1}^n a_{ii} h_i \ge 0$. Then system* (1) *is not controllable in $\overset{\circ}{\mathbf{R}}{}^n_+$.*

Proof. Let L be the hyperplane in $\mathbf{R}^n$ orthogonal to the vector h. We have $L \supset l = \operatorname{span}(b_1,\dots,b_m)$, so we can choose vectors $b_{m+1}, \dots, b_{n-1}$ complementing $b_1, \dots, b_m$ to a basis of L. Let us introduce the diagonal matrices $B_i = \operatorname{diag}(b_i)$ for $i = m+1,\dots,n-1$. Then the system $\dot{x} = (A + \sum_{i=1}^{n-1} u_i B_i)x$ has codimension one and is not controllable by statement 2. of Theorem 3.3. That is why system (1) is not controllable too. □

THEOREM 6.2. *Let $m < n-2$, the matrix A be essentially positive, and there exists a vector $h \in \mathbf{R}^n$, $h \perp span\,(e,l)$, such that for some $i = 1,\dots,n$ we have $h_i > 0$ and $h_j \le 0$ for $j \ne i$. Then system* (1) *is not controllable in $\overset{\circ}{\mathbf{R}}{}^n_+$.*

Proof. With the help of the same argument as in Theorem 6.1 we complement system (1) to a system of codimension two and obtain noncontrollability by Theorem 5.4. □

Acknowledgments. The author thanks Professor A. F. Filippov for his kind attention to this work. The author is also grateful to the anonymous referees for their helpful suggestions on correction of some statements and language of the paper.

REFERENCES

[1] W. M. BOOTHBY, *Some comments on positive orthant controllability of bilinear systems*, SIAM J. Control Optim., 20 (1982), pp. 634–644.
[2] A. BACCIOTTI, *On the positive orthant controllability of two-dimensional bilinear systems*, Systems Control Lett., 3 (1983), pp. 53–55.
[3] Y. L. SACHKOV, *Positive orthant controllability of single-input bilinear systems*, Mat. Zametki, 85 (1995), pp. 419–424 (in Russian, translated into English).
[4] A. BACCIOTTI AND G. STEFANI, *On the relationship between global and local controllability*, Math. Systems Theory, 16 (1983), pp. 79–91.
[5] F. R. GANTMACHER, *The Theory of Matrices*, vol. II, Chelsea, New York, 1959.

SIAM J. CONTROL OPTIM.
Vol. 35, No. 1, pp. 36–55, January 1997

003

LYAPUNOV EXPONENTS FOR FINITE STATE NONLINEAR FILTERING*

RAMI ATAR† AND OFER ZEITOUNI†

Abstract. Consider the Wonham optimal filtering problem for a finite state ergodic Markov process in both discrete and continuous time, and let σ be the noise intensity for the observation. We examine the sensitivity of the solution with respect to the filter's initial conditions in terms of the gap between the first two Lyapunov exponents of the Zakai equation for the unnormalized conditional probability. This gap is studied in the limit as $\sigma \to 0$ by techniques involving considerations of nonlinear filtering and the stochastic Feynman–Kac formula. Conditions are given for the limit to be either negative or $-\infty$. Asymptotic bounds are derived in the latter case.

Key words. Lyapunov exponents, nonlinear filtering, Wonham's equation, Feynman–Kac formula

AMS subject classifications. 93E11, 60J57

PII. S0363012994272046

1. Introduction and statement of results.

Let $\{X_n\}_{n=0}^{\infty}$ denote a finite state space, discrete time homogeneous Markov chain, with transition matrix G and initial distribution p_0. Without loss of generality, we take the state space of the Markov chain to consist of the set $\{1,\dots,d\}$. Denote the law of the chain X_n on $\{1,\dots,d\}^{\mathbb{Z}}$ by P. Throughout this paper, we assume that G leads to an ergodic noncyclic chain. That is, we assume

(A1) there exists a $k \geq 1$ such that $G^k(i,j) > 0$ for all $i,j \in \{1,\dots,d\}$.

We denote by E_s expectations under the unique stationary measure of $\{X_n\}$.

We assume that the Markov chain X_n is observed through the sequence $\{Y_n\}_{n=1}^{\infty}$, where

$$Y_n = \delta h_{X_n} + \sqrt{\delta}\sigma\nu_n\,.$$

Here, $h : \{1,\dots,d\} \to \mathbb{R}$ is the observation function, $\delta > 0$ is a parameter (which, for as long as one deals only with discrete time, may be taken as $\delta = 1$), σ is an observation noise parameter related to the signal-to-noise ratio (SNR), and $\{\nu_n\}_{n=1}^{\infty}$ is a sequence of i.i.d., standard Gaussian random variables.

Let $\mathcal{Y}_n$ denote the σ-algebra generated by the observations $Y_1,\dots,Y_n$. The nonlinear filtering problem consists of computing the conditional law $p_j(n) = P(X_n = j|\mathcal{Y}_n)$. Let D_n denote the diagonal matrix with $D_n(i,i) = \exp[\sigma^{-2}(h_i Y_n - h_i^2\delta/2)]$, and define

$$\rho(n) = D_n G^* \rho(n-1)\,, \tag{1}$$

where G^* denotes the transpose of G, and $\rho(0) = p_0$. It is a straightforward consequence from Bayes's rule (see, e.g., [1, p. 460] and also the continuous time case in [9]) that the vector $p(n) = (p_1(n),\dots,p_d(n))^*$ satisfies $p(n) = \rho(n)/\langle\rho(n),\mathbf{1}\rangle$, where

*Received by the editors July 29, 1994; accepted for publication (in revised form) October 2, 1995. The research of the second author was partially supported by the fund for basic research administered by the Israeli academy of science and by the fund for promotion of research at the Technion.
http://www.siam.org/journals/sicon/35-1/27204.html

†Department of Electrical Engineering, Technion-Israel Institute of Technology, Haifa 32000, Israel (rami@aluf.technion.ac.il, zeitouni@ee.technion.ac.il).

$\rho(n) = (\rho_1(n), \ldots, \rho_d(n))^*$, $\mathbf{1} = (1, \ldots, 1)^*$, and $\langle \cdot, \cdot \rangle$ denotes the standard inner product in $\mathbb{R}^d$.

Often, one has no access to the initial distribution p_0. A common procedure is then to initialize (1) with some initial condition $q_0 \in S^{d-1}$, where S^{d-1} denotes the $(d-1)$-dimensional simplex. Denote by $\rho^{q_0}(n)$ the solution to (1) initialized this way, and denote by $p^{q_0}(n)$ the corresponding normalized (random) probability vector. Natural questions are then, how far is $p^{q_0}(n)$ from $p^{p_0}(n)$, what are the conditions for stability in the sense that $\|p^{q_0}(n) - p^{p_0}(n)\| \to_{n\to\infty} 0$, and under these conditions what is the rate of convergence? We emphasize that we deal here with the dependence of the optimal filter on *its* initial conditions and not with its dependence on perturbations of the initial distribution of the state process $\{X_n\}$. The latter is a different problem which we do not deal with here.

Motivated by the approach taken in [4] (see [6] for a related computation in the continuous time, linear case), we couch the question in terms of Lyapunov exponents. That is, for any two $q_0 \neq q_0' \in S^{d-1}$, define

$$\gamma_\sigma^\delta(q_0, q_0', \omega) = \limsup_{n\to\infty} \frac{1}{n\delta} \log ||p^{q_0}(n) - p^{q_0'}(n)|| \,.$$

Although here and in what follows we take $||\cdot||$ to denote the Euclidean norm, note that the definition does not depend on the precise norm used and, in particular, one could use the variation (ℓ^1) norm here.

We will see that, under mild conditions, $\gamma_\sigma^\delta(q_0, q_0', \omega)$ is almost surely deterministic and $\gamma_\sigma^\delta = \gamma_\sigma^\delta(q_0, q_0', \omega)$ is independent of q_0, q_0' for a.e. q_0, q_0' (when q_0, q_0' are distributed uniformly over the simplex) and is related to the gap between the top two Lyapunov exponents associated with the Zakai equation for the unnormalized conditional probability. The deterministic quantity $-1/\gamma_\sigma^\delta$ can then be interpreted as the "memory length" of the filter. Obviously, this approach is meaningful only if $\gamma_\sigma^\delta < 0$. We will identify below sufficient conditions for this to happen. An analogous continuous time question is examined as well.

We remark that in order to deal with the filter's memory length, we introduce and use tools borrowed from the theory of products of random matrices. Especially, we formulate the (qualitative) question of stability and the (quantitative) question of memory length in terms of Lyapunov exponents of the solution of Zakai's equation. While the question of computing Lyapunov exponents is, in general, difficult, we study the above-mentioned gap in the limiting cases, i.e., the regimes $\sigma \to \infty$ and $\sigma \to 0$. Under appropriate conditions, we obtain the exact order of the memory length as a function of σ in the latter case.

A natural guess is that γ_σ^δ becomes more negative as the SNR increases (i.e., as $\sigma \to 0$). As pointed out in [4] for the continuous time setup, this is not always the case, and one may even have situations where $\lim_{\sigma\to 0} \gamma_\sigma^\delta = 0$ though $\gamma_\sigma^\delta < 0$ for all positive σ. We identify below conditions for the memory length $-1/\gamma_\sigma^\delta$ to remain bounded as a function of σ and conditions for it to decay to zero as $\sigma \to 0$.

The structure of the paper is as follows. In the rest of this section, we describe the results for the memory length in both discrete and continuous time. In particular, in both cases we provide the uniform bounds on γ_σ^δ alluded to above and determine under appropriate conditions the limits of γ_σ^δ under both high and low SNR. Sections 2 and 3, respectively, are devoted to proofs of the discrete and continuous time results.

We begin with the following rather straightforward consequence of Oseledec's theorem (see [2, p. 181] and [3]).

THEOREM 1.1. *Assume* (A1). *Then there exists a deterministic function of* σ *and* δ, γ_σ^δ *which admits the following:*

(1) *Let* q_0, q_0' *be random, uniformly distributed (U) on the simplex* S^{d-1}, *independent of each other, and of the chain* $X_0, \{X_n, Y_n\}_{n=1}^\infty$. *Then*

$$\gamma_\sigma^\delta(q_0, q_0', \omega) = \gamma_\sigma^\delta\,, \quad U \times U \times P - a.s.$$

(2) *For any deterministic* $q_0 \neq q_0'$ *with all entries strictly positive, one has*

$$\gamma_\sigma^\delta(q_0, q_0', \omega) \leq \gamma_\sigma^\delta\,, \quad P - a.s.$$

As is seen in section 2, γ_σ^δ is just δ^{-1} times the difference between the two top Lyapunov exponents of solutions of (1).

We turn to study γ_σ^δ quantitatively. First is a bound which is uniform with respect to σ.

THEOREM 1.2. *Assume that all entries of* G *are strictly positive. Then*

$$\gamma_\sigma^\delta \leq \frac{c}{\delta} < 0$$

for some constant c *independent of* h, σ, δ.

Remark. Actually, one may somewhat relax the condition that all entries of G are positive and still have the conclusion of the theorem. See Theorem 2.3 in section 2 for such a statement and its proof there (which also serves as a proof of Theorem 1.2) for the explicit dependence of c on the matrix G.

While the above bound relies on the nature of the law of $\{X_n\}$ and *its* mixing properties, the next bound relies on the quality of the observation. In fact, it is shown that under a condition on h, the decay rate tends to infinity as the noise parameter tends to zero. The condition required on h is that it possesses one coordinate which differs from the rest (h one to one suffices). For each $i \in \{1, \ldots, d\}$, define the set

$$\mathrm{nbr}(i) = \{j \neq i : |h_i - h_j| = \min_{k \neq i} |h_i - h_k|\}$$

and define $h_{\mathrm{nbr}(i)} = h_j$, where j is one of the members in the set $\mathrm{nbr}(i)$.

THEOREM 1.3. *Assume* (A1). *Then*

$$\limsup_{\sigma \to 0} \sigma^2 \gamma_\sigma^\delta \leq -\frac{1}{2} E_s [h_{X_1} - h_{\mathrm{nbr}(X_1)}]^2\ . \tag{2}$$

If, in addition, $\det(G) \neq 0$, *then*

$$\liminf_{\sigma \to 0} \sigma^2 \gamma_\sigma^\delta \geq -\frac{1}{2} E_s \sum_{i=1}^d [h_{X_1} - h_i]^2\ . \tag{3}$$

Note that while the gap between the upper and the lower bounds increases with the dimension d (and is nonzero as soon as $d > 2$), one may conclude from Theorem 1.3 that $\gamma_\sigma^\delta = \Omega(\sigma^{-2})$ as soon as there exists an i such that the set $\{j : h_j = h_i\}$ consists of a single point. The memory length is thus of the order of σ^2.

In continuous time, the behavior at low SNR ($\sigma \to \infty$) is completely determined by the top, nonzero eigenvalue of G (see [4]). An analogous result is shown here to hold for the discrete time case.

Let τ be the Birkhoff contraction coefficient (see (11) for a definition).

THEOREM 1.4. *Assume* (A1). *Then*

$$\limsup_{\sigma\to\infty}\gamma_\sigma^\delta \le \inf_{m\ge 1}\frac{1}{m\delta}\log\tau(P^m) < 0.$$

In continuous time we prove results analogous to Theorems 1.1, 1.2, and 1.3. Though the statements are similar, the proofs are harder and involve different techniques; in particular, a naive discretization approach fails. Let $\{x_t\}$ denote a Markov chain, with state space $\{1,\dots,d\}$ and transition matrix $\hat{G}$. We assume that $\hat{G}$ leads to an ergodic chain; that is,

(A2) for every $\delta > 0$, $(\exp(\hat{G}\delta))(i,j) > 0$ for all $i,j \in \{1,\dots,d\}$.

The above holds iff all states are communicating. Next, assume that $\{x_t\}$ is observed via

$$dy_t = h_{x_t}dt + \sigma d\nu_t,$$

where ν_t is a standard Wiener process independent of $\{x_t\}$ and h is as in the discrete time case. Let H denote the diagonal matrix with elements $H(i,i) = h_i$; then the Zakai equation for the problem is

$$d\rho_t = \hat{G}^*\rho_t dt + \sigma^{-2}H\rho_t dy_t \tag{4}$$

with $p_t = \rho_t/\langle\rho_t,\mathbf{1}\rangle$. Now define for every $q_0 \ne q_0' \in S^{d-1}$

$$\gamma_\sigma(q_0,q_0',\omega) = \limsup_{t\to\infty}\frac{1}{t}\log\|p_t^{q_0} - p_t^{q_0'}\|;$$

then a result similar to Theorem 1.1 holds.

THEOREM 1.5. *Assume* (A2). *Then there exists a deterministic function of* σ, γ_σ, *which admits the following:*

(1) *Let* q_0, q_0' *be random, uniformly distributed (U) on the simplex* S^{d-1}, *independent of each other and of the chain* $\{x_t,y_t\}_{t=0}^\infty$. *Then*

$$\gamma_\sigma(q_0,q_0',\omega) = \gamma_\sigma\,, \quad U\times U\times P - a.s.$$

(2) *For any deterministic* $q_0 \ne q_0'$, *one has*

$$\gamma_\sigma(q_0,q_0',\omega) \le \gamma_\sigma\,, \quad P - a.s.$$

A result analogous to Theorem 1.2 holds also.

THEOREM 1.6. *Assume* (A2). *Then*

$$\gamma_\sigma \le -2\min_{1\le i,j\le d\,:\,i\ne j}(g_{ij}g_{ji})^{1/2},$$

where $g_{ij} = \hat{G}(i,j)$.

Remark. In [4] it is already proven that $\gamma_\sigma < 0$ under certain conditions, although not uniformly in σ.

Finally, a result analogous to Theorem 1.3 holds.

THEOREM 1.7. *Assume* (A2). *Then*

$$\limsup_{\sigma\to 0}\sigma^2\gamma_\sigma \le -\frac{1}{2}E_s[h_{x_0} - h_{\mathrm{nbr}(x_0)}]^2\,. \tag{5}$$

Moreover,

$$\liminf_{\sigma\to 0}\sigma^2\gamma_\sigma \ge -\frac{1}{2}E_s\sum_{i=1}^d[h_{x_0} - h_i]^2\,. \tag{6}$$

2. Proofs—discrete time. Throughout, we let $T_n = D_n G^*$ and $M_n^\sigma = T_n \cdots T_1$. We denote by $a \wedge b$ the exterior product of two vectors in $\mathbb{R}^d$ and by $A \wedge B$ the exterior product of two subspaces of $\mathbb{R}^d$ (see [2] for definitions of exterior products). For a $d \times d$ matrix A, $||A||$ denotes the operator norm (with respect to the Euclidean norm on $\mathbb{R}^d$). Finally, we use c throughout to denote a constant whose value may change from line to line which is independent of n, σ, δ.

Proof of Theorem 1.1. Note first that it is enough to prove the theorem in the case in which X_0 is distributed according to the stationary distribution of $\{X_n\}$. Indeed, due to (A1), the stationary distribution has all entries strictly positive, and thus all almost sure statements, once proved for X_0 distributed according to the stationary law, must translate to the case where $X_0 = j$ for any $j = 1, \ldots, d$. The case of general initial distributions follows immediately.

We may thus assume that X_0 is distributed according to its stationary law. In that case, the sequence of matrices $\{D_n G^*\}_{n=1}^\infty$ possesses a stationary law, which is also ergodic by (A1). Moreover,

$$E \log^+ ||D_n G^*|| \le cE \max_{i=1}^d \sigma^{-2} \left(Y(n) h_i - \frac{1}{2} h_i^2 \delta \right)^+ < \infty .$$

Hence, we may apply Oseledec's theorem (see, e.g., [2, p. 181]) to conclude that there exists a (random) strict subspace $S_\omega^1 \subset \mathbb{R}^d$ such that if $q_0 \notin S_\omega^1$ then

$$\frac{1}{n} \log ||\rho^{q_0}(n)|| \to \lambda_1^\sigma , \quad P - \text{a.s.} \tag{7}$$

Here and in what follows, λ_i^σ denotes the ith (nonrandom) Lyapunov exponent associated with the product of matrices M_n^σ. As is well known (see [3]), the matrix sequence $((M_n^\sigma)^* M_n^\sigma)^{1/2n}$ has a (random) limit a.s., the eigenvalues of which are $e^{\lambda_i^\sigma}$. Note that $(M_n^\sigma)^* M_n^\sigma$ is a nonnegative matrix, thus by the Perron–Frobenius theorem the eigenvector associated with the highest eigenvalue of $(M_n^\sigma)^* M_n^\sigma$ has all coordinates real and nonnegative. The last property thus holds for $((M_n^\sigma)^* M_n^\sigma)^{1/2n}$, too, and hence also for $\lim_{n\to\infty} (M_n^* M_n)^{1/2n}$. Since S_ω^1 must be orthogonal to the eigenvector associated with the highest eigenvalue of $\lim_{n\to\infty} (M_n^* M_n)^{1/2n}$, it follows that S_ω^1 cannot include any probability vector with all entries strictly positive. As for the case where q_0 does not have all its entries strictly positive, notice that $p^{q_0}(k)$ does (where k is such that $G^k(i,j) > 0$ for all $i, j \in \{1, \ldots, d\}$). Thus (7) really holds for any $q_0 \in S^{d-1}$.

Using again Oseledec's theorem, this time for the $\mathbb{R}^d \wedge \mathbb{R}^d$-valued process $\rho^{q_0}(n) \wedge \rho^{q_0'}(n)$, there exists a (random) strict subspace $S_\omega^2 \subset \mathbb{R}^d \wedge \mathbb{R}^d$ such that if $q_0 \wedge q_0' \notin S_\omega^2$ then

$$\frac{1}{n} \log ||\rho^{q_0}(n) \wedge \rho^{q_0'}(n)|| \to_{n\to\infty} \lambda_1^\sigma + \lambda_2^\sigma , \quad P - \text{a.s.} \tag{8}$$

Furthermore, for $q_0 \wedge q_0' \in S_\omega^2$, Oseledec's theorem implies

$$\limsup_{n\to\infty} \frac{1}{n} \log ||\rho^{q_0}(n) \wedge \rho^{q_0'}(n)|| \le \lambda_1^\sigma + \lambda_2^\sigma , \quad P - \text{a.s.} \tag{9}$$

Next, note that there exists a dimensional constant c_d such that if a, b are two probability vectors in S^{d-1} then

$$\frac{1}{c_d} |\sin(a,b)| \le ||a - b|| \le c_d |\sin(a,b)| ,$$

where (a,b) denotes the angle between the vectors a,b. Since for any two nonzero vectors c,d (not necessarily normalized) one has that $|\sin(c,d)| = ||c \wedge d||/(||c|| \cdot ||d||)$, one may conclude that

$$\limsup_{n\to\infty} \frac{1}{n} \log ||p^{q_0}(n) - p^{q_0'}(n)||$$

$$= \limsup_{n\to\infty} \frac{1}{n} [\log ||\rho^{q_0}(n) \wedge \rho^{q_0'}(n)|| - \log ||\rho^{q_0}(n)|| - \log ||\rho^{q_0'}(n)||] .$$

Combining this and the fact that (7) holds for any probability vectors q_0, q_0' with either (8) or (9) yields both parts of the theorem, with $\gamma_\sigma^\delta = \delta^{-1}(\lambda_2^\sigma - \lambda_1^\sigma)$. □

It is useful to state the last sentence of the proof of Theorem 1.1 as the following.

COROLLARY 2.1.

$$\gamma_\sigma^\delta = \delta^{-1}(\lambda_2^\sigma - \lambda_1^\sigma) . \tag{10}$$

As is clear from [4] (and is evident also in the course of the proof of Theorem 1.1), the gap between the first and the second Lyapunov exponents will play a crucial role in our study of the stability of the nonlinear filter. Before providing the proof of Theorem 1.2, it is useful to recall some definitions and a result of Peres concerning this gap. We follow the notations of [7], [8].

We say that a matrix A possessing nonnegative entries is *allowable* if it contains no columns or rows whose entries are all zero. Let S_+^{d-1} denote those elements of S^{d-1} whose entries are all strictly positive. *Hilbert's projective metric* is the metric $\bar{h}(\cdot,\cdot)$ on $S_+^{d-1} \times S_+^{d-1}$ defined by

$$\bar{h}(x,y) = \log \max_{1\le i,j\le d} \frac{x_i y_j}{x_j y_i} .$$

Every allowable matrix A can be seen, by normalization of the linear action of A, as an operator $A : S_+^{d-1} \to S_+^{d-1}$. We denote by $A.x$ its action on $x \in S_+^{d-1}$. Define now the *Birkhoff contraction coefficient* of an allowable matrix A by

$$\tau(A) = \sup \left\{ \frac{\bar{h}(A.x, A.y)}{\bar{h}(x,y)} \,\middle|\, x,y \in S_+^{d-1}, x \neq y \right\} . \tag{11}$$

LEMMA 2.2 (see Peres [7]). *Let $\{T_n\}_{n\ge1}$ be an ergodic stationary sequence of nonnegative, allowable matrices, such that $E\log^+||T_1|| < \infty$. Let λ_1, λ_2 denote the top two Lyapunov exponents for the random product of the T_i. Then*

$$\lambda_1 - \lambda_2 \ge -E \log \tau(T_1) ,$$

where $\lambda_2 = -\infty$ if the right-hand side is infinite.

Proof. See [7, Prop. 5]. □

We recall from [7] and [8] the following useful properties of the contraction coefficient $\tau(\cdot)$:

Property 1. $\tau(AD) = \tau(DA) = \tau(A)$ for any diagonal matrix D with strictly positive diagonal terms.

Property 2. For any matrix A with strictly positive entries, $\tau(A) < 1$.

Property 3. Let A be allowable and define

$$\psi(A) = \min_{i,j,k,l} \left\{ \frac{a_{ik}a_{jl}}{a_{il}a_{jk}} \middle| a_{il}a_{jk} \neq 0 \right\} . \tag{12}$$

Then

$$\tau(A) = \frac{1-\sqrt{\psi(A)}}{1+\sqrt{\psi(A)}}\,. \tag{13}$$

We are now in a position to state the extension of Theorem 1.2 alluded to in the introduction.

THEOREM 2.3. *Assume that $\tau(G) < 1$. Then*

$$\gamma_\sigma^\delta \le \frac{\log \tau(G)}{\delta} < 0\,.$$

Note that Theorem 1.2 follows at once from Theorem 2.3 by using Property 2 for $\tau(G)$. Moreover, it follows that c may be taken as $c = \log(1-\Psi)/(1+\Psi)$ with $\Psi = \min_{i,j} G_{ij}/\max_{i,j} G_{ij}$.

Proof of Theorem 2.3. Applying Theorem 1.1 and Corollary 2.1 in combination with Lemma 2.2 to the recursion (1), one sees that

$$\gamma_\sigma^\delta \le \delta^{-1} E \log \tau(D_1 G^*) = \delta^{-1} \log \tau(G^*) = \delta^{-1} \log \tau(G) < 0\,,$$

where the first equality follows from Property 1 for $\tau(\cdot)$, the second from Property 3, and the last inequality from the assumption $\tau(G) < 1$. □

Proof of Theorem 1.3. Suppose equation (1) is given two initial conditions q_0, q_0'; denote

$$q_n = p_n^{q_0}, \quad q_n' = p_n^{q_0'}, \quad r_n = q_n - q_n'.$$

Now, $q_n = T_n q_{n-1}/\langle T_n q_{n-1}, \mathbf{1}\rangle$, and subtracting $\langle T_n q_{n-1}', \mathbf{1}\rangle q_n' = T_n q_{n-1}'$ from $\langle T_n q_{n-1}, \mathbf{1}\rangle q_n = T_n q_{n-1}$ one gets

$$\langle T_n q_{n-1}, \mathbf{1}\rangle r_n + \langle T_n r_{n-1}, \mathbf{1}\rangle q_n' = T_n r_{n-1}.$$

Denoting $a_n = \langle T_n q_{n-1}, \mathbf{1}\rangle$ and noticing $a_n > 0$ one obtains

$$\begin{aligned} r_n &= a_n^{-1} T_n r_{n-1} - a_n^{-1} q_n' \langle T_n r_{n-1}, \mathbf{1}\rangle = a_n^{-1}(I - q_n' \mathbf{1}^*) T_n r_{n-1} \\ &= a_n^{-1}(I - q_n' \mathbf{1}^*) D_n G^* r_{n-1}. \end{aligned}$$

The following recursion for r_n then holds:

$$\begin{aligned} r_0 &= q_0 - q_0', \\ r_n &= a_n^{-1} T_n' G^* r_{n-1}, \end{aligned} \tag{14}$$

where we denote

$$T_n' = (I - q_n' \mathbf{1}^*) D_n.$$

In order to estimate the growth rate of r_n one notices

$$\frac{1}{n}\log\|r_n\| \le \frac{1}{n}\sum_{i=1}^n \log a_i^{-1} + \frac{1}{n}\sum_{i=1}^n \log\|T_i'\| + \frac{1}{n}\sum_{i=1}^n \log\|G^*\| + \frac{1}{n}\log\|r_0\|. \tag{15}$$

Since the third term is bounded by zero and the fourth tends to zero, we turn to bound the two first terms. The first term tends for any q_0, a.s., to $-\lambda_1^\sigma$, since

$$\frac{1}{n}\sum_{i=1}^{n}\log a_i^{-1} = -\frac{1}{n}\log\langle T_n T_{n-1}\cdots T_1 q_0, \mathbf{1}\rangle$$

(cf. the discussion following (7) above). Hence, it suffices to compute the limit of the last quantity for $q_0 = p_0$. Denoting the density of $(Y_1,\dots,Y_n)$ by $f_{Y_1^n}(\beta_1^n)$ and the distribution of $(X_1,\dots,X_n)$ by $P((X_1,\dots,X_n)=(\alpha_1,\dots,\alpha_n)) = p_{X_1^n}(\alpha_1^n)$ it follows from Bayes' rule that

$$\begin{aligned}
\frac{1}{n}\sum_{i=1}^{n}\log a_i^{-1} &= -\frac{1}{n}\log\left[f_{Y_1^n}(Y_1^n)(2\pi\sigma^2\delta)^{n/2}\exp\frac{1}{2\sigma^2\delta}\sum_i Y_i^2\right]\\
&= -\frac{1}{n}\log\left[\sum_{\alpha_1^n} p_{X_1^n}(\alpha_1^n)(2\pi\sigma^2\delta)^{-n/2}\exp-\frac{1}{2\sigma^2\delta}\sum_i (Y_i - h_{\alpha_i}\delta)^2\right]\\
&\quad -\frac{1}{2}\log 2\pi\sigma^2\delta - \frac{1}{2n\sigma^2\delta}\sum_i Y_i^2\\
&\le -\frac{1}{n}\log\left[p_{X_1^n}(X_1^n)\exp-\sum_i\frac{1}{2\sigma^2}(\sigma\nu_i)^2\right] - \frac{1}{2\sigma^2\delta}\frac{1}{n}\sum_i Y_i^2\\
&= -\frac{1}{n}\log p_{X_1^n}(X_1^n) - \frac{1}{2\sigma^2}\frac{1}{n}\sum_i[h_{X_i}^2\delta + 2\sqrt{\delta}\sigma\nu_i h_{X_i}]. \qquad (16)
\end{aligned}$$

Next we turn to the second term in the right-hand side of (15). Writing the diagonal terms of D_n as $\Delta_n^i = D_n(i,i)$ we have the following expression for T_n':

$$T_n' = \begin{pmatrix}
\Delta_n^1(1-q_n'^1) & \Delta_n^2(-q_n'^1) & \cdots & \Delta_n^d(-q_n'^1)\\
\Delta_n^1(-q_n'^2) & \Delta_n^2(1-q_n'^2) & \cdots & \Delta_n^d(-q_n'^2)\\
\vdots & \vdots & \ddots & \vdots\\
\Delta_n^1(-q_n'^d) & \Delta_n^2(-q_n'^d) & \cdots & \Delta_n^d(1-q_n'^d)
\end{pmatrix}.$$

It is useful to consider here the operator norm of $T_n' : \ell^1 \to \ell^1$, namely, $\|T_n'\|_1 = \max_k \sum_i |(T_n')_{ik}|$. Fix n and suppose $X_n = j$; then $Y_n = h_j\delta + \sigma\nu_n\sqrt{\delta}$, and

$$\|T_n'\|_1 = \max_i\left\{\Delta_n^i\left[1 - q_n'^i + \sum_{l\ne i} q_n'^l\right]\right\} = 2\max_i \Delta_n^i(1-q_n'^i).$$

Denoting the vector $b_n = (b_n^1,\dots,b_n^d)^* := G^* q_{n-1}'$ it follows that

$$q_n'^j = \frac{b_n^j\Delta_n^j}{\sum_{l=1}^d b_n^l\Delta_n^l},$$

and thus

$$1 - q_n'^j = \frac{\sum_{k\ne j} b_n^k\Delta_n^k}{\sum_{l=1}^d b_n^l\Delta_n^l} \le \min\left(1, \frac{\max_{k\ne j}\Delta_n^k}{b_n^j\Delta_n^j}\right) \le 1_{\{b_n^j<\alpha\}} + 1_{\{b_n^j\ge\alpha\}}\frac{\max_{k\ne j}\Delta_n^k}{\alpha\Delta_n^j}$$

for every fixed $0 < \alpha < 1$. Therefore,

$$2\Delta_n^j(1-q_n'^j) \le 2\Delta_n^j 1_{\{b_n^j<\alpha\}} + \frac{2}{\alpha} 1_{\{b_n^j \ge \alpha\}} \max_{k\neq j} \Delta_n^k$$

and, clearly,

$$\forall i, i \neq j \quad 2\Delta_n^i(1-q_n'^i) \le \frac{2}{\alpha} \max_{k\neq j} \Delta_n^k.$$

Using the equivalence of the norms $||\cdot||$ and $||\cdot||_1$, it follows that there exists a constant c, independent of n such that

$$\|T_n'\| \le \frac{2c}{\alpha}\left[1_{\{b_n^j<\alpha\}} \max_k \Delta_n^k + 1_{\{b_n^j\ge\alpha\}} \max_{k\neq j} \Delta_n^k\right]$$

and thus, defining $h_{\max} = \max_i\{|h_i|\}$,

$$\begin{aligned}
&\frac{1}{n}\sum_{i=1}^n \log\|T_i'\| \\
&\le \log\frac{2c}{\alpha} + \frac{1}{n}\sum_{i=1}^n 1_{\{b_i^{X_i}<\alpha\}}\, \sigma^{-2} \max_k \left(h_k h_{X_i}\delta + h_k\sigma\sqrt{\delta}\nu_i - \frac{1}{2}\, h_k^2\delta\right) \\
&\quad + \frac{1}{n}\sum_{i=1}^n 1_{\{b_i^{X_i}\ge\alpha\}} \sigma^{-2} \max_{k\neq X_i} \left(h_k h_{X_i}\delta + h_k\sigma\sqrt{\delta}\nu_i - \frac{1}{2}\, h_k^2\delta\right) \\
&\le \log\frac{2c}{\alpha} + \frac{1}{n}\sum_{i=1}^n 1_{\{b_i^{X_i}<\alpha\}} (\sigma^{-2}\delta {h_{\max}}^2 + \sigma^{-1}\sqrt{\delta} h_{\max}|\nu_i|) \\
&\quad + \frac{1}{n}\sum_{i=1}^n \sigma^{-2}\delta\left(h_{\mathrm{nbr}(X_i)} h_{X_i} - \frac{1}{2}\, h_{\mathrm{nbr}(X_i)}^2\right) + \sigma^{-1}\sqrt{\delta} h_{\max}|\nu_i|.
\end{aligned}$$

Now, $b_i^{X_i} \ge (G)_{X_{i-1}X_i} q_{i-1}'^{X_{i-1}}$ so choosing $\alpha = \frac{1}{2}\min_{u,v:(G)_{uv}>0}(G)_{uv}$ it follows that $1_{\{b_i^{X_i}<\alpha\}} \le 1_{\{q_{i-1}^{X_{i-1}}<\frac{1}{2}\}}$. Combining this with inequalities (15) and (16) one arrives, after taking expectation and limit, at

$$\begin{aligned}
\lim_{n\to\infty}\frac{1}{n} E\,\log\|r_n\| \le c_1 + c_2\sqrt{\delta}/\sigma - \frac{\delta}{2\sigma^2} E_s\left(h_{X_i} - h_{\mathrm{nbr}(X_i)}\right)^2 \\
+ c_3\sigma^{-2}\lim_{n\to\infty}\frac{1}{n}\sum_{i=1}^n P\left(q_{i-1}^{X_{i-1}} < \frac{1}{2}\right).
\end{aligned}$$

Since $P\left(q_n'^{X_n} < \frac{1}{2}\right) \longrightarrow_{\sigma\to 0} 0$ uniformly in n and since again by Oseledec's theorem (see, e.g., [2, p. 181])

$$\lim\frac{1}{n}\log\|\wedge^r T_n\cdots T_1\| = \lim\frac{1}{n} E\,\log\|\wedge^r T_n\cdots T_1\| \quad \text{a.s.},$$

the first part of the theorem is proved.

The second part easily follows from the following facts. First, the spectrum of the matrix process certainly satisfies

$$\lambda_2^\sigma - \lambda_1^\sigma \ge \lambda_2^\sigma - \lambda_1^\sigma + 2\lambda_1^\sigma + \lambda_3^\sigma + \cdots + \lambda_d^\sigma - d\lambda_1^\sigma = \sum_{i=1}^d \lambda_i^\sigma - d\lambda_1^\sigma.$$

Second, since $\det T_n \cdots T_1 = \det T_n \cdots \det T_1$, the sum of the exponents can be explicitly expressed as

$$\sum_{i=1}^{d} \lambda_i^\sigma = \lim_{n\to\infty} \frac{1}{n} \log |\det T_n \cdots T_1| = E_s \frac{1}{\sigma^2} \left(Y_1 \sum_{i=1}^{d} h_i - \frac{1}{2} \sum_{i=1}^{d} h_i^2 \delta \right) + \log |\det G|$$
$$= \frac{\delta}{\sigma^2} E_s \left(h_{X_1} \sum_{i=1}^{d} h_i - \frac{1}{2} \sum_{i=1}^{d} h_i^2 \right) + \log |\det G| ,$$

while

$$\lambda_1^\sigma \le E \log \| \operatorname{diag}(\Delta_1^i)_{i=1}^d \|_1 + \log \|G\|_1$$
$$\le E \max_i \left[-\frac{\delta}{2\sigma^2} (h_{X_1} - h_i)^2 + \frac{\delta h_{X_1}^2}{2\sigma^2} + h_i \frac{\sqrt{\delta}\nu_1}{\sigma} \right] \le \frac{\delta}{2\sigma^2} E_s\, h_{X_1}^2 + \frac{c\sqrt{\delta}}{\sigma}.$$

Thus we conclude that

$$\liminf_{\sigma\to 0} \sigma^2 \gamma_\sigma^\delta \ge E_s \left[h_{X_1} \sum_{i=1}^{d} h_i - \frac{1}{2} \sum_{i=1}^{d} h_i^2 - \frac{d}{2} h_{X_1}^2 \right] = E_s \left[-\frac{1}{2} \sum_i (h_{X_1} - h_i)^2 \right]. \qquad \square$$

Proof of Theorem 1.4. The last inequality holds, since by the assumption there exists an m_0 such that for all $m \ge m_0$, $G^{m_0} > 0$. As for the first inequality, as in the proof of Theorem 1.1, it suffices to work under the assumption that X_0 is distributed according to the stationary distribution. One may apply Lemma 2.2 for the process of matrices that are derived from $\{T_n\}$ by taking products of blocks at length m, where $m \ge m_0$:

$$T_m T_{m-1} \cdots T_1, \quad T_{2m} T_{2m-1} \cdots T_{m+1}, \quad \ldots.$$

Ergodicity, stationarity, and integrability follow from those of $\{T_n\}$. Since $(G^*)^m$ is positive, that is,

$$\sum_{i_2,\ldots,i_{m-1}} (G^*)_{i_1 i_2} \cdots (G^*)_{i_{m-1} i_m} > 0,$$

it follows that

$$(T_m \cdots T_1)_{i_1 i_m} = \sum_{i_2,\ldots,i_{m-1}} \Delta_i^{i_1} (G^*)_{i_1 i_2} \cdots \Delta_m^{i_m} (G^*)_{i_{m-1} i_m} > 0$$

and allowability follows. The Lyapunov spectrum for this sequence is $\{m\lambda_i^\sigma\}_{i=1}^d$, thus

$$\gamma_\sigma^\delta \le \frac{1}{m\delta} E \log \tau(T_m T_{m-1} \cdots T_1). \tag{17}$$

The diagonal terms Δ_j^i for which $T_j = \operatorname{diag}(\Delta_j^i)_{i=1}^d G^*$ may be expressed as

$$\Delta_j^i = \exp \sigma^{-2} \left(h_i Y_j - \frac{1}{2} h_i^2 \delta \right) = \exp \left[\delta \sigma^{-2} \left(h_i h_{X_j} - \frac{1}{2} h_i^2 \right) + \sqrt{\delta} \sigma^{-1} h_i \nu_j \right] = 1 + \alpha_j^i.$$

Thus

$$T_m T_{m-1} \cdots T_1 = (I + \operatorname{diag}(\alpha_m^i)_{i=1}^d) G^* \cdots (I + \operatorname{diag}(\alpha_1^i)_{i=1}^d) G^* = (G^*)^m + M,$$

where M is a matrix satisfying

$$\|M\| \le \|G\|^m[(1+\|\operatorname{diag}(\alpha_m^i)_{i=1}^d\|)\cdots(1+\|\operatorname{diag}(\alpha_1^i)_{i=1}^d\|) \ -1].$$

Now,

$$\|\operatorname{diag}(\alpha_j^i)_{i=1}^d\| = \max_i \left|\exp\left[\delta\sigma^{-2}\left(h_i h_{X_j} - \frac{1}{2}\,h_i^2\right) + \sqrt{\delta}\sigma^{-1}h_i\nu_j\right] - 1\right| \longrightarrow_{\sigma\to\infty} 0 \quad \text{a.s.;}$$

therefore

$$T_m T_{m-1}\cdots T_1 \longrightarrow_{\sigma\to\infty} (G^*)^m \quad \text{a.s.}$$

Since $(G^*)^m$ is positive, ψ is continuous at $(G^*)^m$ and so is τ; hence,

$$\tau(T_m T_{m-1}\cdots T_1) \longrightarrow_{\sigma\to\infty} \tau((G^*)^m) = \tau(G^m) \quad \text{a.s.}$$

Since $\log(\tau(\cdot)) \le 0$, Fatou's lemma may be applied to get

$$\limsup_{\sigma\to\infty} E\log\tau(T_m T_{m-1}\cdots T_1) \le E\ \limsup_{\sigma\to\infty}\log\tau(T_m T_{m-1}\cdots T_1)$$

$$= E\ \log\tau(G^m) = \log\tau(G^m),$$

and the result follows from inequality (17). □

3. Proofs—continuous time. Throughout this section, c denotes a t-independent deterministic constant (whose value may change from line to line). p^x_{stat} denotes the (unique, by (A2)) stationary law corresponding to $\hat{G}$. We use the notations x_0^t and y_0^t to denote the sub σ-fields generated, respectively, by $\{x_s, 0 \le s \le t\}$ and $\{y_s, 0 \le s \le t\}$. E_0 denotes expectations under the product measure $\mathcal{P}_x \times \mathcal{P}_y$, where $\mathcal{P}_x$ denotes the law of the Markov chain x (under the stationary measure) and $\mathcal{P}_y$ denotes the law of the observation process $\{y_t, 0 \le t < \infty\}$.

Proof of Theorem 1.5. Aside from the conditions needed to meet the assumptions of Oseledec's theorem that are proved below, the proof is identical to that of Theorem 1.1. Notice that equation (4) is bilinear, thus there exists a multiplicative process denoted $U = \{U_t\}_{t\in\mathbb{R}_+}$ such that $\rho_t = U_t\rho_0$. Assuming x_0 is distributed according to its stationary law, the shift transformation θ_t is measure preserving with respect to $\{x, \nu\}$ and thus with respect to U. Ergodicity of U follows from that of $\{x, \nu\}$, and separability follows from continuity. It follows from Theorem 2.1 in [5] that $U_t : \mathbb{R}^d \to \mathbb{R}^d$ is a homeomorphism and thus is invertible for any $t \ge 0$. For Oseledec's theorem to hold, one needs to show also the integrability of (see [2, p. 181])

$$u_1 = \sup_{0\le t\le 1} \log^+\|U_t\|, \quad u_2 = \sup_{0\le t\le 1}\log^+\|U_t^{-1}\|.$$

To show u_1 is integrable, it suffices to show that $\sup_{0\le t\le 1}\|U_t\|$ is. Note that by the Kallianpur–Striebel formula, U_t is a nonnegative matrix for any $t \ge 0$, and hence the unit vector w maximizing $\|U_t w\|$ has nonnegative entries. Thus, it suffices, by considering the projection on w, to show integrability of $\sup_{0\le t\le 1}\|U_t v\|$ for some v, all of whose entries are positive. Under (A2) all entries of p^x_{stat} are positive, so v may be chosen to be p^x_{stat}. Since this vector is also the initial distribution of x, it follows that $\|U_t p^x_{\text{stat}}\|_1 = \langle\rho_t, \mathbf{1}\rangle$. By the Kallianpur–Striebel formula,

$$\langle\rho_t, \mathbf{1}\rangle = E_0\left[\exp\int_0^t\left(h_{x_s}dy_s - \frac{1}{2}\,h_{x_s}^2 ds\right)\Big| y_0^t\right] \le E_0\left[\exp h_{\max}\sum_i|\Delta y_i|\right],$$

where $\Delta y_i = y_{\tau_{i+1}} - y_{\tau_i}, \tau_0 = 0, \tau_i = \min\{t > \tau_{i-1} : x_t \neq x_{\tau_{i-1}}\} \wedge 1$, and integrability follows from the existence of exponential moments of the normal distribution and the exponential law of $\tau_i - \tau_{i-1}$. As for u_2, denote, for a symmetric matrix A, by $\lambda_i(A)$ the ith largest eigenvalue of A, then

$$\begin{aligned}\log^+ \|U_t^{-1}\| \leq \|U_t^{-1}\| &= [\lambda_1(U_t^{-1*}U_t^{-1})]^{1/2} = [\lambda_d(U_tU_t^*)]^{-1/2} \\ &\leq \frac{[\lambda_1(U_tU_t^*)]^{(d-1)/2}}{(|\det U_tU_t^*|)^{1/2}} \leq \frac{\|(U_tU_t^*)\|^{(d-1)/2}}{|\det U_t|} \leq \frac{\|U_t\|^{d-1}}{|\det U_t|}. \end{aligned} \tag{18}$$

Now, U_t solves the Stratonovich equation

$$dU_t = \left(\hat{G}^* - \frac{1}{2}\,\sigma^{-2}H^2\right)U_t dt + \sigma^{-2}HU_t \circ dy_t\,;$$

thus

$$|\det U_t| = \exp\left[\int_0^t \operatorname{trace}\left(\hat{G}^* - \frac{1}{2}\,\sigma^{-2}H^2\right)ds + \int_0^t \operatorname{trace}(\sigma^{-2}H)\circ dy_s\right],$$

and combining this with inequality (18) and the Cauchy–Schwartz inequality, the integrability of u_2 follows. □

A corollary analogous to Corollary 2.1 follows.

COROLLARY 3.1. *Let $\lambda_1^\sigma \geq \lambda_2^\sigma \geq \cdots \geq \lambda_d^\sigma$ denote the Lyapunov exponents associated with the multiplicative process U_t. Then,*

$$\gamma_\sigma = \lambda_2^\sigma - \lambda_1^\sigma\ . \tag{19}$$

Proof of Theorem 1.6. By Oseledec's theorem,

$$\lim_t \frac{1}{t}\log\|\wedge^r U_t\| = \lambda_1^\sigma + \cdots + \lambda_r^\sigma \quad \text{a.s.}$$

This limit equals the limit on the discrete time sequence $\{n\delta\}$ for some $\delta > 0$, so if one looks at the sequence of linear operators $\{\hat{A}_n^\delta\}$ for which

$$\rho^{p_0}(n\delta) = \hat{A}_n^\delta \rho^{p_0}((n-1)\delta)\,, \quad \rho^{p_0}(0) = p_0\,, \tag{20}$$

then

$$\lim_n \frac{1}{n\delta}\log\|\wedge^r \hat{A}_n^\delta \cdots \hat{A}_1^\delta\| = \lambda_1^\sigma + \cdots + \lambda_r^\sigma \quad \text{a.s.}$$

Stationarity, ergodicity, and integrability of $\{\hat{A}_n^\delta\}$ follow from those of the continuous time process, so the assumptions of Oseledec's theorem hold, and there exists a Lyapunov spectrum for (20) denoted $\{\lambda_i^{\sigma,\delta}\}_{i=1}^d$. The relation between the spectra is $\frac{1}{\delta}\lambda_i^{\sigma,\delta} = \lambda_i^\sigma$ and thus by (19),

$$\gamma_\sigma = \frac{1}{\delta}(\lambda_2^{\sigma,\delta} - \lambda_1^{\sigma,\delta}).$$

It is useful to consider here the well-known representation of the solution to the Zakai equation as

$$\rho^{p_0}(t) = L_t f_t,$$

where

$$L_t = \operatorname{diag}\left\{\exp \sigma^{-2}\left(h_i y_t - \frac{1}{2}\, h_i^2 t\right)\right\}_{i=1}^{d}$$

and $f : \mathbb{R}^+ \mapsto \mathbb{R}^d$ is the solution of

$$\begin{cases} \dot{f}_t = L_t^{-1}\hat{G}^* L_t f_t, \\ f_0 = p_0. \end{cases}$$

Denote by $\{A_i^\delta\}$ the matrices for which $f_{n\delta} = A_n^\delta f_{(n-1)\delta}$. Now, by Property 1, $\tau(\hat{A}_1^\delta) = \tau(L_\delta A_1^\delta) = \tau(A_1^\delta)$. As U_t is a homeomorphism, $\hat{A}_1^\delta$ is invertible. Therefore, A_1^δ is invertible and thus also allowable. One therefore has by Lemma 2.2 and Fatou's lemma

$$\gamma_\sigma \le \limsup_{\delta\to 0} \frac{1}{\delta} E \log \tau(A_1^\delta) \le E \limsup_{\delta\to 0} \frac{1}{\delta} \log \tau(A_1^\delta). \tag{21}$$

Since f_t belongs to $C^1[0,\infty)$, it follows that

$$f_\delta = (I + L_0^{-1}\hat{G}^* L_0 \delta + M^\delta) f_0,$$

where M^δ is a $d \times d$ matrix with $\|M^\delta\| = o(\delta)$. It suffices to prove the theorem for $\hat{G}$, for which $\forall i, j,\ i \ne j,\ g_{ij} > 0$. Under this condition, ψ may be expressed as

$$\psi(A_1^\delta) = \psi(I + \hat{G}^*\delta + M^\delta) = \min_{1\le i,j,k,l\le d} \frac{(1_{\{i=j\}} + g_{ji}\delta + m_{ij}^\delta)(1_{\{l=k\}} + g_{kl}\delta + m_{lk}^\delta)}{(1_{\{i=k\}} + g_{ki}\delta + m_{ik}^\delta)(1_{\{l=j\}} + g_{jl}\delta + m_{lj}^\delta)},$$

where $m_{ij}^\delta = (M^\delta)_{ij}$. There exists a δ_0 such that for every $0 < \delta < \delta_0$, the minimum is achieved on $i = k \ne l = j$; thus

$$\psi(A_1^\delta) = \min_{i,j:i\ne j} g_{ij} g_{ji}\delta^2 + o(\delta^2),$$
$$\psi^{1/2}(A_1^\delta) = \min_{i,j:i\ne j} (g_{ij} g_{ji})^{1/2}\delta + o(\delta),$$

and

$$\frac{1}{\delta}\log \tau(A_1^\delta) = \frac{1}{\delta}\log\frac{1-\psi^{1/2}(A_1^\delta)}{1+\psi^{1/2}(A_1^\delta)} = \frac{1}{\delta}\left[-2\min_{i,j:i\ne j}(g_{ij}g_{ji})^{1/2}\delta + o(\delta)\right]$$
$$\longrightarrow_{\delta\to 0} -2\min_{i,j:i\ne j}(g_{ij}g_{ji})^{1/2},$$

and the result follows from inequality (21). □

Proof of Theorem 1.7. It seems natural to approach the continuous time case as a limit of the discrete time problem. Note, however, that a change in the order of limits is needed to carry out this approach, and justifying this change of order seems challenging. We thus take below a different route. Although the general idea is similar to the discrete time case, extra care is needed due to the fact that the trajectories of the x process do not possess positive probability, and an appropriate version of the Feynman–Kac formula is needed.

The first part of the theorem is a direct consequence of the following three lemmas, whose proof is deferred.

LEMMA 3.2. *Assume* (A2) *holds. Then* $\limsup_{\sigma\to 0} \sigma^2 \lambda_1^\sigma \le \frac{1}{2} E h_{x_0}^2$.

LEMMA 3.3. *Assume* (A2) *holds. Then*

$$\limsup_{\sigma\to 0}\sigma^2(\lambda_1^\sigma+\lambda_2^\sigma)\le \frac{1}{2}\,Eh_{x_0}^2+Eh_{x_0}h_{\mathrm{nbr}(x_0)}-\frac{1}{2}\,Eh_{\mathrm{nbr}(x_0)}^2.$$

LEMMA 3.4. *Assume* (A2) *holds. Then* $\liminf_{\sigma\to 0}\sigma^2\lambda_1^\sigma\ge \frac{1}{2}\,Eh_{x_0}^2$.

Given Lemma 3.2 above, the proof of (6) is similar to the proof of (3) in the discrete time setup, with trace $\hat{G}$ playing the role of $\log|\det G|$ there. □

Proof of Lemma 3.2. Using (A2), and denoting by e_i the unit vectors in $\mathbb{R}^d$, it holds that $\cos(e_i,p^x_{\mathrm{stat}})\ge c>0$ for some c independent of t. Therefore, since U_t is nonnegative, and using c_1 to denote another positive deterministic constant independent of t, $\parallel U_tp^x_{\mathrm{stat}}\parallel\ge \min_i\cos(e_i,p^x_{\mathrm{stat}})\max_i\parallel U_te_i\parallel\ge c_1\parallel U_t\parallel$ and

$$\lambda_1^\sigma=\lim_{t\to\infty}\frac{1}{t}E\log\parallel U_t\parallel\le\lim_{t\to\infty}\frac{1}{t}E\log\parallel U_tp^x_{\mathrm{stat}}\parallel\ .$$

Let $\tilde{x}$ denote the realization of the x process under E_0, initialized at the stationary measure. Then, by the Kallianpur–Striebel formula and Oseledec's theorem,

$$\lambda_1^\sigma\le\lim_{t\to\infty}\frac{1}{t}E\log E_0\left[\exp\left(\sigma^{-2}\left(\int_0^t h(\tilde{x}_s)dy_s-\frac{1}{2}\,h^2(\tilde{x}_s)ds\right)\right)\,\middle|\,y_0^t\right]. \tag{22}$$

Fix $\delta>0$ and define $\Delta_iy=y_{(i+1)\delta}-y_{i\delta}$, $|\Delta y|^{i,\delta}_{\max}=\max\{|y_t-y_{t'}|\,:\,t,t'\in[i\delta,(i+1)\delta)\}$. Let $\{\tau_i\}$ be the jumping times of $\{\tilde{x}_t\}$, and let $|\Delta y|^{\tau_i,\delta}_{\max}=\max\{|y_t-y_{t'}|\,:\,t,t'\in[\tau_i-\delta,\tau_i+\delta)\}$. Define similarly $\Delta_i\nu$, $|\Delta\nu|^{i,\delta}_{\max}$, $|\Delta\nu|^{\tau_i,\delta}_{\max}$ and let $h_{\max}=\max_i|h_i|$. Let $i_t=[t/\delta]$ and $N_t=\max\{i:\tau_i\le t\}=\#\{\tau_i\le t\}$. We control the integral in (22) by its discrete time skeleton, with errors occurring only around jump times. That is,

$$\begin{aligned}
&\int_0^t\left(h(\tilde{x}_s)dy_s-\frac{1}{2}\,h^2(\tilde{x}_s)ds\right)\\
&\le\sum_{i=0}^{i_t}\left(h(\tilde{x}_{i\delta})\Delta_iy-\frac{1}{2}\,h^2(\tilde{x}_{i\delta})\delta\right)+\sum_{i=1}^{N_t}(2h_{\max}|\Delta y|^{\tau_i,\delta}_{\max}+{h_{\max}}^2\delta)\\
&\le\frac{1}{2}\sum_{i=0}^{i_t}\frac{(\Delta_iy)^2}{\delta}+\sum_{i=1}^{N_t}\left[2h_{\max}\left(2h_{\max}\delta+\sigma|\Delta\nu|^{\tau_i,\delta}_{\max}\right)+h_m^2\delta\right]\ .
\end{aligned} \tag{23}$$

Thus, by Jensen's inequality,

$$\begin{aligned}
&\frac{1}{t}E\log E_0\left[\exp\left(\sigma^{-2}\int_0^t h(\tilde{x}_s)dy_s-\frac{1}{2}\,h^2(\tilde{x}_s)ds\right)\,\middle|\,y_0^t\right]\\
&\le\frac{1}{t}\frac{1}{2\sigma^2\delta}E\sum_{i=0}^{i_t}(\Delta_iy)^2+\frac{1}{t}\log E_0\left[\exp\left(\sigma^{-2}\sum_{i=1}^{N_t}(5{h_{\max}}^2\delta+2h_{\max}\sigma|\Delta\nu|^{\tau_i,\delta}_{\max})\right)\right].
\end{aligned} \tag{24}$$

On the other hand, using stationarity,

$$\begin{aligned}
E(\Delta_iy)^2&=E\left(\int_0^\delta h(x_s)ds+\sigma\nu_\delta\right)^2=E\left(\int_0^\delta h(x_s)ds\right)^2+\sigma^2\delta\\
&\le E_sh^2(x_0)\delta^2+\sigma^2\delta+E\left[1_{\{x_t\ \text{jumps in}\ [0,\delta)\}}(2h_{\max}\delta)^2\right]\\
&=E_sh^2(x_0)\delta^2+\sigma^2\delta+\delta^2C'_\delta\,,
\end{aligned} \tag{25}$$

with $C'_\delta\underset{\delta\to 0}{\longrightarrow}0$.

Conditioning on N_t, one has

$$\begin{aligned}
&\frac{1}{t}\log E_0 \exp\left(\sigma^{-2}\sum_{i=1}^{N_t}\left(5{h_{\max}}^2\delta + 2h_{\max}\sigma|\Delta\nu|^{\tau_i,\delta}_{\max}\right)\right)\\
\leq&\frac{1}{t}\log E_0 \exp(N_t c\sigma^{-2}\delta)\\
\leq&\frac{1}{t}\log\sum_{n=0}^{\infty}\exp(nc\sigma^{-2}\delta)c\frac{(\mu t)^n}{n!}e^{-\mu t}\\
=&\frac{1}{t}\log c + \mu(e^{c\delta/\sigma^2}-1)\,, \qquad (26)
\end{aligned}$$

where $\mu = \max_i \sum_{j\neq i}\hat{G}_{ij}$. Combining (22), (24), (25), and (26),

$$\sigma^2\lambda_1^\sigma \leq \frac{1}{2}\,E_s h^2(x_0) + \frac{\sigma^2}{2\delta} + \frac{C'_\delta}{2} + \sigma^2\mu\left(e^{c\delta/\sigma^2}-1\right)\,. \qquad (27)$$

Now take $\sigma^2/\delta = \epsilon$ and $\delta, \sigma \to 0$, then take infimum over $\{\epsilon > 0\}$ to get

$$\limsup_{\sigma\to 0}\sigma^2\lambda_1^\sigma \leq \frac{1}{2}\,E_s h^2(x_0)\,. \qquad \square$$

Proof of Lemma 3.3. We use the same notations as in Lemma 3.2. Let $d\rho_t = \hat{G}^*\rho_t dt + \sigma^{-2}H\rho_t dy_t$, $d\eta_t = \hat{G}^*\eta_t dt + \sigma^{-2}H\eta_t dy_t$ (the difference between ρ_t and η_t lies in possibly different initial conditions). In what follows, we suppress the index t. Write $\rho\wedge\eta = \frac{1}{2}(\rho\eta^* - \eta\rho^*)$, then

$$\begin{aligned}
d\rho\eta^* &= \hat{G}^*\rho\eta^* dt + \sigma^{-2}H\rho\eta^* dy_t + \rho\eta^*\hat{G}dt + \sigma^{-2}\rho\eta^* H dy_t + \sigma^{-2}H\rho\eta^* H dt\,, \qquad (28)\\
d(\rho\wedge\eta) &= \left[\hat{G}^*(\rho\wedge\eta) - \left(\hat{G}^*(\rho\wedge\eta)\right)^*\right]dt + \sigma^{-2}\left[H(\rho\wedge\eta) - (H(\rho\wedge\eta))^*\right]dy_t\\
&\quad + \sigma^{-2}H(\rho\wedge\eta)H dt\,.
\end{aligned}$$

Let the $(d-1)d$-dimensional vector $\underline{\alpha} = (\alpha_{12}\ \alpha_{13}\cdots\alpha_{1d}\ \alpha_{21}\ \alpha_{23}\cdots\alpha_{d(d-1)})^*$ be defined by

$$\rho\wedge\eta = \begin{pmatrix} 0 & \alpha_{12}\ \alpha_{13} & \cdots & \alpha_{1d}\\ \alpha_{21} & 0\ \alpha_{23} & \cdots & \alpha_{2d}\\ \vdots & \ddots & & \\ \alpha_{d1} & & & 0 \end{pmatrix}.$$

Then (28) can be written as

$$d\underline{\alpha} = \bar{G}^*\,\underline{\alpha}\,dt + \sigma^{-2}\bar{H}_1\,\underline{\alpha}\,dt + \sigma^{-2}\bar{H}_2\,\underline{\alpha}\,dy_t \qquad (29)$$

with $\bar{G}(i,j) \geq 0$ for all $i \neq j$,

$$\bar{H}_1 = \begin{pmatrix} h_1h_2 & & & \\ & h_1h_3 & & \\ & & \ddots & \\ & & & h_dh_{d-1} \end{pmatrix},$$

$$\bar{H}_2 = \begin{pmatrix} h_1+h_2 & & & \\ & h_1+h_3 & & \\ & & \ddots & \\ & & & h_d+h_{d-1} \end{pmatrix}.$$

We may now regard $\underline{\alpha}$ as a $d(d-1)$-dimensional vector indexed by ij with $i \neq j$. Viewed this way, the matrix $\bar{G}$ has off-diagonal entries

$$\bar{G}_{ij,\ell m} = \begin{cases} \hat{G}_{\ell i} & j = m \neq \ell, \\ \hat{G}_{mj} & i = \ell \neq m, \\ 0 & j \neq m \,\text{and}\, i \neq \ell. \end{cases}$$

The matrix $\bar{G}$ is not necessarily a transition-rate matrix. However, there exists a transition-rate matrix $\tilde{G}$ which is equal to $\bar{G}$ off the diagonal. Thus (29) may be written

$$d\underline{\alpha} = \tilde{G}^* \,\underline{\alpha}\, dt + \sigma^{-2}\tilde{H}_1 \,\underline{\alpha}\, dt + \sigma^{-2}\tilde{H}_2 \,\underline{\alpha}\, dy_t\,, \tag{30}$$

with $\tilde{G}^* + \sigma^{-2}\tilde{H}_1 = \bar{G}^* + \sigma^{-2}\bar{H}_1$, $\tilde{H}_2 = \bar{H}_2$. It follows that

$$\tilde{H}_1 = \begin{pmatrix} h_1h_2+\sigma^2\Delta g_{12} & & & \\ & h_1h_3+\sigma^2\Delta g_{13} & & \\ & & \ddots & \\ & & & h_dh_{d-1}+\sigma^2\Delta g_{d(d-1)} \end{pmatrix}.$$

Note that while we are primarily interested in solutions to (30) which are in the anti-symmetric subspace $\underline{\alpha}_{ij} = -\underline{\alpha}_{ji}$, (30) makes perfect sense for arbitrary vectors in $\mathbb{R}^{d(d-1)}$. This point of view is particularly useful when computing upper bounds on Lyapunov exponents.

We now use $\tilde{h}_i(j\,k)$ ($\bar{h}_i(j\,k)$) to denote the jkth element on the diagonal of $\tilde{H}_i$ (respectively, $\bar{H}_i$), $i = 1, 2$. Let $S = \{jk : j,k \in \{1,\ldots,d\}, j \neq k\}$. Associate to the Markovian generator $\tilde{G}$ the S-valued process $\{\tilde{x}_t\}$, independent of $\{x_0^t\}$ and of $\{y_0^t\}$. We now introduce an auxiliary assumption on $\tilde{G}$, which will later be proved to be implied by (A2).

(A3) Let (A2) hold. In addition, assume that $\tilde{x}$ has no transient states.

Note that (A3) implies that $\tilde{x}$ possesses a stationary distribution (not necessarily unique) with strictly positive components.

By the stochastic Feynman–Kac formula of nonlinear filtering (using, e.g., an argument similar to Lemma 2.1 of [10]), if $\underline{\alpha}_{ij}(t=0) = P(\tilde{x}_0 = (ij))$, then

$$\langle \underline{\alpha}, \mathbf{1} \rangle = \sum_{i,j} \underline{\alpha}_{ij} = E_0\left[\exp\left(\sigma^{-2}\left(\int_0^t \tilde{h}_2(\tilde{x}_s)dy_s - \frac{1}{2}\,\tilde{h}_2^2(\tilde{x}_s)ds + \tilde{h}_1(\tilde{x}_s)ds\right)\right)\,\middle|\, y_0^t\right].$$

Let A_t denote the linear map $\underline{\alpha}_0 \to \underline{\alpha}_t$. Let $p^{\tilde{x}}_{\rm stat}$ denote the stationary distribution of $\tilde{x}$, which by (A3) has all entries strictly positive. Mimicking the argument used in the proof of Lemma 3.2, one has by positivity and Oseledec's theorem that

(31)
$$\begin{aligned}
&\lambda_1^\sigma + \lambda_2^\sigma \\
&= \lim_{t\to\infty} \frac{1}{t} \log \sup_{\{\rho_0,\eta_0:\|\rho_0\wedge\eta_0\|=1\}} \| \rho_t \wedge \eta_t \| \leq \lim_{t\to\infty} \frac{1}{t} E \log \|A_t\| \\
&\leq \lim_{t\to\infty} \frac{1}{t} E \log \| A_t p^{\tilde{x}}_{\rm stat} \| \\
&= \lim_{t\to\infty} \frac{1}{t} E \log E_0 \Big[\exp\Big(\sigma^{-2} \Big(\int_0^t \tilde{h}_2(\tilde{x}_s) dy_s - \frac{1}{2} \int_0^t \tilde{h}^2(\tilde{x}_s) ds + \int_0^t \tilde{h}_1(x_s) ds \Big) \Big) | y_0^t \Big].
\end{aligned}$$

Define $\hat{n} : \mathbb{R} \to \{1,\dots,d\}$ by $\hat{n}(a) = \mathrm{argmin}_i |h_i - a|$. Then there exists a constant $r_0 > 0$ such that if $|h_i - a| \leq r_0$ then $\mathrm{argmin}_{i\neq\hat{n}(a)} |h_i - a| = \mathrm{nbr}(\hat{n}(a))$. Now, denoting $g_m = \max |\Delta g(\cdot)|$,

(32)
$$\begin{aligned}
J_t :=& \int_0^t \tilde{h}_2(\tilde{x}_s) dy_s - \frac{1}{2} \int_0^t \tilde{h}_2^2(\tilde{x}_s) ds + \int_0^t \tilde{h}_1(\tilde{x}_s) ds \\
=& \int_0^t \bar{h}_2(\tilde{x}_s) dy_s + \int_0^t \sigma^2 \Delta g(\tilde{x}_s) dy_s - \int_0^t \frac{1}{2} \bar{h}_2^2(\tilde{x}_s) ds \\
&+ \int_0^t \bar{h}_2(\tilde{x}_s) \sigma^2 \Delta g(\tilde{x}_s) ds - \int_0^t \frac{1}{2} \sigma^4 \Delta g^2(\tilde{x}_s) ds + \int_0^t \bar{h}_1(\tilde{x}_s) ds \\
\leq& \sum_{i=0}^{i_t} \Big\{ \bar{h}_2(\tilde{x}_{i\delta}) \Delta_i y - \frac{1}{2} \bar{h}_2^2(\tilde{x}_{i\delta}) \delta + \bar{h}_1(\tilde{x}_{i\delta}) + \sigma^2 g_m |\Delta y|^{i,\delta}_{\max} + 2 h_{\max} \sigma^2 g_m \delta \Big\} \\
&+ \sum_{i=1}^{N_t} (2 h_{\max} |\Delta y|^{\tau_i,\delta}_{\max} + 2 {h_{\max}}^2 \delta) \\
\leq& J_t^1 + J_t^2,
\end{aligned}$$

where

$$\begin{aligned}
J_t^1 =& \sum_{i=0}^{i_t} \delta \Big\{ \Big[h\Big(\hat{n}\Big(\frac{\Delta_i y}{\delta} \Big) \Big) + h\Big(\mathrm{nbr}\Big(\hat{n}\Big(\frac{\Delta_i y}{\delta} \Big) \Big) \Big) \Big] \frac{\Delta_i y}{\delta} \\
&- \frac{1}{2} h^2\Big(\hat{n}\Big(\frac{\Delta_i y}{\delta} \Big) \Big) - \frac{1}{2} h^2\Big(\mathrm{nbr}\Big(\hat{n}\Big(\frac{\Delta_i y}{\delta} \Big) \Big) \Big) \Big\} \\
&+ \sum_{\{i\leq i_t : |\frac{\Delta_i y}{\delta} - h_j| > r_0 \ \forall j\}} \big\{ 4 h_{\max} |\Delta_i y|^{i,\delta}_{\max} + h_m^2 \delta \big\}
\end{aligned}$$

and

$$\begin{aligned}
J_t^2 =& \sum_{i=1}^{i_t} \big\{ \sigma^2 g_m (2 h_{\max} \delta + \sigma |\Delta \nu|^{i,\delta}_{\max}) + 2 h_{\max} g_m \sigma^2 \delta \big\} \\
&+ \sum_{i=1}^{N_t} \big\{ 2 h_{\max} \big(2 h_{\max} \delta + \sigma |\Delta \nu|^{\tau_i,\delta}_{\max} \big) + 2 {h_{\max}}^2 \delta \big\} \ .
\end{aligned}$$

Having J_t^1 measurable with respect to y_0^t, it follows using Jensen's inequality that

$$\frac{1}{t} E \log E_0 \left[\exp\left(\sigma^{-2} J_t\right) | y_0^t \right] \leq \frac{1}{t} \frac{1}{\sigma^2} E J_t^1 + \frac{1}{t} \log E_0 \exp\left(\sigma^{-2} J_t^2\right). \tag{33}$$

Now,

$$\begin{aligned} \frac{1}{t} E J_t^1 \leq & \frac{1}{\delta} E_s \left\{ (h(x_0) + h(\mathrm{nbr}(x_0)))(\delta h(x_0) + \sigma \Delta\nu) - \frac{\delta}{2} h^2(x_0) - \frac{\delta}{2} h^2(\mathrm{nbr}(x_0)) \right\} \\ & + c\delta + c e^{-r_0^2 \delta / 2\sigma^2}, \end{aligned} \tag{34}$$

where the second term is due to the fact that the probability of having a jump in the x process on any δ-interval is of order δ, and the last term is due to the Gaussian law of ν. Next,

$$\begin{aligned} & \frac{1}{t} \log E_0 \left(\exp \sigma^{-2} J_t^2 \right) \leq \frac{1}{t} \log E_0 \, E_0 \left[\exp(\sigma^{-2} J_t^2) | N_t \right] \\ & \leq \frac{1}{t} \log E_0 \, E_0 \left[\exp \left\{ \sum_{i=1}^{N_t} \left(c\delta + c\sigma |\Delta\nu|_{\max}^{i,\delta} \right) + \sum_{i=1}^{N_t} c\delta\sigma^{-2} + c\sigma^{-1} |\Delta\nu|_{\max}^{\tau_i \delta} \right\} \, | N_t \right] \\ & = \frac{1}{t} \log E_0 \exp \left(\frac{t}{\delta} c\delta + \frac{t}{\delta} c^2\sigma^2\delta + N_t c\delta\sigma^{-2} + N_t c^2 \sigma^{-2} \delta \right) \\ & = c + c^2\sigma^2 + \frac{1}{t} \log E \, \exp c N_t \delta \sigma^{-2} \\ & \leq c + c^2\sigma^2 + \lambda(e^{c\delta/\sigma^2} - 1). \end{aligned} \tag{35}$$

Finally, combining (33), (34), and (35),

$$\begin{aligned} \sigma^2(\lambda_1^\sigma + \lambda_2^\sigma) \leq E_s & \left[(h(x_0) + h(\mathrm{nbr}(x_0)))h(x_0) - \frac{1}{2} h^2(x_0) - \frac{1}{2} h^2(\mathrm{nbr}(x_0)) \right] + c\delta \\ & + c e^{-r_0^2 \delta / 2\sigma^2} + \sigma^2 c + \sigma^4 c^2 + \sigma^2 c (e^{c\delta/\sigma^2} - 1). \end{aligned}$$

Take now $\sigma^2/\delta = \epsilon$, $\delta, \sigma \to 0$, then take $\epsilon \to 0$ to conclude the lemma under (A3).

Although (A2) does not imply that all states of $\tilde{x}$ are communicating (as the example $d = 3, \hat{G}_{13} = \hat{G}_{31} = 0$ shows), we now show that it does imply (A3). It suffices to show that for every two states $ij,\ kl \in S$, if $(\tilde{G}^m)_{ij,\,kl} > 0$ for some m, then there exists an n such that $(\tilde{G}^n)_{kl,\,ij} > 0$, or, in the terminology we use in what follows, if the path $ij \to kl$ exists then the path $kl \to ij$ exists, too. Note next that it suffices to show the above for $j = l$ and for $i, k \neq j$ such that $\hat{G}_{i\,k} > 0$ (that is $i \to k$ in one step). Suppose, then, that $ij \to kj$ in one step. It needs to be shown that $kj \to ij$. If there exists a path $k \to i$ that does not contain j, then the claim is proved. Otherwise, since $\hat{G}$ is communicating, there exists a path $k \to j \to i$ such that $k \to j$ does not contain i and $j \to i$ does not contain k. Thus the following path exists too:

$$kj \to ki \to ji \to jk \to ik \to ij,$$

and it follows that $\tilde{x}$ has no transient states. This concludes the proof. □

Proof of Lemma 3.4. The top Lyapunov exponent satisfies

$$\lambda_1^\sigma = \lim_{t\to\infty} \frac{1}{t} E \log \|U_t\| \geq \lim_{t\to\infty} \frac{1}{t} E \log \|U_t p_{\mathrm{stat}}^x\|.$$

Again, we compute $\|U_t p^x_{\rm stat}\|$ using the Kallianpur–Striebel formula. Fix $\delta > 0$ and let I_i be the interval $[\delta j, \delta(j+1))$ such that $\tau_i \in I_i$. Then denoting, as above, by $\tilde{x}$ a copy of the process x which is independent of y_0^t under E_0,

$$\begin{aligned}
&\frac{1}{t} E \log E_0 \left[\exp\left(\sigma^{-2} \left(\int_0^t h(\tilde{x}_s) dy_s - \frac{1}{2} \int_0^t h^2(\tilde{x}_s) ds \right) \right) | y_0^t \right] \\
&\geq \frac{1}{t} E \log E_0 \left[E_0 \left[1_{\{\tilde{x}_s = x_s \, \forall s \notin \bigcup_i I_i, \, s<t\}} \right. \right. \\
&\qquad \left. \left. \exp\left(\sigma^{-2} \left(\int_0^t h(\tilde{x}_s) dy_s - \frac{1}{2} \int_0^t h^2(\tilde{x}_s) ds \right) \right) | x_0^t, y_0^t \right] | y_0^t \right] \\
&\geq \frac{1}{t} E \log E_0 \left[E_0 \left[1_{\{\tilde{x}_s = x_s \, \forall s \notin \bigcup_i I_i, \, s<t\}} \exp\left(\sigma^{-2} \, B_t(x,y) \right) | x_0^t, y_0^t \right] | y_0^t \right],
\end{aligned}$$

where

$$B_t(x,y) = \int_0^t h(x_s) dy_s - \frac{1}{2} \int_0^t h^2(x_s) ds - \sum_i (c |\Delta y|^{\tau_i, \delta}_{\max} + c\delta).$$

Thus, by Jensen's inequality,

$$\begin{aligned}
\frac{1}{t} E \log \|U_t p^x_{\rm stat}\| &\geq \frac{1}{t} E \log E_0 \left[1_{\{\tilde{x}_s = x_s \, \forall s \notin \bigcup_i I_i, \, s<t\}} \exp\left(\sigma^{-2} B_t(x,y) \right) | x_0^t, y_0^t \right] \\
&= \frac{1}{t} \log E \left[1_{\{\tilde{x}_s = x_s \, \forall s \notin \bigcup_i I_i, \, s<t\}} | x_0^t \right] + \frac{1}{t} E \sigma^{-2} B_t(x,y).
\end{aligned}$$

Now,

$$\begin{aligned}
&\frac{1}{t} E \log E \left[1_{\{\tilde{x}_s = x_s \, \forall s \notin \bigcup_i I_i, \, s<t\}} | x_0^t \right] \\
&= \frac{1}{t} E \log E \left[1_{\{\tilde{x}_{j\delta} = x_{j\delta} \, \forall j < t/\delta \text{ s.t. } j\delta \notin \bigcup_i I_i\}} | x_0^t \right] \\
&\geq \frac{1}{t} E \log E \left[1_{\{\tilde{x}_{j\delta} = x_{j\delta} \, \forall j < t/\delta\}} | x_0^t \right].
\end{aligned}$$

The last quantity tends to $-\frac{1}{\delta} \mathcal{H}(\{x_{j\delta}\})$, where $\mathcal{H}(\{x_{j\delta}\})$ is the entropy rate for $\{x_{j\delta}\}$. Moreover,

$$\lim_{t \to \infty} \frac{1}{t} E B_t(x,y) = \frac{1}{2} E h^2(x_0) - c\delta,$$

and thus we have shown

$$\sigma^2 \lambda_1^\sigma \geq \frac{1}{2} E h^2(x_0) - c\delta - \frac{\sigma^2}{\delta} \log d.$$

Now taking $\sigma \to 0$ and then $\delta \to 0$ yields the result. □

Remark. The vector $\rho \wedge \eta$ has only $(d-1)d/2$ degrees of freedom, the same as the vector $\underline{\alpha}$. We have used a $(d-1)d$ dimensional vector $\underline{\alpha}$ in order to write (29) as a matrix with nonnegative off-diagonal entries. For general (nonfiltering) situations, this might not be possible.

Acknowledgment. We thank P. Bougerol for bringing [7] to our attention.

REFERENCES

[1] R. W. Brockett, *Nonlinear systems and nonlinear estimation theory*, in Stochastic Systems: The Mathematics of Filtering and Identification and Applications, M. Hazwinkel and J. C. Willems, eds., D. Reidel Publishing Company, Dordrecht, the Netherlands, 1981, pp. 441–477.

[2] R. Carmona and J. Lacroix, *Spectral Theory of Random Schrödinger Operators,* Birkhäuser, Zurich, 1990.

[3] J. E. Cohen, H. Kesten, and C. M. Newman, *Oseledec's multiplicative ergodic theorem: A proof,* in Random Matrices and Their Applications, J. E. Cohen, H. Kesten, and C. M. Newman, eds., Am. Math. Soc., Providence, RI, 1986, pp. 23–30.

[4] B. Delyon and O. Zeitouni, *Lyapunov exponents for filtering problems,* in Applied Stochastic Analysis, M. H. A. Davis and R. J. Elliott, eds., Gordon and Breach, London, 1991, pp. 511–521.

[5] H. Kunita, *Stochastic Partial Differential Equations Connected with the Nonlinear Filtering,* Lecture Notes in Mathematics 972, Springer, Berlin, 1981, pp. 101–168.

[6] D. Ocone and E. Pardoux, *Asymptotic stability of the optimal filter with respect to its initial condition*, SIAM J. Control Optim., 34 (1996), pp. 226–243.

[7] Y. Peres, *Domains of analytic continuation for the top Lyapunov exponent,* Ann. Inst. H. Poincaré Probab. Statist., 28 (1992), pp. 131–148.

[8] E. Seneta, *Non-negative Matrices and Markov Chains,* Springer-Verlag, Berlin, New York, 1981.

[9] W. M. Wonham, *Some applications of stochastic differential equations to optimal nonlinear filtering,* SIAM J. Control Optim., 2 (1965), pp. 347–368.

[10] O. Zeitouni and B. Z. Bobrovsky, *On the reference probability approach to the equations of non-linear filtering,* Stochastics, 19 (1986), pp. 133–149.

SIAM J. CONTROL OPTIM.
Vol. 35, No. 1, pp. 56–77, January 1997

004

THE EXTENDED EULER–LAGRANGE CONDITION FOR NONCONVEX VARIATIONAL PROBLEMS*

RICHARD VINTER† AND HARRY ZHENG†

Abstract. This paper provides necessary conditions of optimality for a general variational problem for which the dynamic constraint is a differential inclusion with a possibly nonconvex right side. They take the form of an Euler–Lagrange inclusion involving convexification in only one coordinate, supplemented by the transversality and Weierstrass conditions. It is also shown that for time-invariant, free time problems, the adjoint arc can be chosen so that the Hamiltonian function is constant along the minimizing state arc. The methods used here, based on simple "finite dimensional" nonsmooth calculus, Clarke decoupling, and a rudimentary version of the maximum principle, offer an alternative, and somewhat simpler, derivation of such results to those used by Ioffe and Rockafellar in concurrent research.

Key words. Euler–Lagrange condition, calculus of variations, nonconvex differential inclusion, nonsmooth analysis, limiting subdifferential

AMS subject classification. 49K24

PII. S0363012995283133

1. Introduction.

We consider nonsmooth variational problems of the form (P):

$$\text{(P)}\qquad \begin{aligned} &\text{minimize } l(x(0),x(1)) + \int_0^1 L(t,x(t),\dot{x}(t))dt \\ &\text{over arcs } x \in W^{1,1}([0,1];R^n) \text{ which satisfy} \\ &\dot{x}(t) \in F(t,x(t)) \quad \text{a.e. on } [0,1]. \end{aligned}$$

Here $l : R^n \times R^n \to R \cup \{+\infty\}$ and $L : [0,1] \times R^n \times R^n \to R$ are given functions and $F : [0,1] \times R^n \rightrightarrows R^n$ is a given multifunction.

The minimization is performed over arcs x in $W^{1,1}([0,1];R^n)$ (the space of absolutely continuous R^n-valued functions on $[0,1]$) which satisfy the differential inclusion constraint and for which $L(t,x(t),\dot{x}(t))$ is an integrable function. Let $\bar{x}$ be a minimizer.

Problem (P) provides a framework for dynamic optimization which emphasizes the constraints on allowable velocities $\dot{x}$ for a given time and state (t,x). It covers the formulation traditionally adopted in optimal control theory involving a differential equation parameterized by a control function u

$$\dot{x}(t) = f(t,x(t),u(t)) \quad \text{a.e.}$$

and

$$u(t) \in U(t) \quad \text{a.e.}$$

Here we choose the multifunction to be $F(t,x) \equiv f(t,x(t),U(t))$. (A differential inclusion arising in this way is said to have a "parameterization" (f,U).) But problem

*Received by the editors March 15, 1995; accepted for publication (in revised form) October 9, 1995. This research was supported by the Engineering and Physics Science Research Council.
http://www.siam.org/journals/sicon/35-1/28313.html

†Department of Electrical and Electronic Engineering and Centre for Process Systems Engineering, Imperial College, Exhibition Road, London SW7 2BT, UK (r.vinter@ic.ac.uk, h.zheng@ic.ac.uk).

(P) is a convenient description of dynamic systems incorporating feedback loops, state-dependent control constraints, and other such features for purposes of deriving optimality conditions.

In the case of no dynamical constraints ($F \equiv R^n$), smooth l, and L and when the Lipschitz rank of $(x,u) \to L(t,x,u)$ is suitably bounded, the following classical first-order necessary conditions are satisfied by the minimizer $\bar{x}$: there exists $p \in W^{1,1}$ such that

$$(\dot{p}(t), p(t)) = \nabla L(t, \bar{x}(t), \dot{\bar{x}}(t)) \tag{1}$$

(the Euler–Lagrange condition),

$$p(t)\dot{\bar{x}}(t) - L(t, \bar{x}(t), \dot{\bar{x}}(t)) \geq p(t)v - L(t, \bar{x}(t), v) \quad \forall\ v \in R^n \text{ a.e.} \tag{2}$$

(the Weierstrass condition), and

$$(p(0), -p(1)) \in \nabla l(\bar{x}(0), \bar{x}(1)) \tag{3}$$

(the transversality condition).

Modern developments provide necessary conditions when a dynamic constraint is present ($F \not\equiv R^n$) and when the functions l and L are possibly nonsmooth. There are three main strands to this research. The first is the maximum principle (for parameterized problems) and its nonsmooth counterparts. The second is to replace (1) and (2) with Clarke's Hamiltonian inclusion [2]:

$$(-\dot{p}(t), \dot{\bar{x}}(t)) \in \text{co}\partial H_\lambda(t, \bar{x}(t), p(t)) \tag{4}$$

$$(p(0), -p(1)) \in \partial l(\bar{x}(0), \bar{x}(1)),$$

in which H_λ is the Hamiltonian

$$H_\lambda(t,x,p) := \sup\{pv - \lambda L(t,x,v) : v \in F(t,x)\}.$$

Here ∂H_λ denotes the limiting subdifferential of $H_\lambda(t,\cdot,\cdot)$, defined below (its convex hull $\text{co}\partial H_\lambda$ is the Clarke generalized gradient). The presence of constraints necessitates the introduction of a cost multiplier λ ($\lambda \geq 0$, λ and $p(\cdot)$ not both zero).

We focus attention on the third type of conditions, namely conditions which are nonsmooth analogues of the classical conditions (1)–(3). What form should they take? Notice that $\bar{x}$ is a minimizer for a variational problem in which the dynamic constraint is absorbed into the cost integrand, namely

$$\text{minimize } \int_0^1 (L(t,x,\dot{x}) + \Psi_{\text{Gr}\{F(t,\cdot)\}}(x,\dot{x}))dt.$$

(Here Ψ_A is the indicator function for the set A, which takes value 0 on A and $+\infty$ elsewhere.) This would suggest a nonsmooth version of (1) "$(\dot{p}(t), p(t)) \in \lambda\text{co}\partial(L(t,\cdot,\cdot) + \Psi_{\text{Gr}\{F(t,\cdot)\}})(\bar{x}(t), \dot{\bar{x}}(t))$," from which we might expect to deduce, via a sum rule and a subdifferential calculus for indicator functions

$$(\dot{p}(t), p(t)) \in \lambda\text{co}\partial L(t, \bar{x}(t), \dot{\bar{x}}(t)) + \overline{\text{co}}N_{\text{Gr}\{F(t,\cdot)\}}(\bar{x}(t), \dot{\bar{x}}(t)) \quad \text{a.e.}$$

Here $N_{\text{Gr}\{F(t,\cdot)\}}$ is the limiting normal cone, defined below. (Its closed convex hull $\overline{\text{co}}N_{\text{Gr}F(t,\cdot)}$, featured here, is the Clarke normal cone.) See Clarke's paper [1] for results in this spirit, which we refer to as the Euler–Lagrange inclusion.

The weak convergence techniques used in the proof of necessary conditions of this type make it almost inevitable that the conditions take the form of a convex set inclusion. Nonetheless a research theme of recent years has been to reduce the extent of convexification involved. Under the convexity hypothesis Mordukhovich [11, 12] derived, via discrete approximations, a form of the Euler–Lagrange inclusion which involves convexification with respect to only one coordinate

$$\dot{p}(t) \in \text{co}\{\eta : (\eta, p(t)) \in \lambda \partial L(t, \bar{x}(t), v) + N_{\text{Gr}\{F(t,\cdot)\}}(\bar{x}(t), v) \text{ for some } v \in F(t, \bar{x}(t))$$

$$\text{such that } p(t)v - \lambda L(t, \bar{x}(t), v) = H_\lambda(t, \bar{x}(t), p(t))\} \quad \text{a.e.}$$

The sharpest available conditions along these lines, in which the convex hull is taken just with respect to $v = \dot{\bar{x}}(t)$ rather than all v's achieving the maximum in the Hamiltonian, were obtained by Loewen and Rockafellar [9] with an analysis based on Hamiltonian inclusions and a subdifferential calculus of perturbed Hamiltonian functions for convex differential inclusions. Under the convexity hypothesis they showed that there exists $\lambda \geq 0$ and $p \in W^{1,1}$ (not all zero) such that

$$\dot{p}(t) \in \text{co}\{\eta : (\eta, p(t)) \in \lambda \partial L(t, \bar{x}(t), \dot{\bar{x}}(t)) + N_{\text{Gr}\{F(t,\cdot)\}}(\bar{x}(t), \dot{\bar{x}}(t))\} \tag{5}$$

$$(p(0), -p(1)) \in \lambda \partial l(\bar{x}(0), \bar{x}(1)).$$

(We have been informed that similar results were derived by Smirnov [16] for a narrower class of problems using discrete approximation methods.) We refer to (5) as the extended Euler–Lagrange condition.

Interest in these conditions has been heightened by recent findings of Rockafellar [15] that under the convexity assumption and other mild hypotheses this last condition is equivalent to

$$\dot{p}(t) \in \text{co}\{\eta : (-\eta, \dot{\bar{x}}(t)) \in \partial H_\lambda(t, \bar{x}(t), p(t))\} \quad \text{a.e.},$$

which will be recognized as a sharpened version of the Hamiltonian inclusion (4) involving convexification with respect to only one coordinate. Thus sharpened forms of the Hamiltonian inclusion and extended Euler–Lagrange inclusion coalesce under the convexity hypothesis.

We mention that in addition to the three types of necessary conditions for (P) outlined above there are hybrid conditions due to Kaśkosz and Lojasiewicz [6, 7] and refined by Zhu [18]. These conditions, expressed in terms of a family of "Lipschitz selectors" of the multifunction co$F$, are applicable to variational problems involving general, unparameterized, differential inclusions, yet have the character of the maximum principle.

What necessary conditions are valid for variational problems (P) with general endpoint constraints when the convexity hypothesis is dropped? The maximum principle remains valid and also the maximum principle–like conditions of Kaśkosz, Lojasiewicz, and Zhu. Whether the Hamiltonian inclusion is valid in this situation is a long-standing open question in dynamic optimization.

What about nonsmooth analogues of the classical first-order conditions for nonsmooth, nonconvex problems? The expected conditions are as follows: there exist λ (≥ 0) and $p \in W^{1,1}$, not both zero, such that

$$\dot{p}(t) \in \text{co}\{\eta : (\eta, p(t)) \in \lambda \partial L(t, \bar{x}(t), \dot{\bar{x}}(t)) + N_{\text{Gr}\{F(t,\cdot)\}}(\bar{x}(t), \dot{\bar{x}}(t))\}, \tag{6}$$

$$p(t)\dot{\bar{x}}(t) - \lambda L(t, \bar{x}(t), \dot{\bar{x}}(t)) \geq p(t)v - \lambda L(t, \bar{x}(t), v) \quad \text{for all } v \in F(t, \bar{x}(t)) \text{ a.e.}, \tag{7}$$

and

$$(p(0), -p(1)) \in \lambda \partial l(\bar{x}(0), \bar{x}(1)). \tag{8}$$

Notice that the extended Euler–Lagrange inclusion (6) and transversality conditions (8) have been supplemented by a form of the Weierstrass condition (7) appropriate to problems with dynamic constraints. The Weierstrass condition is superfluous for convex problems (it is implied by the extended Euler–Lagrange inclusion (6)), but for nonconvex problems it is a genuinely independent condition on minimizers which introduces an important global dimension into the optimality conditions regarding treatment of the velocity variable. The methods of Loewen and Rockafellar [9], which depend critically on the convexity of the differential inclusion concerned, give little indication of how such conditions might be derived.

Mordukhovich [14], using a discrete approximation approach, derived the extended Euler–Lagrange inclusion and transversality condition (but not the Weierstrass condition) for the costate arc p in the nonconvex case. However, the dropping of the convexity hypothesis was counterbalanced by imposition of hypotheses stronger than those of Loewen and Rockafellar regarding regularity and boundedness of F and its associated subdifferentials.

In a recent paper [5] Ioffe and Rockafellar established validity of the extended Euler–Lagrange, Weierstrass, and transversality conditions for nonconvex problems in the case $F \equiv R^n$ (no dynamic constraints). For the class of problems considered, the regularity and boundedness hypotheses imposed on the data are considerably weaker than those of [14]. The analysis is based on an analysis of integral functionals and is a showcase of new constructs of infinite dimensional nonsmooth analysis (those associated with fuzzy calculus of approximate subdifferential, etc.).

We come now to the contributions of this paper. These are partly improvements on earlier results concerning extended Euler–Lagrange and related conditions for nonconvex problems with a dynamic constraint and partly methodology. A relatively straightforward derivation of Ioffe and Rockafellar's necessary conditions for nonconvex problems is first given, when $F \equiv R^n$, which just uses the "elementary" finite dimensional calculus of limiting subdifferentials, a simple version of the maximum principle and the most traditional of nonsmooth variational principles, Ekeland's theorem. These results are a stepping stone to our subsequent derivation of related necessary conditions (via an exact penalization technique akin to that used by Clarke [2]) when a (nonconvex) differential inclusion is added to the constraints. The necessary conditions provided here go beyond those for $W^{1,1}$ minimizers in [14], both because they incorporate the Weierstrass condition and because they are derived under boundedness and regularity hypotheses akin to those adopted by Loewen and Rockafellar (in the convex case), which are significantly weaker than those invoked in [14]. In particular the hypotheses concerning a.e. continuity of $F(\cdot, x)$, upper semicontinuous dependence of certain subdifferentials, and uniform boundedness of F are dispensed with. Finally, we examine free time, autonomous problems (problems where neither L nor F depend on t). It is shown here that the earlier conditions can be augmented by a constancy condition on the Hamiltonian and a modified transversality condition.

A word about methodology. This borrows ideas from Clarke's "decoupling" technique [3], whereby a variational problem involving a state-dependent velocity constraint is approximated by one for which the velocity constraint is state-free. The Pontryagin maximum principle provides very precise optimality conditions for this

latter category of variational problems; these are applied to the approximate problems and we then pass to the limit. However, our approximations procedure is based on use of Ekeland's theorem rather than on density properties of proximal subgradients (as with Clarke), which allows us to conclude strong L^1 convergence of velocities. (It would appear that weak L^1 convergence provided by the proximal analysis approach is inadequate for the derivation of optimality conditions in the nonconvex case.)

Consider the $F = R^n$ case. Our simple idea is to replace the variational problem (P) with the "decoupled" problem

$$\text{minimize } l(x(0), x(1)) + \int_0^1 L(t, w(t), v(t))dt + \epsilon^{-1} \int_0^1 k(t)|x(t) - w(t)|^2 dt$$

over $(x, (w, v))$ which satisfies $\dot{x}(t) = v(t)$ for some arbitrarily small $\epsilon > 0$. (Here $k(t)$ is a Lipschitz constant for the data.) (v, w) are treated as control functions. Because of the third (penalty) term in the cost, which forces w to approximate a state trajectory corresponding to v, $\bar{x}$ is an approximate minimizer for this new problem. We then can perturb the approximating problem, with the help of Ekeland's theorem, to guarantee it has a local minimizer converging to $\bar{x}$ as $\epsilon \downarrow 0$, in some sense. Applying the Pontryagin maximum principle to the perturbed problem and passing to the limit as $\epsilon \downarrow 0$ give the desired necessary conditions when $F = R^n$. From this everything follows.

We claim our methods provide a simple derivation of the extended Euler–Lagrange and related conditions. Some readers might object that we call upon the Pontryagin maximum principle, which, in its full generality, has a lengthy proof. However, the special case of it required for application to the decoupled approximation problem, a case treated by Mordukhovich in [10], admits a simple, direct proof because the dynamics and cost integrand are smooth in the state variable and nonsmoothness is confined to the description of the endpoint constraints.

Concurrently, Ioffe [4] too has derived extended Euler–Lagrange and related conditions for nonconvex problems of type (P) with a dynamic constraint. Ioffe improves on the necessary conditions reported in this paper, regarding fixed time problems, by weakening the Lipschitz continuity hypotheses on F under which they apply. Again, the starting point is the necessary conditions of Ioffe and Rockafellar for the case $F \equiv R^n$, but then a more refined penalty function argument than that employed here is used to introduce the dynamic constraint. It would appear, however, that a combination of the proof techniques of this paper based on application of a simple version of the maximum principle to treat the case $F \equiv R^n$ and the penalization arguments of Ioffe to allow for dynamic constraints provides the most straightforward but general derivation of the Euler–Lagrange and related conditions currently available.

The picture that emerges of necessary conditions for optimal control problems involving differential inclusions is one in which the extended Euler–Lagrange condition (coupled with the Weierstrass and transversality conditions) has a pivotal position. As we have noted, under the convexity hypothesis, it is more or less equivalent to the latest refinements of the Hamiltonian inclusion condition, but it also applies to nonconvex problems. Besides, as Ioffe has recently shown [4], the extended Euler–Lagrange condition (with its associated conditions) provides general versions of the Pontryagin-type necessary conditions of a kind previously derived by Kaśkosz and Lojasiewicz [7] as straightforward corollaries. Simple derivations of the extended Euler–Lagrange inclusion, etc., such as we provide have a valuable role then in making more accessible latest developments in the theory of necessary conditions.

The norm on $W^{1,1}$ is taken to be

$$\|x\|_{W^{1,1}} := |x(0)| + \|\dot{x}\|_{L^1}.$$

$|\cdot|$ denotes the Euclidean norm throughout. The Euclidian closed unit ball is written B. $d_C(x)$ denotes the Euclidian distance of the point $x \in R^n$ from the set $C \subset R^n$. $d_{\text{Haus}}(F_1, F_2)$ denotes the Hausdorff distance between two sets. epi$\{f\}$ is the epigraph set of the function f.

The following two constructs from nonsmooth analysis are required.

DEFINITION 1. *Take a closed set $A \subseteq R^k$ and points $x \in A$, $p \in R^k$. We say that p is a* limiting normal *to A at x if and only if there exists $p_i \to p$ and $x_i \to x$ in A such that, for each i, $p_i \cdot (x - x_i) \le o(|x - x_i|)$ for all $x \in A$, in which $o(\alpha)/\alpha \to 0$ as $\alpha \downarrow 0$ (i.e., limiting normals are limits of vectors which support A at points near x to first order). The limiting normal cone to A at x, written $N_A(x)$, is the set of all limiting normals to A at x.*

DEFINITION 2. *Given a lower semicontinuous function $f : R^k \to R \cup \{+\infty\}$ and a point $x \in R^k$ such that $f(x) < +\infty$, the* limiting subdifferential *of f at x, written $\partial f(x)$, is*

$$\partial f(x) := \{\xi : (\xi, -1) \in N_{\text{epi}\{f\}}(x, f(x))\},$$

in which epi$\{f\}$ *denotes the epigraph set $\{(x, \alpha) \in R^k \times R : \alpha \ge f(x)\}$.*

We refer to [8, 13] for expository accounts of the properties of limiting normal cones, limiting subdifferentials, and associated calculus rules.

Finally we mention that necessary conditions are derived throughout this paper for an arc $\bar{x}$ to be a $W^{1,1}$ local minimizer for (P) (or special cases of this general problem). This means that there exists $\eta > 0$ such that $\bar{x}$ is a minimizer with respect to all arcs $x \in W^{1,1}$ which satisfy the constraints of (P) and also

$$\|x - \bar{x}\|_{W^{1,1}} < \eta. \tag{9}$$

The concept of a $W^{1,1}$ local minimizer is less restrictive than that of strong local minimizer (where (9) is replaced by $\|x - \bar{x}\|_{L^\infty} < \eta$). See [14] and [17] for a discussion of this point.

2. The Bolza problem with finite Lagrangian. We begin by deriving necessary conditions for the special case of problem (P) in which $F = R^n$, namely,

$$\text{minimize } J(x) := l(x(0), x(1)) + \int_0^1 L(t, x(t), \dot{x}(t))dt \text{ over arcs } x \in W^{1,1}. \tag{2.1}$$

(This is described as the "finite Lagrangian" case because the extended Lagrangian $L_e(t, x, v)$ which coincides with $L(t, x, v)$ on $F(t, x)$ and takes value $+\infty$ off $F(t, x)$ is finite valued.) The primary role of these necessary conditions is to provide a stepping stone to treat problems with a dynamic constraint. But necessary conditions for finite Lagrangian problems are of independent interest because of the unrestrictive nature of the hypotheses under which they are valid (apart of course from the assumption that $F = R^n$!). These hypotheses, which relate to the local minimizer $\bar{x}$ of interest, are as follows:

(H1) l is lower semicontinuous.

(H2) $L(\cdot, x, \cdot)$ is measurable for each x with respect to the product σ-algebra $\mathcal{L} \times \mathcal{B}$. ($\mathcal{L}$ denotes the Lebesgue subsets of $[0, 1]$ and $\mathcal{B}$ the Borel sets of R^n.) $L(t, \cdot, \cdot)$ is lower semicontinuous for a.e. t.

(H3) For every $K > 0$ there exist $\delta > 0$ and $k \in L^1$ such that

$$|L(t, x', v) - L(t, x, v)| \leq k(t)|x' - x|, \quad L(t, \bar{x}(t), v) \geq -k(t)$$

for all $x', x \in \bar{x}(t) + \delta B$, and $v \in \dot{\bar{x}}(t) + KB$ a.e. $t \in [0, 1]$.

THEOREM 3. *Let $\bar{x}$ be a $W^{1,1}$ local minimizer for (2.1), for which $J(\bar{x}) < \infty$. Then there exists an arc $p \in W^{1,1}$ which satisfies*

(i) *the Euler condition: $\dot{p}(t) \in co\{\eta : (\eta, p(t)) \in \partial L(t, \bar{x}(t), \dot{\bar{x}}(t))\}$ a.e.;*
(ii) *the transversality condition: $(p(0), -p(1)) \in \partial l(\bar{x}(0), \bar{x}(1))$;*
(iii) *the Weierstrass condition: $p(t)\dot{\bar{x}}(t) - L(t, \bar{x}(t), \dot{\bar{x}}(t)) \geq p(t)v - L(t, \bar{x}(t), v)$ for all $v \in R^n$ a.e.*

Remark. If $L(t, x, v)$ is continuous in x, v, then the measurability condition in (H2) can be replaced by the requirement that $L(\cdot, x, v)$ is Lebesgue measurable for each x, v. We make use of this fact later.

Proof. Fix $K > 0$ and let $k(\cdot)$ and δ be the corresponding bounds and constant of (H3). We may of course assume that $k(t) \geq 1$ a.e. Let $\eta > 0$ be a constant such that $\bar{x}$ is a minimizer with respect to all competing arcs which satisfy $\|x - \bar{x}\|_{W^{1,1}} \leq \eta$ and also such that $\|x - \bar{x}\|_{W^{1,1}} \leq \eta$ implies $\|x - \bar{x}\|_{L^\infty} \leq \delta$.

The first step of the proof is to find $p \in W^{1,1}$ which satisfies the conditions of the theorem statement, except that (iii) is replaced by a weaker local version of the condition:

(iii′) $p(t)\dot{\bar{x}}(t) - L(t, \bar{x}(t), \dot{\bar{x}}(t)) \geq p(t)v - L(t, \bar{x}(t), v)$ for all $v \in \dot{\bar{x}}(t) + KB$.

We can assume without loss of generality that (H3) has been strengthened to

(H3′) $|L(t, x', v) - L(t, x, v)| \leq k(t)|x' - x|$ and $L(t, \bar{x}(t), v) \geq -k(t)$ for all $x', x \in R^n$ and $v \in \dot{\bar{x}}(t) + KB$ a.e.

This is because, if (H3′) were not satisfied, we could replace it with

$$L'(t, x, v) := \begin{cases} L(t, x, v) & \text{if } |x - \bar{x}(t)| \leq \delta, \\ L(t, \bar{x}(t) + \delta \frac{x - \bar{x}(t)}{|x - \bar{x}(t)|}, v) & \text{otherwise.} \end{cases}$$

The data (L', l) satisfy (H3′) (in addition to (H1) and (H2)). $\bar{x}$ remains a $W^{1,1}$ local minimizer. If some p satisfies (i), (ii), and (iii′) for data (L', l) at $\bar{x}$, it also satisfies these same conditions for data (L, l) because of their "local" nature. So we may assume that (H3′) is satisfied. Define

$$\tilde{L}(t, w, v) := L(t, \bar{x}(t) + w, \dot{\bar{x}}(t) + v),$$
$$\tilde{l}(x, y) := l(\bar{x}(0) + x, \bar{x}(1) + y).$$

Choose a positive sequence $\epsilon_i \to 0$. Define

$$W := \{(\xi, w, v) \in R^n \times L^1 \times L^1 : |v(t)| \leq K \text{ a.e.}, \|x_{\xi,v}\|_{W^{1,1}} \leq \eta\},$$

where $x_{\xi,v}(t) = \xi + \int_0^t v(s)ds$,

$$\|(\xi, w, v)\|_k := |\xi| + \|kw\|_{L^1} + \|kv\|_{L^1}$$

and, for each i,

$$\tilde{J}_i(\xi, w, v) := \tilde{l}(x_{\xi,v}(0), x_{\xi,v}(1)) + \int_0^1 \tilde{L}(t, w(t), v(t))dt + \epsilon_i^{-1} \int_0^1 k(t)|x_{\xi,v}(t) - w(t)|^2 dt.$$

LEMMA 4. *For each i, $(W, \|\cdot\|_k)$ is a complete metric space and $\tilde{J}_i$ is lower semicontinuous on $(W, \|\cdot\|_k)$. There exists a positive sequence $\alpha_i \to 0$ such that, for each i, $\tilde{J}_i(0,0,0) \le \inf_W \tilde{J}_i(\xi, w, v) + \alpha_i^2$.*

Proof. Obviously, W is a subset of the Banach space $\{(\xi, w, v) \in R^n \times L^1 \times L^1 : \|(\xi, w, v)\|_k < \infty\}$ with norm $\|\cdot\|_k$. We show that it is strongly closed and, for each i, $\tilde{J}_i$ is lower semicontinuous on W. Take an arbitrary sequence $(\xi_j, w_j, v_i) \to (\xi, w, v)$ in $(W, \|\cdot\|_k)$. Write $x_j = x_{\xi_j, v_j}$. Then $x_j \to x_{\xi,v}$ in $W^{1,1}$. Restricting attention to a subsequence, we have $w_j(t) \to w(t)$ and $v_j(t) \to v(t)$ a.e. So $|v(t)| \le K$ and $\|x_{\xi,v}\|_{W^{1,1}} \le \eta$. The limit point (ξ, w, v) then satisfies the conditions confirming membership of W, so W is strongly closed. This establishes that $(W, \|\cdot\|_k)$ is complete.

Next we show that $\tilde{J}_i$ is lower semicontinuous. Again take an arbitrary sequence $(\xi_j, w_j, v_j) \to (\xi, w, v)$ in $(W, \|\cdot\|_k)$. By hypothesis (H3′), the sequence $\tilde{L}(t, w_j(t), v_j(t)) + k(t)|w_j(t) - w(t)|$ is bounded below by the integrable function $-k(t) - k(t)|w(t)|$. We may therefore deduce from the lower semicontinuity of $\tilde{L}$ and Fatou's lemma that

$$\begin{aligned}\liminf_{j\to\infty} \int_0^1 \tilde{L}(t, w_j(t), v_j(t))dt &= \liminf_{j\to\infty} \int_0^1 (\tilde{L}(t, w_j(t), v_j(t)) + k(t)|w_j(t) - w(t)|)dt \\ &\ge \int_0^1 \liminf_{j\to\infty} (\tilde{L}(t, w_j(t), v_j(t)) + k(t)|w_j(t) - w(t)|)dt \\ &\ge \int_0^1 \tilde{L}(t, w(t), v(t))dt.\end{aligned}$$

Since $x_j \to x_{\xi,v}$ uniformly, lower semicontinuity of $\tilde{l}$ gives

$$\liminf_{j\to\infty} \tilde{l}(x_j(0), x_j(1)) \ge \tilde{l}(x_{\xi,v}(0), x_{\xi,v}(1)).$$

It follows that

$$\begin{aligned}\liminf_{j\to\infty} \tilde{J}_i(\xi_j, w_j, v_j) &\ge \liminf_{j\to\infty} \tilde{l}(x_j(0), x_j(1)) + \liminf_{j\to\infty} \int_0^1 \tilde{L}(t, w_j(t), v_j(t))dt \\ &\quad + \liminf_{j\to\infty} \int_0^1 \epsilon_i^{-1} k(t)|x_j(t) - w_j(t)|^2 dt \\ &\ge \tilde{l}(x_{\xi,v}(0), x_{\xi,v}(1)) + \int_0^1 \tilde{L}(t, w(t), v(t))dt \\ &\quad + \int_0^1 \epsilon_i^{-1} k(t)|x_{\xi,v}(t) - w(t)|^2 dt \\ &= \tilde{J}_i(\xi, w, v).\end{aligned}$$

We conclude that $\tilde{J}_i$ is lower semicontinuous.

Define

$$\alpha_i^2 := \tilde{J}_i(0,0,0) - \inf_W \tilde{J}_i(\xi, w, v).$$

Since $(0,0,0) \in W$, $\alpha_i^2 \ge 0$. For arbitrary $(\xi, w, v) \in W$,

$$\begin{aligned}\tilde{J}_i(\xi, w, v) &= \tilde{J}_i(\xi, x_{\xi,v}, v) + \int (\tilde{L}(t, w(t), v(t)) - \tilde{L}(t, x_{\xi,v}(t), v(t)))dt + \epsilon_i^{-1} c^2 \\ &\ge \tilde{J}_i(0,0,0) - \int k(t)|x_{\xi,v}(t) - w(t)|dt + \epsilon_i^{-1} c^2\end{aligned}$$

(by the minimizing properties of $(0,0,0)$)

$$\begin{aligned}
&\geq \tilde{J}_i(0,0,0) - cd + \epsilon_i^{-1}c^2 \\
&= \tilde{J}_i(0,0,0) + \epsilon_i^{-1}(c - \epsilon_i d/2)^2 - \epsilon_i d^2/4 \\
&\geq \tilde{J}_i(0,0,0) - \epsilon_i d^2/4,
\end{aligned}$$

where $c^2 := \int_0^1 k(t)|w(t) - x_{\xi,v}(t)|^2 dt$ and $d^2 := \int_0^1 k(t)dt$. So we have

$$0 \leq \alpha_i^2 \leq \epsilon_i \int_0^1 k(t)dt/4.$$

Since the right side converges to 0 as $i \to \infty$, the lemma is proved. □

We have shown that for each i, $(0,0,0)$ is an "α_i^2 minimizer" for $\tilde{J}_i$ over W. By Ekeland's variational principle there exists $(\xi_i, w_i, v_i) \in W$ which minimizes

$$J_i(\xi, w, v) := \tilde{J}_i(\xi, w, v) + \alpha_i \|(\xi, w, v) - (\xi_i, w_i, v_i)\|_k$$

over W. Also

$$\|(\xi_i, w_i, v_i)\|_k \leq \alpha_i.$$

Write $x_i = x_{\xi_i, v_i}$. This last property implies that, for some subsequence, $(w_i, v_i) \to 0$ in L^1 and a.e. and $x_i \to 0$ uniformly.

Since $\tilde{J}_i(0,0,0) + \alpha_i\|(\xi_i, w_i, v_i)\|_k \geq \tilde{J}_i(\xi_i, w_i, v_i)$, we have

$$\begin{aligned}
\tilde{J}_i(0,0,0) &= \tilde{l}(0,0) + \int_0^1 \tilde{L}(t,0,0)dt \\
&\geq \limsup_{i\to\infty} \tilde{J}_i(\xi_i, w_i, v_i) \\
&\geq \limsup_{i\to\infty} \tilde{l}(x_i(0), x_i(1)) + \liminf_{i\to\infty} \int_0^1 \tilde{L}(t, w_i(t), v_i(t))dt
\end{aligned}$$

But $\liminf_{i\to\infty} \tilde{l}(x_i(0), x_i(1)) \geq \tilde{l}(0,0)$ and $\liminf_{i\to\infty} \int_0^1 \tilde{L}(t, w_i(t), v_i(t))dt \geq \int_0^1 \tilde{L}(t,0,0)dt$. It follows from these relationships that

$$\lim_{i\to\infty} \tilde{l}(x_i(0), x_i(1)) = \tilde{l}(0,0)$$

and

$$\liminf_{i\to\infty} \int_0^1 (\tilde{L}(t, w_i(t), v_i(t)) - \tilde{L}(t,0,0))dt = 0.$$

But as before we can use Fatou's lemma to deduce that

$$\int_0^1 (\liminf_{i\to\infty} \tilde{L}(t, w_i(t), v_i(t)) - \tilde{L}(t,0,0))dt \leq \liminf_{i\to\infty} \int_0^1 (\tilde{L}(t, w_i(t), v_i(t)) - \tilde{L}(t,0,0))dt = 0.$$

Since $\liminf_{i\to\infty} \tilde{L}(t, w_i(t), v_i(t)) \geq \tilde{L}(t,0,0)$ a.e. We conclude that

$$\liminf_{i\to\infty} \tilde{L}(t, w_i(t), v_i(t)) = \tilde{L}(t,0,0) \quad \text{a.e.}$$

We pause to sharpen this last relationship.

LEMMA 5. *We may arrange by subsequence extraction that*

$$\lim_{i\to\infty} \tilde{L}(t, w_i(t), v_i(t)) = \tilde{L}(t,0,0) \quad a.e.$$

Proof. Write $\Delta_i(t) := \tilde{L}(t, w_i(t), v_i(t)) - \tilde{L}(t, 0, 0) + k(t)|w_i(t)|$. Note that the functions $\Delta_i(t)$ are bounded below by an integral function $k_1(t) := -k(t) - \tilde{L}(t, 0, 0)$. This fact is required for application of Fatou's lemma below. Since $kw_i \to 0$ in L^1 and a.e., we have from the preceding analysis that $\liminf_{i\to\infty} \Delta_i(t) \to 0$ a.e. After extracting a subsequence we also have

$$\lim_{i\to\infty} \int_0^1 (\tilde{L}(t, w_i(t), v_i(t)) - \tilde{L}(t, 0, 0))dt = 0.$$

This implies that $\int_0^1 \Delta_i(t)dt \to 0$ as $i \to \infty$.

We show $\Delta_i \to 0$ in measure. This will imply that, for a subsequence, $\Delta_i(t) \to 0$ a.e.

Suppose that it is not true; then there exist two positive numbers ϵ and δ and a subsequence of $\{\Delta_i\}$ such that

$$m(\{t : |\Delta_i(t)| > \epsilon\}) > \delta, \tag{10}$$

where m is Lebesgue measure on $[0, 1]$. Write $A_i^\epsilon = \{t : \Delta_i(t) > \epsilon\}$ and $B_i^r = \{t : \Delta_i(t) < -r\}$. Here r is a positive number whose value will be set presently. Note that

$$\liminf_{i\to\infty} \Delta_i(t)\chi_{B_i^r}(t) = 0,$$

where $\chi_{B_i^r}(t)$ equals 1 if $t \in B_i^r$ and 0 otherwise. We have that

$$\begin{aligned}\liminf_{i\to\infty}(-rm(B_i^r)) &\geq \liminf_{i\to\infty} \int_0^1 \Delta_i(t)\chi_{B_i^r}(t)dt \\ &\geq \int_0^1 \liminf_{i\to\infty} \Delta_i(t)\chi_{B_i^r}(t)dt = 0.\end{aligned}$$

Hence $\limsup_{i\to\infty} m(B_i^r) = 0$. By (10) however, $m(A_i^\epsilon) > \delta$ for i sufficiently large. Now choose $r > 0$ and an integer N such that

$$r + \int_{B_i^r} k_1(t)dt < \epsilon\delta/2 \quad \text{for} \quad i \geq N.$$

We have

$$\begin{aligned}\int_0^1 \Delta_i(t)dt &= \int_{A_i^\epsilon} \Delta_i(t)dt + \int_{B_i^r} \Delta_i(t)dt + \int_{\{-r\leq\Delta_i(t)\leq\epsilon\}} \Delta_i(t)dt \\ &\geq \epsilon\delta - \int_{B_i^r} k_1(t)dt - r > \epsilon\delta/2 \quad \text{for} \quad i \geq N.\end{aligned}$$

This contradicts $\int_0^1 \Delta_i(t)dt \to 0$ as $i \to \infty$. So $\Delta_i \to 0$ in measure. It follows that $\Delta_i(t) \to 0$ a.e. along a subsequence. □

We can summarize the above discussion in control theoretic terms. Define

$$\begin{aligned}\tilde{L}_i(t, x, w, v) &:= \tilde{L}(t, w, v) + \alpha_i k(t)|w - w_i(t)| + \alpha_i k(t)|v - v_i(t)| + \epsilon_i^{-1} k(t)|x - w|^2, \\ \tilde{l}_i(x, y) &:= \tilde{l}(x, y) + \alpha_i|x - x_i(0)|.\end{aligned}$$

The minimizing property of (ξ_i, v_i, w_i) can be expressed as follows: $((x_i, y_i, z_i), (v_i, w_i))$ is a minimizer for the problem

$$\text{minimize } z(1) + \int_0^1 \tilde{L}_i(t, x(t), w(t), v(t))dt$$

subject to

$$\begin{cases} \dot{x}(t) = v(t), \dot{z}(t) = 0, \\ w(t) \in R^n, v(t) \in KB, \\ (x(0), x(1), z(1)) \in \text{epi}\{\tilde{l}_i\} \end{cases}$$

and

$$\begin{cases} \dot{y}(t) = |v(t)|, \\ y(0) = 0,\ |x(0)| + y(1) \le \eta. \end{cases}$$

Here $x_i(t) := x_{\xi_i, v_i}(t)$, $y_i(t) := \int_0^t |v_i(s)|ds$, and $z_i(t) := \tilde{l}_i(x_i(0), x_i(1))$. We have shown $v_i, w_i \to 0$ in L^1 and a.e., $x_i, y_i \to 0$ uniformly and

$$\tilde{L}(t, w_i(t), v_i(t)) \to \tilde{L}(t, 0, 0),$$
$$\tilde{l}(x_i(0), x_i(1)) \to \tilde{l}(0, 0).$$

This is an optimal control problem to which the maximum principle in [2, Thm. 5.2.1] is applicable, with transversality conditions refined as indicated in [9]. We observe that the differential equation constraint has a right side which is independent of the state variable. Also, since $k(t)|x_i(t) - w_i(t)|^2$ is an L^1 function there exist a function $c : [0, 1] \times R^n \times R^n \to R^+$ and $\eta > 0$ such that (a) the cost integrand $x \to \tilde{L}_i(t, x, v, w)$ is Lipschitz continuous on $x_i(t) + \eta B$ with rank $c(t, v, w)$ for all $v \in KB$, $w \in R^n$ and a.e. t and (b) $c(t, x_i(t), v_i(t), w_i(t))$ is integrable in accordance with the Lipschitz continuity hypotheses which must be checked for application of [2, Thm. 5.2.1] with modified transversality condition. We take advantage then of the unrestrictive nature of the hypotheses under which this version of the maximum principle applies. (The hypothesis that $c(t, x, v, w) = \tilde{c}(t)$ for some integrable function $\tilde{c}$, invoked elsewhere in the necessary conditions literature, is violated.)

Notice that because the right endpoint constraint on y is inactive at $y \equiv y_i$ and because of the "decoupled" structure of the cost and dynamics in y and (x, z), the costate arc component associated with y must be zero; we therefore drop it from the relationships. The optimality conditions tell us that there exist $p_i \in W^{1,1}$ and $\lambda_i \ge 0$ such that

(A) $-\dot{p}_i(t) = -2\lambda_i \epsilon_i^{-1} k(t)(x_i(t) - w_i(t))$,
(B) $(p_i(0), -p_i(1), -\lambda_i) \in N_{\text{epi}\{\tilde{l}_i\}}(x_i(0), x_i(1), \tilde{l}_i(x_i(0), x_i(1)))$,
(C) $(w, v) \mapsto p_i(t)v - \lambda_i \tilde{L}_i(t, x_i(t), w, v)$ achieves its maximum at $(w_i(t), v_i(t))$ over all $(w, v) \in R^n \times KB$ for almost every t,
(D) $\|p_i\|_\infty + \lambda_i = 1$.

Condition (C) implies

$$(\dot{p}_i(t), p_i(t)) \in \lambda_i \partial[\tilde{L}(t, w_i(t), v_i(t)) + \Psi_{KB}(v_i(t))] + \lambda_i \alpha_i k(t)(B \times B). \tag{11}$$

Fix $v = v_i(t)$, then $w \mapsto p_i(t)v_i(t) - \lambda_i \tilde{L}_i(t, x_i(t), w, v_i(t))$ achieves its maximum at $w_i(t)$ over all $w \in R^n$. This implies

$$\dot{p}_i(t) \in \lambda_i \partial_w \tilde{L}(t, w_i(t), v_i(t)) + \lambda_i \alpha_i k(t) B. \tag{12}$$

Fix $w = w_i(t)$, then $v \mapsto p_i(t)v - \lambda_i \tilde{L}_i(t, x_i(t), w_i(t), v)$ achieves its maximum at $v_i(t)$ over $v \in KB$. This implies

(13)
$$p_i(t)(v - v_i(t)) \le \lambda_i \tilde{L}(t, w_i(t), v) - \lambda_i \tilde{L}(t, w_i(t), v_i(t)) + \lambda_i \alpha_i k(t)|v - v_i(t)| \quad \forall v \in KB \text{ a.e.}$$

Since $\tilde{L}(t, \cdot, v)$ is Lipschitz with rank $k(t)$ for all $v \in KB$, (12) implies $|\dot{p}_i(t)| \le 2k(t)$. Since the p_i's are uniformly bounded (see (D)), we can arrange, by limiting attention to a subsequence, that $p_i \to p$ uniformly and $\dot{p}_i \to \dot{p}$ weakly in L^1 for some $p \in W^{1,1}$. We can also ensure that $\lambda_i \to \lambda$ for some $\lambda \ge 0$ such that

$$\|p\|_{L^\infty} + \lambda = 1. \tag{14}$$

(13) implies in the limit that

$$p(t)v \le \lambda \tilde{L}(t, 0, v) - \lambda \tilde{L}(0, 0, 0) \text{ a.e. for } v \in KB.$$

Observe now that $\lambda > 0$, since otherwise this last relation implies $p(t) \equiv 0$, which contradicts (D). Since $\lambda > 0$, (B) implies

$$(p_i(0), -p_i(1)) \in \lambda_i \partial \tilde{l}_i(x_i(0), x_i(1)),$$

and we conclude that

$$(p(0), -p(1)) \in \lambda \partial \tilde{l}(0, 0).$$

We next verify the Euler–Lagrange inclusion in the limit.

By Mazur's theorem there exists for each i an integer $N_i \ge i$ and a convex combination $\{\lambda_{i1}, \ldots, \lambda_{iN_i}\}$ such that if we write

$$q_i(t) = \sum_{j=i}^{N_i} \lambda_{ij} \dot{p}_j(t),$$

then

$$q_i(t) \to \dot{p}(t) \quad \text{strongly in } L^1.$$

Appealing to Carathéodory's theorem, we deduce that for each i and t there exists a convex combination $\{\alpha_{i0}(t), \ldots, \alpha_{in}(t)\}$ and integers $0 \le k_{i0}(t) < \cdots < k_{in}(t)$ such that

$$q_i(t) = \sum_{j=0}^{n} \alpha_{ij}(t) \dot{p}_{i+k_{ij}(t)}(t).$$

A subsequence $\{q_i\}_{i \in S}$ can be chosen (S denotes the index values which are retained) such that

$$q_i(t) \xrightarrow{S} \dot{p}(t) \quad \text{a.e.}$$

Write $A \subset [0,1]$ for the set of full measure:

$$A := \{t : q_i(t) \overset{S}{\to} \dot{p}(t), |\dot{p}_i(t)| \leq 2k(t) \forall i, k(t) < \infty \text{ and } (w_i(t), v_i(t)) \to (0,0)\}.$$

Fix $t \in A$. For any $j \in \{0, \ldots, n\}$ we note that $\{\alpha_{ij}(t)\}_{i=1,2,\ldots}$ and $\{\dot{p}_{i+k_{ij}(t)}(t)\}_{i=1,2,\ldots}$ are bounded sequences. Consequently we may choose subsequences with index values the set $S' \subset S$ such that as $i \overset{S'}{\to} \infty$,

$$\alpha_{ij}(t) \to \alpha_j(t) \text{ and } \dot{p}_{i+k_{ij}}(t) \to q_j(t) \text{ for } j = 0, \ldots, n$$

for some convex combination $\{\alpha_j(t) : j = 0, \ldots, n\}$ and $q_j(t) \in R^n$. Since $(w_i(t), v_i(t)) \to (0,0)$, $\tilde{L}_i(t, w_i(t), v_i(t)) + \Psi_{KB}(v_i(t)) \to \tilde{L}(t,0,0)$ and $p_i(t) \to p(t)$ as $i \to \infty$, we deduce from (11) that

$$(q_j(t), p(t)) \in \lambda \partial \tilde{L}(t,0,0) \text{ for } j = 0, \ldots, n.$$

It follows that

$$\dot{p}(t) = \sum_{j=0}^{n} \alpha_j q_j(t) \in \text{co}\{\eta : (\eta, p(t)) \in \lambda \partial \tilde{L}(t,0,0)\} \quad \text{a.e.}$$

These are precisely the assertions of Theorem 3, except that they are expressed in terms of a cost multiplier λ which is possibly not equal to 1 and that, in the last condition, the inequality holds only for $v \in \dot{\bar{x}}(t) + KB$.

Take $K_i \to \infty$. Let p_i denote the adjoint arc and $\lambda_i > 0$ the cost multiplier when $K = K_i$. We deduce from the Euler condition that the $\dot{p}_i$'s are uniformly integrably bounded. Of course the p_i's are uniformly bounded. Now extract subsequences to arrange that p_i converges uniformly to some limit p, $\dot{p}_i$ converges weakly to $\dot{p}$, and $\lambda_i \to \lambda$ for some λ such that $\|p\|_{L^\infty} + \lambda = 1$. Arguing as before, we arrive at our earlier conclusions, but the Weierstrass condition is now satisfied globally. From the Weierstrass condition, however, $\lambda = 0$ implies $p(t) \equiv 0$, which is not possible. So $\lambda > 0$. The final touch is to scale p and λ so that $\lambda = 1$. The theorem is proved. □

3. Problems with dynamic and endpoint constraints. We now derive necessary conditions for a version of (P) which allow for dynamic constraints $\dot{x} \in F$, with possibly $F \neq R^n$, and endpoint constraints. The problem (which is labeled (Q)) is

$$\text{(Q)} \qquad \begin{aligned} &\text{minimize } l(x(0), x(1)) + \int_0^1 L(t, x(t), \dot{x}(t)) dt \\ &\text{over arcs } x \in W^{1,1} \text{ which satisfy} \\ &\dot{x}(t) \in F(t, x(t)), \quad (x(0), x(1)) \in C. \end{aligned}$$

$l : R^n \times R^n \to R$ and $L : [0,1] \times R^n \times R^n \to R$ are given functions. $F : [0,1] \times R^n \to R^n$ is a given multifunction. $C \subset R^n \times R^n$ is a given closed set.

Notice that in this formulation l is everywhere finite. Endpoint constraints are specified directly via the constraint set C rather than implicitly in terms of the effective domain of l.

The hypotheses which will be invoked are now listed. They involve $\delta \in (0, \infty)$ and $k_l \in (0, \infty)$ (for convenience we assume $k_l \geq 1$) and nonnegative, measurable functions k_F and k_L which satisfy

$$k_F \in L^1, \quad k_L \in L^1, \quad k_F k_L \in L^1.$$

(G1) $|l(x,y)-l(x',y')| \le k_l|(x,y)-(x',y')| \forall (x,y),(x',y') \in (\bar{x}(0),\bar{x}(1))+\delta(B\times B)$.

(G2) $F(t,x)$ is nonempty and closed for each (t,x). $F(t,x)$ is measurable in t for fixed x and

$$F(t,x') \subset F(t,x) + k_F(t)|x'-x|B \quad \forall x,x' \in \bar{x}(t)+\delta B \text{ a.e.}$$

(G3) $L(t,x,v)$ is measurable in t for fixed (x,v) and $|L(t,x,v) - L(t,x',v')| \le k_L(t)|(x,v)-(x',v')| \forall (x,v),(x',v') \in (\bar{x}(t)+\delta B)\times R^n$.

THEOREM 6. *Let $\bar{x}$ be a local minimizer for* (Q)*, for which hypotheses* (G1)–(G3) *above are satisfied for some $\delta > 0$. Then there exist $p \in W^{1,1}$ and $\lambda \ge 0$, not both zero, satisfying*

(i) *the Euler condition:*

$$\dot{p}(t) \in co\{\eta : (\eta, p(t)) \in \lambda \partial L(t,\bar{x}(t),\dot{\bar{x}}(t)) + N_{\mathrm{Gr}\{F(t,\cdot)\}}(\bar{x}(t),\dot{\bar{x}}(t))\} \quad a.e.;$$

(ii) *the transversality condition:*

$$(p(0),-p(1)) \in \lambda \partial l(\bar{x}(0),\bar{x}(1)) + N_C(\bar{x}(0),\bar{x}(1));$$

(iii) *the Weierstrass condition:*

$$p(t)\dot{\bar{x}}(t) - \lambda L(t,\bar{x}(t),\dot{\bar{x}}(t)) \ge p(t)v - \lambda L(t,\bar{x}(t),v) \forall v \in F(t,\bar{x}(t)) \quad a.e.$$

We begin by proving the theorem under the condition $L = 0$. We may assume without loss of generality that F is globally Lipschitz continuous in x (i.e., the first condition in hypothesis (G2) is satisfied for all $x,x' \in R^n \times R^n$.) This is because we can derive necessary conditions for the "truncated" differential inclusion

$$\tilde{F}(t,x) := \begin{cases} F(t,x) & \text{if } x \in \bar{x}(t)+\delta B, \\ F\left(t, \bar{x}(t) + \delta \frac{x-\bar{x}(t)}{|x-\bar{x}(t)|}\right) & \text{otherwise,} \end{cases}$$

which is globally Lipschitz continuous, and the necessary conditions are the same.

We may arrange, by reducing δ if necessary, that $\bar{x}$ is a minimizer with respect to competing arcs which satisfy $\|x-\bar{x}\|_{W^{1,1}} \le \delta/2$.

Denote by W the subset of $R^n \times W^{1,1}$:

$$W = \{(e,x) \in R^n \times W^{1,1} : (x(0),e) \in C, \dot{x}(t) \in F(t,x(t)), \|x-\bar{x}\|_{W^{1,1}} \le \delta/2\}.$$

Here W is equipped with norm

$$\|(e,x)\|_W := |e| + |x(0)| + \|\dot{x}\|_{L^1}.$$

We must show that $(W, \|\cdot\|_W)$ is complete. Consider a Cauchy sequence (e_j,x_j) in $(W,\|\cdot\|_W)$. Then $e_j \to e$, $x_j(0) \to \eta$, and $\dot{x}_j \to \xi$ in L^1 for some $e \in R^n$, $\eta \in R^n$, and $\xi \in L^1$. We see that $x_j(t) - \int_0^t \xi(s)ds - \eta = \int_0^t (\dot{x}_j(s) - \xi(s))ds + (x_j(0)-\eta)$ tends to zero uniformly in t. So in fact $x_j(t)$ converges in $W^{1,1}$ to the absolutely continuous function $x(t) := \eta + \int_0^t \xi(s)ds$. It follows we can identify $\xi(t)$ with $\dot{x}(t)$ and $(e,x(0)) \in C$ since C is closed. Along a subsequence then $\dot{x}_j \to \dot{x}$ a.e. By the continuity of F in x, $\dot{x}(t) \in F(t,x(t))$ a.e. we see that $(e,x) \in W$, so $(W,\|\cdot\|_W)$ is complete.

For each i define

$$l_i(x,y,x',y') := \max\{l(x,y) - l(\bar{x}(0),\bar{x}(1)) + \epsilon_i^2, |x'-y'|\}.$$

Notice that l_i is Lipschitz continuous with rank at most k_l on $\delta(B \times B) \times \delta(B \times B)$. (We use here $k_l \geq 1$.)

Now consider the minimization problem

$$\text{minimize } \{l_i(x(0), e, x(1), e) : (e, x) \in W\}.$$

The functional $(e, x) \to l_i(x(0), e, x(1), e)$ is continuous on $(W, \|\cdot\|_W)$. Notice also that l_i is nonnegative valued and

$$l_i(\bar{x}(0), \bar{x}(1), \bar{x}(1), \bar{x}(1)) = \epsilon_i^2.$$

So $(\bar{x}(1), \bar{x})$ is an "ϵ_i^2 minimizer." According to Ekeland's principle, there exists $(e_i, x_i) \in W$ such that

$$l_i(x_i(0), e_i, x_i(1), e_i) \leq l_i(x(0), e, x(1), e) + \epsilon_i \|(e, x) - (e_i, x_i)\|_W \tag{15}$$

for all $(e, x) \in W$ and

$$\|(e_i, x_i) - (\bar{x}(1), \bar{x})\|_W \leq \epsilon_i \tag{16}$$

for all i. Take y_i to be the constant arc

$$y_i(t) \equiv e_i.$$

(16) implies that $y_i(1) \to \bar{x}(1)$, $x_i(0) \to \bar{x}(0)$, and also (following extraction of subsequences) $\dot{x}_i \to \dot{\bar{x}}$ both in L^1 and a.e. We have then that $x_i \to \bar{x}$. We can arrange, by eliminating initial terms of the sequence, that $\|x_i - \bar{x}\|_{W^{1,1}} < \delta/2$ for all i; i.e., the state constraint $\|x - \bar{x}\|_{W^{1,1}} \leq \delta/2$ implicit in the definition of W is nonactive and can be disregarded since we are interested only in local minimizing properties of (e_i, x_i).

Define

$$\tilde{l}_i(x, y, x', y') := l_i(x, y, x', y') + \epsilon_i |x - x_i(0)| + \epsilon_i |y - y_i(0)|.$$

(15) implies that for each i the arc (x_i, y_i) is a local minimizer for the variational problem:

$$\begin{aligned} &\text{minimize } J_i(x, y) := \tilde{l}_i(x(0), y(0), x(1), y(1)) + \epsilon_i \int_0^1 |\dot{x}(t) - \dot{x}_i(t)| dt \\ &\text{over arcs } (x, y) \in W^{1,1} \text{ which satisfy} \\ &(\dot{x}(t), \dot{y}(t)) \in F(t, x(t)) \times \{0\} \quad \text{a.e.}, \quad (x(0), y(0)) \in C. \end{aligned} \tag{17}$$

The following lemma justifies replacing (17) by another variational problem in which the dynamic constraint is represented by a term in the cost function ("exact penalization," cf. [2, Chap. 3]). This involves the function

$$\rho(t, x, y, v, w) := d_{F(t,x) \times \{0\}}(v, w) = d_{F(t,x)}(v) + |w|.$$

LEMMA 7. *For each i, (x_i, y_i) is a local minimizer for*

$$\begin{aligned} &\textit{minimize } J_i(x, y) + M \int_0^1 \rho(t, x(t), y(t), \dot{x}(t), \dot{y}(t)) dt \\ &\textit{over arcs } (x, y) \textit{ which satisfy } (x(0), y(0)) \in C. \end{aligned}$$

Here M is any constant satisfying $M > (k_l + 3)K$, where $K = \exp(\int_0^1 k_F(t) dt)$.

Proof. Suppose that the assertions of Lemma 7 are false. Then there must exist a sequence $\{(\tilde{x}_j, \tilde{y}_j)\}$ in $W^{1,1}$ such that $\|(\tilde{x}_j, \tilde{y}_j) - (x_i, y_i)\|_{W^{1,1}} \to 0$ and, for each j, $(\tilde{x}_j(0), \tilde{y}_j(0)) \in C$ and

$$J_i(\tilde{x}_j, \tilde{y}_j) + M \int_0^1 \rho(t, \tilde{x}_j(t), \tilde{y}_j(t), \dot{\tilde{x}}_j(t), \dot{\tilde{y}}_j(t))dt < J_i(x_i, y_i).$$

Since $J_i(\tilde{x}_j, \tilde{y}_j) \geq l_i(\tilde{x}_j(0), \tilde{y}_j(0), \tilde{x}_j(1), \tilde{y}_j(1))$ for each j and the right side of this inequality tends to $J_i(x_i, y_i)$ as $j \to \infty$, we deduce that

$$\int_0^1 \rho(t, \tilde{x}_j(t), \tilde{y}_j(t), \dot{\tilde{x}}_j(t), \dot{\tilde{y}}_j(t))dt \to 0.$$

For sufficiently large j, the successive approximations theorem [2, Thm. 3.1.6] yields an arc $(\hat{x}_j, \hat{y}_j)$ satisfying the differential inclusion and endpoint constraints of problem (17) and for which

$$(\hat{x}_j(0), \hat{y}_j(0)) = (\tilde{x}_j(0), \tilde{y}_j(0))$$

and

$$\begin{aligned}\|(\hat{x}_j, \hat{y}_j) - (\tilde{x}_j, \tilde{y}_j)\|_{W^{1,1}} &\leq \int_0^1 |(\dot{\hat{x}}_j(t), \dot{\hat{y}}_j(t)) - (\dot{\tilde{x}}_j(t), \dot{\tilde{y}}_j(t)|dt \\ &\leq K \int_0^1 \rho(t, \tilde{x}_j(t), \tilde{y}_j(t), \dot{\tilde{x}}_j(t), \dot{\tilde{y}}_j(t))dt.\end{aligned}$$

Since $(\tilde{x}_j, \tilde{y}_j) \to (x_i, y_i)$ uniformly, those relationships tell us that $(\hat{x}_j, \hat{y}_j) \to (x_i, y_i)$ uniformly. For sufficiently large j therefore we have from the fact that (x_i, y_i) is a local minimizer for (17),

$$J_i(x_i, y_i) \leq J_i(\hat{x}_j, \hat{y}_j).$$

Also, we can deduce from the Lipschitz continuity properties of the data that

$$\begin{aligned}J_i(\hat{x}_j, \hat{y}_j) &\leq J_i(\tilde{x}_j, \tilde{y}_j) + (k_l + 3\epsilon_i)K \int_0^1 \rho(t, \tilde{x}_j(t), \tilde{y}_j(t), \dot{\tilde{x}}_j(t), \dot{\tilde{y}}_j(t))dt \\ &\leq J_i(\tilde{x}_j, \tilde{y}_j) + M \int_0^1 \rho(t, \tilde{x}_j(t), \tilde{y}_j(t), \dot{\tilde{x}}_j(t), \dot{\tilde{y}}_j(t))dt \\ &< J_i(x_i, y_i).\end{aligned}$$

This contradicts the preceding inequality, and Lemma 7 is proved. □

LEMMA 8. *For any $x, v \in R^n \times R^n$ and $t \in [0, 1]$*

$$d_{F(t,x)}(v) \leq (1 + k_F(t))d_{\mathrm{Gr}\{F(t,\cdot)\}}(x, v).$$

Proof. Because $F(t, \cdot)$ has closed values and is assumed to be globally Lipschitz continuous, $\mathrm{Gr}\{F(t, \cdot)\}$ is closed and there exists $(y, w) \in \mathrm{Gr}\{F(t, \cdot)\}$ such that $d_{\mathrm{Gr}\{F(t,\cdot)\}}(x, v) = |(x, v) - (y, w)|$. By the Lipschitz continuity of $F(t, \cdot)$, however,

$$\begin{aligned}d_{F(t,x)}(v) &\leq d_{F(t,y)}(v) + d_{\mathrm{Haus}}(F(t, x), F(t, y)) \\ &\leq |v - w| + k_F(t)|x - y| \\ &\leq (1 + k_F(t))|(x, v) - (y, w)| \\ &\leq (1 + k_F(t))d_{\mathrm{Gr}\{F(t,\cdot)\}}(x, v).\end{aligned}$$

(Here $d_{\text{Haus}}(F_1, F_2)$ is the Hausdorff distance function.) This result is the desired inequality. □

Lemma 8 tells us that the penalty term $M\int_0^1 \rho(t,x,y,\dot{x},\dot{y})dt$ satisfies

$$M\int_0^1 \rho(t,x,y,\dot{x},\dot{y})dt \le M\int_0^1 ((1+k_F)d_{\mathrm{Gr}\{F(t,\cdot)\}}(x,\dot{x}) + |\dot{y}|)dt$$

for all $(x,y) \in W^{1,1}$. Of course, because $(\dot{x}_i, \dot{y}_i) \in F(t,x_i) \times \{0\}$,

$$0 = M\int_0^1 \rho(t,x_i,y_i,\dot{x}_i,\dot{y}_i)dt = M\int_0^1 (1+k_F)d_{\mathrm{Gr}\{F(t,\cdot)\}}(x_i,\dot{x}_i)dt.$$

These relationships combine with the assertions of the lemma to give the following: for each i, (x_i, y_i) is a local minimizer for

$$\text{minimize } J_i(x,y) + M\int_0^1 ((1+k_F)d_{\mathrm{Gr}\{F(t,\cdot)\}}(x,\dot{x}) + |\dot{y}|)dt$$
$$\text{over arcs } (x,y) \in W^{1,1} \text{ which satisfy } (x(0),y(0)) \in C.$$

Our findings up to this point can be expressed as follows: for each i, (x_i, y_i) is a minimizer for

$$\text{minimize } \tilde{l}_i(x(0),y(0),x(1),y(1)) + \int_0^1 \tilde{L}_i(t,x(t),y(t),\dot{x}(t),\dot{y}(t))dt$$
$$\text{over arcs } (x,y) \text{ which satisfy } (x(0),y(0)) \in C.$$

Here, we recall

$$\tilde{l}_i(x,y,x',y') := \max\{l(x,y) - l(\bar{x}(0),\bar{x}(1)) + \epsilon_i^2, |x'-y'|\} + \epsilon_i|x - x_i(0)| + \epsilon_i|y - y_i(0)|,$$

and $\tilde{L}_i$ is taken to be the function

$$\tilde{L}_i(t,x,y,v,w) := \epsilon_i|v - \dot{x}_i(t)| + M(1+k_F(t))d_{\mathrm{Gr}\{F(t,\cdot)\}}(x,v) + M|w|.$$

The hypotheses under which Theorem 3 applies are satisfied, to give the following information about the minimizer (x_i, y_i). (See Remark after Theorem 3.) There exist arcs p_i and q_i such that

(A′) $(\dot{p}_i(t), \dot{q}_i(t)) \in \mathrm{co}\{(\eta,\xi) : (\eta,\xi,p_i(t),q_i(t)) \in \partial\tilde{L}_i(t,x_i(t),y_i(t),\dot{x}_i(t),0)\}$;

(B′) $(p_i(0),q_i(0),-p_i(1),-q_i(1)) \in \partial\tilde{l}_i(x_i(0),y_i(0),x_i(1),y_i(1)) + N_C(x_i(0),y_i(0)) \times \{(0,0)\}$;

(C′) $p_i(t)\dot{x}_i(t) \ge p_i(t)v + q_i(t)w - \tilde{L}_i(t,x_i(t),y_i(t),v,w) \quad \forall v,w \in R^n$ a.e.

Condition (A′) implies that $q_i(t)$ is a constant (we write it q_i); $|q_i| \le M$, $|\dot{p}_i(t)| \le M(1+k_F(t))$; and

$$\begin{aligned}\dot{p}_i(t) &\in \mathrm{co}\{\eta : (\eta,p_i(t)) \in \{0\} \times \epsilon_i B + M(1+k_F(t))\partial d_{\mathrm{Gr}\{F(t,\cdot)\}}(x_i(t),\dot{x}_i(t))\} \\ &\subseteq \mathrm{co}\{\eta : (\eta,p_i(t)) \in \{0\} \times \epsilon_i B + N_{\mathrm{Gr}\{F(t,\cdot)\}}(x_i(t),\dot{x}_i(t))\}.\end{aligned} \tag{18}$$

From (B′),

$$\begin{aligned}(p_i(0),q_i,-p_i(1),-q_i) \in \partial l_i(x_i(0),y_i(0),x_i(1),y_i(1)) \\ + \epsilon_i(B\times B)\times\{(0,0)\} + N_C(x_i(0),y_i(0))\times\{(0,0)\}.\end{aligned} \tag{19}$$

We observe that $l_i(x_i(0), y_i(0), x_i(1), y_i(1)) > 0$ for all i. Otherwise it is zero for some i which implies $y_i(1) = x_i(1)$, $(x_i(0), x_i(1)) \in C$ and $l(x_i(0), x_i(1)) \le l(\bar{x}(0), \bar{x}(1)) - \epsilon_i^2$. But then x_i would satisfy the constraints of (Q) and also $\|x_i - \bar{x}\|_{W^{1,1}} \le \delta/2$ and have lower cost than $\bar{x}$; this contradicts the optimality properties of $\bar{x}$.

It is now claimed that there exist $\lambda_i \ge 0$ and $e_i \in R^n$ such that $\lambda_i + |e_i| = 1$ and

$$\partial l_i(x_i(0), y_i(0), x_i(1), y_i(1)) \subseteq \lambda_i \partial l(x_i(0), y_i(0)) \times \{(0,0)\} + \{(0, 0, e_i, -e_i)\}. \tag{20}$$

Fix i. There are two cases to consider:

(i) $x_i(1) = y_i(1)$. In this case $l_i(x, y, x', y') = l(x, y) - l(\bar{x}(0), \bar{x}(1)) + \epsilon_i^2$ on a neighborhood of $(x_i(0), y_i(0), x_i(1), y_i(1))$, so

$$\partial l_i(x_i(0), y_i(0), x_i(1), y_i(1)) = \partial l(x_i(0), y_i(0)) \times \{(0,0)\}.$$

This implies (20) with $\lambda_i = 1$ and $e_i = 0$.

(ii) $x_i(1) \ne y_i(1)$. In this case the chain rule gives $\partial\{(x, y) \to |x - y|\}|_{x_i(1), y_i(1)} = (\tilde{e}_i, -\tilde{e}_i)$ for some unit vector $\tilde{e}_i$. We may therefore deduce from the calculus rules concerning function defined by the pointwise-maximum operation (see, e.g., [13, Thm. 6.9]) that, for some $\lambda_i \in [0, 1]$,

$$\partial l_i(x_i(0), y_i(0), x_i(1), y_i(1)) \subseteq \lambda_i \partial l(x_i(0), y_i(0)) \times \{(0,0)\} + (1 - \lambda_i)\{(0, 0, \tilde{e}_i, -\tilde{e}_i)\}.$$

Now set $e_i = (1 - \lambda_i)\tilde{e}_i$. We see that $\lambda_i + |e_i| = 1$ and (20) is satisfied.

We may now deduce from (19) that $-p_i(1) = e_i = q_i$, and

$$|p_i(1)| + \lambda_i = 1, \tag{21}$$

$$(p_i(0), -p_i(1)) \in \lambda_i \partial l(x_i(0), y_i(0)) + \epsilon_i(B \times B) + N_C(x_i(0), y_i(0)). \tag{22}$$

Condition (C′) implies

$$p_i(t)\dot{x}_i(t) \ge p_i(t)v - \epsilon_i|v - \dot{x}_i(t)| \quad \forall v \in F(t, x_i(t)). \tag{23}$$

Recall that $|\dot{p}_i(t)| \le M(1 + k_F(t))$ and $|p_i(1)| \le 1$. We may therefore arrange by subsequence extraction that $\dot{p}_i \to \dot{p}$ weakly in L^1 and $p_i \to p$ uniformly for some arc p. We have shown also that $\dot{x}_i \to \dot{\bar{x}}$ strongly in L^1 and a.e., $y_i(0) \to \bar{x}(1)$, and $x_i \to \bar{x}$ uniformly. Straightforward limit-taking arguments permit us now to deduce all the assertions of Theorem 6 from (18), (21)–(23). (Analysis of (18) in the limit is based on the upper semicontinuous properties of the limiting normal cone and an appeal to Carathéodory's theorem.)

This completes the proof in the case $L = 0$.

Suppose now the integral term is present in the cost of (Q). We reformulate (Q) as

$$\begin{aligned} &\text{minimize } l(x(0), x(1)) + z(1) \\ &\text{over arcs } (x, z) \text{ which satisfy} \\ &(\dot{x}(t), \dot{z}(t)) \in E(t, x(t)) \\ &(x(0), x(1)) \in C, z(0) = 0. \end{aligned} \tag{24}$$

Here

$$E(t, x) := \{(v, w) : v \in F(t, x), w \ge L(t, x, v)\}.$$

For reasons similar to those previously given, we may assume that the multifunctions $F(t,\cdot)$ and the functions $L(t,\cdot,\cdot)$ are Lipschitz continuous on R^n and $R^n \times (\dot{\bar{x}}(t) + rB)$ with Lipschitz ranks at most $k_F(t)$ and $k_L(t)$, respectively. We claim that

$$E(t,x) \subset E(t,x') + k(t)|x - x'|B \quad \forall x, x' \in R^n \text{ a.e.},$$

where $k(t) := k_F(t) + k_L(t) + k_F(t)k_L(t)$ which, under the hypotheses, is an integrable function. To verify this take any $x', x \in R^n$ and $(v', w') \in E(t,x')$. Then $v' \in F(t,x')$ and $w' \geq L(t,x',v')$. The Lipschitz continuity of $F(t,\cdot)$ implies the existence of some $v \in F(t,x)$ such that $|v' - v| \leq k_F(t)|x' - x|$. Define $w := L(t,x,v) - L(t,x',v') + w'$. Obviously, $(v,w) \in E(t,x)$. We have by the Lipschitz continuity properties of L

$$|w' - w| = |L(t,x',v') - L(t,x,v)| \leq k_L(t)|(x',v') - (x,v)|.$$

But then

$$\begin{aligned} |(v',w') - (v,w)| &\leq (k_F^2|x' - x|^2 + k_L^2(|x' - x|^2 + k_F^2|x' - x|^2))^{1/2} \\ &\leq (k_F + k_L + k_L k_F)|x' - x| = k(t)|x' - x|. \end{aligned}$$

Set $\bar{z}(t) = \int_0^t L(s,\bar{x}(s),\bar{x}'(s))ds$. Then, since $\bar{x}$ is a minimizer for (Q), $(\bar{x},\bar{z})$ is a minimizer for (24), a problem of the special kind ("$L \equiv 0$") for which necessary conditions have already been derived. We have already checked that $E(t,\cdot)$ satisfies the Lipschitz continuity hypothesis for application of Theorem 4, the other hypotheses on the data for (24) are easily verified. From Theorem 6 we know that there exist $p \in W^{1,1}$, $q \in R^n$, and $\lambda \geq 0$, not all zero, such that

(A″) $\dot{p}(t) \in \text{co}\{\eta : (\eta, p(t), q) \in N_{\text{Gr}\{E(t,\cdot)\}}(\bar{x}(t), \dot{\bar{x}}(t), L(t,\bar{x}(t),\dot{\bar{x}}(t)))\}$,

(B″) $(p(0), -p(1), q, -q) \in \lambda[\partial l(\bar{x}(0),\bar{x}(1)) \times \{(0,1)\}] + N_C(\bar{x}(0),\bar{x}(1)) \times R \times \{0\}$,

(C″) $p(t)\dot{\bar{x}}(t) + q\dot{\bar{z}}(t) \geq p(t)v + qw \quad \forall (v,w) \in E(t,\bar{x}(t))$.

Condition (B″) implies $q = -\lambda$, so in particular p and λ are not both zero, and

$$(p(0), -p(1)) \in \lambda \partial l(\bar{x}(0),\bar{x}(1)) + N_C(\bar{x}(0),\bar{x}(1)).$$

Now $\text{Gr}\{E(t,\cdot)\} = \text{epi}\{(x,v) \to \Psi_{\text{Gr}E(t,\cdot)}(x,v) + L(t,x,v)\}$. Furthermore, since $\text{Gr}\{E(t,\cdot)\}$ is closed, the normal cone of this epigraph set at $(\bar{x}(t), \dot{\bar{x}}(t), L(t,\bar{x}(t),\dot{\bar{x}}(t)))$ is known to be contained in

$$\{(\xi, -\alpha) : \alpha \geq 0, \xi \in \alpha \partial_{x,v} L(t,\bar{x}(t),\dot{\bar{x}}(t)) + N_{\text{Gr}\{F(t,\cdot)\}}(\bar{x}(t),\dot{\bar{x}}(t))\}.$$

It follows therefore from (A″) that we must choose $\alpha = -q = \lambda$ and

$$\dot{p}(t) \in \text{co}\{\eta : (\eta, p(t)) \in \lambda \partial L(t,\bar{x}(t),\dot{\bar{x}}(t)) + N_{\text{Gr}\{F(t,\cdot)\}}(\bar{x}(t),\dot{\bar{x}}(t))\}.$$

Finally from condition (C″), since $(v, L(t,\bar{x}(t),v)) \in E(t,\bar{x}(t))$ for any $v \in F(t,\bar{x}(t))$, we get

$$p(t)\dot{\bar{x}}(t) - \lambda L(t,\bar{x}(t),\dot{\bar{x}}(t)) \geq p(t)v - \lambda L(t,\bar{x}(t),v) \quad \forall v \in F(t,\bar{x}(t)).$$

Proof of the theorem is complete. □

4. Autonomous problems. Consider a modification of problem (Q), in which the dynamics and the cost integrand are independent of time, the terminal time is a choice variable, and the terminal time, together with the end states, are constrained to lie in a specified set:

$$\text{minimize } l(x(0),x(T),T)+\int_0^T L(x(t),\dot{x}(t))dt$$

over terminal times $T>0$ and arcs $x\in W^{1,1}$ satisfying

$$\dot{x}(t)\in F(x(t)),\quad (x(0),x(T),T)\in D.$$

The data here are functions $l: R^n\times R^n\times R\to R$, $L: R^n\times R^n\to R$, a multifunction $F: R^n\rightrightarrows R^n$, and a closed set $D\subset R^n\times R^n\times R$.

A pair $(\bar{T},\bar{x})$ is said to be a local minimizer for the free time problem with autonomous dynamics if there exists $\epsilon>0$ such that it is minimizing over all (T,x)'s such that

$$\mathrm{Gr}\{x\}\subset \mathrm{Gr}\{\bar{x}\}+\epsilon B. \tag{25}$$

Necessary conditions are now derived with the help of transformation of the independent variable techniques previously used, for example, in [2, p. 151 et seq.].

THEOREM 9. *Let $(\bar{T},\bar{x})$ be a local minimizer for the free time problem with autonomous dynamics, with $\bar{T}>0$. Assume that for some $\delta\in(0,\infty)$ and some nonnegative measurable functions k_F and k_L such that $k_F\in L^1$, $k_L\in L^1$, and $k_Fk_L\in L^1$, we have the following:*

(GF1) *l is Lipschitz continuous on a neighborhood of $(\bar{x}(0),\bar{x}(\bar{T}),\bar{T})$.*

(GF2) *$F(x)$ is nonempty and closed for each x,*

$$F(x')\subset F(x)+k_F(t)|x'-x|B\quad \forall x,x'\in\bar{x}(t)+\delta B\quad a.e.$$

(GF3) $|L(x,v)-L(x',v')|\le k_L(t)|(x,v)-(x',v')|\forall (x,v),(x',v')\in(\bar{x}(t)+\delta)\times R^n$.

Then there exist $p\in W^{1,1}([0,\bar{T}];R^n)$ and $\lambda\ge 0$, not both zero, such that

$$\dot{p}(t)\in \mathrm{co}\{\eta:(\eta,p(t))\in\lambda\partial L(\bar{x}(t),\dot{\bar{x}}(t))+N_{\mathrm{Gr}\{F(\cdot)\}}(\bar{x}(t),\dot{\bar{x}}(t))\},$$

$$p(t)\dot{\bar{x}}(t)-\lambda L(\bar{x}(t),\dot{\bar{x}}(t))\ge p(t)v-\lambda L(\bar{x}(t),v)\forall v\in F(\bar{x}(t)),$$

$$p(t)\dot{\bar{x}}(t)-\lambda L(\bar{x}(t),\dot{\bar{x}}(t))=h\quad \text{a.e. for some constant } h, \text{ and}$$

$$(p(0),-p(1),h)\in\lambda\partial l(\bar{x}(0),\bar{x}(\bar{T}),\bar{T})+N_D(\bar{x}(0),\bar{x}(\bar{T}),\bar{T}).$$

Proof. We deal first with the case $L\equiv 0$. Consider the fixed time problem:

$$\text{minimize } l(x(0),x(\bar{T}),y(\bar{T}))$$

subject to

$$(\dot{x}(t),\dot{y}(t))\in\tilde{F}(x(t))$$

$$(x(0),x(\bar{T}),y(\bar{T}))\in D, y(0)=0$$

in which $\tilde{F}(x):=\{(\alpha v,\alpha):\alpha\in[0.5,1.5],v\in F(x)\}$.

Take any $\epsilon'>0$ and any feasible arc (x,y) for this problem satisfying $|(x(s),y(s))-(\bar{x}(s),\bar{y}(s)=s)|\le\epsilon'$ for all $s\in[0,\bar{T}]$. Then there exist measurable function α :

$[0,\bar{T}] \to [0.5, 1.5]$ and an integrable function $\phi(s)$ such that $\dot{x}(s) = \alpha(s)\phi(s)$, $\phi(s) \in F(x(s))$, and $\dot{y}(s) = \alpha(s)$ a.e. Now consider the strictly increasing function $\theta(s) = \int_0^s \alpha(\sigma)d\sigma$. We can use this to define a change of variables $t = \theta(s)$. Write $\tilde{x}(t) = x(\theta^{-1}(t)), 0 \le t \le T$, where $T = \theta(\bar{T})$. We can show, with the help of Fabini's theorem, that $\tilde{x}$ is absolutely continuous, $\dot{\tilde{x}}(t) = \phi(\theta^{-1}(t)) \in F(\tilde{x}(t))$ a.e. We note also that

$$(\tilde{x}(0), \tilde{x}(T), T) = (x(0), x(\bar{T}), y(\bar{T})).$$

Hence $l(\tilde{x}(0), \tilde{x}(T), T) = l(x(0), x(\bar{T}), y(\bar{T}))$, and $(\tilde{x}(0), \tilde{x}(T), T) \in D$. We see that $(T, \tilde{x})$ is feasible for the original free time problem and has the same cost as that of (x, y) for the fixed time problem. It is also the case that we can arrange, by choosing ϵ' sufficiently small that $\mathrm{Gr}\{x\} \subset \mathrm{Gr}\{\bar{x}\} + \epsilon B$, where ϵ is the parameter in (25) associated with $(\bar{T}, \bar{x})$ as local minimizer. Since $(\bar{x}(s), \bar{y}(s) = s)$ is feasible for the fixed time problem and has cost $l(\bar{x}(0), \bar{x}(\bar{T}), \bar{T})$ we conclude that $(\bar{x}, \bar{y})$ is a local minimizer. The hypotheses are satisfied under which Theorem 6 is available to give information about the local minimizer $(\bar{x}, \bar{y})$: there exist $p \in W^{1,1}$, $q \in R$, and $\lambda \ge 0$, not all zero, such that

$$\dot{p}(s) \in \mathrm{co}\{\eta : (\eta, p(s), q) \in N_{\mathrm{Gr}\{\tilde{F}\}}(\bar{x}(s), \dot{\bar{x}}(s), 1)\},$$

$$(p(0), -p(\bar{T}), -q) \in \lambda \partial l(\bar{x}(0), \bar{x}(\bar{T}), \bar{T}) + N_D(\bar{x}(0), \bar{x}(\bar{T}), \bar{T}),$$

$$p(s)\dot{\bar{x}}(s) + q \ge (p(s)v + q)\alpha \quad \forall \alpha \in [0.5, 1.5], v \in F(\bar{x}(s)).$$

This last relation tells us that

$$-q = p(s)\dot{\bar{x}}(s) \quad \text{a.e.} \quad s \in [0, \bar{T}],$$

$$p(s)\dot{\bar{x}}(s) \ge p(s)v \quad \forall v \in F(\bar{x}(s)).$$

To analyze the implications of the costate equation above, note that neighboring points to $(\bar{x}(s), \dot{\bar{x}}(s), 1)$ in the graph of $\tilde{F}$ can be expressed as $(x', \alpha' v', \alpha')$ for some x', v', α' close to $(\bar{x}(s), \dot{\bar{x}}(s), 1)$ and that, if (η', p', q') is a proximate normal to $\mathrm{Gr}\{\tilde{F}\}$ for some such $(x', \alpha' v', \alpha')$, then

$$\eta'(x - x') + p'(\alpha v - \alpha' v') + q'(\alpha - \alpha') \le M(|(x, \alpha v, \alpha) - (x', \alpha' v', \alpha')|^2)$$

for all $x \in R^n$, $v \in F(x)$, and α near (x', v', α'). Setting $\alpha = \alpha'$, we deduce from the inequality that $(\eta', \alpha' p') \in N_{\mathrm{Gr}F}(x', v')$. The usual limit-taking procedure now permits us to deduce from $(\eta, p(s), q) \in N_{\mathrm{Gr}\{\tilde{F}\}}(\bar{x}(s), \dot{\bar{x}}(s), 1)$ that

$$(\eta, p(s)) \in N_{\mathrm{Gr}\{F\}}(\bar{x}(s), \dot{\bar{x}}(s)).$$

Assembling all the above relationships and defining $h = -q$ we arrive at the theorem statement. (Note that $q = 0$ if $p = 0$, so that p and λ cannot both be zero.)

To conclude, we allow an integral cost term. The free time problem is reformulated as an integral cost free problem by state augmentation (the dynamics remain autonomous). The desired necessary conditions for the original problem are now obtained by applying the special case of the necessary conditions (already proved) to the "augmented" problem. (The arguments required to show that the relevant hypotheses are satisfied for application of the special case of the necessary conditions are in essential respects the same as those appearing in the proof of Theorem 6.) □

Acknowledgment. Helpful comments on a preliminary version of this paper provided by F. H. Clarke, A. D. Ioffe, and B. S. Mordukhovich are gratefully acknowledged.

REFERENCES

[1] F. H. Clarke, *The generalized problem of Bolza*, SIAM J. Control Optim., 14 (1976), pp. 469–478.
[2] F. H. Clarke, *Optimization and Nonsmooth Analysis*, John Wiley, New York, 1983.
[3] F. H. Clarke, *Methods of Dynamic and Nonsmooth Optimization*, CBMS-NSF Regional Conf. Ser. in Appl. Math. 57, SIAM, Philadelphia, PA, 1989.
[4] A. D. Ioffe, *Euler Lagrange and Hamiltonian formalisms in dynamic optimization*, Trans. Amer. Math. Soc., to appear.
[5] A. D. Ioffe and R. T. Rockafellar, *The Euler and Weierstrass conditions for nonsmooth variational problems*, Calc. Var., 4 (1996), pp. 59–87.
[6] B. Kaśkosz and S. Lojasiewicz, Jr., *A maximum principle for generalized control systems*, Nonlinear Anal.: Theory, Methods Appl., 9 (1985), pp. 109–130.
[7] B. Kaśkosz and S. Lojasiewicz, Jr., *Lagrange-type extremal trajectories in differential inclusions*, Systems Control Lett., 19 (1992), pp. 241–247.
[8] P. D. Loewen, *Optimal Control Via Nonsmooth Analysis*, CRM Proc. Lecture Notes, AMS, Providence, RI, 1993.
[9] P. D. Loewen and R. T. Rockafellar, *Optimal control of unbounded differential inclusions*, SIAM J. Control Optim., 32 (1994), pp. 442–470.
[10] B. S. Mordukhovich, *Maximum principle in problems of time optimal control with nonsmooth constraints*, J. Appl. Math. Mech., 40 (1976), pp. 960–969.
[11] B. S. Mordukhovich, *Metric approximations and necessary optimality conditions for general classes of extremal problems*, Dokl. Akad. Nauk, 22 (1980), pp. 526–530.
[12] B. S. Mordukhovich, *Approximation Methods in Problems of Optimization and Control*, Nauka, Moscow, 1988 (in Russian).
[13] B. S. Mordukhovich, *Generalized differential calculus for nonsmooth and set valued mappings*, J. Math. Anal. Appl., 183 (1994), pp. 250–288.
[14] B. S. Mordukhovich, *Discrete approximations and refined Euler Lagrange conditions for nonconvex differential inclusions*, SIAM J. Control Optim., 33 (1995), pp. 882–915.
[15] R. T. Rockafellar, *Equivalent subgradient versions of Hamiltonian and Euler–Lagrange equations in variational analysis*, SIAM J. Control Optim., 34 (1996), pp. 1300–1314.
[16] G. V. Smirnov, *Discrete approximations and optimal solutions to differential inclusions*, Kibernetika, 6 (1991), pp. 76–79 (in Russian).
[17] R. B. Vinter and P. D. Woodford, *On the occurence of intermediate local minimizers that are not strong local minimizers*, Systems Control Lett., to appear.
[18] Q. Zhu, *Necessary Optimality Conditions for Nonconvex Differential Inclusions with Endpoint Constraints*, Western Michigan University, Kalamazoo, MI, 1994, preprint.

SIAM J. CONTROL OPTIM.
Vol. 35, No. 1, pp. 78–116, January 1997

005

LOW-GAIN CONTROL OF UNCERTAIN REGULAR LINEAR SYSTEMS*

HARTMUT LOGEMANN† AND STUART TOWNLEY‡

Abstract. It is well known that closing the loop around an exponentially stable, finite-dimensional, linear, time-invariant plant with square transfer-function matrix $\mathbf{G}(s)$ compensated by a controller of the form $(k/s)\Gamma_0$, where $k \in \mathbb{R}$ and $\Gamma_0 \in \mathbb{R}^{m\times m}$, will result in an exponentially stable closed-loop system which achieves tracking of arbitrary constant reference signals, provided that (i) all the eigenvalues of $\mathbf{G}(0)\Gamma_0$ have positive real parts and (ii) the gain parameter k is positive and sufficiently small.

In this paper we consider a rather general class of infinite-dimensional linear systems, called regular systems, for which convenient representations are known to exist, both in time and in frequency domain. The purpose of the paper is twofold: (i) we extend the above result to the class of exponentially stable regular systems and (ii) we show how the parameters k and Γ_0 can be tuned adaptively. The resulting adaptive tracking controllers are not based on system identification or parameter estimation algorithms, nor is the injection of probing signals required.

Key words. regular infinite-dimensional systems, integral controllability, robust tracking, adaptive tracking, state-space methods, frequency-domain methods

AMS subject classifications. 93C20, 93C25, 93C40, 93D09, 93D15, 93D21, 93D25

PII. S0363012994275920

1. Introduction.

The synthesis of low-gain I and PI-controllers for uncertain stable plants has received considerable attention in the past 20 years. Let $\mathbf{G}$ be a stable proper rational transfer function matrix. The main existence result on robust low-gain I-control says that for any matrix Γ_0 satisfying

$$\text{spectrum}(\mathbf{G}(0)\Gamma_0) \subset \{s \in \mathbb{C} \mid \operatorname{Re} s > 0\}, \tag{1.1}$$

there exists $k^* > 0$ such that for all $k \in (0, k^*)$ the controller $(1/s)k\Gamma_0$ stabilizes $\mathbf{G}$ and the resulting closed-loop system asymptotically tracks arbitrary constant reference signals. This result has been proved by Davison [4][1] and Lunze [18] using state-space methods and by Grosdidier, Morari, and Holt [5] and Morari [25] using frequency-domain methods (see also the book by Lunze [20, Chapter 10], and the textbook by Morari and Zafiriou [26, p. 362]). There are consequently two parts to the design of low-gain tracking controllers: choosing Γ_0 and tuning k. Such a controller design approach, called "tuning regulator theory" [4], has been successfully applied to industrial control problems; see Coppus, Sha, and Wood [2] and Lunze [19].

In the case that $\mathbf{G}$ is square, $\mathbf{G}^{-1}(0)$ would be a natural choice for Γ_0, but in the presence of uncertainty, $\mathbf{G}(0)$ might not be known exactly. However, an estimate

*Received by the editors October 24, 1994; accepted for publication (in revised form) October 16, 1995. This research was supported by the Human Capital and Mobility programme (project CHRX-CT93-0402) and the British Council/DAAD (ARC project 464).

http://www.siam.org/journals/sicon/35-1/27592.html

†School of Mathematical Sciences, University of Bath, Claverton Down, Bath BA2 7AY, UK (hl@maths.bath.ac.uk).

‡Department of Mathematics and Centre for Systems and Control Engineering, Department of Engineering, University of Exeter, North Park Road, Exeter EX4 4QE, UK (townley@maths.exeter.ac.uk).

[1] In [4] the result is proven for the special choice $\Gamma_0 = \mathbf{G}^{-1}(0)$. However, an inspection of the Lyapunov argument in the proof of lemma 3 in [4] shows that it can be easily extended to the more general case when Γ_0 satisfies (1.1) (simply replace the identity I in equation (28) in [4] by N, where N is the positive definite solution of the Lyapunov equation $(\mathbf{G}(0)\Gamma_0)^T N + N(\mathbf{G}(0)\Gamma_0) = -I$).

G_0 of $\mathbf{G}(0)$ can be obtained, in principle, by performing step response experiments on the plant. In this case the matrix Γ_0 is then chosen such that (1.1) holds with $\mathbf{G}(0)$ replaced by G_0. Although Mustafa [28] has recently derived a formula for the maximal k^* in terms of a minimal realization (A, B, C, D) of $\mathbf{G}$, in the presence of uncertainty there are only crude methods available for determining a number $k^* > 0$ such that all gain parameters $k \in (0, k^*)$ will lead to a stable closed-loop system; see, e.g., Lunze [18] and Owens and Chotai [29]. Methods for tuning Γ_0 and k by means of experiments and simulation have been developed and discussed in many places; we mention only [4], [18], [20], [29], and the paper by Penttinen and Koivo [31].

The above-mentioned tuning regulator result has been extended by Pohjolainen [32], [33], Pohjolainen and Lätti [34], Logemann and Owens [15] and Logemann, Bontsema, and Owens [11] to various classes of (abstract) infinite-dimensional systems and by Koivo and Pohjolainen [9] and Jussila and Koivo [8] to differential delay systems.

If the plant uncertainty is large and/or if reliable plant step data are not available, then the parameters k and Γ_0 need to be tuned adaptively. It turns out that, once the tuning problem for k is solved, the tuning of Γ_0 can be achieved by applying the spectrum unmixing techniques used in multivariable high-gain adaptive stabilization, Mårtensson [21], [22]. Low-gain universal adaptive controllers which achieve asymptotic tracking of constant reference signals for finite-dimensional linear stable plants have been presented by Cook [1] and Miller and Davison [23], [24].[2] By "universal" we mean that the controllers are not based on system identification or parameter estimation algorithms. The controller given in [1] is smooth, while the control laws derived in [23], [24] are "piecewise constant." The controller given in [24] satisfies a control input constraint.

In this paper we consider the problem of low-gain I-control for the class of exponentially stable, linear, regular infinite-dimensional systems introduced and studied by Weiss; see [44], [45], [46], [47], [48], [49]. This class is rather general and includes all distributed parameter systems and all time-delay systems (retarded and neutral) which are of interest in applications. In particular, it includes the classes of infinite-dimensional systems considered in the references [8], [9], [15], [11], [32], [33], [34] mentioned earlier and the well-known class of Pritchard–Salamon systems; see Pritchard and Salamon [35], [36] and Curtain et al. [3]. Although there exist well-posed infinite-dimensional systems which are not regular, the authors believe that any physically motivated well-posed linear time-invariant control system is regular.

In section 2 we provide the necessary background on regular systems which will be needed in sections 3–5. With one exception, all the results in section 2 are due to Weiss [44], [45], [46], [47], [48], [49], the exception being a nonlinear existence result which is required for adaptive low-gain control. The proof of this result is relegated to an appendix.

Section 3 is devoted to nonadaptive low-gain control of regular systems. We first prove a frequency-domain result on the existence of low-gain tuning regulators of the form $(1/s)k\Gamma_0$ for all square transfer function matrices $\mathbf{G}$ which are holomorphic and bounded on some right-half plane $\operatorname{Re} s > \alpha$ for some $\alpha = \alpha(\mathbf{G}) < 0$ and satisfy $\det \mathbf{G}(0) \neq 0$. This result is then applied to regular state-space systems, and it is shown that for all sufficiently small k the closed-loop system will achieve asymptotic

[2] Surprisingly, the low-gain adaptive tracking problem has received less attention than its high-gain counterpart; see Ilchmann [7], Logemann and Ilchmann [12], Ryan [38], and the references therein.

tracking of constant reference signals, provided that the initial state of the open-loop system is sufficiently "smooth."

In sections 4 and 5 we consider the adaptive low-gain tracking problem for regular infinite-dimensional systems. While the problem of universal adaptive stabilization for infinite-dimensional systems has received some attention in recent years (see Logemann [10], Logemann and Mårtensson [13], Logemann and Owens [14], Logemann and Zwart [17], and Townley [41]), very little work has been done on adaptive tracking (see, however, the paper by Logemann and Ilchmann [12] on a high-gain adaptive servomechanism for a class of infinite-dimensional systems). In particular, it seems that so far no research has been carried out on the adaptive low-gain control problem in an infinite-dimensional setting. We mention that the main result in Cook [1] (at least as we understand it) relies on the Kalman–Yakubovich lemma. A straightforward extension of the approach in [1] to regular infinite-dimensional systems is not possible, since the existence of an appropriate infinite-dimensional version of the Kalman–Yakubovich lemma is a difficult open problem. The (discontinuous) piecewise constant controllers presented in Miller and Davison [23], [24] seem unnecessarily complicated and would not generalize to the infinite-dimensional case either. Section 4 is restricted to the case when the steady-state gain matrix $\mathbf{G}(0)$ is sign definite; i.e., $\mathbf{G}(0)$ is either positive or negative definite. We first give an alternative proof of the finite-dimensional result obtained by Cook [1]. Our proof illustrates certain special system theoretic properties of the low-gain problem, properties which can even be exploited in the infinite-dimensional case. The basic idea in [1] is to set the integrator gain k equal to $\mathcal{K}(\gamma)$, where $\mathcal{K}$ is a function, the so-called tuning function, and γ is a parameter which is adjusted by a suitable adaptation law. The class of tuning functions $\mathcal{K}$ given in [1] exploits the low-gain nature of the problem in the sense that $\mathcal{K}(\gamma) \to 0$ as $\gamma \to \infty$. We then prove the main result in section 4, a low-gain adaptive tuning regulator result for infinite-dimensional regular systems. The choice of tuning functions is more constrained than in the finite-dimensional case, although we can still work with functions $\mathcal{K}$ satisfying that $\mathcal{K}(\gamma) \to 0$ as $\gamma \to \infty$. In the sign-indefinite case, which is treated in section 5, we have to resort to tuning functions which oscillate smoothly between 0 and an arbitary positive number.

We illustrate our results by a number of examples and simulations in section 6.

Notation.

- For $\alpha \in \mathbb{R}$ set $\mathbb{C}_\alpha := \{s \in \mathbb{C} \,|\, \operatorname{Re} s > \alpha\}$.
- For $\alpha \in \mathbb{R}$ and H a Hilbert space we define the exponentially weighted L^2-space $L^2_\alpha(\mathbb{R}_+, H) := \{f \in L^2_{loc}(\mathbb{R}_+, H) \,|\, f(\cdot)\exp(-\alpha\,\cdot) \in L^2(\mathbb{R}_+, H)\}$.
- If A is a linear operator, then the domain, spectrum, and resolvent set of A are denoted by $D(A)$, $\sigma(A)$, and $\varrho(A)$, respectively.
- The Laplace transform is denoted by $\mathbb{L}$.

2. Preliminaries on abstract linear systems. In this section we give some background on abstract linear systems. Apart from Proposition 2.4 almost all the results are due to Weiss [44], [45], [46], [47], [48], [49].

First we introduce some notation. For any Hilbert space H and any $\tau \geq 0$, $\mathbf{R}_\tau$ and $\mathbf{L}_\tau$ will denote the right-shift by τ and the left-shift by τ on $L^2_{loc}(\mathbb{R}_+, H)$, respectively. The truncation operator $\mathbf{P}_\tau : L^2_{loc}(\mathbb{R}_+, H) \to L^2(\mathbb{R}_+, H)$ is given by

$$(\mathbf{P}_\tau u)(t) = \begin{cases} u(t) & \text{if} \quad t \in [0, \tau], \\ 0 & \text{if} \quad t > \tau. \end{cases}$$

For $u, v \in L^2_{loc}(\mathbb{R}_+, H)$ and $\tau \geq 0$, the *τ-concatenation* $u \overset{\tau}{\diamondsuit} v$ is defined by

$$u \overset{\tau}{\diamondsuit} v = \mathbf{P}_\tau u + \mathbf{R}_\tau v\,.$$

The following concept was introduced by Weiss [46]. An equivalent definition can be found in Salamon [39].

DEFINITION 2.1. *Let U, X, and Y be real Hilbert spaces. An* abstract linear system *with* state-space X, input-space U, *and* output-space Y *is a quadruple* $\mathbf{\Sigma} = (\mathbf{T}, \mathbf{\Phi}, \mathbf{\Psi}, \mathbf{F})$, *where*

(i) $\mathbf{T} = (\mathbf{T}_t)_{t\geq 0}$ *is a C_0-semigroup of bounded linear operators on X;*

(ii) $\mathbf{\Phi} = (\mathbf{\Phi}_t)_{t\geq 0}$ *is a family of bounded linear operators from $L^2(\mathbb{R}_+, U)$ to X such that*

$$\mathbf{\Phi}_{\tau+t}(u \overset{\tau}{\diamondsuit} v) = \mathbf{T}_t \mathbf{\Phi}_\tau u + \mathbf{\Phi}_t v$$

for all $u, v \in L^2(\mathbb{R}_+, U)$ and all $\tau, t \geq 0$;

(iii) $\mathbf{\Psi} = (\mathbf{\Psi}_t)_{t\geq 0}$ *is a family of bounded linear operators from X to $L^2(\mathbb{R}_+, Y)$ such that*

$$\mathbf{\Psi}_{\tau+t} x_0 = \mathbf{\Psi}_\tau x_0 \overset{\tau}{\diamondsuit} \mathbf{\Psi}_t \mathbf{T}_\tau x_0$$

for all $x_0 \in X$ and all $\tau, t \geq 0$, and $\mathbf{\Psi}_0 = 0$;

(iv) $\mathbf{F} = (\mathbf{F}_t)_{t\geq 0}$ *is a family of bounded linear operators from $L^2(\mathbb{R}_+, U)$ to $L^2(\mathbb{R}_+, Y)$ such that*

$$\mathbf{F}_{\tau+t}(u \overset{\tau}{\diamondsuit} v) = \mathbf{F}_\tau u \overset{\tau}{\diamondsuit} (\mathbf{\Psi}_t \mathbf{\Phi}_\tau u + \mathbf{F}_t v)\,,$$

$u, v \in L^2(\mathbb{R}_+, U)$ and all $\tau, t \geq 0$, and $\mathbf{F}_0 = 0$.

It follows easily from the definition that $\mathbf{\Phi}_0 = 0$ and that for any $\tau \geq 0$, $x_0 \in X$, and $u \in L^2_{loc}(\mathbb{R}_+, U)$

$$(\mathbf{\Psi}_\tau x_0)(t) = (\mathbf{F}_\tau u)(t) = 0 \quad \text{for a.e. } t \geq \tau\,.$$

Let an input $u \in L^2_{loc}(\mathbb{R}_+, U)$ and an initial state $x_0 \in X$ be given. The state $x(t) = x(t; x_0, u)$ of $\mathbf{\Sigma}$ at time $t \geq 0$ and the output $y(\cdot) = y(\cdot\,; x_0, u)$ of $\mathbf{\Sigma}$ are defined by

$$x(t) = \mathbf{T}_t x_0 + \mathbf{\Phi}_t \mathbf{P}_t u\,, \tag{2.1a}$$

$$\mathbf{P}_t y = \mathbf{\Psi}_t x_0 + \mathbf{F}_t \mathbf{P}_t u\,. \tag{2.1b}$$

The state trajectory $x(\cdot)$ is continuous from $\mathbb{R}_+ \to X$, and the output $y(\cdot)$ is in $L^2_{loc}(\mathbb{R}_+, Y)$. Furthermore, if $t \geq \tau \geq 0$, then the functions $x(\cdot)$ and $y(\cdot)$ defined by (2.1) satisfy

$$x(t) = \mathbf{T}_{t-\tau} x(\tau) + \mathbf{\Phi}_{t-\tau} \mathbf{L}_\tau \mathbf{P}_t u\,, \tag{2.2a}$$

$$\mathbf{L}_\tau \mathbf{P}_t y = \mathbf{\Psi}_{t-\tau} x(\tau) + \mathbf{F}_{t-\tau} \mathbf{L}_\tau \mathbf{P}_t u\,. \tag{2.2b}$$

The equations (2.2) express the *time-invariance* of $\mathbf{\Sigma}$. They follow in a straightforward way from Definition 2.1. We say that $\mathbf{\Sigma}$ is *exponentially stable* if the semigroup $\mathbf{T}$ is exponentially stable, i.e.,

$$\omega(\mathbf{T}) := \lim_{t\to\infty} \frac{1}{t} \log \|\mathbf{T}_t\| < 0\,.$$

It is clear that there exist unique operators $\mathbf{\Psi}_\infty : X \to L^2_{loc}(\mathbb{R}_+, Y)$ and $\mathbf{F}_\infty : L^2_{loc}(\mathbb{R}_+, U) \to L^2_{loc}(\mathbb{R}_+, Y)$ such that for all $\tau \geq 0$

$$\mathbf{\Psi}_\tau = \mathbf{P}_\tau \mathbf{\Psi}_\infty \,, \quad \mathbf{F}_\tau = \mathbf{P}_\tau \mathbf{F}_\infty \,.$$

The generator of $\mathbf{T}$ is denoted by A. Let X_1 be the space $D(A)$ endowed with the graph norm, and let X_{-1} be the completion of X with respect to the norm $\|x\|_{-1} = \|(\lambda I - A)^{-1}x\|$, where $\lambda \in \varrho(A)$ is fixed. We have $X_1 \subset X \subset X_{-1}$ and the canonical injections are bounded and dense. The semigroup $\mathbf{T}$ can be restricted to a C_0-semigroup on X_1 and extended to a C_0-semigroup on X_{-1}. The exponential growth constant is the same on all three spaces. The generator on X_1 is the restriction of A to $D(A^2)$, and the generator on X_{-1} is an extension of A to X (which is bounded as an operator from X to X_{-1}). We shall use the same symbols for the original semigroup and its generator and the corresponding restrictions and extensions.

By a representation theorem due to Salamon [39] (see also Weiss [44], [45]) there exist unique operators $B \in \mathcal{L}(U, X_{-1})$ and $C \in \mathcal{L}(X_1, Y)$ (the *control operator* and the *observation operator* of $\mathbf{\Sigma}$, respectively) such that for all $t \geq 0$, all $u \in L^2_{loc}(\mathbb{R}_+, U)$, and all $x_0 \in X_1$

$$\mathbf{\Phi}_t \mathbf{P}_t u = \int_0^t \mathbf{T}_{t-\xi} B u(\xi)\, d\xi \quad \text{and} \quad (\mathbf{\Psi}_\infty x_0)(t) = C\mathbf{T}_t x_0 \ .$$

B is called *bounded* if $B \in \mathcal{L}(U, X)$ (and *unbounded* otherwise), whereas C is called *bounded* if it can be extended continuously to X (and *unbounded* otherwise).

The *Lebesgue extension* of C was introduced in [45] and is defined by

$$C_L x_0 = \lim_{t \to 0} C \frac{1}{t} \int_0^t \mathbf{T}_\xi x_0 \, d\xi \,,$$

where $D(C_L)$ is equal to the set of all those $x_0 \in X$ for which the above limit exists. Clearly $X_1 \subset D(C_L) \subset X$, and for any $x_0 \in X$ we have that $\mathbf{T}_t x_0 \in D(C_L)$ for almost every $t \geq 0$. Furthermore,

$$(\mathbf{\Psi}_\infty x_0)(t) = C_L \mathbf{T}_t x_0 \quad \text{for a.e. } t \geq 0 \,.$$

Let Ω be a subset of $\mathbb{C}$. A function $\mathbf{H} : \Omega \to \mathcal{L}(U, Y)$ is called *well posed* if there exists $\alpha \in \mathbb{R}$ such that $\mathbb{C}_\alpha \subset \Omega$ and $\mathbf{H}$ is holomorphic and bounded on $\mathbb{C}_\alpha$. It can be shown (see Weiss [47]) that if $\alpha > \omega(\mathbf{T})$ and if $u \in L^2_\alpha(\mathbb{R}_+, U)$, then $\mathbf{F}_\infty u \in L^2_\alpha(\mathbb{R}_+, Y)$ and there exists a unique well-posed function $\mathbf{G} : \mathbb{C}_{\omega(\mathbf{T})} \to \mathcal{L}(U, Y)$ such that

$$\mathbf{G}(s)(\mathbb{L}u)(s) = [\mathbb{L}(\mathbf{F}_\infty u)](s) \quad \forall\, s \in \mathbb{C}_\alpha \,.$$

In particular, $\mathbf{G}$ is holomorphic on $\mathbb{C}_{\omega(\mathbf{T})}$ and bounded on $\mathbb{C}_\alpha$ for all $\alpha > \omega(\mathbf{T})$. The function $\mathbf{G}$ is called the *transfer function* of $\mathbf{\Sigma}$. Conversely, due to a result by Salamon [39], any well-posed function can be realized by an abstract linear system in the sense of Definition 2.1.

The following lemma will be needed in section 3. Certainly, it should be well known. However, since we could not find it in the literature, we include the proof.

LEMMA 2.2. *Suppose that* $\mathbf{\Sigma} = (\mathbf{T}, \mathbf{\Phi}, \mathbf{\Psi}, \mathbf{F})$ *is exponentially stable. For any* $x_0 \in X$ *and any* $u \in L^2(\mathbb{R}_+, U)$, *the functions* $x(\cdot)$ *and* $y(\cdot)$ *defined by* (2.1) *satisfy*

$$x \in L^2(\mathbb{R}_+, X) \,, \quad y \in L^2(\mathbb{R}_+, Y) \,.$$

Proof. Since $x(t) = \mathbf{T}_t x_0 + \int_0^t \mathbf{T}_{t-\xi} B u(\xi)\, d\xi$, it follows from the exponential stability of $\mathbf{T}$ that $x \in L^2(\mathbb{R}_+, X)$ if and only if the function $\bar{x} : t \mapsto \int_0^t \mathbf{T}_{t-\xi} B u(\xi)\, d\xi$ is in $L^2(\mathbb{R}_+, X)$. Let $H^2(\mathbb{C}_0, X)$ denote the usual Hardy space of holomorphic functions defined on $\mathbb{C}_0$ with values in X. Appealing to the Paley–Wiener theorem, it follows that $\bar{x} \in L^2(\mathbb{R}_+, X)$ if we can show that $\mathbb{L}\bar{x} \in H^2(\mathbb{C}_0, X)$. To this end set $\omega_0 := \omega(\mathbf{T})$ and recall from [48] that for any $\omega > \omega_0$ there exists $M_\omega > 0$ such that

$$\|(sI - A)^{-1} B\|_{\mathcal{L}(U,X)} \leq \frac{M_\omega}{\sqrt{\operatorname{Re} s - \omega}} \quad \forall\, s \in \mathbb{C}_\omega\,. \tag{2.3}$$

(In particular, $(sI - A)^{-1} B \in \mathcal{L}(U, X)$ for all $s \in \mathbb{C}_\omega$.) Moreover, it is routine to show that the function

$$\mathbb{C}_{\omega_0} \to \mathcal{L}(U, Y)\,, \quad s \mapsto (sI - A)^{-1} B$$

is holomorphic. Finally, the Laplace transform of $\bar{x}$ is given by

$$(\mathbb{L}\bar{x})(s) = (sI - A)^{-1} B (\mathbb{L}u)(s) \quad \forall\, s \in \mathbb{C}_{\omega_0}\,, \tag{2.4}$$

and by hypothesis, $\omega_0 < 0$ and $\mathbb{L}u \in H^2(\mathbb{C}_0, X)$. Therefore, combining (2.3) and (2.4) we obtain that $\mathbb{L}\bar{x} \in H^2(\mathbb{C}_0, X)$.

In order to prove that $y \in L^2(\mathbb{R}_+, Y)$, write y in the form

$$y = \mathbf{\Psi}_\infty x_0 + \mathbf{F}_\infty u.$$

Using the remarks preceding the lemma, it follows from the hypothesis that $\mathbf{F}_\infty u \in L^2(\mathbb{R}_+, Y)$. It remains to show that $\mathbf{\Psi}_\infty x_0 \in L^2(\mathbb{R}_+, Y)$. By the exponential stability of $\mathbf{T}$ it follows in a straightforward way from condition (iii) in Definition 2.1 that there exists a constant $\gamma > 0$ such that

$$\|\mathbf{\Psi}_\tau x_0\|_{L^2(\mathbb{R}_+, Y)} \leq \gamma \|x_0\| \quad \forall\, \tau \geq 0 \;\; \forall\, x_0 \in X\,.$$

Hence

$$\|\mathbf{P}_\tau \mathbf{\Psi}_\infty x_0\|_{L^2(\mathbb{R}_+, Y)} = \|\mathbf{\Psi}_\tau x_0\|_{L^2(\mathbb{R}_+, Y)} \leq \gamma \|x_0\| \quad \forall\, \tau \geq 0\,,\; \forall\, x_0 \in X\,,$$

which implies that $\mathbf{\Psi}_\infty x_0 \in L^2(\mathbb{R}_+, Y)$. □

$\mathbf{\Sigma}$ and its transfer function $\mathbf{G}$ are called *regular* if for any $u \in U$ the limit

$$\lim_{s \to \infty,\, s \in \mathbb{R}} \mathbf{G}(s) u = Du$$

exists. It follows from the principle of uniform boundedness that $D \in \mathcal{L}(U, Y)$. The operator D is called the *feedthrough operator* of $\mathbf{\Sigma}$. If $\mathbf{\Sigma}$ is regular, then for any $x_0 \in X$ and $u \in L^2_{loc}(\mathbb{R}_+, U)$, the functions $x(\cdot)$ and $y(\cdot)$, defined by (2.1), satisfy the equations

$$\dot{x}(t) = Ax(t) + Bu(t)\,, \tag{2.5a}$$

$$y(t) = C_L x(t) + Du(t) \tag{2.5b}$$

for a.e. $t \geq 0$ (in particular $x(t) \in D(C_L)$ for a.e. $t \geq 0$). The derivative on the left-hand side of (2.5a) has of course to be understood in X_{-1}. Moreover, as has been shown in [47], if $\mathbf{\Sigma}$ is regular, then $(sI - A)^{-1} BU \subset D(C_L)$ for all $s \in \varrho(A)$ and the transfer function $\mathbf{G}$ can be expressed in the following way:

$$\mathbf{G}(s) = C_L (sI - A)^{-1} B + D \quad \forall\, s \in \mathbb{C}_{\omega(\mathbf{T})}\,,$$

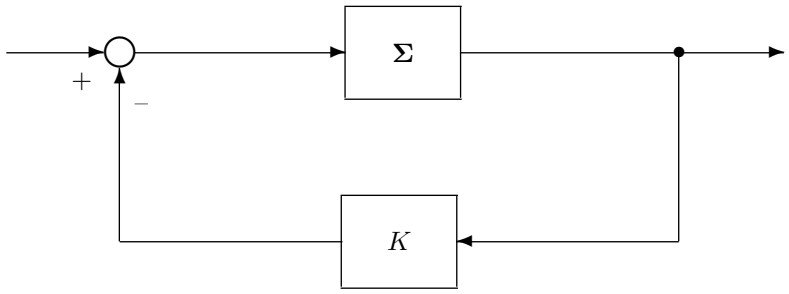


FIG. 2.1. *Static output feedback.*

which is familiar from finite-dimensional systems theory. The operators A, B, C, and D are called the *generating operators* of $\mathbf{\Sigma}$.

Finally, we review some of the results on static output feedback for abstract linear systems which have been recently obtained by Weiss [49]. Consider the feedback system shown in Figure 2.1.

An operator $K \in \mathcal{L}(Y,U)$ is called an *admissible feedback operator* for $\mathbf{\Sigma}$ if $I + K\mathbf{G}$ has a well-posed inverse, i.e., if there exists a well-posed transfer function $\mathbf{J}$ such that

$$\mathbf{J}(s)(I + K\mathbf{G}(s)) = (I + K\mathbf{G}(s))\mathbf{J}(s) = I \quad \forall\, s \in \mathbb{C}_\alpha$$

for some $\alpha \in \mathbb{R}$. It is easy to see that $I + K\mathbf{G}$ has a well-posed inverse if and only if $I + \mathbf{G}K$ has. If $\mathbf{\Sigma}$ is regular and if $K \in \mathcal{L}(Y,U)$ is an admissible feedback operator for $\mathbf{\Sigma}$, then $I + DK$ (and hence also $I + KD$) is left invertible. In particular, if U or Y is finite-dimensional, then $I + DK$ (and hence also $I + KD$) is invertible.

The next result shows that if K is an admissible feedback operator for $\mathbf{\Sigma}$, then there exists a unique abstract linear system $\mathbf{\Sigma}^K$ representing the feedback system shown in Figure 2.1.

THEOREM 2.3. *Let $\mathbf{\Sigma} = (\mathbf{T}, \mathbf{\Phi}, \mathbf{\Psi}, \mathbf{F})$ be an abstract linear system, let $\mathbf{G}$ denote its transfer function and let $K \in \mathcal{L}(Y,U)$ be an admissible feedback operator for $\mathbf{\Sigma}$. Then the following statements are true:*

(i) *There exists a unique abstract linear system $\mathbf{\Sigma}^K = (\mathbf{T}^K, \mathbf{\Phi}^K, \mathbf{\Psi}^K, \mathbf{F}^K)$ such that, when we denote*

$$\mathbf{\Sigma}_\tau = \begin{pmatrix} \mathbf{T}_\tau & \mathbf{\Phi}_\tau \\ \mathbf{\Psi}_\tau & \mathbf{F}_\tau \end{pmatrix}, \quad \mathbf{\Sigma}^K_\tau = \begin{pmatrix} \mathbf{T}^K_\tau & \mathbf{\Phi}^K_\tau \\ \mathbf{\Psi}^K_\tau & \mathbf{F}^K_\tau \end{pmatrix}$$

($\tau \geq 0$), we have

$$\mathbf{\Sigma}^K_\tau = \mathbf{\Sigma}_\tau - \mathbf{\Sigma}_\tau \begin{pmatrix} 0 & 0 \\ 0 & K \end{pmatrix} \mathbf{\Sigma}^K_\tau \quad \textit{and} \quad \mathbf{\Sigma}_\tau = \mathbf{\Sigma}^K_\tau + \mathbf{\Sigma}^K_\tau \begin{pmatrix} 0 & 0 \\ 0 & K \end{pmatrix} \mathbf{\Sigma}_\tau \quad \forall\, \tau \geq 0\,. \tag{2.6}$$

The transfer function $\mathbf{G}^K$ of $\mathbf{\Sigma}^K$ is given by $\mathbf{G}^K = \mathbf{G}(I + K\mathbf{G})^{-1}$. Moreover, $L \in \mathcal{L}(Y,U)$ is an admissible feedback operator for $\mathbf{\Sigma}^K$ if and only if $K+L$ is an admissible feedback operator for $\mathbf{\Sigma}$. If this is the case, then

$$(\mathbf{\Sigma}^K)^L = \mathbf{\Sigma}^{K+L}\,. \tag{2.7}$$

(ii) *Under the extra assumptions that* $\mathbf{\Sigma}$ *is regular and that* $I + DK$ *is invertible, it follows that* $\mathbf{\Sigma}^K$ *is regular, and the generating operators* A^K, B^K, C^K, *and* D^K *of* $\mathbf{\Sigma}^K$ *are given by*

$$A^K = A - BK(I + DK)^{-1}C_L\,,\ C^K = (I + DK)^{-1}C_L\,,\ B^K = B(I + KD)^{-1}\,,$$
$$\text{and}\quad D^K = (I + DK)^{-1}D\,,$$

where $D(A^K) = \{x \in D(C_L) \mid (A - BK(I + DK)^{-1}C_L)x \in X\}$.

For $x_0 \in X$ and $u \in L^2_{loc}(\mathbb{R}_+, U)$ define the functions $x(\cdot)$ and $y(\cdot)$ by (2.1). The second equation in (2.6) then implies for $t \geq 0$

$$x(t) = \mathbf{T}^K_t x_0 + \mathbf{\Phi}^K_t \mathbf{P}_t(Ky + u), \tag{2.8a}$$

$$\mathbf{P}_t y = \mathbf{\Psi}^K_t x_0 + \mathbf{F}^K_t \mathbf{P}_t(Ky + u)\,. \tag{2.8b}$$

Moreover, for $t \geq \tau \geq 0$ we have that

$$x(t) = \mathbf{T}^K_{t-\tau} x(\tau) + \mathbf{\Phi}^K_{t-\tau} \mathbf{L}_\tau \mathbf{P}_t(Ky + u), \tag{2.9a}$$

$$\mathbf{L}_\tau \mathbf{P}_t y = \mathbf{\Psi}^K_{t-\tau} x(\tau) + \mathbf{F}^K_{t-\tau} \mathbf{L}_\tau \mathbf{P}_t(Ky + u)\,. \tag{2.9b}$$

The above formulas (2.8) and (2.9) will turn out to be very useful in sections 4 and 5.

Finally, consider the nonlinear system given by

$$\dot\gamma(t) = \|v(t)\|^2\,, \quad k(0) = k_0 \in \mathbb{R}, \tag{2.10a}$$

$$w(t) = \mathcal{K}(\gamma(t))v(t)\,, \quad t \geq 0\,, \tag{2.10b}$$

where $v \in L^2_{loc}(\mathbb{R}_+, \mathbb{R}^m)$ is the input and w denotes the output. The function $\mathcal{K} : \mathbb{R} \to \mathbb{R}$ is assumed to be locally Lipschitz.

For sections 4 and 5 we need a well-posedness result for the feedback interconnection of $\mathbf{\Sigma}$ and (2.10). More precisely, consider the feedback system given by (2.1), (2.10), and the interconnection equations

$$v = y\,, \quad u = -w$$

(where, of course, we assume that $U = Y = \mathbb{R}^m$). The closed-loop equations for y and γ then take the following form:

$$y(t) = (\mathbf{\Psi}_\infty x_0)(t) - (\mathbf{F}_\infty \mathcal{K}(\gamma) y)(t)\,, \tag{2.11a}$$

$$\gamma(t) = \gamma_0 + \int_0^t \|y(\xi)\|^2\, d\xi\,. \tag{2.11b}$$

Let $\tau \in (0, \infty]$. A function $(y, \gamma) : [0, \tau) \to \mathbb{R}^m \times \mathbb{R}$ is called a *solution* of (2.11) on $[0, \tau)$ if

(i) $(y, \gamma) \in L^2([0, \tau'], \mathbb{R}^m) \times AC([0, \tau'], \mathbb{R})$ for all $\tau' \in [0, \tau)$, where $AC([0, \tau'], \mathbb{R})$ denotes the real-valued absolutely continuous functions defined on $[0, \tau']$.

(ii) (y, γ) satisfies (2.11) almost everywhere on $[0, \tau)$.

If (2.11) has a solution (y, γ) on $[0, \tau)$, then the corresponding state trajectory of $\mathbf{\Sigma}$ is given by

$$x(t) = \mathbf{T}_t x_0 - \mathbf{\Phi}_t(\mathbf{P}_t \mathcal{K}(\gamma) y) \quad \forall\, t \in [0, \tau)\,.$$

PROPOSITION 2.4. *Suppose that* $U = Y = \mathbb{R}^m$ *and that* $\mathbb{L}^{-1}\mathbf{G} \in L^1_{loc}(\mathbb{R}_+, \mathbb{R}^{m\times m})$. *Then for any* $(x_0, \gamma_0) \in X \times \mathbb{R}$ *there exists a maximal solution of* (2.11). *To be more*

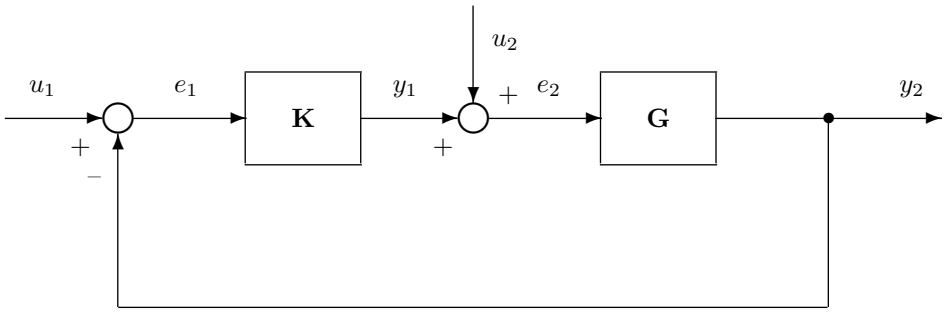


FIG. 3.1. *Closed-loop system* $\mathcal{F}(\mathbf{G},\mathbf{K})$.

precise, there exists $\tau_{max} \in (0,\infty]$ *such that* (2.11) *has a unique solution* (y_{max},γ_{max}) *on* $[0,\tau_{max})$*, and moreover*

$$\tau_{max} < \infty \quad \Longrightarrow \quad \int_0^{\tau_{max}} \|y_{max}(t)\|^2\, dt = \infty\ .$$

The proof of Proposition 2.4 is given in the appendix.

3. Nonadaptive low-gain control. For $\alpha \in \mathbb{R}$ let $\mathcal{M}_\alpha$ denote the field of all meromorphic functions defined on $\mathbb{C}_\alpha$. The algebra of all bounded holomorphic functions defined on $\mathbb{C}_\alpha$ will be denoted by H_α^∞. The symbol H_α^2 stands for the vector space of all holomorphic functions $f : \mathbb{C}_\alpha \to \mathbb{C}$ such that $\sup_{\xi>\alpha} \int_{-\infty}^{\infty} |f(\xi+\imath\omega)|^2\, d\omega < \infty$. Moreover, we define

$$\mathcal{M}_- := \bigcup_{\alpha<0} \mathcal{M}_\alpha\,, \quad H_-^\infty := \bigcup_{\alpha<0} H_\alpha^\infty\,, \quad H_-^2 := \bigcup_{\alpha<0} H_\alpha^2\,.$$

Let $\mathbf{G} \in \mathcal{M}_-^{m\times m}$ and $\mathbf{K} \in \mathcal{M}_-^{m\times m}$ be square transfer-function matrices, and consider the feedback system shown in Figure 3.1, which will be denoted by $\mathcal{F}(\mathbf{G},\mathbf{K})$. We shall call the feedback system $\mathcal{F}(\mathbf{G},\mathbf{K})$ input-output stable if every transfer function $u_i \mapsto y_j$ that occurs around the loop has all its entries in H_-^∞. More precisely, we make the following definition.

DEFINITION 3.1. *Let* $\mathbf{G} \in \mathcal{M}_-^{m\times m}$ *and* $\mathbf{K} \in \mathcal{M}_-^{m\times m}$*. The feedback system* $\mathcal{F}(\mathbf{G},\mathbf{K})$ *is called* input-output stable *if* $\det(I+\mathbf{G}(s)\mathbf{K}(s)) \not\equiv 0$ *and*

$$F(\mathbf{G},\mathbf{K}) := \begin{pmatrix} \mathbf{K}(I+\mathbf{G}\mathbf{K})^{-1} & -\mathbf{K}\mathbf{G}(I+\mathbf{K}\mathbf{G})^{-1} \\ \mathbf{G}\mathbf{K}(I+\mathbf{G}\mathbf{K})^{-1} & \mathbf{G}(I+\mathbf{K}\mathbf{G})^{-1} \end{pmatrix} \in H_-^{\infty\ 2m\times 2m}\ .$$

We say that $\mathbf{K}$ stabilizes $\mathbf{G}$ *if* $\mathcal{F}(\mathbf{G},\mathbf{K})$ *is input-output stable.*

Note that the above concept of input-output stability is stronger than L^2-stability, which is equivalent to $F(\mathbf{G},\mathbf{K}) \in H_0^{\infty\ 2m\times 2m}$. However, Definition 3.1 has the advantage that it guarantees the analyticity of the closed-loop transfer function on $\mathbb{C}_\alpha$ for some $\alpha < 0$, a property which will be needed in the following.

Remark 3.2. (i) It is trivial that $\mathbf{K}$ stabilizes $\mathbf{G}$ if and only if $\mathbf{G}$ stabilizes $\mathbf{K}$.

(ii) Let $\mathcal{Q}(H_-^\infty)$ denote the quotient field of H_-^∞, i.e., $\mathcal{Q}(H_-^\infty) = \{n/d \,|\, n,d \in H_-^\infty,\ d(s) \not\equiv 0\}$. If $\mathcal{F}(\mathbf{G},\mathbf{K})$ is input-output stable, then $\mathbf{G} \in \mathcal{Q}(H_-^\infty)^{m\times m}$ and $\mathbf{K} \in \mathcal{Q}(H_-^\infty)^{m\times m}$.

(iii) If $\mathbf{G} \in H_-^{\infty\ m\times m}$, then $\mathcal{F}(\mathbf{G},\mathbf{K})$ is input-output stable if and only if $\det(I+\mathbf{G}(s)\mathbf{K}(s)) \not\equiv 0$ and $\mathbf{K}(I+\mathbf{G}\mathbf{K})^{-1}$ is in $H_-^{\infty\ m\times m}$.

(iv) A *left coprime factorization* of $\mathbf{G}$ over H_-^∞ is a pair $(\mathbf{D},\mathbf{N}) \in {H_-^\infty}^{m\times m} \times {H_-^\infty}^{m\times m}$ such that $\det \mathbf{D} \not\equiv 0$, $\mathbf{G} = \mathbf{D}^{-1}\mathbf{N}$ and there exist $\mathbf{X}, \mathbf{Y} \in {H_-^\infty}^{m\times m}$ satisfying $\mathbf{DX}+\mathbf{NY} = I$. *Right coprime factorizations* over H_-^∞ are defined in an analogous way. It follows from Smith [40] that $\mathbf{G}$ and $\mathbf{K}$ admit left and right coprime factorizations over H_-^∞ if $\mathcal{F}(\mathbf{G},\mathbf{K})$ is input-output stable.

PROPOSITION 3.3. *Let* $\mathbf{G} \in \mathcal{M}_-^{m\times m}$ *and* $\mathbf{K} \in \mathcal{M}_-^{m\times m}$. *If* $\mathbf{K}$ *stabilizes* $\mathbf{G}$ *and if*

$$\lim_{\operatorname{Re} s\to\infty} \mathbf{K}(s) = 0\,,$$

then $\mathbf{G}$ *is well posed.*

Proof. By Remark 3.2 (ii) we have that $\mathbf{G},\mathbf{K} \in \mathcal{Q}(H_-^\infty)^{m\times m}$, and hence, by Remark 3.2 (iv), there exists a right coprime factorization $(\mathbf{N_G},\mathbf{D_G})$ of $\mathbf{G}$ over H_-^∞ and a left coprime factorization $(\mathbf{D_K},\mathbf{N_K})$ of $\mathbf{K}$ over H_-^∞. By a standard result in fractional representation theory (cf. Vidyasagar, Schneider, and Francis [42]) the input-output stability of the closed-loop system is equivalent to

$$\inf_{s\in\mathbb{C}_0^{cl}} |\det[\mathbf{N_K}(s)\mathbf{N_G}(s) + \mathbf{D_K}(s)\mathbf{D_G}(s)]| > 0\,. \tag{3.1}$$

Seeking a contradiction, suppose that $\mathbf{G}$ is not well posed. Then there exists a sequence $(s_n)_{n\in\mathbb{N}} \subset \mathbb{C}_0^{cl}$ with $\lim_{n\to\infty} \operatorname{Re} s_n = \infty$ and such that $\lim_{n\to\infty} \|\mathbf{G}(s_n)\| = \infty$. As a consequence

$$\lim_{n\to\infty} \det \mathbf{D_G}(s_n) = 0\,. \tag{3.2}$$

On the other hand $\lim_{n\to\infty} \mathbf{K}(s_n) = 0$, and hence

$$\lim_{n\to\infty} \mathbf{N_K}(s_n) = 0\,. \tag{3.3}$$

Combining (3.2) and (3.3) shows that

$$\lim_{n\to\infty} \det[\mathbf{N_K}(s_n)\mathbf{N_G}(s_n) + \mathbf{D_K}(s_n)\mathbf{D_G}(s_n)] = 0\,,$$

contradicting (3.1). □

Since in this paper we will be mainly concerned with controllers of the form $\mathbf{K}(s) = (1/s)\Gamma$, where $\Gamma \in \mathbb{R}^{m\times m}$, the following definition will turn out to be useful.

DEFINITION 3.4. *A transfer function matrix* $\mathbf{G} \in \mathcal{M}_-^{m\times m}$ *is called* integral stabilizable *if there exists* $\Gamma \in \mathbb{R}^{m\times m}$ *such that the controller* $\mathbf{K}(s) = (1/s)\Gamma$ *stabilizes* $\mathbf{G}$. *If the extra condition*

$$[\mathbf{GK}(I+\mathbf{GK})^{-1}](0) = I \tag{3.4}$$

is satisfied, then $\mathbf{G}$ *is called* integral controllable.

A controller of the form $(1/s)\Gamma$ is called an *integrator*. It is a trivial consequence of Proposition 3.3 that if a transfer-function matrix in $\mathcal{M}_-^{m\times m}$ is integral stabilizable, then necessarily it is well posed.

In the following let $\theta(\cdot)$ denote the Heaviside step function, i.e.,

$$\theta(t) = \begin{cases} 1 & \text{if} \quad t > 0, \\ 0 & \text{if} \quad t \le 0, \end{cases}$$

As usual, convolution will be denoted by $\star$. The next result shows that condition (3.4) is closely related to the asymptotic tracking of constant reference signals.

PROPOSITION 3.5. *Suppose that* $\mathbf{G} \in \mathcal{M}_-^{m\times m}$ *is integral stabilizable, and let* $\mathbf{K}(s) = (1/s)\Gamma$, *where* $\Gamma \in \mathbb{R}^{m\times m}$, *be a stabilizing integrator. Then*

$$\lim_{t\to\infty}[\mathbb{L}^{-1}(\mathbf{GK}(I+\mathbf{GK})^{-1}) \star \theta r](t) = r$$

for all $r \in \mathbb{R}^m$ *if and only if* (3.4) *holds.*

For the proof of the above proposition we need the following lemma, which is a special case of the main result in Mossaheb [27].

LEMMA 3.6. *Suppose that* h *is a holomorphic function defined on* $\mathbb{C}_\alpha$ *such that the function* $s \mapsto sh(s)$ *is in* H_α^∞. *Then there exists a measurable function* $f : \mathbb{R}_+ \to \mathbb{C}$ *with* $f(\cdot)\exp(-\beta\,\cdot) \in L^1(\mathbb{R}_+, \mathbb{C})$ *for all* $\beta > \alpha$ *and such that*

$$(\mathbb{L}f)(s) = h(s) \quad \forall\, s \in \mathbb{C}_\alpha\,.$$

Proof of Proposition 3.5. By assumption we have that

$$\mathbf{H} := (I+\mathbf{GK})^{-1}\mathbf{G} = \mathbf{G}(I+\mathbf{KG})^{-1} \in H_-^{\infty\,m\times m}\,,$$

and hence

$$s[\mathbf{GK}(I+\mathbf{GK})^{-1}](s) = s[(I+\mathbf{GK})^{-1}\mathbf{GK}](s) = \mathbf{H}(s)\Gamma \in H_-^{\infty\,m\times m}\,.$$

Thus, by Lemma 3.6

$$\mathbb{L}^{-1}[\mathbf{GK}(I+\mathbf{GK})^{-1}] \in L^1(\mathbb{R}_+, \mathbb{C}^{m\times m})\,.$$

Therefore

$$\begin{aligned}\lim_{t\to\infty}[\mathbb{L}^{-1}(\mathbf{GK}(I+\mathbf{GK})^{-1}) \star \theta r](t) &= \lim_{t\to\infty}\left(\int_0^t [\mathbb{L}^{-1}(\mathbf{GK}(I+\mathbf{GK})^{-1})](\tau)\,d\tau\right) r \\ &= [\mathbf{GK}(I+\mathbf{GK})^{-1}](0)r\,,\end{aligned}$$

which yields the claim. □

The next result gives a necessary condition for integral controllability. It shows that an integral controllable transfer function does not have any transmission zeros at 0.

PROPOSITION 3.7. *Suppose that* $\mathbf{G} \in \mathcal{M}_-^{m\times m}$ *is integral controllable. Then there exists a left coprime factorization* $(\mathbf{D}, \mathbf{N})$ *of* $\mathbf{G}$ *over* H_-^∞, *and the numerator* $\mathbf{N}$ *in any such factorization satisfies*

$$\det \mathbf{N}(0) \neq 0\,.$$

Proof. It follows from Remark 3.2 (iv) that there exists a left coprime factorization $(\mathbf{D}, \mathbf{N})$ of $\mathbf{G}$ over H_-^∞. Let $\Gamma \in \mathbb{R}^{m\times m}$ be such that $\mathbf{K}(s) = (1/s)\Gamma$ stabilizes $\mathbf{G}$ and (3.4) is satisfied. Define

$$\mathbf{H} := \mathbf{GK}(I+\mathbf{GK})^{-1}\,, \quad \Delta := \lim_{s\to 0}[\mathbf{K}(I+\mathbf{GK})^{-1}](s)\,.$$

Then $\mathbf{DH} = \mathbf{NK}(I+\mathbf{GK})^{-1}$. Moreover, letting $s \to 0$ and using (3.4) yield $\mathbf{D}(0) = \mathbf{N}(0)\Delta$. Since $\mathbf{D}$ and $\mathbf{N}$ are left coprime over H_-^∞, it follows that

$$\operatorname{rank} \mathbf{N}(0)(\Delta, I) = \operatorname{rank}[\mathbf{D}(0), \mathbf{N}(0)] = m\,.$$

Therefore $\operatorname{rank} \mathbf{N}(0) = m$, and hence $\det \mathbf{N}(0) \neq 0$. □

The following theorem is the main input-output result of this section.

THEOREM 3.8. *Suppose that* $\mathbf{G} \in H_-^{\infty\, m\times m}$ *and that* $\mathbf{G}(0)$ *is real. Then* $\mathbf{G}$ *is integral controllable if and only if*

$$\det \mathbf{G}(0) \neq 0\,. \tag{3.5}$$

If (3.5) *holds, then there exists* $\Gamma_0 \in \mathbb{R}^{m\times m}$ *such that*

$$\sigma(\mathbf{G}(0)\Gamma_0) \subset \mathbb{C}_0\,, \tag{3.6}$$

and for any $\Gamma_0 \in \mathbb{R}^{m\times m}$ *satisfying* (3.6)*, there exists* $k^* > 0$ *such that for all* $k \in (0, k^*)$

$$F(\mathbf{G}, \mathbf{K}_k) \in H_-^{\infty\, 2m\times 2m} \quad \textit{and} \quad [\mathbf{G}\mathbf{K}_k(I + \mathbf{G}\mathbf{K}_k)^{-1}](0) = I\,, \tag{3.7}$$

where $\mathbf{K}_k(s) := (1/s)k\Gamma_0$*. Moreover, setting* $\mathbf{E}_k(s) = (1/s)(I + \mathbf{G}\mathbf{K}_k)^{-1}(s)$*, we have that* $\mathbf{E}_k \in H_-^{2\ m\times m}$ *for all* $k \in (0, k^*)$*.*

The result shows in particular that there exist low-gain integral controllers which achieve stability and asymptotic tracking of constant reference signals. Since for constant reference signals $r\theta(t)$, the error signal $e(t)$ of the feedback system is given by $(\mathbb{L}e)(s) = \mathbf{E}_k(s)r$, it follows from the last statement of Theorem 3.8 via the Paley–Wiener theorem that $e \in L^2(\mathbb{R}_+, \mathbb{R}^m)$ for all $k \in (0, k^*)$. In order to apply Theorem 3.8, we have to know only that the plant is stable and that (3.5) holds. Estimates of G_0 of $\mathbf{G}(0)$ can be obtained from step response data. An obvious choice for the gain matrix Γ_0 is $\Gamma_0 = G_0^{-1}$. Once a Γ_0 satisfying (3.6) has been found, the solution of the tracking problem reduces to the tuning of the gain parameter k.

Proof of Theorem 3.8. The necessity of (3.5) for integral controllability follows from Proposition 3.7 and from the hypothesis that $\mathbf{G} \in H_-^{\infty\, m\times m}$. In order to prove sufficiency, define $\Gamma_0 := \mathbf{G}^{-1}(0)$. Then, trivially, (3.6) is satisfied. Moreover, as in Logemann and Owens [15, pp. 17, 18], it can be shown that there exists a number $k^* > 0$ such that for all $k \in (0, k^*)$ the controller $\mathbf{K}_k$ stabilizes $\mathbf{G}$, i.e.,

$$F(\mathbf{G}, \mathbf{K}_k) \in H_-^{\infty\, 2m\times 2m}\,.$$

Next observe that by the invertibility of $k\mathbf{G}(0)\Gamma_0$

$$\lim_{s\to 0}[\mathbf{G}\mathbf{K}_k(I + \mathbf{G}\mathbf{K}_k)^{-1}](s) = \lim_{s\to 0} \mathbf{G}(s)k\Gamma_0(sI + k\mathbf{G}(s)\Gamma_0)^{-1} = I\,,$$

which yields the second equation in (3.7). Finally, $\mathbf{E}_k = k^{-1}\Gamma_0^{-1}\mathbf{K}_k(I + \mathbf{G}\mathbf{K}_k)^{-1}$, and therefore $\mathbf{E}_k \in H_-^{\infty\, m\times m}$ for all $k \in (0, k^*)$. Since for all such k the transfer function matrix $(I + \mathbf{G}\mathbf{K}_k)^{-1}$ is in $H_-^{\infty\, m\times m}$ as well, we see that $\mathbf{E}_k \in H_-^{2\ m\times m}$ for all $k \in (0, k^*)$. □

For Hermitian matrices $M, N \in \mathbb{C}^{m\times m}$, in the following we write $M \prec N$ if $N - M$ is positive definite and $M \succ N$ if $N - M$ is negative definite. Similarly, we write $M \preceq N$ if $N - M$ is positive semidefinite and $M \succeq N$ if $N - M$ is negative semidefinite. Moreover, for a complex matrix M let M^H denote the conjugate transpose of M.

The next result will be an important tool in section 4, although it is interesting in its own right.

PROPOSITION 3.9. *Let* $\mathbf{G} \in H_-^{\infty\, m\times m}$ *and suppose that* $\det \mathbf{G}(0) \neq 0$*. Setting* $\tilde{\mathbf{G}}(s) := (1/s)\mathbf{G}(s)$ *and using the notation of Theorem* 2.3 *we write*

$$\tilde{\mathbf{G}}^k(s) = \tilde{\mathbf{G}}(s)(I + k\tilde{\mathbf{G}}(s))^{-1} = \frac{1}{s}\mathbf{G}(s)\left(I + \frac{k}{s}\mathbf{G}(s)\right)^{-1}\,,$$ [3]

[3] By slight abuse of notation we write $\tilde{\mathbf{G}}^k$ instead of $\tilde{\mathbf{G}}^{kI}$.

where $k \in \mathbb{R}$. Under these conditions there exists $k^* > 0$ such that for all $k \in (0, k^*)$

$$\|\tilde{\mathbf{G}}^k\|_\infty = \frac{1}{k} \tag{3.8}$$

if and only if $\mathbf{G}(0) \succ 0$. Moreover, the claim remains true if we replace k with $-k$ in (3.8) *and $\mathbf{G}(0) \succ 0$ by $\mathbf{G}(0) \prec 0$.*

As usual, the H^∞-norm in (3.8) is defined to be the supremum over $\mathbb{C}_0$ of $\sigma_{max}(\tilde{\mathbf{G}}^k(s))$, the largest singular value of $\tilde{\mathbf{G}}^k(s)$. For the single-input single-output case it follows that if $\mathbf{G}(0) \neq 0$ and if $\mathbf{G}(0) \in \mathbb{R}$, then there exists $k^* > 0$ such that $\|\tilde{\mathbf{G}}^k\|_\infty = 1/|k|$ for all $k \in \mathbb{R}$ satisfying $|k| \in (0, k^*)$ and $k\mathbf{G}(0) > 0$.

Proposition 3.9 is an immediate consequence of the following lemma.

LEMMA 3.10. *Let $\mathbf{G} \in H_-^{\infty\, m\times m}$. Using the notation of Proposition* 3.9*, the following statements hold:*

(i) *Suppose that $\det \mathbf{G}(0) \neq 0$ and $k \neq 0$. Then* (3.8) (*with k replaced by $|k|$*) *is true if and only if $I + k\tilde{\mathbf{G}}(s) + k\tilde{\mathbf{G}}^H(s) \succeq 0$ for all $s \in \mathbb{C}_0$.*

(ii) *There exists $k^* > 0$ such that $I + k\tilde{\mathbf{G}}(s) + k\tilde{\mathbf{G}}^H(s) \succeq 0$ for all $s \in \mathbb{C}_0$ and for all $k \in (0, k^*)$ if and only if $\mathbf{G}(0) \succeq 0$.*

Note that if $\mathbf{G}(s) \in \mathbb{R}^{\,m\times m}$ for all $s \in (0, \infty)$, then $I + k\tilde{\mathbf{G}}(s) + k\tilde{\mathbf{G}}^H(s) \succeq 0$ for all $s \in \mathbb{C}_0$ if and only if $(1/2)I + k\tilde{\mathbf{G}}(s)$ is positive real.

Proof of Lemma 3.10. (i) By assumption, $\mathbf{G}^{-1}(0)$ exists, and thus $\sigma_{max}(\tilde{\mathbf{G}}^k(0)) = 1/k$. Therefore (3.8) holds if and only if

$$\sigma_{max}(\tilde{\mathbf{G}}^k(s)) \leq \frac{1}{k} \quad \forall\, s \in \mathbb{C}_0\,,$$

or equivalently

$$(I + k\tilde{\mathbf{G}}(s))^{-1}\tilde{\mathbf{G}}(s)\tilde{\mathbf{G}}^H(s)(I + k\tilde{\mathbf{G}}^H(s))^{-1} \preceq \frac{1}{k^2} I \quad \forall\, s \in \mathbb{C}_0\,,$$

or equivalently

$$k^2\tilde{\mathbf{G}}(s)\tilde{\mathbf{G}}^H(s) \preceq (I + k\tilde{\mathbf{G}}(s))(I + k\tilde{\mathbf{G}}^H(s)) \quad \forall\, s \in \mathbb{C}_0\,,$$

which in turn is equivalent to the positive semidefiniteness of $I + k\tilde{\mathbf{G}}(s) + k\tilde{\mathbf{G}}^H(s)$ for all $s \in \mathbb{C}_0$.

(ii) Since $\mathbf{G}$ is holomorphic at 0, we can write

$$\mathbf{G}(s) = \mathbf{G}(0) + \sum_{i=1}^{\infty} G_i s^i\,, \tag{3.9}$$

where $G_i \in \mathbb{C}^{\,m\times m}$ and the power series in (3.9) converges in some disc Δ_ε centred at 0 and with radius $\varepsilon > 0$. Consequently,

$$I + k\tilde{\mathbf{G}}(s) + k\tilde{\mathbf{G}}^H(s) = I + \frac{k}{s}\mathbf{G}(0) + \frac{k}{\bar{s}}\mathbf{G}^H(0) + k\mathbf{H}(s) \quad \forall\, s \in \Delta_\varepsilon\,, \tag{3.10}$$

where

$$\mathbf{H}(s) := \sum_{i=1}^{\infty} G_i s^{i-1} + \sum_{i=1}^{\infty} G_i^H \bar{s}^{\,i-1}\,.$$

Moreover, since $\tilde{\mathbf{G}}(s)$ is bounded on $\mathbb{C}_0 \setminus \Delta_\varepsilon$, there exists $k_1 > 0$ such that

$$I + k\tilde{\mathbf{G}}(s) + k\tilde{\mathbf{G}}^H(s) \succeq 0 \quad \forall\, s \in \mathbb{C}_0 \setminus \Delta_\varepsilon\,,\ \forall\, k \in (0, k_1)\,. \tag{3.11}$$

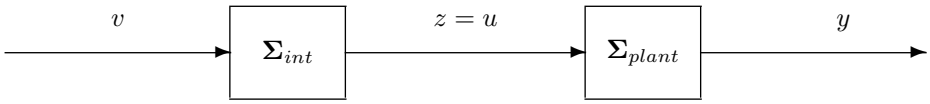


FIG. 3.2. *Cascade $\tilde{\mathbf{\Sigma}}$ with input v and output y.*

Suppose first that $\mathbf{G}(0) \succeq 0$. Then, using (3.10) and the boundedness of $\mathbf{H}(s)$ on Δ_ε, it follows that there exists $k_2 > 0$ such that

$$I + k\tilde{\mathbf{G}}(s) + k\tilde{\mathbf{G}}^H(s) \succeq 0 \quad \forall\, s \in \mathbb{C}_0 \cap \Delta_\varepsilon\,,\ \forall\, k \in (0, k_2)\,. \tag{3.12}$$

Setting $k^* := \min(k_1, k_2)$ we obtain from (3.11) and (3.12) that

$$I + k\tilde{\mathbf{G}}(s) + k\tilde{\mathbf{G}}^H(s) \succeq 0 \quad \forall\, s \in \mathbb{C}_0\,,\ \forall\, k \in (0, k^*)\,. \tag{3.13}$$

Conversely, suppose that (3.13) holds. Then, by (3.10), we obtain for any $\xi \in \mathbb{C}^m$ that

$$2\mathrm{Re}\left\langle \xi, \frac{k}{s}\mathbf{G}(0)\xi \right\rangle + \|\xi\|^2 + k\langle \xi, \mathbf{H}(s)\xi\rangle \geq 0 \quad \forall\, s \in \mathbb{C}_0 \cap \Delta_\varepsilon\,,\ \forall\, k \in (0, k^*)\,,$$

and hence it follows that for all $s \in \mathbb{C}_0 \cap \Delta_\varepsilon$ and all $k \in (0, k^*)$

$$\frac{2k}{|s|^2}\left(\mathrm{Re}\, s\, \mathrm{Re}\,\langle \xi, \mathbf{G}(0)\xi\rangle - \mathrm{Im}\, s\, \mathrm{Im}\,\langle \xi, \mathbf{G}(0)\xi\rangle\right) + \|\xi\|^2 + k\langle \xi, \mathbf{H}(s)\xi\rangle \geq 0\,.$$

Therefore, using the boundedness of $\mathbf{H}(s)$ on Δ_ε, we may conclude that for all $\xi \in \mathbb{C}^m$, $\mathrm{Im}\,\langle \xi, \mathbf{G}(0)\xi\rangle = 0$ and $\mathrm{Re}\,\langle \xi, \mathbf{G}(0)\xi\rangle \geq 0$, which in turn implies that $\mathbf{G}(0) \succeq 0$. □

In the following we will apply Theorem 3.8 to regular linear state space-systems. Since this additional assumption of regularity does not exclude any physically motivated well-posed system, the following results are as general as can be expected. For the rest of the section let $\mathbf{\Sigma}_{plant} = (\mathbf{T}, \mathbf{\Phi}, \mathbf{\Psi}, \mathbf{F})$ be an exponentially stable abstract linear regular system with generating operators (A, B, C, D), state space X, input space $U = \mathbb{R}^m$, output space $Y = \mathbb{R}^m$, and transfer function $\mathbf{G}$. Clearly, by exponential stability, $\mathbf{G} \in H_-^{\infty\, m\times m}$. If $u \in L^2_{loc}(\mathbb{R}_+, \mathbb{R}^m)$ denotes the input and $x_0 \in X$ denotes the initial state, then the state $x(\cdot)$ and the output $y(\cdot)$ are given by (2.1). Moreover, let $\mathbf{\Sigma}_{int}$ denote the integrator described by

$$z(t) = z_0 + \int_0^t v(\tau)\, d\tau\,, \quad z_0 \in \mathbb{R}^m\,,$$

where $v \in L^2_{loc}(\mathbb{R}_+, \mathbb{R}^m)$ is the integrator input.

We will consider the series connection $\tilde{\mathbf{\Sigma}}$ of $\mathbf{\Sigma}_{int}$ followed by $\mathbf{\Sigma}_{plant}$ with input v and output y (cf. Figure 3.2).

In order to show that $\tilde{\mathbf{\Sigma}}$ is again an abstract linear regular system, we introduce an extra external input $w \in L^2_{loc}(\mathbb{R}_+, \mathbb{R}^m)$ and consider the cascade interconnection $\hat{\mathbf{\Sigma}}$ with input (w, v) and output (y, z) obtained by setting $u = z + w$ (cf. Figure 3.3).

We claim that $\hat{\mathbf{\Sigma}}$ is an abstract linear regular system. To this end consider the parallel interconnection $\mathbf{\Sigma}_{par}$ of $\mathbf{\Sigma}_{int}$ and $\mathbf{\Sigma}_{plant}$ shown in Figure 3.4.

Clearly, $\mathbf{\Sigma}_{par}$ is an abstract linear regular system, and the matrix J given by

$$J = \begin{pmatrix} 0 & -I \\ 0 & 0 \end{pmatrix}$$

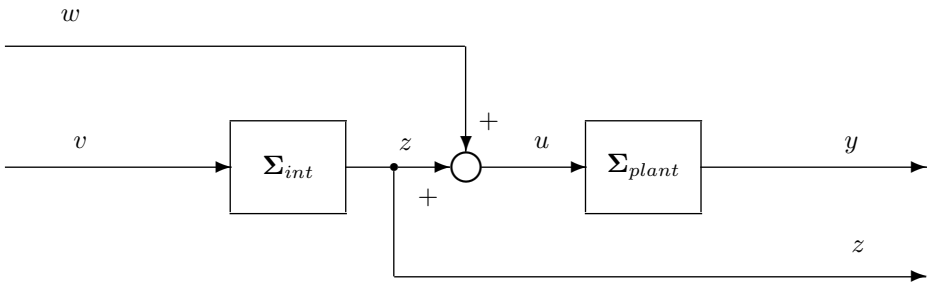


FIG. 3.3. *Cascade* $\hat{\mathbf{\Sigma}}$ *with input* (w, v) *and output* (y, z).

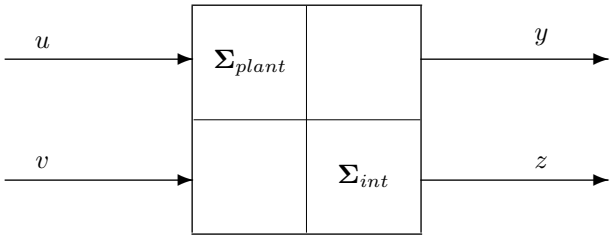


FIG. 3.4. *Parallel interconnection* $\mathbf{\Sigma}_{par}$.

is an admissible feedback operator for $\mathbf{\Sigma}_{par}$. Using the notation of section 2, we have that $\hat{\mathbf{\Sigma}} = (\mathbf{\Sigma}_{par})^J$, and hence it follows from Theorem 2.3 that $\hat{\mathbf{\Sigma}}$ is an abstract linear regular system. Writing $\hat{\mathbf{\Sigma}} = (\hat{\mathbf{T}}, \hat{\mathbf{\Phi}}, \hat{\mathbf{\Psi}}, \hat{\mathbf{F}})$, we see that $\tilde{\mathbf{\Sigma}} = (\tilde{\mathbf{T}}, \tilde{\mathbf{\Phi}}, \tilde{\mathbf{\Psi}}, \tilde{\mathbf{F}})$, where

$$\tilde{\mathbf{T}} = \hat{\mathbf{T}}\,, \quad \tilde{\mathbf{\Phi}} = \hat{\mathbf{\Phi}}\begin{pmatrix} 0 \\ I \end{pmatrix}, \quad \tilde{\mathbf{\Psi}} = (I, 0)\hat{\mathbf{\Psi}}\,, \quad \tilde{\mathbf{F}} = (I, 0)\hat{\mathbf{F}}\begin{pmatrix} 0 \\ I \end{pmatrix}.$$

Therefore $\tilde{\mathbf{\Sigma}}$ is an abstract linear regular system whose state, input, and output spaces are given by $X \times \mathbb{R}^m$, $U = \mathbb{R}^m$, and $Y = \mathbb{R}^m$, respectively. Denoting the generating operators of $\tilde{\mathbf{\Sigma}}$ by $\tilde{A}$, $\tilde{B}$, $\tilde{C}$, and $\tilde{D}$ it follows from Theorem 2.3 (ii) that

$$\tilde{A} = \begin{pmatrix} A & B \\ 0 & 0 \end{pmatrix}, \quad \tilde{B} = \begin{pmatrix} 0 \\ I \end{pmatrix}, \quad \tilde{C} = (C_L, D), \quad \tilde{D} = 0, \tag{3.14}$$

where the domain $D(\tilde{A})$ of $\tilde{A}$ is given by

$$D(\tilde{A}) = \{(x, u) \in D(C_L) \times \mathbb{R}^m \,|\, Ax + Bu \in X\}\,.$$

If B is bounded, then it follows easily that $D(\tilde{A}) = D(A) \times \mathbb{R}^m$. Note that any unboundedness of B is absorbed into the unboundedness of $\tilde{A}$ and hence the control operator $\tilde{B}$ of $\tilde{\mathbf{\Sigma}}$ is bounded. Trivially, the function $\tilde{\mathbf{G}}(s) := (1/s)\mathbf{G}(s)$ is the transfer function of $\tilde{\mathbf{\Sigma}}$.

LEMMA 3.11. *Every* $\Gamma \in \mathbb{R}^{m \times m}$ *is an admissible feedback operator for* $\tilde{\mathbf{\Sigma}}$ *and* (*using the notation of section* 2) *we have that for all* $\Gamma \in \mathbb{R}^{m \times m}$

$$D(\tilde{A}^\Gamma) = D(\tilde{A}) = \{(x, u) \in X \times \mathbb{R}^m \,|\, Ax + Bu \in X\}\,. \tag{3.15}$$

Proof. Since $\tilde{\mathbf{G}}(s) = (1/s)\mathbf{G}(s)$ and $\mathbf{G} \in H_\alpha^\infty$ for some $\alpha < 0$, it follows from section 2 that any $\Gamma \in \mathbb{R}^{m \times m}$ is an admissible feedback operator for $\tilde{\mathbf{\Sigma}}$.

We show first that the second equality in (3.15) holds. It is clear that

$$D(\tilde{A}) \subset \{(x,u) \in X \times \mathbb{R}^m \,|\, Ax + Bu \in X\} =: \mathcal{D}\,,$$

and it only remains to prove that $\mathcal{D} \subset D(\tilde{A})$. To this end define

$$W := D(A) + (\lambda I - A)^{-1} B\mathbb{R}^m\,,$$

where $\lambda \in \varrho(A)$. Since $D(A) \subset D(C_L)$ and, by regularity, $(\lambda I - A)^{-1}B\mathbb{R}^m \subset D(C_L)$, it follows that $W \subset D(C_L)$.

Let $(x,u) \in \mathcal{D}$. Then $\xi := (\lambda I - A)x - Bu \in X$, and hence

$$x = (\lambda I - A)^{-1}\xi + (\lambda I - A)^{-1}Bu \in W\,.$$

It follows that $x \in D(C_L)$ and therefore $(x,u) \in D(\tilde{A})$.

In order to show that the first equality in (3.15) is true, recall from section 2 that

$$\tilde{A}^\Gamma(x,u) = (\tilde{A} - \tilde{B}\Gamma\tilde{C}_L)(x,u)$$

for all $(x,u) \in D(\tilde{A}^\Gamma)$, where $D(\tilde{A}^\Gamma)$ is given by

$$D(\tilde{A}^\Gamma) = \{(x,u) \in D(\tilde{C}_L) \,|\, (\tilde{A} - \tilde{B}\Gamma\tilde{C}_L)(x,u) \in X \times \mathbb{R}^m\}\,.$$

Moreover, using (3.14), we see that for all $(x,u) \in D(\tilde{A}^\Gamma)$

$$\tilde{A}^\Gamma(x,u) = (Ax + Bu, -\Gamma\tilde{C}_L(x,u))\,.$$

This shows that

$$D(\tilde{A}^\Gamma) = \{(x,u) \in D(\tilde{C}_L) \,|\, Ax + Bu \in X\}\,. \tag{3.16}$$

Since $D(\tilde{C}_L) \subset X \times \mathbb{R}^m$, it follows from (3.16) that $D(\tilde{A}^\Gamma) \subset \mathcal{D} = D(\tilde{A})$.

To prove that $D(\tilde{A}) \subset D(\tilde{A}^\Gamma)$, let $(x,u) \in D(\tilde{A})$. Then $(x,u) \in D(\tilde{C}_L)$ and $Ax + Bu \in X$, and hence, by (3.16), $(x,u) \in D(\tilde{A}^\Gamma)$. □

In the following we endow $D(\tilde{A}^\Gamma)$ with its graph norm. The resulting complete space will be denoted by $\tilde{X}_1^\Gamma$.

PROPOSITION 3.12. *Let $\Gamma \in \mathbb{R}^{m\times m}$ and suppose that $\det\Gamma \neq 0$. If the integrator $\mathbf{K}(s) = (1/s)\Gamma$ stabilizes $\mathbf{G}$ (in the sense of Definition 3.1), then the following statements hold:*

(i) *The closed-loop semigroup $\tilde{\mathbf{T}}^\Gamma$ is exponentially stable.*

(ii) *$\tilde{C}^\Gamma = \tilde{C}$ and there exist $M > 0$ and $\omega > 0$ such that for all $(x_0,u_0) \in D(\tilde{A})$*

$$\|\tilde{C}\tilde{\mathbf{T}}_t^\Gamma(x_0,u_0)\| \le Me^{-\omega t}\|(x_0,u_0)\|_{\tilde{X}_1^\Gamma} \quad \forall\, t \ge 0\,.$$

If the observation operator C is bounded, then for any $(x_0,u_0) \in X \times \mathbb{R}^m$

$$\|\tilde{C}\tilde{\mathbf{T}}_t^\Gamma(x_0,u_0)\| \le Me^{-\omega t}\|(x_0,u_0)\|_{X\times\mathbb{R}^m} \quad \forall\, t \ge 0\,.$$

Proof. (i) The semigroup $\tilde{\mathbf{T}}^\Gamma$ describes the dynamics of the feedback system shown in Figure 3.5. Note that the state of $\mathbf{\Sigma}_{int}$ and the input of $\mathbf{\Sigma}_{plant}$ are identical. Therefore we denote both by the same symbol $u(\cdot)$.

The state $(x(t),u(t)) \in X \times \mathbb{R}^m$ at time $t \ge 0$ is given by

$$(x(t),u(t)) = \tilde{\mathbf{T}}_t^\Gamma(x_0,u_0)\,,$$

where $(x_0,u_0) := (x(0),u(0)) \in X \times \mathbb{R}^m$. Defining

$$y_0(t) := C_L\mathbf{T}_t x_0\,, \quad t \ge 0\,,$$

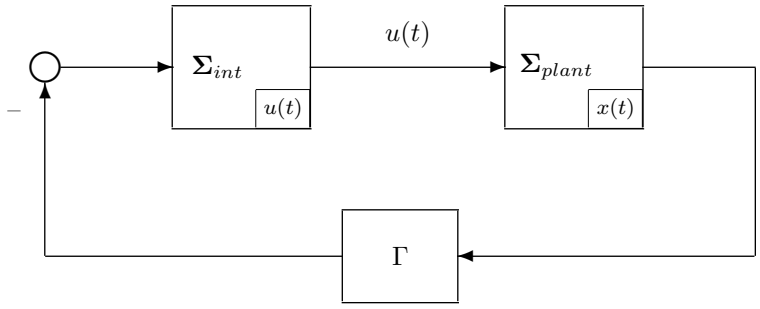


FIG. 3.5. *Internal dynamics of the closed loop.*

it follows from the exponential stability of $\mathbf{T}$ that $y_0 \in L^2(\mathbb{R}_+, \mathbb{R}^m)$. The Laplace transform of $u(\cdot)$ is then given by

$$(\mathbb{L}u)(s) = \frac{1}{s}u_0 - \mathbf{K}(s)[(\mathbb{L}y_0)(s) + \mathbf{G}(s)(\mathbb{L}u)(s)];$$

cf. Figure 3.5. It follows that

$$\mathbb{L}u = (I + \mathbf{KG})^{-1}\mathbf{K}\Gamma^{-1}u_0 - (I + \mathbf{KG})^{-1}\mathbf{K}\mathbb{L}y_0\,. \tag{3.17}$$

By assumption the closed-loop system is input-output stable, and so $(I + \mathbf{KG})^{-1}\mathbf{K}$, $(I + \mathbf{KG})^{-1} \in H_-^{\infty\ m\times m}$. Using the fact that $\mathbf{K}(s) = (1/s)\Gamma$ we see that

$$(I + \mathbf{KG})^{-1}\mathbf{K} \in H_-^{2\ m\times m}\,,$$

and thus, since $\mathbb{L}y_0 \in H_0^{2\ m}$, we obtain from (3.17) that $\mathbb{L}u \in H_0^{2\ m}$. Hence, by the Paley–Wiener theorem, $u \in L^2(\mathbb{R}_+, \mathbb{R}^m)$. Moreover, $\mathbf{\Sigma}_{plant}$ is exponentially stable and driven by u, and therefore by Lemma 2.2, $x \in L^2(\mathbb{R}_+, X)$. Thus, we see that for all $(x_0, u_0) \in X \times \mathbb{R}^m$

$$t \mapsto \tilde{\mathbf{T}}_t^{\Gamma}(x_0, u_0) \in L^2(\mathbb{R}_+, X \times \mathbb{R}^m)\,.$$

By a well-known result on the stability of C_0-semigroups (cf. Pazy [30, p. 116]) it follows that the semigroup $\tilde{\mathbf{T}}^{\Gamma}$ is exponentially stable.

(ii) Since $\tilde{D} = 0$, it follows from Theorem 2.3 (ii) that

$$\tilde{C}^{\Gamma}(x, u) = \tilde{C}_L(x, u) \qquad \forall\,(x, u) \in D(\tilde{A}^{\Gamma})\,.$$

An application of Lemma 3.11 shows that $\tilde{C}^{\Gamma} = \tilde{C}$.

Let $(x_0, u_0) \in D(\tilde{A})$. Then, by Lemma 3.11, $(x_0, u_0) \in \tilde{X}_1^{\Gamma}$. By part (i) the semigroup $\tilde{\mathbf{T}}^{\Gamma}$ is exponentially stable on $\tilde{X} = X \times \mathbb{R}^m$, and hence it is also exponentially stable on $\tilde{X}_1^{\Gamma}$. Since $\tilde{C}^{\Gamma} \in \mathcal{L}(\tilde{X}_1^{\Gamma}, \mathbb{R}^m)$, it follows from the above that $\tilde{C} \in \mathcal{L}(\tilde{X}_1^{\Gamma}, \mathbb{R}^m)$ as well. As a consequence there exist $M, \omega > 0$ such that

$$\|\tilde{C}\tilde{\mathbf{T}}_t^{\Gamma}(x_0, u_0)\| \le Me^{-\omega t}\|(x_0, u_0)\|_{\tilde{X}_1^{\Gamma}} \qquad \forall\, t \ge 0\,.$$

The last statement of part (ii) follows from the fact that the boundedness of the observation operator C implies the boundedness of the observation operator $\tilde{C}$. ∎

Remark 3.13. Part (i) of Proposition 3.12 shows that in our special situation (i.e., the plant is exponentially stable and the controller is an integrator) input-output stability implies exponential stability. Using a result by Rebarber [37], it can be

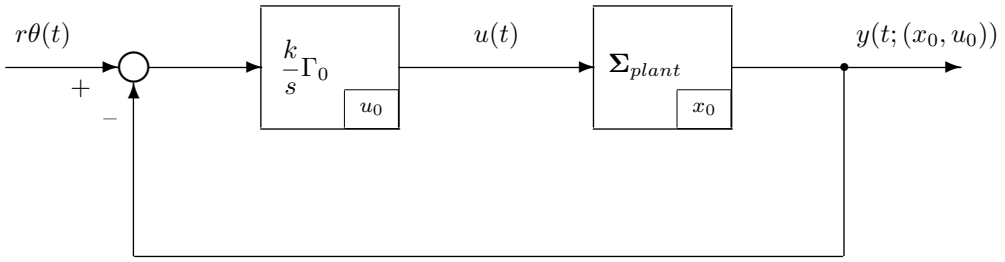


FIG. 3.6. *Low-gain control system.*

shown (Weiss [50]) that under suitable stabilizability and detectability assumptions the feedback interconnection of any two linear regular systems is exponentially stable if it is input-output stable. Since this result is not yet available in the literature (not even in form of a preprint), we have included a proof of Proposition 3.12 (i).

We are now in the position to prove the main result of this section, an internal version of Theorem 3.8 which applies to abstract linear regular state-space systems. Consider the feedback system in Figure 3.6, where $r \in \mathbb{R}^m$, $\Gamma_0 \in \mathbb{R}^{m\times m}$, $k > 0$, and $(x_0, u_0) \in X \times \mathbb{R}^m$. The output $y(\cdot\,;(x_0,u_0))$ can be written in the form

$$y(t;(x_0,u_0)) = \tilde{C}_L^{k\Gamma_0}\tilde{\mathbf{T}}_t^{k\Gamma_0}(x_0,u_0) + y(t;(0,0))\,. \tag{3.18}$$

Moreover, we define the corresponding error by

$$e(t;(x_0,u_0)) = r\theta(t) - y(t;(x_0,u_0))\,.$$

THEOREM 3.14. *Let $r \in \mathbb{R}^m$. Suppose that $\det \mathbf{G}(0) \neq 0$ and let $\Gamma_0 \in \mathbb{R}^{m\times m}$ be such that $\sigma(\mathbf{G}(0)\Gamma) \subset \mathbb{C}_0$. Then there exists $k^* > 0$ such that for any $k \in (0,k^*)$ the closed-loop semigroup $\tilde{\mathbf{T}}^{k\Gamma_0}$ is exponentially stable and $e(\cdot\,;(x_0,u_0)) \in L^2(\mathbb{R}_+,\mathbb{R}^m)$ for all $(x_0,u_0) \in X \times \mathbb{R}^m$. Furthermore,*

$$\lim_{t\to\infty} e(t;(x_0,u_0)) = 0 \quad \forall\,(x_0,u_0) \in D(\tilde{A})\,.$$

If the observation operator C is bounded, then the above equation holds for all $(x_0,u_0) \in X \times \mathbb{R}^m$.

Remark 3.15. If $(x_0,u_0) \notin D(\tilde{A})$, then in general $e(t) := e(t;(x_0,u_0))$ will not converge to 0 as $t \to \infty$. (In fact $e(\cdot)$ does not even make sense pointwise.) However, by Theorem 3.14, we still have that $e \in L^2(\mathbb{R}_+,\mathbb{R}^m)$, which implies that $e(t)$ *converges to* 0 *in measure* as $t \to \infty$ in the sense that for any $\varepsilon > 0$ and any $\delta > 0$ there exists $T = T(\varepsilon,\delta) > 0$ such that

$$\boldsymbol{\lambda}(\{t \in [\tau,\infty)\,|\,|e(t)| > \delta\} < \varepsilon \quad \forall\,\tau \geq T\,,$$

where $\boldsymbol{\lambda}$ denotes the Lebesgue measure.

Proof of Theorem 3.14. As in Theorem 3.8 we set $\mathbf{E}_k(s) = (1/s)(I+\mathbf{G}\mathbf{K}_k)^{-1}(s)$. By Theorem 3.8 there exists a $k^* > 0$ such that for all $k \in (0,k^*)$ the compensator $\mathbf{K}_k(s) = (1/s)k\Gamma_0$ stabilizes $\mathbf{G}$, and furthermore

$$\mathbf{E}_k \in H_-^{2\ m\times m} \quad \text{and} \quad [\mathbf{G}\mathbf{K}_k(I+\mathbf{G}\mathbf{K}_k)^{-1}](0) = I \qquad \forall\,k \in (0,k^*)\,. \tag{3.19}$$

In particular it follows from Proposition 3.12(i) that the semigroup $\tilde{\mathbf{T}}^{k\Gamma_0}$ is exponentially stable for all $k \in (0,k^*)$. Moreover, we have that

$$e(\cdot\,;(0,0)) = \mathbb{L}^{-1}(\mathbf{E}_k r)\,, \quad y(\cdot\,;(0,0)) = \mathbb{L}^{-1}[\mathbf{G}\mathbf{K}_k(I+\mathbf{G}\mathbf{K}_k)^{-1} \star \theta r]\,,$$

and therefore we obtain, using (3.19) and Proposition 3.5,

$$e(\cdot\,;(0,0)) \in L^2(\mathbb{R}_+,\mathbb{R}^m) \quad \text{and} \quad \lim_{t\to\infty} e(t;(0,0)) = 0\,, \tag{3.20}$$

provided that $k \in (0,k^*)$. Since the function

$$y_0(t;(x_0,u_0)) := \tilde{C}_L^{k\Gamma_0}\tilde{\mathbf{T}}_t^{k\Gamma_0}(x_0,u_0)$$

is the output of an exponentially stable regular system, it follows from Lemma 2.2 that $y_0(\cdot\,;(x_0,u_0)) \in L^2(\mathbb{R}_+,\mathbb{R}^m)$ for all $(x_0,u_0) \in X \times \mathbb{R}^m$ and all $k \in (0,k^*)$. Now, by (3.18),

$$e(t;(x_0,u_0)) = e(t;(0,0)) - y_0(t;(x_0,u_0))\,,$$

and thus, using (3.20), we obtain

$$e(\cdot\,;(x_0,u_0)) \in L^2(\mathbb{R}_+,\mathbb{R}^m) \quad \forall\,(x_0,u_0) \in X \times \mathbb{R}^m\,,$$

provided that $k \in (0,k^*)$. Finally, let $(x_0,u_0) \in D(\tilde{A})$. Then, by Proposition 3.12(ii), we may conclude that

$$\lim_{t\to\infty} y_0(t;(x_0,u_0)) = \lim_{t\to\infty} \tilde{C}_L^{k\Gamma_0}\tilde{\mathbf{T}}_t^{k\Gamma_0}(x_0,u_0) = \lim_{t\to\infty} \tilde{C}\tilde{\mathbf{T}}_t^{k\Gamma_0}(x_0,u_0) = 0\,. \tag{3.21}$$

Using (3.18), (3.20), and (3.21) we obtain that

$$\lim_{t\to\infty} e(t;(x_0,u_0)) = 0\,. \tag{3.22}$$

It follows from Proposition 3.12 (ii) that (3.22) holds for all $(x_0,u_0) \in X \times \mathbb{R}^m$ if the observation operator C is bounded. □

We close this section with a lemma which will be needed in section 4 in order to reformulate adaptive tracking problems as adaptive stabilization problems.

LEMMA 3.16. *For any $r \in \mathbb{R}^m$ there exists $(x_r,u_r) \in D(\tilde{A})$ such that*

$$\tilde{C}\tilde{\mathbf{T}}_t(x_r,u_r) = r \quad \forall\, t \geq 0\,.$$

Proof. For given $r \in \mathbb{R}^m$ define

$$x_r := -A^{-1}B\mathbf{G}^{-1}(0)r\,, \quad u_r := \mathbf{G}^{-1}(0)r\,.$$

Then $(x_r,u_r) \in X \times \mathbb{R}^m$, and moreover $Ax_r + Bu_r = 0$. It follows that $(x_r,u_r) \in \{(x,u) \in X \times \mathbb{R}^m \,|\, Ax + Bu \in X\} = D(\tilde{A})$, and by (3.14), $\tilde{A}(x_r,u_r) = 0$. We therefore easily conclude that $\tilde{\mathbf{T}}_t(x_r,u_r) = (x_r,u_r)$ for all $t \geq 0$. Finally, since $\mathbf{G}(0) = D - C_LA^{-1}B$, we see that for all $t \geq 0$

$$\tilde{C}\tilde{\mathbf{T}}_t(x_r,u_r) = C_Lx_r + Du_r = (\mathbf{G}(0) - D)\mathbf{G}^{-1}(0)r + D\mathbf{G}^{-1}(0)r = r\,. \qquad \square$$

4. Adaptive low-gain control of multivariable systems with sign-definite steady-state gain. In this section we consider the adaptive low-gain control of systems with *sign-definite* steady-state gains $\mathbf{G}(0)$, that is where either $\mathbf{G}(0) \succ 0$ or $\mathbf{G}(0) \prec 0$. This situation arises most naturally in the single-input single-output case where we need to assume only that the steady-state gain is nonzero.[4] In the multivariable case the situation of significance is when the steady-state gain is positive definite (see Propositions 4.4 and 4.6).

[4] Of course, we also need that $\mathbf{G}(0)$ is real. This will always be the case if $\mathbf{G}$ is the transfer function of a regular system, which is real by definition.

Consider the control law given by

$$\dot{u}(t) = \mathcal{K}(\gamma(t))e(t)\,, \quad u(0) = u_0\,, \tag{4.1a}$$

$$\dot{\gamma}(t) = \|e(t)\|^2\,, \quad \gamma(0) = \gamma_0 > a \geq -\infty\,, \tag{4.1b}$$

where $\mathcal{K} : (a, \infty) \to \mathbb{R}$ is locally Lipschitz. In the following $\mathcal{K}$ will be called a *tuning function*. Choosing $a = 0$ and

$$\mathcal{K}(\gamma) = \sin(\gamma^q)/\gamma^p\,, \quad 0 < q < p < 1 - q, \tag{4.2}$$

Cook [1] has shown that (4.1) is a universal adaptive, low-gain tracking controller for the class of single-input single-output, exponentially stable, finite-dimensional, linear systems with transfer function $\mathbf{G}$, input function $u(\cdot)$, output function $y(\cdot)$, and constant reference signal $r\theta(t), r \in \mathbb{R}$, in the sense that (i) $e(t) = (r - y(t)) \to 0$ as $t \to \infty$ and (ii) state and input functions remain bounded, independently of initial data, provided that $\mathbf{G}(0) \neq 0$. It is also shown in [1] that if $\mathbf{G}(0) > 0$, then $\mathcal{K}$ in (4.2) can be replaced by $\mathcal{K}(\gamma) = \gamma^{-p}$, $0 < p < 1$. The main tool in [1] is the fact that the return difference function is positive real for all k small enough and of the correct sign. It is clear, using Lemma 3.10, that these results extend to the multivariable case provided that $\mathbf{G}(0)$ is sign-definite.

In this section we prove that with different, suitably chosen tuning functions $\mathcal{K}$, these results extend to the case when the system is infinite-dimensional, regular, and exponentially stable. However, first we give alternative proofs of the finite-dimensional results in [1].

The finite-dimensional case. Our approach is based on Proposition 3.9, i.e., the fact that the H^∞-norm of the closed-loop transfer function $\tilde{\mathbf{G}}^k$ equals $1/|k|$ for all small enough k of the correct sign, and on the connection between this result and the existence of solutions to certain algebraic Riccati equations which arise in the characterization of the complex stability radius given in Hinrichsen and Pritchard [6]. We note that whilst neither this approach based on the algebraic Riccati equation nor the approach based on positive realness of the return difference equation and associated Lur'e equations extends to general regular systems, Proposition 3.9 will remain a crucial tool in the infinite-dimensional case.

LEMMA 4.1. *There exists $k^* > 0$ such that for any k with $|k| < k^*$ and $k\mathbf{G}(0) \succ 0$ the Riccati equation*

$$(\tilde{A} - k\tilde{B}\tilde{C})^T Z + Z(\tilde{A} - k\tilde{B}\tilde{C}) - k^2\tilde{C}^T\tilde{C} - Z\tilde{B}\tilde{B}^T Z = 0, \tag{4.3}$$

where $\tilde{A}, \tilde{B}$, and $\tilde{C}$ are given by (3.14)*, has a unique solution $\tilde{P}_k = \tilde{P}_k^T \preceq 0$.*

Proof. An application of Theorem 3.8 and Proposition 3.9 shows the existence of a constant $k^* > 0$ such that for any k with $|k| < k^*$ and $k\mathbf{G}(0) \succ 0$ the matrix $\tilde{A} - k\tilde{B}\tilde{C}$ is exponentially stable and $\|\tilde{\mathbf{G}}^k\|_\infty = 1/|k|$. Therefore the existence of a unique $\tilde{P}_k = \tilde{P}_k^T \preceq 0$ satisfying (4.3) is guaranteed by Hinrichsen and Pritchard [6, pp. 107–109]. □

The above lemma can now be used to give an alternative proof of the main result in [1].

THEOREM 4.2. *Let*

$$\dot{x}(t) = Ax(t) + Bu(t)\,, \quad x(0) = x_0 \in \mathbb{R}^n\,, \tag{4.4a}$$

$$y(t) = Cx(t) + Du(t) \tag{4.4b}$$

be any finite-dimensional, m-input m-output, exponentially stable system with sign-definite steady-state gain $\mathbf{G}(0)$. *Moreover, let* $\mathcal{K} : (a, \infty) \to \mathbb{R}$, *where* $a \geq -\infty$, *be locally Lipschitz and bounded with* $\mathcal{K} \in L^2(b, \infty; \mathbb{R})$ *for some* $b > a$ *and such that*

$$\liminf_{\gamma\to\infty} \int_b^\gamma \mathcal{K}(\xi)\, d\xi = -\infty\,, \quad \limsup_{\gamma\to\infty} \int_b^\gamma \mathcal{K}(\xi)\, d\xi = +\infty\,. \tag{4.5}$$

If $r\theta(t)$, $r \in \mathbb{R}^m$, *is any constant reference signal and* $u(t)$ *is defined by* (4.1), *with* $e(t) = r - y(t)$, *then for each* $\gamma_0 > a$, $x_0 \in \mathbb{R}^n$, *and* $u_0 \in \mathbb{R}^m$ *the following statements hold:*

(i) $\lim_{t\to\infty} \gamma(t) = \gamma_\infty < \infty$;
(ii) $\|x(t)\|$ *and* $\|u(t)\|$ *remain bounded as* $t \to \infty$;
(iii) $\lim_{t\to\infty} y(t) = r$.

Proof. The first step is to realize the reference signal $r\theta$ as an unforced motion of the series connection of the integrator $1/s$ followed by (4.4). By Lemma 3.16, applied in this simple finite-dimensional context, there exists $(x_r, u_r) \in \mathbb{R}^n \times \mathbb{R}^m$ such that $r = \tilde{C}e^{\tilde{A}t}(x_r, u_r)$ for all $t \geq 0$. It follows that

$$e(t) = r - y(t) = \tilde{C}\tilde{x}(t)\,, \tag{4.6}$$

where $\tilde{x}(t)$ is given by

$$\tilde{x}(t) = e^{\tilde{A}t}(x_r, u_r) - (x(t), u(t))\,. \tag{4.7}$$

Clearly, $\tilde{x}(\cdot)$ satisfies

$$\dot{\tilde{x}}(t) = \tilde{A}\tilde{x}(t) - \mathcal{K}(\gamma(t))\tilde{B}\tilde{C}\tilde{x}(t) \tag{4.8}$$

$$= (\tilde{A} - k\tilde{B}\tilde{C})\tilde{x}(t) - (\mathcal{K}(\gamma(t) - k)\tilde{B}e(t)\,, \tag{4.9}$$

where $k \in \mathbb{R}$ is arbitrary. Now the right-hand sides of (4.1b) and (4.8) are locally Lipschitz in $\tilde{x}$ and γ so that $\tilde{x}(t)$ and $\gamma(t)$ are uniquely determined on a maximal interval of existence—say, $[0, \tau)$. We now invoke Lemma 4.1 and define

$$V(t) = -\langle \tilde{x}(t), \tilde{P}_k\tilde{x}(t)\rangle,$$

where $\tilde{P}_k = \tilde{P}_k^T \preceq 0$ is the unique solution of (4.3), with $|k|$ small enough and $k\mathbf{G}(0) \succ 0$. Differentiating V along solutions of (4.1b) and (4.8) gives

$$\begin{aligned}
\dot{V} &= -k^2\|\tilde{C}\tilde{x}\|^2 - \|\tilde{B}^T\tilde{P}_k\tilde{x}\|^2 - 2(\mathcal{K}(\gamma) - k)\langle\tilde{C}\tilde{x}, \tilde{B}^T\tilde{P}_k\tilde{x}\rangle \\
&= -(k^2 - (\mathcal{K}(\gamma) - k)^2)\|\tilde{C}\tilde{x}\|^2 - \|(\mathcal{K}(\gamma) - k)\tilde{C}\tilde{x} + \tilde{B}^T\tilde{P}_k\tilde{x}\|^2 \\
&\leq \mathcal{K}(\gamma)(\mathcal{K}(\gamma) - 2k)\|\tilde{C}\tilde{x}\|^2\,.
\end{aligned}$$

Integrating this inequality from t_0 to t, where $0 \leq t_0 < t < \tau$, and using (4.1b) and (4.6) yield

$$-\infty < -V(t_0) \leq V(t) - V(t_0) \leq \int_{\gamma(t_0)}^{\gamma(t)} \mathcal{K}(\xi)(\mathcal{K}(\xi) - 2k)\, d\xi\,. \tag{4.10}$$

Seeking a contradiction, assume that $\lim_{t\to\tau} \gamma(t) = \infty$. Then, using (4.5) and exploiting the assumption that $\mathcal{K} \in L^2(b, \infty; \mathbb{R})$ we obtain

$$\lim_{n\to\infty} \int_{\gamma(t_0)}^{\gamma(t_n)} \mathcal{K}(\xi)(\mathcal{K}(\xi) - 2k)\, d\xi = -\infty$$

for some sequence $(t_n)_{n\in\mathbb{N}}$ with $\gamma(t_0) = b$ and $\lim_{n\to\infty} t_n = \tau$. Since this contradicts (4.10), it follows that $\gamma(t)$ is bounded on $[0,\tau)$ and consequently $\tau = \infty$, which establishes (i).

In order to prove statements (ii) and (iii), choose k in (4.9) such that $\tilde{A} - k\tilde{B}\tilde{C}$ is exponentially stable (this is possible by Theorem 3.8). Trivially, by (i), $e \in L^2(\mathbb{R}_+, \mathbb{R}^m)$, and so it follows from the boundedness of $\mathcal{K}$ that the forcing term on the right-hand side of (4.9) is in $L^2(\mathbb{R}_+, \mathbb{R}^m)$. Therefore $\tilde{x}(t)$ is the state of an exponentially stable system driven by an L^2-input, and consequently $\lim_{t\to\infty} \tilde{x}(t) = 0$. Statements (ii) and (iii) follow now from (4.7) and (4.6), respectively. □

Remark 4.3. Whilst the property of symmetry for a general $m \times m$ matrix is non-generic in that symmetry is destroyed by arbitrarily small perturbations, symmetry of $\mathbf{G}(0)$ is a direct consequence of, for example,

$$A = A^T\,, \quad B = C^T, \quad \text{and} \quad D = D^T\,.$$

If additionally, $D \succeq 0$, then positive definiteness of $\mathbf{G}(0)$ follows, since A is exponentially stable and $\mathbf{G}(0)$ is invertible.

It is not difficult to show that the function given in (4.2) satisfies the conditions imposed on $\mathcal{K}$ in Theorem 4.2. Notice that in general these conditions do not imply that $\lim_{\gamma\to\infty} \mathcal{K}(\gamma) = 0$.

PROPOSITION 4.4. *Suppose* $\mathbf{G}(0) \succ 0$. *With the tuning function* $\mathcal{K}(\gamma) = \gamma^{-p}$, $0 < p < 1$, *and* $\gamma_0 > 0$ *statements* (i)–(iii) *of Theorem* 4.2 *hold.*

Proof. The proof is the same as that of Theorem 4.2 up to (4.10). By the special choice of $\mathcal{K}$, (4.10) implies that $\gamma(\cdot)$ is bounded. The remainder of the proof is the same as that of Theorem 4.2. □

In Proposition 4.4 we may replace γ^{-p} by any function $\mathcal{K}$ which satisfies

$$\int_{\gamma_0}^{\infty} \mathcal{K}(\xi)(\mathcal{K}(\xi) - 2k)d\xi = -\infty$$

for some stabilizing gain $k > 0$.

The infinite-dimensional case. For the rest of this paper we will let $\mathbf{\Sigma}_{plant} = (\mathbf{T}, \mathbf{\Phi}, \mathbf{\Psi}, \mathbf{F})$ be an exponentially stable regular system with transfer function $\mathbf{G}$. Let A, B, C, and D denote the generating operators of $\mathbf{\Sigma}_{plant}$. As in section 3 we denote the series connection of the integrator $1/s$ followed by $\mathbf{\Sigma}_{plant}$ by $\tilde{\mathbf{\Sigma}} = (\tilde{\mathbf{T}}, \tilde{\mathbf{\Phi}}, \tilde{\mathbf{\Psi}}, \tilde{\mathbf{F}})$. It was shown in section 3 that the system $\tilde{\mathbf{\Sigma}}$ is regular. Let $\tilde{A}$, $\tilde{B}$, and $\tilde{C}$ denote the corresponding generating operators (trivially, $\tilde{D} = 0$), and let $\tilde{\mathbf{G}}(s) = (1/s)\mathbf{G}(s)$ denote the transfer function of $\tilde{\mathbf{\Sigma}}$.

We were not able to extend the proofs of Theorem 4.2 and Proposition 4.4 to the infinite-dimensional setting outlined in section 2. The problem is caused by the fact that Lemma 4.1 does not hold in the infinite-dimensional case, unless very strong and unnatural controllability assumptions are imposed. As already mentioned in the introduction, the approach in Cook [1] does not carry over to infinite-dimensional systems either. Nevertheless, it will turn out that in the infinite-dimensional situation we can still use tuning functions $\mathcal{K}$ satisfying $\mathcal{K}(\gamma) \to 0$ as $\gamma \to \infty$.

THEOREM 4.5. *Let* $\mathbf{\Sigma}_{plant}$ *be a* m*-input* m*-output exponentially stable regular system given by* (2.1). *Suppose that the transfer function* $\mathbf{G}$ *of* $\mathbf{\Sigma}_{plant}$ *is such that* $\mathbf{G}(0)$ *is sign definite. Let* $r\,\theta(t)$, $r \in \mathbb{R}^m$, *be an arbitrary constant vector-valued reference signal, and consider the control law*

$$u(t) = u_0 + \int_0^t \log^{-p}\gamma(\xi)\cos(\log^q\gamma(\xi))e(\xi)\,d\xi\,, \tag{4.11}$$

$$\dot{\gamma}(t) = \|e(t)\|^2\,, \quad \gamma(0) = \gamma_0\,, \tag{4.12}$$

where $e(t) = r - y(t)$ *and* $p \geq 0$, $q > 0$, *and* $q + 2p < 1$. *Then for all* $(x_0, u_0) \in X \times \mathbb{R}^m$ *and* $\gamma_0 > 1$, *where* X *denotes the state space of* $\mathbf{\Sigma}_{plant}$, *the following statements hold true:*

(i) $\lim_{t\to\infty} \gamma(t) = \gamma_\infty < \infty$;
(ii) $\|x(t)\|$ *and* $\|u(t)\|$ *remain bounded as* $t \to \infty$;
(iii) $e(\cdot) \in L^2(\mathbb{R}_+, \mathbb{R}^m)$.

Moreover, if $(x_0, u_0) \in D(\tilde{A})$, *then*

$$\lim_{t\to\infty} y(t) = r\,. \tag{4.13}$$

If the observation operator C *of* $\mathbf{\Sigma}_{plant}$ *is bounded, then* (4.13) *is true for all* $(x_0, u_0) \in X \times \mathbb{R}^m$.

Proof. We assume throughout the proof that $p > 0$. The case $p = 0$ can be proven using the techniques in the proof of Theorem 5.1. The first step is to convert the tracking problem ($r \neq 0$) into a stabilization problem ($r = 0$). By Lemma 3.16 there exists $(x_r, u_r) \in D(\tilde{A})$ so that $r = \tilde{C}\tilde{\mathbf{T}}_t(x_r, u_r)$ for all $t \geq 0$. Therefore, setting $\mathcal{K}(\gamma) = \log^{-p}\gamma \cos(\log^q \gamma)$ and using (4.11), it follows that

$$e = r\theta - y = \tilde{\mathbf{\Psi}}_\infty(x_r - x_0, u_r - u_0) - \tilde{\mathbf{F}}_\infty(\mathcal{K}(\gamma)e)\,. \tag{4.14}$$

The nonlinear closed-loop system given by (4.14) and (4.12) is in a form so that Proposition 2.4 is applicable. Let $[0, \tau)$ be the maximal interval of existence for solutions (e, γ) of (4.14) and (4.12) as guaranteed by Proposition 2.4. We know that $\tau < \infty$ only if $\lim_{t\to\tau} \gamma(t) = \infty$. We will prove that $\gamma(t)$ is bounded on $[0, \tau)$.

Let $(\rho_i)_{i\in\mathbb{N}}$, with $\rho_0 \geq \gamma_0$, be a strictly increasing sequence converging to ∞ and satisfying

$$\operatorname{sign}(\mathbf{G}(0)) \cos(\log^q \rho_{2i}) = 1 \quad \text{and} \quad \mathcal{K}(\rho_{2i+1}) = \mathcal{K}(\rho_{2i})/2\,, \quad i = 0, 1, 2, \ldots\,,$$

where $\operatorname{sign}(\mathbf{G}(0)) = \pm 1$, depending on whether $\mathbf{G}(0)$ is positive or negative definite. Choosing ρ_0 sufficiently large, it follows from Theorem 3.8 that the numbers

$$k_i := \mathcal{K}(\rho_{2i})$$

are stabilizing gains for $\tilde{\mathbf{G}}(s) = (1/s)\mathbf{G}(s)$; i.e., the integrators k_i/s stabilize $\mathbf{G}$ in the sense of Definition 3.1. Note that $(\rho_i)_{i\in\mathbb{N}}$ can be chosen so that

$$|\mathcal{K}(\gamma)| \in (|k_i|/2, |k_i|) \quad \text{and} \quad k_i\mathcal{K}(\gamma) > 0 \qquad \forall\, \gamma \in (\rho_{2i}, \rho_{2i+1})$$

and that $|k_i| \searrow 0$ as $i \to \infty$. Moreover, by applying Proposition 3.9 we can always choose ρ_0 sufficiently large so that

$$\|\tilde{\mathbf{G}}^{k_i}\|_\infty = \frac{1}{|k_i|} \tag{4.15}$$

for all i.

Seeking a contradiction, suppose that $\gamma(t)$ is unbounded on $[0, \tau)$. Then we can find a sequence of times $t_0 < t_1 < \cdots < \tau$ with

$$\gamma(t_i) = \rho_i.$$

We now use these observations combined with estimates we obtain from contraction-mapping–type arguments. Using (2.9b) on each interval $[t_{2i}, t_{2i+1}]$ we can write the error $e(\cdot)$ as

$$\mathbf{L}_{t_{2i}}\mathbf{P}_{t_{2i+1}}e = \tilde{\mathbf{\Psi}}^{k_i}_{t_{2i+1}-t_{2i}}(\tilde{x}(t_{2i})) - \tilde{\mathbf{F}}^{k_i}_{t_{2i+1}-t_{2i}}(\mathbf{L}_{t_{2i}}\mathbf{P}_{t_{2i+1}}(\mathcal{K}(\gamma) - k_i)e)\,, \tag{4.16}$$

[5]

[5] By slight abuse of notation we write $\tilde{\mathbf{\Psi}}^{k_i}_{t_{2i+1}-t_{2i}}$ instead of $\tilde{\mathbf{\Psi}}^{k_i I}_{t_{2i+1}-t_{2i}}$, etc.

where

$$\tilde{x}(t) = \tilde{\mathbf{T}}_t(x_r - x_0, u_r - u_0) - \tilde{\mathbf{\Phi}}_t(\mathbf{P}_t\mathcal{K}(\gamma)e)\,.$$

By using (2.8a) we can express $\tilde{x}(t)$ as

$$\tilde{x}(t) = \tilde{\mathbf{T}}^{k_0}_t(x_r - x_0, u_r - u_0) - \tilde{\mathbf{\Phi}}^{k_0}_t(\mathbf{P}_t(\mathcal{K}(\gamma) - k_0)e)\,. \tag{4.17}$$

Using (2.7) and (2.8b), with $u = 0$ and $K = k_0 - k_i$, we obtain

$$\tilde{\mathbf{\Psi}}^{k_i}_t z = \tilde{\mathbf{\Psi}}^{k_0}_t z - \tilde{\mathbf{F}}^{k_0}_t((k_i - k_0)\tilde{\mathbf{\Psi}}^{k_i}_t z) \quad \forall\, t \geq 0\,,\ \forall\, z \in X \times \mathbb{R}^m\,. \tag{4.18}$$

Now for all $t \in [t_{2i}, t_{2i+1}]$ we have

$$|\mathcal{K}(\gamma(t)) - k_i| \leq \frac{|k_i|}{2}.$$

Moreover, $\|\tilde{\mathbf{F}}^{k_i}_\infty\| = \|\tilde{\mathbf{G}}^{k_i}\|_\infty$, and hence it follows from (4.15) that

$$\|\tilde{\mathbf{F}}^{k_i}_{t_{2i+1}-t_{2i}}\| \leq \|\tilde{\mathbf{F}}^{k_i}_\infty\| = \frac{1}{|k_i|}\,, \quad \text{whilst} \quad \|\tilde{\mathbf{F}}^{k_0}_t\| \leq \|\tilde{\mathbf{F}}^{k_0}_\infty\| = \frac{1}{|k_0|}\,.$$

Therefore integrating (4.16) from 0 to $t_{2i+1} - t_{2i}$ and taking estimates we have

$$\|e\|_{L^2(t_{2i},t_{2i+1})} \leq \frac{1}{1 - \|\tilde{\mathbf{F}}^{k_i}_\infty\|\,\|\mathcal{K}(\gamma) - k_i\|_{L^\infty(t_{2i},t_{2i+1})}}\|\tilde{\mathbf{\Psi}}^{k_i}_{t_{2i+1}-t_{2i}}(\tilde{x}(t_{2i}))\|_{L^2(0,t_{2i+1}-t_{2i})}$$

$$\leq 2\|\tilde{\mathbf{\Psi}}^{k_i}_{t_{2i+1}-t_{2i}}(\tilde{x}(t_{2i}))\|_{L^2(0,t_{2i+1}-t_{2i})}\,. \tag{4.19}$$

Since $\tilde{\mathbf{G}}^{k_i} \in H^{\infty\ m\times m}_-$, an application of Theorem 3.14 yields that the closed-loop semigroup $\tilde{\mathbf{T}}^{k_i}$ is exponentially stable. It follows that $\tilde{\mathbf{\Psi}}^{k_i}_\infty z \in L^2(\mathbb{R}_+, \mathbb{R}^m)$ for all $z \in X \times \mathbb{R}^m$. As a consequence, integrating in (4.18) from 0 to ∞ and taking estimates gives

$$\|\tilde{\mathbf{\Psi}}^{k_i}_\infty(\tilde{x}(t_{2i}))\|_{L^2(0,\infty)} \leq \frac{1}{1 - \|\tilde{\mathbf{F}}^{k_0}_\infty\||k_0 - k_i|}\|\tilde{\mathbf{\Psi}}^{k_0}_\infty(\tilde{x}(t_{2i}))\|_{L^2(0,\infty)}$$

$$= \frac{k_0}{k_i}\|\tilde{\mathbf{\Psi}}^{k_0}_\infty(\tilde{x}(t_{2i}))\|_{L^2(0,\infty)}\,. \tag{4.20}$$

Combining (4.19) and (4.20) and using the definition of $\gamma(t)$, we obtain

$$\sqrt{\rho_{2i+1} - \rho_{2i}} = \|e\|_{L^2(t_{2i},t_{2i+1})} \leq 2\frac{k_0}{k_i}\|\tilde{\mathbf{\Psi}}^{k_0}_\infty\tilde{x}(t_{2i})\|_{L^2(0,\infty)} \leq \frac{c_0}{|k_i|}\|\tilde{x}(t_{2i})\|\,, \tag{4.21}$$

where $c_0 > 0$ is a constant obtained from the exponential stability of the semigroup $\tilde{\mathbf{T}}^{k_0}$. Setting $t = t_{2i}$ in (4.17) and taking estimates yield

$$\|\tilde{x}(t_{2i})\| \leq c_1 + c_2\sqrt{\rho_{2i} - \gamma_0} \tag{4.22}$$

for suitable constants $c_1 > 0$ and $c_2 > 0$. Combining (4.21) and (4.22) and using the fact that $k_i = \mathcal{K}(\rho_{2i})$, we have

$$\sqrt{\rho_{2i+1} - \rho_{2i}} \leq \frac{c_0}{|\mathcal{K}(\rho_{2i})|}(c_1 + c_2\sqrt{\rho_{2i} - \gamma_0})\,. \tag{4.23}$$

Now, by the mean value theorem, there exists $\xi_{2i} \in (\rho_{2i}, \rho_{2i+1})$ such that

$$-\frac{1}{\mathcal{K}'(\xi_{2i})} = \frac{\rho_{2i+1} - \rho_{2i}}{\mathcal{K}(\rho_{2i}) - \mathcal{K}(\rho_{2i+1})} = 2\,\frac{\rho_{2i+1} - \rho_{2i}}{\mathcal{K}(\rho_{2i})}$$

so that (4.23) becomes

$$\sqrt{\frac{-\mathcal{K}^3(\rho_{2i})}{2\mathcal{K}'(\xi)}} \le c_0(c_1 + c_2\sqrt{\rho_{2i} - \gamma_0})\,. \tag{4.24}$$

Using the specific form of $\mathcal{K}$ we have

$$\mathcal{K}'(\xi) = -\frac{p\cos(\log^q \xi) + q\log^q \xi \sin(\log^q \xi)}{\xi \log^{1+p} \xi}\,,$$

which on substituting in (4.24) and rearranging yields

$$1 \le 2\,\left[c_0(c_1 + c_2\sqrt{\rho_{2i} - \gamma_0})\right]^2 \frac{1}{|\mathcal{K}(\rho_{2i})|^3}\,\frac{|p\cos(\log^q \xi_{2i}) + q\log^q \xi_{2i}\sin(\log^q \xi_{2i})|}{\xi_{2i}\log^{1+p}\xi_{2i}}\,. \tag{4.25}$$

Using the fact that $\mathcal{K}(\rho_{2i}) = |\log^{-p} \rho_{2i}|$ and gathering dominant terms in (4.25) lead to

$$1 \le c_3\left(p(\log \rho_{2i})^{2p-1} + q(\log \rho_{2i})^{2p+q-1}\right) \tag{4.26}$$

for some constant $c_3 > 0$. But $p, q > 0$ and $q + 2p < 1$ so that the right-hand side of (4.26) approaches zero for $\rho_{2i} \to \infty$, which is in contradiction to (4.26). Hence $\gamma(\cdot)$ is bounded, which establishes statements (i) and (iii). Boundedness of $\tilde{x}(t)$ and therefore part (ii) follows directly from (4.17), the exponential stability of $\tilde{\mathbf{T}}^{k_0}$, and statements (i) and (iii).

To prove the last statement in the theorem let $(x_0, u_0) \in D(\tilde{A})$. Then $\tilde{x}_0 := (x_r - x_0, u_r - u_0) \in D(\tilde{A})$, and from (4.14) and (2.8b) we obtain

$$e(t) = \tilde{C}^{k_0}\tilde{\mathbf{T}}_t^{k_0}\tilde{x}_0 - (\tilde{\mathbf{F}}_\infty^{k_0}[(\mathcal{K}(\gamma) - k_0)e])(t) \quad \forall\, t \ge 0\,. \tag{4.27}$$

By Lemma 3.11, $\tilde{x}_0 \in D(\tilde{A}^{k_0})$, and hence it follows from the exponential stability of the semigroup $\tilde{\mathbf{T}}^{k_0}$ that the first term on the right-hand side of (4.27) tends exponentially to 0 as $t \to \infty$. In order to show that the second term converges to 0 as $t \to \infty$ set $v(t) = (\mathcal{K}(\gamma(t)) - k_0)e(t)$, and realize that, by statements (i) and (iii), $v \in L^2(\mathbb{R}_+, \mathbb{R}^m)$. Clearly,

$$(\mathbb{L}(\tilde{\mathbf{F}}_\infty^{k_0} v))(s) = \tilde{\mathbf{G}}(s)(1 + k_0\tilde{\mathbf{G}}(s))^{-1}(\mathbb{L}v)(s) = \tilde{\mathbf{G}}^{k_0}(s)(\mathbb{L}v)(s)\,.$$

Since $|k_0|$ is sufficiently small, it follows from Theorem 3.8 that $\tilde{\mathbf{G}}^{k_0} \in H_-^{2\ m\times m}$. (Note that using the notation in Theorem 3.8 we have $\tilde{\mathbf{G}}^{k_0} = \mathbf{G}\mathbf{E}_{k_0}$.) Therefore, by the Paley–Wiener theorem, $\tilde{\mathbf{F}}_\infty^{k_0}$ is a convolution operator with a matrix-valued kernel whose entries are L^2-functions. Now it is well known that the convolution of two functions belonging to $L^2(\mathbb{R}_+, \mathbb{R})$ converges to 0 as $t \to \infty$, and hence $\lim_{t\to\infty}(\tilde{\mathbf{F}}_\infty^{k_0} v)(t) = 0$. Finally, if C is bounded, then $\tilde{C}$ is bounded, i.e., $\tilde{C} \in \mathcal{L}(X \times \mathbb{R}^m, \mathbb{R}^m)$. Furthermore, by Proposition 3.12, $\tilde{C}^{k_0} = \tilde{C}$, and therefore the first term on the right-hand side of (4.27) converges (exponentially) to 0 as $t \to \infty$ for all $(x_0, u_0) \in X \times \mathbb{R}^m$. □

Note that the condition $(x_0, u_0) \in D(\tilde{A})$ in statement (iv) of Theorem 4.5 required in proving that $\lim_{t\to 0} e(t) = 0$ is system dependent. This is a little disturbing

since, from the outset, we assume that the specific system to be controlled is unknown. However, in most cases, the initial states will be sufficiently smooth so that the condition $(x_0, u_0) \in D(\tilde{A})$ is satisfied. Note that if $(x_0, u_0) \notin D(\tilde{A})$, then $e(\cdot)$ will in general not make sense pointwise and cannot be expected to converge to 0 in the usual sense (see, however, Remark 3.15).

Note that in the infinite-dimensional case the tuning function $\mathcal{K}(\gamma)$ decays to 0 like a fractional power of $\log \gamma$ as $\gamma \to \infty$, whereas in the finite-dimensional case it decays to 0 like a fractional power of γ. However, in the case when it is known that $\mathbf{G}(0) \succ 0$, we can use tuning functions which decay to 0 like a fractional power, although more slowly than in the finite-dimensional case.

PROPOSITION 4.6. *Suppose that the conditions of Theorem* 4.5 *hold and that additionally* $\mathbf{G}(0) \succ 0$. *If*

$$u(t) = u_0 + \int_0^t \gamma^{-p}(\xi) e(\xi)\, d\xi\,, \tag{4.28}$$

$$\dot{\gamma}(t) = \|e(t)\|^2\,, \quad \gamma(0) = \gamma_0 > 0\,, \tag{4.29}$$

and $0 < p < \frac{1}{2}$, *then the conclusions of Theorem* 4.5 *hold.*

Proof. It is sufficient to show that $\gamma(\cdot)$ is bounded. Let $[0, \tau)$ be the maximal interval of existence. If $\gamma(\cdot)$ is unbounded on $[0, \tau)$, then there exists $t_1 \geq 0$ such that with $\gamma_1 = \gamma(t_1)$, $k_1 = \gamma_1^{-p}$ is a stabilizing gain. For any $t \in (t_1, \tau)$ we have, as in the proof of Theorem 4.5, that on $[t_1, t]$

$$\mathbf{L}_{t_1} \mathbf{P}_t e = \tilde{\mathbf{\Psi}}^{k_1}_{t-t_1}(\tilde{x}(t_1)) - \tilde{\mathbf{F}}^{k_1}_{t-t_1}(\mathbf{L}_{t_1} \mathbf{P}_t (\mathcal{K}(\gamma) - k_1) e)\,. \tag{4.30}$$

We can assume that k_1 is small enough so that, using Proposition 3.9 and estimating, we obtain

$$\sqrt{\gamma(t) - \gamma_1} \leq c \gamma^p(t)$$

for some $c > 0$ and all $t \in [t_1, \tau)$. This inequality clearly contradicts the unboundedness of $\gamma(\cdot)$ and the assumption that $p < 1/2$. ∎

The condition $\mathbf{G}(0) \succ 0$ is satisfied for a large class of exponentially stable infinite-dimensional systems with self-adjoint generator A, co-located control and observation and positive semidefinite feedthrough (cf. Remark 4.3).

5. Adaptive low-gain control of multivariable systems with sign-indefinite steady-state gain. In this section we consider the adaptive low-gain tracking problem, for stable regular systems with square $m \times m$ transfer functions $\mathbf{G}(s)$ and invertible steady-state gain. In section 4, under the assumption that $\mathbf{G}(0)$ is sign definite, we could exploit the fact that for all gains k having the "correct" sign and with $|k|$ sufficiently small, $\|\tilde{\mathbf{G}}^k\|_\infty = 1/|k|$ (see Proposition 3.9). If $\mathbf{G}(0)$ is sign indefinite or even nonsymmetric, then, again by Proposition 3.9, we no longer have this result. To overcome this problem we do not use a tuning function $\mathcal{K}$ reflecting the low-gain nature of the problem in the sense that $\lim_{\gamma \to \infty} \mathcal{K}(\gamma) = 0$ but instead resort to a gain which oscillates smoothly between 0 and 2. (In fact, 2 could be replaced by any positive number δ.)

As in the previous sections let $u(\cdot)$ and $y(\cdot)$ denote the plant input and plant output, respectively, and set $e(\cdot) = r - y(\cdot)$, where $r \in \mathbb{R}^m$ is a demand vector. Modulo certain technicalities involving "spectrum unmixing" of $\mathbf{G}(0)$ (to be made precise) we show that

$$u(t) = u_0 + \int_0^t [1 + \cos(\log^q \gamma(\xi))] e(\xi)\, d\xi\,, \quad \text{where } 0 < q < 1\,, \tag{5.1}$$

$$\dot{\gamma}(t) = \|e(t)\|^2\,, \quad \gamma(0) = \gamma_0, \tag{5.2}$$

is a universal adaptive tracking controller.

We assume throughout that $\mathbf{\Sigma}_{plant}$ is an m-input m-output, exponentially stable, regular system given by (2.1). We will consider two cases. In the first one we assume that the spectrum of $\mathbf{G}(0)$ is unmixed in the sense that $\sigma(\mathbf{G}(0)) \subset \mathbb{C}_0$. In the second case the a priori knowledge about $\mathbf{G}(0)$ guarantees only that $\mathbf{G}(0)$ is invertible.

THEOREM 5.1. *Assume that $\sigma(\mathbf{G}(0)) \subset \mathbb{C}_0$. Let $r \in \mathbb{R}^m$ be an arbitrary demand vector. If $u(t)$ is given by* (5.1)*, with gain adaptation* (5.2)*, then for each $(x_0, u_0) \in X \times \mathbb{R}^m$ and $\gamma_0 > 1$ we have*

(i) $\lim_{t\to\infty} \gamma(t) = \gamma_\infty < \infty$;

(ii) *$\|x(t)\|$ and $\|u(t)\|$ remain bounded as $t \to \infty$;*

(iii) $e(\cdot) \in L^2(\mathbb{R}_+, \mathbb{R}^m)$.

Moreover, if $(x_0, u_0) \in D(\tilde{A})$, then

$$\lim_{t\to\infty} y(t) = r\,. \tag{5.3}$$

If the observation operator C is bounded, then (5.3) *holds for all $(x_0, u_0) \in X \times \mathbb{R}^m$.*

In the proof of this result we do not have to be so careful with the estimates, since we need only to work in a neighborhood of a stabilizing integral gain and do not need to account for the possibility of the feedback gain approaching 0.

Proof. The first step is to convert the tracking problem ($r \neq 0$) into a stabilization problem ($r = 0$). Let $r \in \mathbb{R}^m$, $(x_0, u_0) \in X \times \mathbb{R}^m$ be given and set $\mathcal{K}(\gamma) := 1 + \cos(\log^q \gamma(t))$. By Lemma 3.16 there exists $\tilde{x}_0 \in X \times \mathbb{R}^m$ such that

$$e = \tilde{\mathbf{\Psi}}_\infty \tilde{x}_0 - \tilde{\mathbf{F}}_\infty(\mathcal{K}(\gamma)e)\,. \tag{5.4}$$

Moreover, if $(x_0, u_0) \in D(\tilde{A})$, then $\tilde{x}_0 \in D(\tilde{A})$. The closed-loop system given by (5.4) and (5.2) is in a form so that Proposition 2.4 is applicable.

By Theorem 3.8 there exists $k \in (0, 1)$ for which $\tilde{\mathbf{G}}^k \in H_-^{\infty\, m\times m}$. Consequently, by Theorem 3.14, $\tilde{\mathbf{T}}^k$ is an exponentially stable semigroup on $X \times \mathbb{R}^m$. As in the sign-definite case, seeking a contradiction, suppose that $\gamma(t)$ is unbounded on the maximal interval of existence $[0, \tau)$. To this end choose $\varepsilon \in (0, k)$ such that $k + \varepsilon < 1$ and let $(\rho_i)_{i\in\mathbb{N}}$ be a sequence with

$$\rho_i \nearrow \infty\,, \quad \rho_0 \geq \gamma_0\,, \quad \mathcal{K}(\rho_{2i}) = k - \varepsilon\,, \quad \mathcal{K}(\rho_{2i+1}) = k + \varepsilon$$

and such that

$$\mathcal{K}(\gamma) \in (k - \varepsilon, k + \varepsilon) \quad \forall\, \gamma \in (\rho_{2i}, \rho_{2i+1})\,.$$

Exploiting the unboundedness of $\gamma(t)$ we can find a sequence of times $t_0 < t_1 < \cdots < \tau$ so that $\gamma(t_i) = \rho_i$. Using (2.9b) we obtain

$$\mathbf{L}_{t_{2i}}\mathbf{P}_{t_{2i+1}} e = \tilde{\mathbf{\Psi}}^k_{t_{2i+1}-t_{2i}} \tilde{x}(t_{2i}) - \tilde{\mathbf{F}}^k_{t_{2i+1}-t_{2i}}(\mathbf{L}_{t_{2i}}\mathbf{P}_{t_{2i+1}}(\mathcal{K}(\gamma) - k)e)\,, \tag{5.5}$$

where

$$\tilde{x}(t) = \tilde{\mathbf{T}}_t \tilde{x}_0 - \tilde{\mathbf{\Phi}}_t \mathbf{P}_t \mathcal{K}(\gamma) e\,.$$

Integrating from 0 to $t_{2i+1} - t_{2i}$ in (5.5) and taking estimates yield

$$\|e\|_{L^2(t_{2i},t_{2i+1})} \leq \frac{1}{1 - \|\tilde{\mathbf{F}}^k_\infty\| \|\mathcal{K}(\gamma) - k\|_{L^\infty(t_{2i},t_{2i+1})}} \|\tilde{\mathbf{\Psi}}^k_\infty \tilde{x}(t_{2i})\|_{L^2(0,\infty)}$$

$$\leq c_0 \|\tilde{x}(t_{2i})\| \tag{5.6}$$

for some suitable $c_0 > 0$, provided that ε is small enough (for example, $\|\tilde{\mathbf{F}}^k_\infty\|\varepsilon = 1/2$).

Applying the input-state variation of parameters formula (2.8a) to $\tilde{\mathbf{\Sigma}}$ with $K = kI$ and $u = \mathcal{K}(\gamma)e$ it follows from the exponential stability of $\tilde{\mathbf{T}}^k$ and (5.2) that

$$\|\tilde{x}(t_{2i})\| \le c_1 + c_2\sqrt{\rho_{2i} - \gamma_0} \tag{5.7}$$

for some constants $c_1 > 0$ and $c_2 > 0$. Combining (5.6) and (5.7) we have

$$\sqrt{\rho_{2i+1} - \rho_{2i}} \le c_0(c_1 + c_2\sqrt{\rho_{2i} - \gamma_0})\,. \tag{5.8}$$

Clearly,

$$\rho_{2i+1} - \rho_{2i} = 2\varepsilon/\mathcal{K}'(\xi_{2i})$$

for some $\xi_{2i} \in (\rho_{2i}, \rho_{2i+1})$. Combining this with (5.8) leads to

$$1 \le \frac{1}{2\varepsilon}[c_0(c_1 + c_2\sqrt{\rho_{2i} - \gamma_0})]^2\mathcal{K}'(\xi_{2i})\,. \tag{5.9}$$

Now

$$\mathcal{K}'(\xi) = -q\sin(\log^q \xi)(\log^{q-1} \xi)/\xi,$$

and $0 < q < 1$, and we see that the right-hand side of (5.9) converges to 0 as $i \to \infty$, which yields a contradiction. It follows that $\gamma(\cdot)$ is bounded, showing that (i) and (iii) hold true. The remaining claims follow readily, using the same techniques as in the proof of Theorem 4.5. □

Specialized to the case when $\mathbf{G}(0) \succ 0$, it is natural to compare the control law in Theorem 5.1 to the one in Proposition 4.6. Intuitively it should be advantageous to use the controller in Proposition 4.6, since in this case the gain passes rapidly into the "correct" parameter region once and remains there, whereas the gain in the controller in Theorem 5.1 oscillates slowly and may pass in and out of the "correct" region several times before converging. Moreover, small output disturbances could lead to further cycles in the gain adaptation.

In Theorem 5.1 we assumed that $\sigma(\mathbf{G}(0)) \subset \mathbb{C}_0$. We now consider the case when we know only that $\det \mathbf{G}(0) \neq 0$. In the context of high-gain adaptive stabilization Mårtensson [21], [22] has shown that there exists a finite set $\{\Gamma_1, \dots, \Gamma_\ell\}$ so that given any invertible $m \times m$ matrix M there exists $\nu \in \{1, 2, \dots, \ell\}$ such that $\sigma(M\Gamma_\nu) \subset \mathbb{C}_0$. We now use this result in order to unmix the spectrum of $\mathbf{G}(0)$. Consider the feedback law

$$u(t) = u_0 + \int_0^t [1 + \cos(\log^q \gamma(\xi))]\Gamma_{S(\gamma(\xi))}e(\xi)\,d\xi\,, \tag{5.10}$$

where $0 < q < 1$ and

$$S(\gamma) = j \quad \text{if} \quad (2\pi)^{-1}\log^q \gamma \in [p\ell + j, p\ell + j + 1) \quad \text{for some} \quad p \in \mathbb{N}.$$

Note that the feedback gain matrix in (5.10) is piecewise smooth but discontinuous whenever $(2\pi)^{-1}\log^q \gamma$ takes on integer values, so Proposition 2.4 is no longer valid. However, these discontinuities in the gain are easily handled by a minor modification to the proof of Proposition 2.4.

THEOREM 5.2. *Assume that* $\det \mathbf{G}(0) \neq 0$. *Let* $r \in \mathbb{R}^m$ *be an arbitrary demand vector. If* $u(t)$ *is given by* (5.10), *with adaptation* (5.2), *then for each* $(x_0, u_0) \in X \times \mathbb{R}^m$ *and* $\gamma_0 > \exp(\sqrt[q]{2\pi})$ *we have*[6]

[6]Note that $S(\gamma)$ is defined only for $\gamma \ge \exp(\sqrt[q]{2\pi})$.

(i) $\lim_{t\to\infty}\gamma(t)=\gamma_\infty<\infty$;
(ii) $\|x(t)\|$ *and* $\|u(t)\|$ *remain bounded as* $t\to\infty$;
(iii) $e(\cdot)\in L^2(\mathbb{R}_+,\mathbb{R}^m)$.

Moreover, if $(x_0,u_0)\in D(\tilde{A})$, *then* (5.3) *holds. If the observation operator* C *is bounded, then* (5.3) *holds for all* $(x_0,u_0)\in X\times\mathbb{R}^m$.

Proof. Let $\nu\in\{1,2,\dots,\ell\}$ be such that $\sigma(\mathbf{G}(0)\Gamma_\nu)\in\mathbb{C}_0$. By Theorem 3.8 there exists $k\in(0,1)$ such that the integrator $(k/s)\Gamma_\nu$ stabilizes $\mathbf{G}$. Consequently, by Theorem 3.14, the semigroup $\tilde{\mathbf{T}}^{k\Gamma_\nu}$ is exponentially stable. As in the proof of Theorem 5.1 set $\mathcal{K}(\gamma)=1+\cos(\log^q\gamma)$. By Lemma 3.16 there exists $\tilde{x}_0\in X\times\mathbb{R}^m$ such that

$$e=\tilde{\mathbf{\Psi}}_\infty\tilde{x}_0-\tilde{\mathbf{F}}_\infty(\mathcal{K}(\gamma)\Gamma_{S(\gamma)}e)\,. \tag{5.11}$$

Let $[0,\tau)$ be the maximal interval of existence for the solution (e,γ) of the closed-loop system given by (5.11) and (5.2). Seeking a contradiction, suppose that $\lim_{t\to\tau}\gamma(t)=\infty$. Choose $\varepsilon\in(0,k)$ such that $\varepsilon+k<1$. Then there exists a sequence $0\le t_0<t_1<\cdots<\tau$ with

$$\mathcal{K}(\gamma(t_{2i}))=k-\varepsilon\,,\quad \mathcal{K}(\gamma(t_{2i+1}))=k+\varepsilon$$

and such that

$$\mathcal{K}(\gamma(t))\in(k-\varepsilon,k+\varepsilon)\quad\text{and}\quad S(\gamma(t))=\nu\qquad\forall\,t\in[t_{2i},t_{2i+1}]\,.$$

As in the proof of Theorem 5.1, we can use (2.9b) to obtain

$$\mathbf{L}_{t_{2i}}\mathbf{P}_{t_{2i+1}}e=\tilde{\mathbf{\Psi}}^{k\Gamma_\nu}_{t_{2i+1}-t_{2i}}\tilde{x}(t_{2i})-\tilde{\mathbf{F}}^{k\Gamma_\nu}_{t_{2i+1}-t_{2i}}(\mathbf{L}_{t_{2i}}\mathbf{P}_{t_{2i+1}}(\mathcal{K}(\gamma)-k)\Gamma_\nu e)\,.$$

The remainder of the proof follows closely that of Theorem 5.1 and is omitted. ∎

The control law given by (5.10) and (5.2) depends crucially on the unmixing set $\{\Gamma_1,\Gamma_2,\dots,\Gamma_\ell\}$. Clearly, if $m=1$, then $\{1,-1\}$ is an unmixing set. For the case $m=2$ an unmixing set of cardinality 6 is given in Mårtensson [21], [22]. Zhu [51] has constructed an unmixing set having cardinality 32 for the case $m=3$. Unfortunately, the cardinality of the unmixing sets given by the general construction in [22] is far too large than would be convenient for applications.

6. Examples and simulations. The results of sections 3–5 apply to the general class of regular linear systems. For the purpose of illustration we consider two simple examples: finite-dimensional systems with output delays and a damped wave equation in a single spatial variable with boundary control and observation. In all of the simulations we used Simulink in Matlab. Note that the reference signals to be tracked are stepped, with nonzero step time.

Example 6.1. Systems with output delays:

We consider a class

$$\dot{x}(t)=Ax(t)+Bu(t)\,,\quad y(t)=Cx(t-h) \tag{6.1}$$

of systems with output delay, where $A\in\mathbb{R}^{n\times n}$, $B\in\mathbb{R}^{n\times m}$, $C\in\mathbb{R}^{m\times n}$, and $h>0$. The system (6.1) can be represented as a so-called Pritchard–Salamon system with state space $\mathbb{R}^n\times L^2(-h,0;\mathbb{R}^n)$; see, e.g., Pritchard and Salamon [35, 36]. Since Pritchard–Salamon systems are regular in the sense of section 2, it follows that the results of sections 3–5 can be applied to (6.1), provided that $\sigma(A)\subset\mathbb{C}_0$ and $\det CA^{-1}B\neq0$. We consider three particular cases.

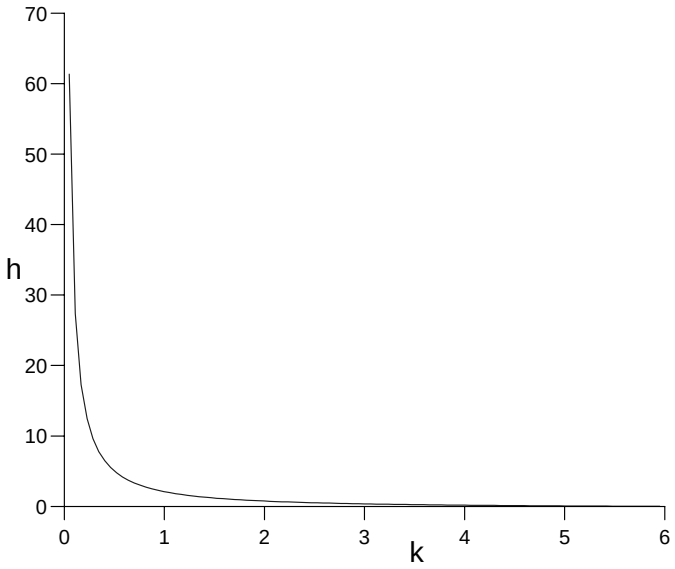


FIG. 6.1. *Tolerable delay as a function of k.*

(a) $m = 1$, $n = 2$, and

$$A = \begin{pmatrix} 0 & 1 \\ -2 & -3 \end{pmatrix}, \quad B = \begin{pmatrix} 0 \\ 1 \end{pmatrix}, \quad C = (1, 0)\,.$$

If $h = 0$, then $\dot{u} = -ky$ stabilizes (6.1) for all $k \in (0, 6)$. Using a stability window analysis (Walton and Marshall [43]), we can compute for each $k \in (0, 6)$ the range of $h \in (0, h(k))$ for which $\dot{u} = -ky$ stabilizes (6.1). In Figure 6.1, $h(k)$ is plotted against k for k in the range $(0, 6)$. Figure 6.2 shows a plot of $y(t)$, $r(t)$, and $\mathcal{K}(\gamma(t))$ against t for (4.28) with $p = 0.4$ when $h = 4$, $x(0) = (-1 \;\; 3)^T$, $u(0) = -1$, and $y(t) = -4$ for $t < 0$. Note in this case that the integrator gain can take values in $(0, 0.6)$ and that $\mathcal{K}(\gamma(\infty)) = 0.07$.

(b) We now consider two cases with $m = 2$, $n = 3$. In the first case $\mathbf{G}(0)$ is sign definite and in the second $\mathbf{G}(0)$ is sign indefinite.

(i) In this example we take

$$A = \begin{pmatrix} -1 & 0 & -2 \\ 0 & -1 & -3 \\ -2 & -3 & -14 \end{pmatrix}, \quad B = \begin{pmatrix} 1 & 0 \\ 0 & 1 \\ 0 & 0 \end{pmatrix}, \quad C = \begin{pmatrix} 1 & 0 & 0 \\ 0 & 1 & 0 \end{pmatrix}$$

so that

$$\mathbf{G}(0) = \begin{pmatrix} 7 & 6 \\ 6 & 11 \end{pmatrix} \succ 0\,.$$

We assume that this knowledge of the sign of the steady-state gain is available and use (4.28) with $p = 0.15$.

Figure 6.3 shows plots of $y(t)$, $r(t)$, and $\mathcal{K}(\gamma(t))$ for the case $h = 1$ with $y(\cdot) = 0$ on $[-1, 0)$, $x(0) = (0.4, 0.3, 0.25)^T$, and $u(0) = (1.5, 1)^T$, with the reference signal $r(t) = \theta(t)(5, 0)^T + \theta(t - 20)(5, 3)^T$.

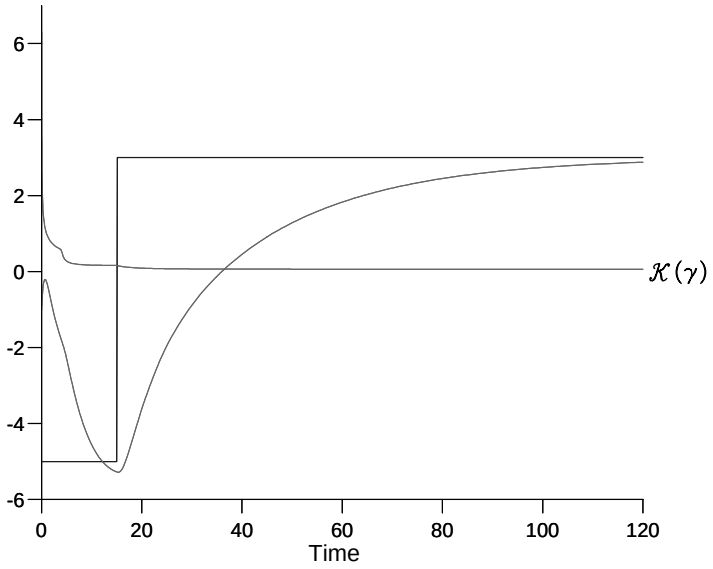


FIG. 6.2. *Simulation with* $\mathcal{K}(\gamma) = \gamma^{-0.4}$.

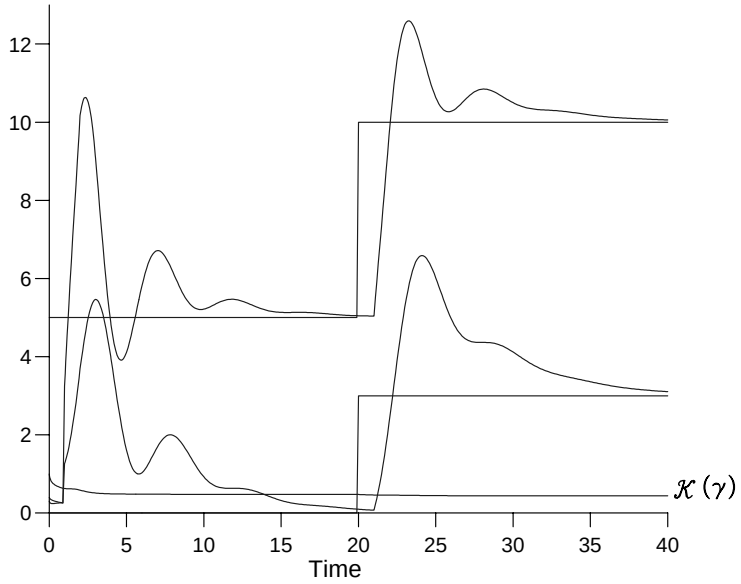


FIG. 6.3. *Simulation with* $\mathcal{K}(\gamma) = \gamma^{-0.15}$.

(ii) In this example we take

$$A = \begin{pmatrix} 0 & 1 & 0 \\ 0 & 0 & 1 \\ -6 & -11 & -6 \end{pmatrix}, \quad B = \begin{pmatrix} 0 & 0 \\ 1 & 0 \\ 0 & 1 \end{pmatrix}, \quad C = \begin{pmatrix} 1 & 0 & 0 \\ 0 & 1 & 0 \end{pmatrix}$$

so that

$$\mathbf{G}(0) = \begin{bmatrix} 3 & 0.1667 \\ 0 & 1 \end{bmatrix}.$$

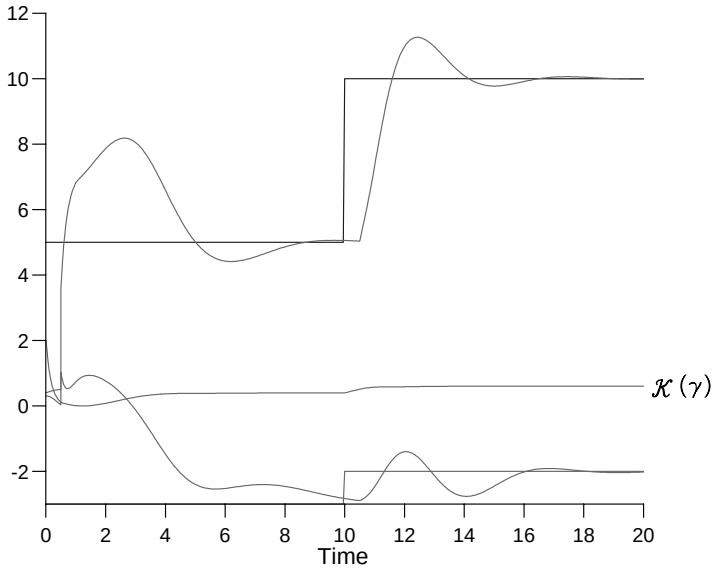


FIG. 6.4. *Simulation with* $\mathcal{K}(\gamma) = 1 + \cos(\log^{0.95}\gamma)$.

Clearly $\sigma(\mathbf{G}(0)) \subset \mathbb{C}_0$. We assume that this knowledge is available and use (5.1) with $q = 0.95$.

Figure 6.4 shows plots of $y(t)$, $r(t)$, and $\mathcal{K}(\gamma(t))$ for the case $h = 0.5$ with $y(\cdot) = 0$ on $[-0.5, 0)$, $x(0) = (0.4, 0.3, 0.25)^T$ and $u(0) = (1.5, 1)^T$ with the reference signal $r(t) = \theta(t)(5, -3)^T + \theta(t-10)(5, 1)^T$.

Example 6.2. A wave equation with boundary control and observation: We consider the damped wave equation

$$\frac{\partial^2 w}{\partial t^2}(z,t) = \frac{\partial^2 w}{\partial z^2}(z,t) - 2a\frac{\partial w}{\partial t}(z,t) - a^2 w(z,t)\,, \qquad t > 0\,,\ z \in (0,1)\,, \tag{6.2}$$

with boundary conditions

$$w(0,t) = 0\,, \quad \frac{\partial w}{\partial z}(1,t) = u(t)$$

and boundary observation

$$y(t) = \frac{\partial w}{\partial t}(1,t) + bw(1,t)\,,$$

where $a > 0$ and $b \neq 0$. This system has a regular, exponentially stable realization on the state space

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

Moreover, $\mathbf{G}(s) = \frac{s+b}{s+a}\frac{\sinh(s+a)}{\cosh(s+a)}$ so that $\mathbf{G}(0) = \frac{b\sinh(a)}{a\cosh(a)} \neq 0$. We assume that $a = \frac{1}{2}\log 0.3$ and $b = 0.3$. For purposes of illustration we assume that $\operatorname{sign}(\mathbf{G}(0))$ is unknown so that we use (4.11) with $p = 0$ and $q = 0.9$ and the initial conditions are equal to zero.

Figure 6.5 shows $y(t)$, $r(t)$, and $\mathcal{K}(\gamma(t))$, whilst Figure 6.6 shows $y(t)$, $r(t)$, and $\mathcal{K}(\gamma(t))$ when the sign of $\mathbf{G}(0)$ is switched. Note that whilst (6.2) gives a partial differential equation realization of $\mathbf{G}(s)$, for the simulations we exploited the fact

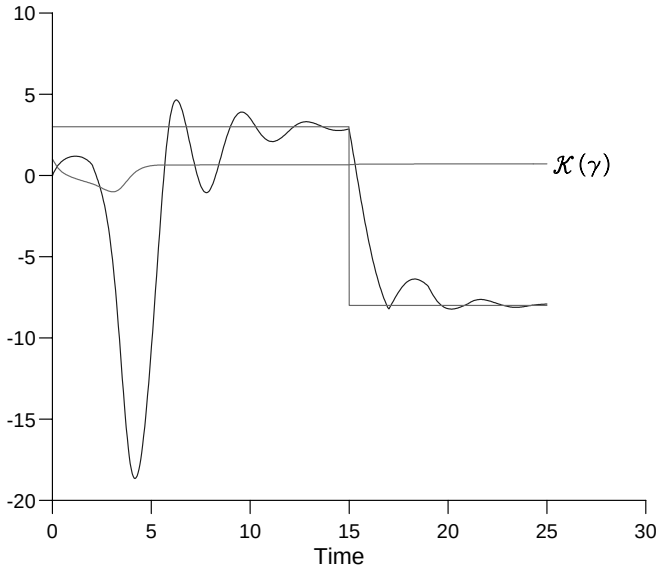


FIG. 6.5. *Simulation with* $\mathcal{K}(\gamma) = \cos(\log^{0.9}\gamma)$ *and* $\mathbf{G}(0) > 0$.

that the input-ouput behavior of (6.2), with zero initial conditions, is the same as that for the series connection of $\frac{s+b}{s+a}$ with the functional difference equation

$$y(t) = -e^{-2a}y(t-2) + u(t) - e^{-2a}u(t-2)\,. \tag{6.3}$$

The system given by (6.3) is easily realized using Simulink in Matlab.

In comparing Figures 6.5 and 6.6, we note that in the former, the gain function $\mathcal{K}(\gamma)$ undergoes two switches in sign before reaching a positive limit and in the latter switches sign only once before reaching a negative limit. The simulations are consistent with the fact that $\mathbf{G}(0) > 0$ in Figure 6.5 and $\mathbf{G}(0) < 0$ in Figure 6.6.

7. Concluding remarks. In this paper we have obtained results on nonadaptive and adaptive low-gain control of square regular systems for tracking step reference signals. It is possible to extend some of the results to nonsquare systems and sinusoidal reference signals. Finally, in [16] we have obtained discrete-time versions of the results in sections 3 and 4, with applications to sampled-data control of regular systems.

Appendix.

Proof of Proposition 2.4. For $a < b \le \infty$ we define $L(a,b) := L^2(a,b;\mathbb{R}^m) \times L^\infty(a,b;\mathbb{R})$ and $L_{loc}(a,\infty) := L^2_{loc}(a,\infty;\mathbb{R}^m) \times L^\infty_{loc}(a,\infty;\mathbb{R})$. We define a norm on $L(a,b)$ by setting $\|(f_1,f_2)\|_{(a,b)} := \|f_1\|_{L^2(a,b)} + \|f_2\|_{L^\infty(a,b)}$. In order to prove Proposition 2.4 we shall first consider an initial value problem which contains (2.11) as a special case.

Let $T \ge 0$, $(y^0,\gamma^0) \in L_{loc}(T,\infty)$ and $(f,g) \in L(0,T)$ be given, and suppose that $F \in L^1_{loc}(\mathbb{R}_+, \mathbb{R}^{m\times m})$ and $\mathcal{K} : \mathbb{R} \to \mathbb{R}$ is a locally Lipschitz function. For $\tau > T$ define the operator $\mathbf{N}_\tau : L(0,\tau) \to L(0,\tau)$ by

$$\text{(A.1a)}\quad \mathbf{N}_\tau \begin{pmatrix} y \\ \gamma \end{pmatrix}(t) = \begin{pmatrix} f(t) \\ g(t) \end{pmatrix}, \quad t \in [0,T]\,,$$

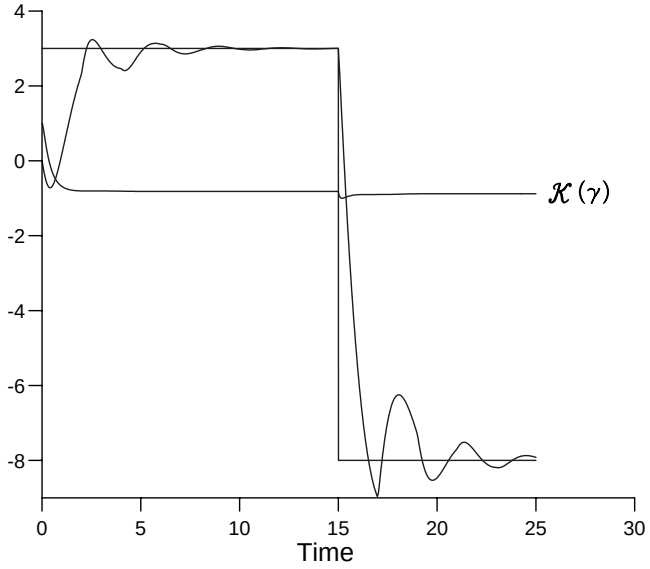


FIG. 6.6. *Simulation with* $\mathcal{K}(\gamma) = \cos(\log^{0.9}\gamma)$ *and* $\mathbf{G}(0) < 0$.

$$\text{(A.1b)}\quad \mathbf{N}_\tau \begin{pmatrix} y \\ \gamma \end{pmatrix}(t) = \begin{pmatrix} y^0(t) \\ \gamma^0(t) \end{pmatrix} + \int_0^t \begin{pmatrix} F(t-\xi) & 0 \\ 0 & 1 \end{pmatrix} \begin{pmatrix} \mathcal{K}(\gamma(\xi))y(\xi) \\ \|y(\xi)\|^2 \end{pmatrix} d\xi\,, \quad t \geq T\,.$$

For $\rho > 0$ and $\tau > T$, let $B_{\rho,\tau}$ denote the closed ball in $L(T,\tau)$ of radius ρ with center in $(y^0|_{[T,\tau]}, \gamma^0|_{[T,\tau]} + \|f\|^2_{L^2(0,T)})$. Finally define

$$\mathfrak{M}_{\rho,\tau} := \{(y,\gamma) \in L(0,\tau) \,|\, (y,\gamma)|_{[T,\tau]} \in B_{\rho,\tau},\ (y,\gamma)|_{[0,T]} = (f,g)\}\,.$$

Endowed with the metric

$$d[(y_1,\gamma_1),(y_2,\gamma_2)] = \|(y_1 - y_2, \gamma_1 - \gamma_2)\|_{(T,\tau)} = \|(y_1 - y_2, \gamma_1 - \gamma_2)\|_{(0,\tau)}\,,$$

$\mathfrak{M}_{\rho,\tau}$ becomes a complete metric space.

The following lemma will be the key tool for the proof of Proposition 2.4.

LEMMA A.1. *Let* $\rho \in (0, 1/2)$. *Then there exists a* $T^* > T$ *such that for all* $\tau \in (T, T^*)$ *the operator* $\mathbf{N}_\tau$ *is a contraction on* $\mathfrak{M}_{\rho,\tau}$, *i.e.,* (i) $\mathbf{N}_\tau \mathfrak{M}_{\rho,\tau} \subset \mathfrak{M}_{\rho,\tau}$ *and* (ii) *there exists* $\delta_\tau \in (0,1)$ *such that for all* $(y_1,\gamma_1), (y_2,\gamma_2) \in \mathfrak{M}_{\rho,\tau}$

$$\|\mathbf{N}_\tau(y_1,\gamma_1) - \mathbf{N}_\tau(y_2,\gamma_2)\|_{(T,\tau)} \leq \delta_\tau \|(y_1,\gamma_1) - (y_2,\gamma_2)\|_{(T,\tau)}\,.$$

In particular, for all τ *as above,* $\mathbf{N}_\tau$ *has a unique fixed point in* $\mathfrak{M}_{\rho,\tau}$.

Proof. Let $\mathbf{\Pi}_i$, $i = 1, 2$, denote the operator on $L(0,\tau)$ defined by $\mathbf{\Pi}_i(f_1, f_2) = f_i$, and let $\tau^* > T$ be fixed.

(i) Setting $\eta(t) := \int_0^T F(t-\xi)\mathcal{K}(g(\xi))f(\xi)\,d\xi$, it is clear that $\eta \in L^2_{loc}(\mathbb{R}_+, \mathbb{R}^m)$. For all $\tau \in (T, \tau^*)$ and all $(y,\gamma) \in \mathfrak{M}_{\rho,\tau}$ it follows that

$$\begin{aligned}
&\|\mathbf{\Pi}_1 \mathbf{N}_\tau(y,\gamma) - y^0\|^2_{L^2(T,\tau)} \\
&= \int_T^\tau \|\eta(t) + \int_0^t (\mathbf{P}_{\tau-T}F)(t-\xi)[(I - \mathbf{P}_T)\mathcal{K}(\gamma)y](\xi)\,d\xi\|^2\,dt \\
&\leq 2\left(\|\eta\|^2_{L^2(T,\tau)} + \left(\int_0^\tau \|(\mathbf{P}_{\tau-T}F)(\xi)\|\,d\xi\right)^2 \int_T^\tau \|\mathcal{K}(\gamma(\xi))y(\xi)\|^2\,d\xi\right)
\end{aligned}$$

$$\leq 2\left(\|\eta\|^2_{L^2(T,\tau)} + K^2\left(\int_0^{\tau-T}\|F(\xi)\|\,d\xi\right)^2\left(\rho^2+\int_T^\tau\|y^0(\xi)\|^2\,d\xi\right)\right), \tag{A.2}$$

where $K>0$ is such that $|\mathcal{K}(\kappa)|\leq K$ for all $\kappa\in\mathbb{R}$ with $|\kappa|\leq\rho+\|\gamma^0\|_{L^\infty(T,\tau^*)}+\|f\|^2_{L^2(0,T)}$. It follows from (A.2) that there exists $T_1\in(T,\tau^*)$ such that for all $(y,\gamma)\in\mathfrak{M}_{\rho,\tau}$ and for all $\tau\in(T,T_1)$

$$\|\mathbf{\Pi}_1\mathbf{N}_\tau(y,\gamma)-y^0\|^2_{L^2(T,\tau)}\leq\frac{\rho^2}{4}\,. \tag{A.3}$$

Moreover, we have that for $(y,\gamma)\in\mathfrak{M}_{\rho,\tau}$

$$\|\mathbf{\Pi}_2\mathbf{N}_\tau(y,\gamma)-\gamma^0(\cdot)-\|f\|^2_{L^2(0,T)}\|_{L^\infty(T,\tau)}=\int_T^\tau\|y(\xi)\|^2\,d\xi\leq\rho^2+\|y^0\|^2_{L^2(T,\tau)}\,. \tag{A.4}$$

Since $\rho<\frac{1}{2}$, it follows that $\rho^2<\rho/2$, and hence we obtain by using (A.4) that there exists $T_2>T$ such that for all $(y,\gamma)\in\mathfrak{M}_{\rho,\tau}$ and for all $\tau\in(T,T_2)$

$$\|\mathbf{\Pi}_2\mathbf{N}_\tau(y,\gamma)-\gamma^0(\cdot)-\|f\|^2_{L^2(0,T)}\|_{L^\infty(T,\tau)}<\frac{\rho}{2}\,. \tag{A.5}$$

Combining (A.3) and (A.5), we see that

$$\mathbf{N}_\tau\mathfrak{M}_{\rho,\tau}\subset\mathfrak{M}_{\rho,\tau}\quad\forall\,\tau\in(T,\min(T_1,T_2))\,. \tag{A.6}$$

(ii) For any $\tau\in(T,\tau^*)$ and any $(y_1,\gamma_1),(y_2,\gamma_2)\in\mathfrak{M}_{\rho,\tau}$ the following estimates hold:

$$\begin{aligned}
&\|\mathbf{\Pi}_1\mathbf{N}_\tau(y_1,\gamma_1)-\mathbf{\Pi}_1\mathbf{N}_\tau(y_2,\gamma_2)\|^2_{L^2(T,\tau)}\\
&\quad=\int_0^\tau\left(\int_0^t(\mathbf{P}_{T-\tau}F)(t-\xi)(\mathcal{K}(\gamma_1(\xi))y_1(\xi)-\mathcal{K}(\gamma_2(\xi))y_2(\xi))\,d\xi\right)^2dt\\
&\quad\leq\left(\int_0^\tau\|(\mathbf{P}_{\tau-T}F)(\xi)\|\,d\xi\right)^2\int_0^\tau\|\mathcal{K}(\gamma_1(\xi))y_1(\xi)-\mathcal{K}(\gamma_1(\xi))y_2(\xi)\\
&\qquad\qquad+\mathcal{K}(\gamma_1(\xi))y_2(\xi)-\mathcal{K}(\gamma_2(\xi))y_2(\xi)\|^2\,d\xi\\
&\quad\leq 2\left(\int_0^{\tau-T}\|F(\xi)\|\,d\xi\right)^2\left(K^2\int_T^\tau\|y_1(\xi)-y_2(\xi)\|^2\,d\xi\right.\\
&\qquad\qquad\left.+\;L^2\left(\int_0^\tau\|y_2(\xi)\|^2\,d\xi\right)\|\gamma_1-\gamma_2\|^2_{L^\infty(T,\tau)}\right),
\end{aligned} \tag{A.7}$$

where we have chosen $K>0$ and $L>0$ in such a way that for all real numbers κ, κ_1 and κ_2 with $|\kappa|,|\kappa_1|,|\kappa_2|\leq\max(\|g\|_{L^\infty(0,T)},\rho+\|\gamma^0\|_{L^\infty(T,\tau^*)}+\|f\|^2_{L^2(0,T)})$

$$\mathcal{K}(\kappa)\leq K\quad\text{and}\quad|\mathcal{K}(\kappa_1)-\mathcal{K}(\kappa_2)|\leq L|\kappa_1-\kappa_2|\,.$$

Realizing that

$$\int_0^\tau\|y_2(\xi)\|^2\,d\xi\leq\|f\|^2_{L^2(0,T)}+\|y^0\|^2_{L^2(T,\tau)}+\rho^2\,,$$

it follows from (A.7) that there exists $M > 0$ such that for all $\tau \in (T, \tau^*)$ and all $(y_1, \gamma_1), (y_2, \gamma_2) \in \mathfrak{M}_{\rho,\tau}$

$$\|\mathbf{\Pi}_1 \mathbf{N}_\tau(y_1, \gamma_1) - \mathbf{\Pi}_1 \mathbf{N}_\tau(y_2, \gamma_2)\|^2_{L^2(T,\tau)} \\ \leq M \left(\int_0^{\tau - T} \|F(\xi)\| \, d\xi \right)^2 (\|y_1 - y_2\|^2_{L^2(T,\tau)} + \|\gamma_1 - \gamma_2\|^2_{L^\infty(T,\tau)}).$$

Defining

$$\delta'_\tau := \sqrt{M} \int_0^{\tau - T} \|F(\xi)\| \, d\xi \,, \tag{A.8}$$

we obtain that for all $\tau \in (T, \tau^*)$ and all $(y_1, \gamma_1), (y_2, \gamma_2) \in \mathfrak{M}_{\rho,\tau}$

$$\|\mathbf{\Pi}_1 \mathbf{N}_\tau(y_1, \gamma_1) - \mathbf{\Pi}_1 \mathbf{N}_\tau(y_2, \gamma_2)\|_{L^2(T,\tau)} \leq \delta'_\tau (\|y_1 - y_2\|_{L^2(T,\tau)} + \|\gamma_1 - \gamma_2\|_{L^\infty(T,\tau)}) \,. \tag{A.9}$$

Furthermore, we have that for all $(y_1, \gamma_1), (y_2, \gamma_2) \in \mathfrak{M}_{\rho,\tau}$

$$\begin{aligned} &\|\mathbf{\Pi}_2 \mathbf{N}_\tau(y_1, \gamma_1) - \mathbf{\Pi}_2 \mathbf{N}_\tau(y_2, \gamma_2)\|_{L^\infty(T,\tau)} \\ &\qquad = \sup_{t \in [T,\tau]} \left| \int_T^t \|y_1(\xi)\|^2 \, d\xi - \int_T^t \|y_2(\xi)\|^2 \, d\xi \right| \\ &\qquad \leq \int_T^\tau (\|y_1(\xi)\| + \|y_2(\xi)\|) \|y_1(\xi) - y_2(\xi)\| \, d\xi \\ &\qquad \leq (\|y_1\|_{L^2(T,\tau)} + \|y_2\|_{L^2(T,\tau)}) \|y_1 - y_2\|_{L^2(T,\tau)} \\ &\qquad \leq 2(\rho + \|y^0\|_{L^2(T,\tau)}) \|y_1 - y_2\|_{L^2(T,\tau)} \,. \end{aligned} \tag{A.10}$$

Setting

$$\delta''_\tau := 2(\rho + \|y^0\|_{L^2(T,\tau)}) \,, \tag{A.11}$$

it follows from (A.10) that for all $(y_1, \gamma_1), (y_2, \gamma_2) \in \mathfrak{M}_{\rho,\tau}$

$$\|\mathbf{\Pi}_2 \mathbf{N}_\tau(y_1, \gamma_1) - \mathbf{\Pi}_2 \mathbf{N}_\tau(y_2, \gamma_2)\|_{L^\infty(T,\tau)} \leq \delta''_\tau \|y_1 - y_2\|_{L^2(T,\tau)} \,. \tag{A.12}$$

Clearly, since $\rho < \frac{1}{2}$ and by (A.8) and (A.11), there exists $T_3 \in (T, \tau^*)$ such that $\delta_\tau := \max(\delta'_\tau, \delta''_\tau) < 1$ for all $\tau \in (T, T_3)$. Setting $T^* = \min(T_1, T_2, T_3)$, we see that $T^* > T$, $\delta_\tau < 1$ for all $\tau \in (T, T^*)$, and moreover, by (A.6), (A.9), and (A.12), we have that for all $\tau \in (T, T^*)$ and all $(y_1, \gamma_1), (y_2, \gamma_2) \in \mathfrak{M}_{\rho,\tau}$

$$\mathbf{N}_\tau \mathfrak{M}_{\rho,\tau} \subset \mathfrak{M}_{\rho,\tau} \,, \quad \|\mathbf{N}_\tau(y_1, \gamma_1) - \mathbf{N}_\tau(y_2, \gamma_2)\|_{(T,\tau)} \leq \delta_\tau \|(y_1, \gamma_1) - (y_2, \gamma_2)\|_{(T,\tau)} \,.$$

Finally, it follows from Banach's contraction mapping theorem that for all τ as above $\mathbf{N}_\tau$ has a unique fixed point in $\mathfrak{M}_{\rho,\tau}$. □

Proof of Proposition 2.4. We proceed in several steps.

Step 1 (existence and uniqueness on a small interval). An application of Lemma A.1 to the case where $T = 0$, $y^0 = \mathbf{\Psi}_\infty x_0$, $\gamma^0(t) \equiv \gamma_0$ and $F = -\mathbb{L}^{-1}\mathbf{G}$ shows that for all sufficiently small $\tau > 0$ the operator $\mathbf{N}_\tau$ has a unique fixed point in $\mathfrak{M}_{\rho,\tau}$ and hence there exists $\tau^* > 0$ such that (2.11) has a unique solution (y^*, γ^*) on $[0, \tau^*)$.

Step 2 (continuation of solutions). If $\|y^*\|_{L^2(0,\tau^*)} = \infty$, then $\tau_{max} = \tau^*$ and $(y_{max}, \gamma_{max}) = (y^*, \gamma^*)$, and we are finished. Thus, let us suppose that $\|y^*\|_{L^2(0,\tau^*)} <$

∞. We claim that then the solution (y^*, γ^*) can be extended beyond τ^*. To this end we apply Lemma A.1 to the case where $T = \tau^*$, $(f, g) = (y^*, \gamma^*)$, $y^0 = (\mathbf{\Psi}_\infty x_0)|_{[\tau^*,\infty)}$, $\gamma^0(t) \equiv \gamma_0$, and $F = -\mathbb{L}^{-1}\mathbf{G}$. It follows that there exist $\tau^{**} > \tau^*$ and $(y^{**}, \gamma^{**}) \in L_{(0,\tau^{**})}$ such that $(y^{**}, \gamma^{**})|_{[0,\tau^*]} = (y^*, \gamma^*)$, and moreover (y^{**}, γ^{**}) solves (2.11) on $[0, \tau^{**})$.

Step 3 (extended uniqueness). Let (y_1, γ_1) and (y_2, γ_2) be two solutions of (2.11) on $[0, \tau_1)$ and $[0, \tau_2)$, respectively, where $\tau_2 \geq \tau_1 > 0$. We claim that

$$(y_2(t), \gamma_2(t)) = (y_1(t), \gamma_1(t)) \quad \text{for a.e. } t \in [0, \tau_1)\,. \tag{A.13}$$

For $\tau \in [0, \tau_1)$ define

$$\Omega_\tau := \{t \in [0, \tau] \,|\, (y_1(t), \gamma_1(t)) \neq (y_2(t), \gamma_2(t))\}\,,$$

and set

$$\hat{\tau} := \inf\{\tau \in [0, \tau_1) \,|\, \boldsymbol{\lambda}(\Omega_\tau) > 0\}\,,$$

where $\boldsymbol{\lambda}$ denotes the Lebesgue measure. It is clear that (A.13) is equivalent to $\hat{\tau} = \tau_1$. Seeking a contradiction, assume that $\hat{\tau} < \tau_1$. Let $t_n \in (0, \hat{\tau})$ with $\lim_{n\to\infty} t_n = \hat{\tau}$. (Recall that by Step 1, $\hat{\tau} > 0$.) Obviously,

$$\Omega_{\hat{\tau}} \setminus \{\hat{\tau}\} = \bigcup_{n\in\mathbb{N}} \Omega_{t_n}\,.$$

Now $\boldsymbol{\lambda}(\Omega_{t_n}) = 0$ for all $n \in \mathbb{N}$, and thus $\boldsymbol{\lambda}(\Omega_{\hat{\tau}}) = 0$, which in turn implies that for a.e. $t \in [0, \hat{\tau}]$

$$(y_1(t), \gamma_1(t)) = (y_2(t), \gamma_2(t)) =: (\hat{y}(t), \hat{\gamma}(t))\,.$$

An application of Lemma A.1 to the case where $T = \hat{\tau}$, $(f, g) = (\hat{y}, \hat{\gamma})$, $y^0 = (\mathbf{\Psi}_\infty x_0)|_{[\hat{\tau},\infty)}$, $\gamma^0(t) \equiv \gamma_0$, and $F = -\mathbb{L}^{-1}\mathbf{G}$ shows that there exists $t^* \in (\hat{\tau}, \tau_1)$ such that the operator $\mathbf{N}_{t^*}$ has a unique fixed point in $\mathfrak{M}_{\rho,t^*}$. Since the restrictions of (y_1, γ_1) and of (y_2, γ_2) to $[0, t^*]$ are both fixed points of $\mathbf{N}_{t^*}$, we see that $(y_1, \gamma_1)|_{[0,t^*]} = (y_2, \gamma_2)|_{[0,t^*]}$, which is in contradiction to the definition of $\hat{\tau}$.

Step 4 (existence of a maximal solution). Define

$$\mathfrak{T} := \{\tau > 0 \,|\, (2.11) \text{ has a solution on } [0, \tau)\}\,.$$

Set $\tau_{max} := \sup \mathfrak{T}$ and let $\tau_n \in \mathfrak{T}$ be such that $\tau_n \nearrow \tau_{max}$ as $n \to \infty$. Let (y_n, γ_n) denote the unique (by Step 3) solution of (2.11) on $[0, \tau_n)$. Using Step 3 again it is clear that $(y_n, \gamma_n)_{[0,\tau_m]} = (y_m, \gamma_m)$ for all $m, n \in \mathbb{N}$ with $n > m$. Therefore, we obtain a well-defined function (y_{max}, γ_{max}) on $[0, \tau_{max})$ by setting

$$(y_{max}(t), \gamma_{max}(t)) = (y_n(t), \gamma_n(t)) \quad \text{if } t \in [0, \tau_n)\,.$$

By construction (y_{max}, γ_{max}) is a solution of (2.11) on $[0, \tau_{max})$, which, by Step 3, is unique. Finally, it follows from Step 2 and the definition of τ_{max} that

$$\tau_{max} < \infty \quad \Longrightarrow \quad \int_0^{\tau_{max}} \|y_{max}(\xi)\|^2\, d\xi = \infty\,. \qquad \square$$

REFERENCES

[1] P. A. Cook, *Controllers with universal tracking properties*, in Proc. Int. IMA Conf. on Control: Modelling, Computation, Information, Manchester, 1992.

[2] G. W. M. Coppus, S. L. Sha, and R. K. Wood, *Robust multivariable control of a binary distillation column*, IEE Proc., Pt. D, 130 (1983), pp. 201–208.
[3] R. F. Curtain, H. Logemann, S. Townley, and H. Zwart, *Well-posedness, stabilizability and admissibility for Pritchard-Salamon systems*, J. Math. Systems, Estim. Control, 4 (1994), summary pp. 493–496; full electronic manuscript = 38 pp.; retrieval code 85404[7]; full hard copy to appear.
[4] E. J. Davison, *Multivariable tuning regulators: The feedforward and robust control of a general servomechanism problem*, IEEE Trans. Automat. Control, 21 (1976), pp. 35–47.
[5] P. Grosdidier, M. Morari, and B. R. Holt, *Closed-loop properties from steady-state gain information*, Ind. Eng. Chem. Fundam., 24 (1985), pp. 221–235.
[6] D. Hinrichsen and A. J. Pritchard, *Stability radius for structured perturbations and the algebraic Riccati equation*, Systems Control Lett., 8 (1986), pp. 105–113.
[7] A. Ilchmann, *Non-Identifier-Based High-Gain Adaptive Control*, Springer-Verlag, London, 1993.
[8] T. T. Jussila and H. N. Koivo, *Tuning of multivariable PI-controllers for unknown delay-differential systems*, IEEE Trans. Automat. Control, 32 (1987), pp. 364–368.
[9] H. N. Koivo and S. Pohjolainen, *Tuning of multivariable PI-controllers for unknown systems with input delay*, Automatica, 21 (1985), pp. 81–91.
[10] H. Logemann, *Adaptive exponential stabilization for a class of nonlinear retarded processes*, Math. Control Signals Systems, 3 (1990), pp. 255–269.
[11] H. Logemann, J. Bontsema, and D. H. Owens, *Low-gain control of distributed parameter systems with unbounded control and observation*, Control Theory Adv. Tech., 4 (1988), pp. 429–446.
[12] H. Logemann and A. Ilchmann, *An adaptive servomechanism for a class of infinite-dimensional systems*, SIAM J. Control Optim., 32 (1994), pp. 917–936.
[13] H. Logemann and B. Mårtensson, *Adaptive stabilization of infinite-dimensional systems*, IEEE Trans. Automat. Control, 37 (1992), pp. 1869–1883.
[14] H. Logemann and D. H. Owens, *Input-output theory of high-gain adaptive stabilization of infinite-dimensional systems with nonlinearities*, International Journal of Adaptive Control and Signal Processing, 2 (1988), pp. 193–216.
[15] H. Logemann and D. H. Owens, *Low-gain control of unknown infinite-dimensional systems: A frequency-domain approach*, Dynamics Stability Systems, 4 (1989), pp. 13–29.
[16] H. Logemann and S. Townley, *Discrete-Time Low-Gain Control of Uncertain Infinite-Dimensional Systems*, Mathematics preprint 95/06, University of Bath, 1995,[8] IEEE Trans. Automat. Control, to appear.
[17] H. Logemann and H. Zwart, *Some remarks on adaptive stabilization of infinite-dimensional systems*, Systems Control Lett., 16 (1991), pp. 199–207.
[18] J. Lunze, *Determination of robust multivariable I-controllers by means of experiments and simulation*, Syst. Anal. Modelling Simulation, 2 (1985), pp. 227–249.
[19] J. Lunze, *Experimentelle Erprobung einer Einstellregel für PI-Mehrgrößenregler bei der Herstellung von Ammoniumnitrat-Harnstoff-Lösung*, Messen Steuern Regeln, 30 (1987), pp. 2–6.
[20] J. Lunze, *Robust Multivariable Feedback Control*, Prentice-Hall, London, 1988.
[21] B. Mårtensson, *Adaptive Stabilization*, Ph.D. thesis, Lund Institute of Technology, Dept. of Automatic Control, 1986.
[22] B. Mårtensson, *The unmixing problem*, IMA J. Math. Control Inform., 8 (1991), pp. 367–377.
[23] D. E. Miller and E. J. Davison, *The self-tuning robust servomechanism problem*, IEEE Trans. Automat. Control, 34 (1989), pp. 511–523.
[24] D. E. Miller and E. J. Davison, *An adaptive tracking problem with a control input constraint*, Automatica, 29 (1993), pp. 877–887.
[25] M. Morari, *Robust stability of systems with integral control*, IEEE Trans. Automat. Control, 30 (1985), pp. 574–577.
[26] M. Morari and E. Zafiriou, *Robust Process Control*, Prentice-Hall, Englewood Cliffs, NJ, 1989.
[27] S. Mossaheb, *On the existence of right-coprime factorizations for functions meromorphic in a half-plane*, IEEE Trans. Automat. Control, 25 (1980), pp. 550–551.
[28] D. Mustafa, *How much integral action can a control system tolerate?*, Linear Algebra Appl., 205/206 (1994), pp. 965–970.

[7] Full paper is available by anonymous ftp from trick.ntp.springer.de in the directory /jmsec.
[8] Preprints in this series are available by anonymous ftp from ftp.maths.bath.ac.uk in the directory /pub/preprints.

[29] D. H. Owens and A. Chotai, *A simulation aid to gain estimates for robust tuning regulators*, IEEE Trans. Automat. Control, 30 (1985), pp. 177–179.
[30] A. Pazy, *Semigroups of Linear Operators and Applications to Partial Differential Equations*, Springer-Verlag, New York, 1983.
[31] J. Penttinen and H. N. Koivo, *Multivariable tuning regulators for unknown systems*, Automatica, 16 (1980), pp. 393–398.
[32] S. Pohjolainen, *Robust multivariable PI-controllers for infinite-dimensional systems*, IEEE Trans. Automat. Control, 27 (1982), pp. 17–30.
[33] S. Pohjolainen, *Robust controller for systems with exponentially stable strongly continuous semigroups*, J. Math. Anal. Appl., 111 (1985), pp. 622–636.
[34] S. Pohjolainen and I. Lätti, *Robust controller for boundary control systems*, Internat. J. Control, 38 (1983), pp. 1189–1197.
[35] A. J. Pritchard and D. Salamon, *The linear-quadratic control problem for retarded systems with delays in control and observation*, IMA J. Math. Control Inform., 2 (1985), pp. 335–362.
[36] A. J. Pritchard and D. Salamon, *The linear-quadratic control problem for infinite-dimensional systems with unbounded input and output operators*, SIAM J. Control Optim., 25 (1987), pp. 121–144.
[37] R. Rebarber, *Conditions for the equivalence of internal and external stability for distributed parameter systems*, IEEE Trans. Automat. Control, 38 (1993), pp. 994–998.
[38] E. P. Ryan, *A nonlinear universal servomechanism*, IEEE Trans. Automat. Control, 39 (1994), pp. 753–761.
[39] D. Salamon, *Realization theory in Hilbert space*, Math. Systems Theory, 21 (1989), pp. 147–164.
[40] M. C. Smith, *On stabilization and the existence of coprime factorizations*, IEEE Trans. Automat. Control, 34 (1989), pp. 1005–1007.
[41] S. Townley, *Simple adaptive stabilization of output feedback stabilizable distributed parameter systems*, Dynamics Control, 5 (1995), pp. 107–123.
[42] M. Vidyasagar, H. Schneider, and B. A. Francis, *Algebraic and topological aspects of feedback stabilization*, IEEE Trans. Automat. Control, 27 (1982), pp. 880–894.
[43] K. Walton and J. E. Marshall, *Direct method for TDS stability analysis*, IEE Proc., Pt. D, 134 (1987), pp. 101–107.
[44] G. Weiss, *Admissibility of unbounded control operators*, SIAM J. Control Optim., 27 (1989), pp. 527–545.
[45] G. Weiss, *Admissible observation operators for linear semigroups*, Israel J. Math., 65 (1989), pp. 17–43.
[46] G. Weiss, *The representation of regular linear systems on Hilbert spaces*, in Distributed Parameter System, F. Kappel, K. Kunisch, and W. Schappacher, eds., Birkhäuser Verlag, Basel, 1989, pp. 401–416.
[47] G. Weiss, *Transfer functions of regular linear systems, part* I: *Characterization of regularity*, Trans. Amer. Math. Soc., 342 (1994), pp. 827–854.
[48] G. Weiss, *Two conjectures on the admissibility of control operators*, in Control and Estimation of Distributed Parameter Systems, F. Kappel, K. Kunisch, and W. Schappacher, eds., Birkhäuser Verlag, Basel, 1991. pp. 367–378.
[49] G. Weiss, *Regular linear systems with feedback*, Math. Control Signals Systems, 7 (1994), pp. 23–57.
[50] G. Weiss, Private communication, 1994.
[51] X.-J. Zhu, *A finite spectrum unmixing set for* $GL(3,\mathbb{R})$, in Computation and Control, K. Bowers and J. Lund, eds., Birkhäuser Verlag, Boston, 1989, pp. 403–410.

SIAM J. CONTROL OPTIM.
Vol. 35, No. 1, pp. 117–133, January 1997

006

REGULARIZATION OF LINEAR DESCRIPTOR SYSTEMS WITH VARIABLE COEFFICIENTS*

RALPH BYERS†, PETER KUNKEL‡, AND VOLKER MEHRMANN§

Abstract. We study linear descriptor control systems with rectangular variable coefficient matrices. We introduce condensed forms for such systems under equivalence transformations and use these forms to detect whether the system can be transformed to a uniquely solvable closed loop system via state or derivative feedback. We show that under some mild assumptions every such system consists of an underlying square subsystem that behaves essentially like a standard state space system, plus some solution components that are constrained to be zero.

Key words. Descriptor systems, differential algebraic equations, condensed forms, smooth singular value decomposition, strangeness index, regularization, linear feedback

AMS subject classifications. 93C50, 34H05, 34A40, 93B10, 93B11, 93B40

PII. S0363012994278936

1. Introduction.

In this paper we study linear variable coefficient descriptor systems

$$E(t)\dot{x}(t) = A(t)x(t) + B(t)u(t) \tag{1}$$

in the interval $[t_0, t_1] \subset \mathcal{R}$ together with an initial condition

$$x(t_0) = x_0. \tag{2}$$

If we denote by $C^r([t_0, t_1], \mathcal{C}^{n,\ell})$ the set of r-times continuously differentiable functions from the interval $[t_0, t_1]$ to the vector space $\mathcal{C}^{n,\ell}$ of complex $n \times \ell$ matrices, then we assume that

$$\begin{array}{ll} E(t), A(t) \in C([t_0, t_1], \mathcal{C}^{n,\ell}), & \\ B(t) \in C([t_0, t_1], \mathcal{C}^{n,m}), & \\ x(t) \in C([t_0, t_1], \mathcal{C}^{\ell}) & \text{is the state of the system,} \\ u(t) \in C([t_0, t_1], \mathcal{C}^{m}) & \text{is the control of the system.} \end{array} \tag{3}$$

Descriptor systems of the form (1) are used in modeling control problems for mechanical multibody systems [32, 30, 31] or electrical circuits [16]. They are also obtained as linearizations of general nonlinear systems along trajectories [6].

In order to study the properties of such systems one needs an understanding of the behavior of the corresponding differential algebraic equations (DAEs). However,

*Received by the editors December 21, 1994; accepted for publication (in revised form) October 17, 1995.

http://www.siam.org/journals/sicon/35-1/27893.html

† Department of Mathematics, University of Kansas, Lawrence, KS 66045 (byers@ariel.math.ukans.edu). The research of this author was supported in part by National Science Foundation grants INT-8922444 and CCR-9404425 and University of Kansas General Research Allocation 3514-20-0038.

‡ Fachbereich Mathematik, Carl von Ossietzky Universität, Postfach 2503, D-26111 Oldenburg, Germany (kunkel@orion.math.uni-oldenburg.de). The research of this author was supported by DFG research grant Me 790/5-1 Differentiell algebraische Gleichungen.

§ Fakultät für Mathematik, Technische Universität Chemnitz-Zwickau, D-09107 Chemnitz, Germany (mehrmann@mathematik.tu-chemnitz.de). The research of this author was supported by DFG research grant Me 790/5-1 Differentiell algebraische Gleichungen.

fundamentally different definitions of solvability, index, etc., appear in the literature. See, for example, [1, 18]. These different definitions lead to different results, although only a few have been achieved so far [7, 8]. In particular, the latter results use the solvability concepts for differential algebraic equations as described in [1, 9, 10, 6].

In recent papers, Kunkel and Mehrmann have discussed a more general solvability concept and developed canonical forms [18, 20], existence and uniqueness theorems [18, 21], and numerical methods [22, 23] for linear variable coefficient DAEs. Extensions to this approach have recently been given by Rabier and Rheinboldt [27].

Several generalizations of the concept of index have also been discussed in the literature. The approach in [7, 8] is based on the solvability concept in [9]. Another approach based on generic solvability is discussed in [14].

It is our ultimate goal to develop numerical methods that allow the computation of the invariants in finite precision arithmetic. We do not discuss the generic approach here, because it is better suited for computer algebra computation. Instead, we will briefly discuss the two different solvability concepts of [9] and [18] in section 2 and give some extensions of solvability results discussed in [18, 20, 27].

We then show in section 3 that analogous methods can be used to study the properties of linear descriptor systems with variable coefficients. We obtain condensed forms for linear systems which display properties of the system.

In section 4 we show that under some mild assumptions every rectangular variable coefficient descriptor system has an underlying square system which in principle behaves like a standard linear state space system, together with some purely algebraic equations and some solution components which are constrained to be zero.

In section 5 we study the question whether the solvability properties of descriptor systems can be improved by different types of feedback, i.e., whether appropriate linear time-varying feedbacks can be chosen, so that the closed loop system is uniquely solvable for all consistent initial conditions. This topic has been discussed for linear, constant coefficient descriptor systems in [5]. There, it is shown (in the square case $n = \ell$) that uncontrollable higher index modes are constrained to be zero. Thus, regularizable descriptor systems consist of a subsystem that can be made index one via feedback together with some zero components of the state. In this paper we come to essentially the same conclusion for time-varying descriptor systems despite the fact that the transformations are more complex.

2. Existence and uniqueness of solutions of linear time-varying DAEs. We begin our analysis of the descriptor system (1), (2) with the following definition from [18, 20] on the solvability of linear variable coefficient DAEs of the form

$$E(t)\dot{x}(t) = A(t)x(t) + f(t), \quad t \in [t_0, t_1] \subset \mathcal{R} \tag{4}$$

with initial condition (2), E, A as in (3), and $f \in C([t_0, t_1], \mathcal{C}^n)$.

DEFINITION 2.1. *A function $x : [t_0, t_1] \to \mathcal{C}^\ell$ is a* solution *of* (4) *if $x \in C^1([t_0, t_1], \mathcal{C}^\ell)$ and x satisfies* (4) *pointwise. It is a* solution of the initial value problem (4), (2) *if x is solution of* (4) *and x satisfies* (2). *An initial condition* (2) *is called* consistent *if the corresponding initial value problem is solvable, i.e., has at least one solution.*

The definition of solvability is still a subject of discussion in the literature. Often it is required that a solution exists for all sufficiently differentiable inhomogeneities [1, p. 22]. Alternatively, only a well-behaved manifold of solutions is required [28]. Also, unique dependence on initial conditions is sometimes incorporated in the definition of solvability [1, 7].

The difficulty is illustrated by the following examples.

Example 1. Consider the DAE

$$\begin{bmatrix} -t & t^2 \\ -1 & t \end{bmatrix} \dot{x}(t) = \begin{bmatrix} -1 & 0 \\ 0 & -1 \end{bmatrix} x(t), \qquad t \in [-1, 1].$$

By Definition 2.1

$$x(t) = c(t) \begin{bmatrix} t \\ 1 \end{bmatrix}$$

is a solution for all $c \in C^1([-1, 1], \mathcal{C})$. Nevertheless, this DAE is not solvable in the sense of [1].

Example 2. The linear DAE

$$\begin{bmatrix} 0 & t \\ 0 & 0 \end{bmatrix} \begin{bmatrix} \dot{x}_1(t) \\ \dot{x}_2(t) \end{bmatrix} = \begin{bmatrix} 1 & 0 \\ 0 & 1 \end{bmatrix} \begin{bmatrix} x_1(t) \\ x_2(t) \end{bmatrix} + \begin{bmatrix} f_1(t) \\ f_2(t) \end{bmatrix}$$

has the solution $x_1(t) = -t\dot{f}_2(t) - f_1(t)$ and $x_2(t) = -f_2(t)$. Although the matrix E changes rank, no singularity appears in the solution. This DAE does not satisfy the hypothesis of the solvability theorem in [18], but as shown in [27], the requirement of constant rank for E can be relaxed in this case and an extension of the solvability theorem under weaker assumptions is true. This system also satisfies the hypothesis of the solvability theorem in [1, p. 30].

Example 3. A special case of (4) is the purely algebraic equation $0 = A(t)x(t) + f(t)$. If $A(t)$ is nonsingular, then there is a unique solution $x(t) = -A(t)^{-1}f(t)$ regardless of the smoothness of $f(t)$. However, this DAE is not solvable in the sense of [1] unless $f(t)$ is differentiable, because the equivalent ODE is $A(t)\dot{x}(t) = -\dot{A}(t)x(t) + \dot{f}(t)$. Although this system is also not solvable in the sense of [18], the normal form given in [18] exists and suggests the introduction of a weaker solution concept; see also [15]. If $A(t)$ drops rank for some t, then one has to apply the extension of the theory in [18, 20] given in [27] to show solvability. The weakness in [1], which is used in the context of control problems in [7, 8], is that it requires differentiability of all components of the inhomogeneity which is usually not the case in the applications from control. Another weakness of this concept is that it does not apply to rectangular systems. The concept introduced in [18] with the extensions given in [20, 27] is more general, is better suited to control problems, and applies to rectangular systems and distributional solutions. This is the reason why we prefer [18] to [1]. In this paper we will, however, discuss only classical solutions in the sense of Definition 2.1.

Remark 1. A simple but useful trick that removes some but not all of the discussed difficulties with the solvability concept is the following. If we add the term $\dot{E}(t)x(t)$ on both sides of (4), we obtain

$$\frac{d}{dt}(E(t)x(t)) = (A(t) + \dot{E}(t))x(t) + f(t). \tag{5}$$

In this form we would have to require sufficient smoothness of $x(t)$ only in the range of $E(t)^T$. This would allow weaker differentiability assumptions for x at the cost of smoothness assumptions for E. Such an approach would be more suitable for index one problems in particular, since exactly the differentiability that is needed is displayed. But as is shown in [18] for higher index problems, it is still not possible to identify the exact differentiability conditions without going to a canonical form. In

order to avoid confusion with the existing literature we therefore use the solvability condition introduced in Definition 2.1.

Another modification of the solvability concept that has been used frequently with constant coefficient systems is to restrict the initial conditions to the range of E by requiring

$$E(t_0)x(t_0) = E(t_0)x_0. \tag{6}$$

This would be in line with (5). We will determine in general what the exact consistency conditions for the initial values are in the following. Since these include not only modifications like (6) but also others, we will use the more general condition in Definition 2.1.

The standard variable coefficient transformations that can be applied to linear DAEs are changes of bases, i.e., $x(t) = Q(t)y(t)$, and premultiplication of (1) by $P(t)$. Under these transformations (4) transforms to

$$P(t)E(t)Q(t)\dot{y}(t) = (P(t)A(t)Q(t) - P(t)E(t)\dot{Q}(t))y(t) + P(t)f(t). \tag{7}$$

DEFINITION 2.2. *Two pairs of matrix functions* $(E_i(t),\ A_i(t))$, $E_i,\ A_i \in C([t_0, t_1], \mathcal{C}^{n,\ell})$, $i = 1, 2$ *are* equivalent *if there exist* $P \in C([t_0, t_1], \mathcal{C}^{n,n})$ *and* $Q \in C^1([t_0, t_1], \mathcal{C}^{\ell,\ell})$ *with* $P(t)$, $Q(t)$ *nonsingular for all* $t \in [t_0, t_1]$ *such that*

$$(E_2(t), A_2(t)) = P(t)(E_1(t), A_1(t)) \begin{bmatrix} Q(t) & -\dot{Q}(t) \\ 0 & Q(t) \end{bmatrix}. \tag{8}$$

Based on suitable equivalence transformations we will now extend the solvability theorems of [18, 20, 27].

To simplify the notation in the condensed forms, we denote in the following by $\Sigma_j(t)$ a square diagonal matrix valued function of dimension $j \times j$ which is invertible for all but finitely many $t \in [t_0, t_1]$. We also denote blocks of a matrix which are not specifically needed but which are not necessarily identically zero by $*$ and zero blocks of all dimensions by 0.

We construct the condensed form via smooth unitary equivalence transformations. This displays the invariants of the system but does not produce a canonical form but rather a condensed staircase form in the sense of Van Dooren [33] from which invariants can be read off. To do the transformations we use smooth singular value decompositions as they were introduced in [2] and for which several numerical methods are available [2, 25, 34]. To apply these transformations, we need derivatives of the right transformations. These can be obtained numerically from the original matrix function $E(t)$ and its derivatives using the method described in [17]. Note, however, that the following theorem and the construction procedure in the proof cannot be applied as a practical numerical algorithm. Nevertheless, it gives an indication how the computation of the invariants can be carried out locally. (For a discussion of this topic for DAEs see [18]. The extension to descriptor systems is currently being investigated.) One step of the construction given in the proof of the following theorem can be carried out locally. If more steps are needed, then local computation is not applicable in the given form. For the use of determining the required information on the global invariants via local computation see [22].

THEOREM 2.3. *Given analytic matrix-valued functions* $E(t)$, $A(t)$ *as in* (3) *there exist unitary analytic matrix-valued functions* $P(t)$, $Q(t)$, *as in* (8) *such that the*

matrices $P(t)E(t)Q(t)$, $P(t)A(t)Q(t) - P(t)E(t)\dot{Q}(t)$ have the following form:

$$\begin{array}{c} d_\mu \\ a_\mu \\ u^l_\mu \end{array} \begin{bmatrix} E_{11}(t) & E_{12}(t) & 0 \\ 0 & E_{22}(t) & 0 \\ 0 & E_{32}(t) & 0 \end{bmatrix}, \begin{bmatrix} A_{11}(t) & A_{12}(t) & A_{13}(t) \\ 0 & A_{22}(t) & 0 \\ 0 & A_{32}(t) & 0 \end{bmatrix}, \tag{9}$$

where $E_{11}(t)$ is diagonal and nonsingular except for isolated points, $E_{22}(t)$, $E_{32}(t)$, and $A_{32}(t)$ are block upper triangular with zero diagonal blocks, and $A_{22}(t)$ is block upper triangular with diagonal blocks which are nonsingular except for isolated points. The block columns have sizes d_μ, a_μ, u^r_μ. (Note that 0×0 matrices are diagonal and invertible, e.g., [11].*)*

Proof. The proof is constructive using a sequence of analytic singular value decompositions (ASVDs); see [2]. In the following we drop the dependence on t in the formulas. Consider the following recursive procedure:

Begin: Let $E_0 = E$, $A_0 = A$ and set $j = 0$, $n_j = n$, $\ell_j = \ell$.

(1) Let P_1, Q_1 be unitary matrices of appropriate dimensions that produce an ASVD of E_0 such that

$$E_1 := P_1^* E_0 Q_1 = \begin{bmatrix} \Sigma_{r_j} & 0 \\ 0 & 0 \end{bmatrix}.$$

Set

$$A_1 := P_1^* A_0 Q_1 - P_1^* E_0 \dot{Q}_1 = \begin{bmatrix} A_{11} & A_{12} \\ A_{21} & A_{22} \end{bmatrix}.$$

If $r_j = n_j$, i.e., E_1 has full row rank except for isolated points, then we **STOP** the process here.

(2) Let $\tilde{P}_2, Q_2$ be unitary matrices that produce a permuted ASVD of $\begin{bmatrix} A_{21} & A_{22} \end{bmatrix}$:

$$\tilde{P}_2^* \begin{bmatrix} A_{21} & A_{22} \end{bmatrix} Q_2 = \begin{bmatrix} 0 & \Sigma_{t_j} \\ 0 & 0 \end{bmatrix}.$$

Set

$$P_2 := \begin{bmatrix} I_{r_j} & 0 \\ 0 & \tilde{P}_2 \end{bmatrix},$$

and set

$$E_2 := P_2^* E_1 Q_2 = \left[\begin{array}{c|c} \tilde{E}_{11} & \tilde{E}_{12} \\ \hline 0 & 0 \\ 0 & 0 \end{array}\right], \; A_2 := P_2^* A_1 Q_2 - P_2^* E_1 \dot{Q}_2 = \left[\begin{array}{c|c} \tilde{A}_{11} & \tilde{A}_{12} \\ \hline 0 & \Sigma_{t_j} \\ 0 & 0 \end{array}\right].$$

The row dimensions are now $r_j, t_j, p_j = n_j - r_j - t_j$ and the column dimensions are $\ell_j - t_j, t_j$.

We then set $n_{j+1} := r_j$, $\ell_{j+1} := \ell_j - t_j$, $E_0 := \tilde{E}_{11}$, $A_0 := \tilde{A}_{11}$, and $j = j + 1$ and repeat the process from step (1) by applying transformations always to the complete system via an appropriate embedding of the transformation matrices.

end

It is clear that the procedure is finite, since n_j decreases in each step. At the end of this recursion we have the form

$$P(t)E(t)Q(t) = \left[\begin{array}{c|c|c|c|c|c} \Sigma_{n_\mu}(t) & 0 & * & * & \dots & * \\ \hline 0 & 0 & 0 & * & \dots & * \\ 0 & 0 & 0 & * & \dots & * \\ \hline & & & 0 & \ddots & \vdots \\ & & & & & \\ & & & 0 & & \vdots \\ \hline & & & & \ddots & \vdots \\ \hline & & & & & 0 \\ & & & & & 0 \end{array}\right] \begin{array}{l} n_\mu \\ \hline t_{\mu-1} \\ p_{\mu-1} \\ \hline t_{\mu-2} \\ \\ p_{\mu-2} \\ \hline \vdots \\ \hline t_0 \\ p_0 \end{array}, \tag{10}$$

$$P(t)A(t)Q(t) - P(t)E(t)\dot{Q}(t)$$

$$= \left[\begin{array}{c|c|c|c|c|c} \hat{A}_{11}(t) & \hat{A}_{12}(t) & * & * & \dots & * \\ \hline 0 & 0 & \Sigma_{t_{\mu-1}}(t) & * & \dots & * \\ 0 & 0 & 0 & * & \dots & * \\ \hline & & & \Sigma_{t_{\mu-2}}(t) & \dots & * \\ & & & 0 & \dots & * \\ \hline & & & & \ddots & \vdots \\ \hline & & & & & \Sigma_{t_0}(t) \\ & & & & & 0 \end{array}\right] \begin{array}{l} n_\mu \\ \hline t_{\mu-1} \\ p_{\mu-1} \\ \hline t_{\mu-2} \\ p_{\mu-2} \\ \hline \vdots \\ \hline t_0 \\ p_0 \end{array}. \tag{11}$$

We then set $E_{11}(t) = \Sigma_{n_\mu}(t)$ and permute the second block column to the end and the block rows in the order $1, 2, 4, 6, \dots, 3, 5, 7, \dots$ to obtain the final form with $d_\mu = n_\mu$, $a_\mu = t_0 + \cdots + t_\mu$, $u_\mu^l = n - a_\mu - d_\mu$, and $u_\mu^r = \ell - a_\mu - d_\mu$. □

Remark 2. Note that the analyticity assumption on the coefficient matrices can be relaxed to the condition that all smooth singular value decompositions exist and that rank changes in the factored matrices occur only at isolated points. This property is hard to quantify, however, since already infinite differentiability of the coefficient matrices may not be enough to guarantee the existence of such a decomposition with once-differentiable unitary factors; see [2]. The construction given in the proof of Theorem 2.3 is similar to the construction given in [18], but it needs fewer assumptions; in particular, no constant rank assumptions are needed.

Remark 3. The block sizes $t_{\mu-1}, \dots, t_0, p_{\mu-1}, \dots, p_0$ can be combined to determine the invariants of the equivalence transformation. However, (10) does not display all invariants, so it is a condensed form—not a canonical form. The quantities d_μ, a_μ, u_μ^ℓ, u_μ^r are invariants (see [18]), and they determine existence and uniqueness, as is shown in the following corollary. The condensed form is analogous to the staircase form of Van Dooren [33], which is a condensed form for constant matrix pencils that displays some of the invariants of the Kronecker canonical. Such condensed forms are useful, because they allow one to compute the relevant information via unitary transformations, which can be implemented in a numerically stable way. The quantity μ is called the *strangeness index* of the DAE, and it is a generalization of the differentiation index, e.g., [1, 18, 27, 21]. A variant of the solvability theorem of [20] is then as follows.

COROLLARY 2.4. *Let $(E(t), A(t))$ be analytic matrix-valued functions as in* (3) *and $f \in C^\mu([t_0, t_1], \mathcal{C}^n)$. Then,* (4) *is equivalent to a DAE of the form*

$$
\begin{aligned}
&\text{(a)} \quad \Sigma_{d_\mu}(t)\dot{x}_1(t) + E_{12}(t)\dot{x}_2(t) \\
&\qquad\qquad = A_{11}(t)x_1(t) + A_{12}(t)x_2(t) + A_{13}(t)x_3(t) + f_1(t), \\
&\text{(b)} \quad E_{22}(t)\dot{x}_2(t) = A_{22}(t)x_2(t) + f_2(t), \\
&\text{(c)} \quad E_{32}(t)\dot{x}_2(t) = A_{32}(t)x_2(t) + f_3(t),
\end{aligned} \tag{12}
$$

where the inhomogeneity is determined by $f^{(0)}, \ldots, f^{(\mu)}$ *and* $E_{22}(t)$, $E_{32}(t)$, $A_{22}(t)$, *and* $A_{32}(t)$ *are as in* (9). *In particular,* d_μ, a_μ, u_μ^r *are the numbers of differential, algebraic, and undetermined components of the unknown* x *in* (a), (b), *while* u_μ^l *is the number of conditions in* (c). *In particular, if in addition* $f \in C^{\mu+1}([t_0, t_1], \mathcal{C}^n)$, *then equation* (4) *is solvable if and only if the following properties hold:*

(i) *At isolated points where the diagonal matrix* $\Sigma_{d_\mu}(t)$ *or the diagonal blocks of* $A_{22}(t)$ *are singular, the components in* $x_3(t)$ *(if they occur) can be chosen so that the solution can be completed in a continuously differentiable way.*

(ii) *The conditions in* (12(b)) *are satisfied at the initial point.*

(iii) *The* u_μ^l *functional consistency conditions in* (12(c)) *are satisfied for* $x_2(t)$, *which is fixed by* (12(b)). (*Observe that* (12(b)) *fixes* $x_2(t)$ *by recursive insertion and differentiation except for points where the diagonal blocks of* $A_{22}(t)$ *are singular, using the nilpotency structure of* $E_{22}(t)$.)

An initial condition (2) *is consistent if and only if the* a_μ *conditions*

$$
E_{22}(t_0)\dot{x}_2(t_0) = A_{22}(t_0)x_2(t_0) + f_2(t_0) \tag{13}
$$

yield an $x_2(t_0)$ *which coincides with solution of* (12(b)) *at* t_0.

The initial value problem (1), (2) *is uniquely solvable if we also have*

$$
u_\mu^r = 0. \tag{14}
$$

Otherwise, we can choose $x_3(t) \in C^1([t_0, t_1], \mathcal{C}^{u_\mu^r})$ *arbitrarily.*

Proof. The proof follows directly from Theorem 2.3. Considering the second block equation, we obtain from the form of $E_{22}(t), A_{22}(t)$ that we can recursively solve for the solution components. The diagonal blocks of $A_{22}(t)$ are diagonal matrices which are invertible except possibly at isolated points. In these points we have to be able to complete the solution in a smooth way, since these components then have to be differentiated to continue the solution process. There are μ differentiations necessary to completely solve for the second block. Inserting $\dot{x}_2$ and x_2 and choosing x_3 we can solve equation (12(a)) except at points where the matrix $\Sigma_{d_\mu}(t)$ is singular. The same argument as above applies in these points. The remaining assertions are straightforward. □

In this section we have given a generalization of the condensed form and the solvability results of [18, 20, 27]. We do not need to apply constant rank assumptions, but we need assumptions that guarantee the existence of all ASVDs according to Remark 2. Here we could generalize the construction to weak solvability, which would allow us to drop some further smoothness assumptions; see [27]. In the next section we perform the corresponding construction for linear systems.

3. Condensed forms for linear descriptor systems. In this section we discuss the set of equivalence transformations that we will apply to variable coefficient descriptor systems and canonical forms under these transformations. Using these forms, we obtain information about the system properties. For constant coefficient systems such forms have been studied for general transformations in [24] and for unitary transformations in [3, 4, 5]. The results that we give here generalize the results for the unitary case even for constant coefficient systems.

Observe that we cannot apply directly the solvability result of section 2, since *usually we cannot assume* that the input functions $u(t)$ are sufficiently differentiable. In principle we can apply differentiation of components only in the uncontrollable subspace, i.e., the part of the system operating in the left nullspace of $B(t)$. Note that this is a major difference to the approach in [7, 8], where it is assumed that the input functions are sufficiently smooth.

We use the following global equivalence transformations for the triple of matrix valued functions $(E(t), A(t), B(t))$.

DEFINITION 3.1. *Two triples of matrix functions* $(E_i(t), A_i(t), B_i(t))$, $E_i, A_i \in C([t_0, t_1], \mathcal{C}^{n,\ell})$, $B_i \in C([t_0, t_1], \mathcal{C}^{n,m}), i = 1, 2$ *are called* equivalent *if there exist* $P \in C([t_0, t_1], \mathcal{C}^{n,n})$, $Q \in C^1([t_0, t_1], \mathcal{C}^{\ell,\ell})$ *and* $W \in C([t_0, t_1], \mathcal{C}^{m,m})$ *with* $P(t), Q(t), W(t)$ *nonsingular for all* $t \in [t_0, t_1]$ *such that*

$$(E_2(t), A_2(t), B_2(t)) = P(t)(E_1(t), A_1(t), B_1(t)) \begin{bmatrix} Q(t) & -\dot{Q}(t) & 0 \\ 0 & Q(t) & 0 \\ 0 & 0 & W(t) \end{bmatrix}. \tag{15}$$

It is easily checked that the above transformations describe equivalence transformations.

We obtain the following condensed form.

THEOREM 3.2. *Given analytic matrix-valued functions* $E(t)$, $A(t)$, $B(t)$ *as in* (3) *there exist unitary analytic matrix-valued functions* $P(t)$, $Q(t)$, $W(t)$ *as in* (15) *such that the three matrices* $P(t)E(t)Q(t)$, $P(t)A(t)Q(t) - P(t)E(t)\dot{Q}(t)$, *and* $P(t)B(t)W(t)$ *have the following form:*

$$\begin{array}{c} \begin{array}{c} d_\nu \\ v_\nu \\ \hline s_\nu \\ \hline \tilde{u}_\nu^l \\ \hline u_\nu^l \end{array} \left[\begin{array}{cc|c|c} \Sigma_{d_\nu}(t) & 0 & E_{13}(t) & 0 \\ 0 & 0 & E_{23}(t) & 0 \\ \hline 0 & 0 & E_{33}(t) & 0 \\ \hline 0 & 0 & E_{43}(t) & 0 \\ \hline 0 & 0 & E_{53}(t) & 0 \end{array}\right], \\ \left[\begin{array}{cc|c|c} A_{11}(t) & A_{12}(t) & A_{13}(t) & A_{14}(t) \\ A_{21}(t) & A_{22}(t) & A_{23}(t) & A_{24}(t) \\ \hline 0 & 0 & A_{33}(t) & 0 \\ \hline A_{41}(t) & A_{42}(t) & A_{43}(t) & A_{44}(t) \\ \hline 0 & 0 & A_{53}(t) & 0 \end{array}\right], \quad \left[\begin{array}{ccc} B_{11}(t) & B_{12}(t) & B_{13}(t) \\ \Sigma_{v_\nu}(t) & 0 & 0 \\ \hline 0 & 0 & 0 \\ \hline 0 & \Sigma_{\tilde{u}_\nu^l}(t) & 0 \\ \hline 0 & 0 & 0 \end{array}\right] \end{array} \tag{16}$$

with $E_{33}(t)$ *block upper triangular with zero diagonal blocks, and* $A_{33}(t)$ *block upper triangular with diagonal blocks which are diagonal matrices that are nonsingular except for isolated points. The block columns in* E, A *have sizes* d_ν, v_ν, s_ν, u_ν^r.

Proof. The proof is constructive using again a sequence of analytic singular value decompositions (ASVDs). We again drop the dependence on t in the formulas. Consider the following recursive procedure:

Begin:

Let $E_0 = E$, $A_0 = A$, $B_0 = B$ and set $j = 0$, $n_j = n$, $\ell_j = \ell$.

(1) Let P_1, Q_1 be unitary matrices of appropriate dimensions that produce an ASVD of E_0 such that

$$E_1 := P_1^* E_0 Q_1 = \begin{bmatrix} \Sigma_{d_j} & 0 \\ 0 & 0 \end{bmatrix}.$$

Set

$$A_1 := P_1^* A_0 Q_1 - P_1^* E_0 \dot{Q}_1 = \begin{bmatrix} A_{11} & A_{12} \\ A_{21} & A_{22} \end{bmatrix}, \qquad B_1 := P_1 B_0 = \begin{bmatrix} B_{11} \\ B_{21} \end{bmatrix}.$$

(2) Let $\tilde{P}_2, W_2$ be unitary matrices of appropriate dimensions that produce an ASVD of B_{21}. Set

$$\tilde{P}_2^* B_{21} W_2 = \begin{bmatrix} \Sigma_{c_j} & 0 \\ 0 & 0 \end{bmatrix}, \; P_2 := \begin{bmatrix} I_{d_j} & 0 \\ 0 & \tilde{P}_2 \end{bmatrix},$$

and

$$E_2 := P_2^* E_1 = \begin{bmatrix} \Sigma_{d_j} & 0 \\ 0 & 0 \\ 0 & 0 \end{bmatrix}, \; A_2 := P_2^* A_1 = \begin{bmatrix} \tilde{A}_{11} & \tilde{A}_{12} \\ \tilde{A}_{21} & \tilde{A}_{22} \\ \tilde{A}_{31} & \tilde{A}_{32} \end{bmatrix},$$
$$B_2 := P_2^* B_1 W_2 = \begin{bmatrix} \tilde{B}_{11} & \tilde{B}_{12} \\ \Sigma_{c_j} & 0 \\ 0 & 0 \end{bmatrix}.$$

If B_{21} has full row rank except at isolated points, then we **STOP** the process here.

(3) Let $\tilde{P}_3, Q_3$ be unitary matrices that produce a permuted ASVD of $\begin{bmatrix} \tilde{A}_{31} & \tilde{A}_{32} \end{bmatrix}$. Set

$$\tilde{P}_3^* \begin{bmatrix} \tilde{A}_{31} & \tilde{A}_{32} \end{bmatrix} Q_3 = \begin{bmatrix} 0 & \Sigma_{k_j} \\ 0 & 0 \end{bmatrix}, \qquad P_3 := \begin{bmatrix} I_{d_j} & 0 & 0 \\ 0 & I_{c_j} & 0 \\ 0 & 0 & \tilde{P}_3 \end{bmatrix}$$

and

$$E_3 := P_3^* E_2 Q_3 = \left[\begin{array}{c|c} \hat{E}_{11} & \hat{E}_{12} \\ 0 & 0 \\ \hline 0 & 0 \\ 0 & 0 \end{array}\right], \qquad A_3 := P_3^* A_2 Q_3 - P_3^* E \dot{Q}_3 = \left[\begin{array}{c|c} \hat{A}_{11} & \hat{A}_{12} \\ \hat{A}_{21} & \hat{A}_{22} \\ \hline 0 & \Sigma_{k_j} \\ 0 & 0 \end{array}\right],$$
$$B_3 := P_3^* B_2 = \left[\begin{array}{c|c} \hat{B}_{11} & \hat{B}_{12} \\ \Sigma_{c_j} & 0 \\ \hline 0 & 0 \\ 0 & 0 \end{array}\right].$$

(17)

with block rows of sizes $d_j, c_j, k_j, q_j = n_j - d_j - c_j - k_j$ and block columns of sizes $\ell_j - k_j, k_j$ in E_3, A_3, and $c_j, m - c_j$ in B_3.

Now we set $n_{j+1} = d_j + c_j$, $\ell_{j+1} = \ell_j - k_j$, and we set E_0, A_0 to be the $n_{j+1} \times \ell_{j+1}$ upper left submatrices of E_3, A_3 and B_0 to be the upper $n_{j+1} \times m$ submatrix of B_3. Set $j = j + 1$, and repeat the process from step (1) by applying transformations always to the complete system via an appropriate embedding of the transformation matrices.

end

It is clear that the procedure is finite, since n_j decreases in each step. At the end of this recursion we have the following forms for the transformed E, A, B:

$$\left[\begin{array}{cc|c|c|c|c}
\Sigma_{d_\nu}(t) & 0 & * & * & \dots & * \\
0 & 0 & * & * & \dots & * \\
\hline
 & & 0 & * & \dots & * \\
 & & 0 & * & \dots & * \\
\hline
 & & & \ddots & \vdots & \\
\hline
 & & & & 0 & * \\
 & & & & 0 & * \\
\hline
 & & & & & 0 \\
 & & & & & 0
\end{array}\right]
\begin{array}{l}
d_\nu \\ c_\nu \\ \hline k_{\nu-1} \\ q_{\nu-1} \\ \hline \vdots \\ \hline k_1 \\ q_1 \\ \hline k_0 \\ q_0
\end{array},$$

$$\left[\begin{array}{cc|c|c|c|c}
A_{11}(t) & A_{12}(t) & * & * & \dots & * \\
A_{21}(t) & A_{22}(t) & * & * & \dots & * \\
\hline
 & & \Sigma_{k_{\nu-1}}(t) & * & \ddots & \vdots \\
 & & 0 & * & & \vdots \\
\hline
 & & & \ddots & \ddots & \vdots \\
\hline
 & & & & \Sigma_{k_1}(t) & * \\
 & & & & 0 & * \\
\hline
 & & & & & \Sigma_{k_0}(t) \\
 & & & & & 0
\end{array}\right]
\begin{array}{l}
d_\nu \\ c_\nu \\ \hline k_{\nu-1} \\ q_{\nu-1} \\ \hline \vdots \\ \hline k_1 \\ q_1 \\ \hline k_0 \\ q_0
\end{array},$$

$$\left[\begin{array}{cc}
B_{11}(t) & B_{12}(t) \\
\Sigma_{c_\nu}(t) & 0 \\
\hline
0 & 0 \\
0 & 0 \\
\hline
\vdots & \vdots \\
\hline
0 & 0 \\
0 & 0 \\
\hline
0 & 0 \\
0 & 0
\end{array}\right]
\begin{array}{l}
d_\nu \\ c_\nu \\ \hline k_{\nu-1} \\ q_{\nu-1} \\ \hline \vdots \\ \hline k_1 \\ q_1 \\ \hline k_0 \\ q_0
\end{array}.$$

The columns of E, A have sizes $d_\nu, \ell - k_{\nu-1} - \cdots - k_0, k_{\nu-1}, \dots, k_0$.

We now split the second block row and column further so that we obtain a square diagonal block of size $v_\nu = \min(c_\nu, \ell - k_{\nu-1} - \cdots - k_0)$ in the (2,2) position. Set $u_\nu^r := \ell - k_{\nu-1} - \cdots - k_0 - v_\nu$, and $u_\nu^l := c_\nu - v_\nu$. The final form is then obtained by a block permutation which moves the new third block column to the end and permutes

the block rows in the order $1, 2, 3, 5, \ldots, 4, 6, \ldots$. It is as follows:

$$P(t)E(t)Q(t) = \left[\begin{array}{cc|cccc|c} \Sigma_{d_\nu} & 0 & * & * & \ldots & * & 0 \\ 0 & 0 & 0 & * & \ldots & * & 0 \\ \hline & & 0 & * & \ldots & * & 0 \\ & & & 0 & \ddots & \vdots & 0 \\ & & & & \ddots & * & 0 \\ & & & & & 0 & 0 \\ \hline & & * & * & \ldots & * & 0 \\ \hline & & 0 & * & \ldots & * & 0 \\ & & & 0 & \ddots & \vdots & 0 \\ & & & & \ddots & * & 0 \\ & & & & & 0 & 0 \end{array}\right] \begin{array}{l} d_\nu \\ v_\nu \\ \hline k_{\nu-1} \\ k_{\nu-2} \\ \vdots \\ k_0 \\ \hline \tilde{u}_\nu^l \\ \hline q_{\nu-1} \\ \vdots \\ q_1 \\ q_0 \end{array},$$

$$P(t)A(t)Q(t) - P(t)E(t)\dot{Q}(t)$$

$$= \left[\begin{array}{cc|cccc|c} A_{11}(t) & A_{12}(t) & * & * & \ldots & * & * \\ A_{21}(t) & A_{22}(t) & * & * & \ldots & * & 0 \\ \hline & & \Sigma_{k_{\nu-1}}(t) & * & \ddots & * & 0 \\ & & & \ddots & \ddots & \vdots & \vdots \\ & & & & \Sigma_{k_1}(t) & * & 0 \\ & & & & & \Sigma_{k_0}(t) & 0 \\ \hline * & * & * & \ldots & \ldots & * & * \\ \hline & & 0 & * & \ldots & * & 0 \\ & & & 0 & \ddots & \vdots & \vdots \\ & & & & 0 & * & 0 \\ & & & & & 0 & 0 \end{array}\right] \begin{array}{l} d_\nu \\ v_\nu \\ \hline k_{\nu-1} \\ k_{\nu-2} \\ \vdots \\ k_0 \\ \hline \tilde{u}_\nu^l \\ \hline q_{\nu-1} \\ \vdots \\ q_1 \\ q_0 \end{array},$$

$$P(t)B(t)W(t) = \left[\begin{array}{ccc} * & * & * \\ \Sigma_{v_\nu}(t) & 0 & 0 \\ \hline 0 & 0 & 0 \\ \vdots & \vdots & \vdots \\ \vdots & \vdots & \vdots \\ 0 & 0 & 0 \\ \hline 0 & \Sigma_{\tilde{u}_\nu^l} & 0 \\ \hline 0 & 0 & 0 \\ \vdots & \vdots & \vdots \\ \vdots & \vdots & \vdots \\ 0 & 0 & 0 \end{array}\right] \begin{array}{l} d_\nu \\ v_\nu \\ \hline k_{\nu-1} \\ k_{\nu-2} \\ \vdots \\ k_0 \\ \hline \tilde{u}_\nu^l \\ \hline q_{\nu-1} \\ \vdots \\ q_1 \\ q_0 \end{array},$$

where the widths of the block columns in the transformed E, A are d_ν, v_ν, $k_{\nu-1}$, $\ldots$, k_0, u_ν^r and v_ν, $\tilde{u}_\nu^l$, $m - v_\nu - \tilde{u}_\nu^l$ in the transformed B. We then combine all the

k_j-blocks and the q_j-blocks to make blocks of sizes $s_\nu = \sum_{j=0}^{\nu} k_j$, $u_\nu^r = \sum_{j=0}^{\nu} q_j$, respectively. □

Note again that the analyticity assumption on the coefficient matrices can be relaxed to the condition that all smooth singular value decompositions exist and that rank changes in the factored matrices occur only at isolated points; see Remark 2. If only one step of the procedure given in the proof of Theorem 3.2 needs to be performed, i.e., $\nu = 0$, then the invariant quantities can be obtain via local rank computations.

Remark 4. Like Theorem 2.3, the condensed form of Theorem 3.2 is not a canonical form, but it does display the relevant information. The corresponding canonical form, using nonunitary transformations has subsequently been developed by Rath [29].

In this section we have determined invariants of the system under global equivalence transformations. We will apply these results in the next section to analyze system properties.

4. The square subsystem. We will now analyze the system (1) after transformation to the condensed form (16).

$$\begin{array}{c} d_\nu \\ v_\nu \\ \hline s_\nu \\ \hline \tilde{u}_\nu^l \\ \hline u_\nu^l \end{array} \left[\begin{array}{cc|c|c} \Sigma_{d_\nu}(t) & 0 & E_{13}(t) & 0 \\ 0 & 0 & E_{23}(t) & 0 \\ \hline 0 & 0 & E_{33}(t) & 0 \\ \hline 0 & 0 & E_{43}(t) & 0 \\ \hline 0 & 0 & E_{54}(t) & 0 \end{array}\right] \left[\begin{array}{c} \dot{x}_1 \\ x_2 \\ \hline x_3 \\ \hline x_4 \end{array}\right]$$

$$= \left[\begin{array}{cc|c|c} A_{11}(t) & A_{12}(t) & A_{13}(t) & A_{14}(t) \\ A_{21}(t) & A_{22}(t) & A_{23}(t) & A_{24}(t) \\ \hline 0 & 0 & A_{33}(t) & 0 \\ \hline A_{41}(t) & A_{42}(t) & A_{43}(t) & A_{44}(t) \\ \hline 0 & 0 & A_{53}(t) & 0 \end{array}\right] \left[\begin{array}{c} x_1 \\ x_2 \\ \hline x_3 \\ \hline x_4 \end{array}\right] + \left[\begin{array}{ccc} B_{11}(t) & B_{12}(t) & B_{13}(t) \\ \Sigma_{v_\nu}(t) & 0 & 0 \\ \hline 0 & 0 & 0 \\ \hline 0 & \Sigma_{\tilde{u}_\nu^l}(t) & 0 \\ \hline 0 & 0 & 0 \end{array}\right] \left[\begin{array}{c} u_1 \\ u_2 \\ u_3 \end{array}\right], \tag{18}$$

with columns of sizes d_ν, v_ν, s_ν, u_r^ν in E, A and columns of sizes v_ν, $\tilde{u}_\nu^l$, $m - u_\nu - \tilde{u}_\nu^l$ in B. We immediately make the following observations:

(1) From the third block equation we obtain by recursive substitution that $x_3(t) = 0$ almost everywhere, and since we want a smooth solution, we obtain $x_3(t) \equiv 0$.

(2) Since $x_3(t) \equiv 0$, the equations given by the last block row are fulfilled trivially. So we might leave these equations off altogether.

Thus we may consider the subsystem

$$\begin{array}{c} d_\nu \\ v_\nu \\ \hline \tilde{u}_\nu^l \end{array} \left[\begin{array}{cc|c} \Sigma_{d_\nu}(t) & 0 & 0 \\ 0 & 0 & 0 \\ \hline 0 & 0 & 0 \end{array}\right] \left[\begin{array}{c} \dot{x}_1 \\ x_2 \\ \hline x_4 \end{array}\right]$$

$$= \left[\begin{array}{cc|c} A_{11}(t) & A_{12}(t) & A_{14}(t) \\ A_{21}(t) & A_{22}(t) & A_{24}(t) \\ \hline A_{41}(t) & A_{42}(t) & A_{44}(t) \end{array}\right] \left[\begin{array}{c} x_1 \\ x_2 \\ \hline x_4 \end{array}\right] + \left[\begin{array}{ccc} B_{11}(t) & B_{12}(t) & B_{13}(t) \\ \Sigma_{v_\nu}(t) & 0 & 0 \\ \hline 0 & \Sigma_{\tilde{u}_\nu^l}(t) & 0 \end{array}\right] \left[\begin{array}{c} u_1 \\ u_2 \\ u_3 \end{array}\right], \tag{19}$$

with columns of sizes d_ν, v_ν, u_r^ν in E, A and columns of sizes v_ν, $\tilde{u}_\nu^l$, $m - u_\nu - \tilde{u}_\nu^l$ in B.

If in this reduced form $u_r^\nu = \tilde{u}_{\nu^l}$ then this pencil is square and if we would compute the condensed form for this subsystem via Theorem 3.2, we would obtain $\nu = 0$, $s_0 = 0$, $\tilde{u}_\nu^l = u_\nu^l$.

If $u_r^\nu \neq \tilde{u}_{\nu^l}$, then we have many possibilities to reconsider this system. Solution component $x_4(t)$ can be chosen arbitrarily and hence could be viewed as an *extra input* to the system. Compare this with the results concerning generalized inverses of differential algebraic operators in [19]. Also, if $\Sigma_{\tilde{u}_\nu^l}$ is pointwise nonsingular, we could use a nonunitary equivalence transformation to eliminate the block $B_{12}(t)$ in B. If we do this and choose some x_4, u_1, u_3, then the first two equations form a subsystem independent from the rest and if its solution is computed, then u_2 is fixed. Thus we could interpret u_2 as an *extra state variable* rather than a control variable. Combining these ideas we could replace (18) with the system

$$
\begin{array}{l} d_\nu \\ v_\nu \\ s_\nu \\ \tilde{u}_\nu^l \end{array}
\begin{bmatrix} \Sigma_{d_\nu} & 0 & E_{13} & 0 \\ 0 & 0 & E_{23} & 0 \\ 0 & 0 & E_{33} & 0 \\ 0 & 0 & E_{43} & 0 \end{bmatrix}
\begin{bmatrix} \dot{x}_1 \\ x_2 \\ x_3 \\ u_2 \end{bmatrix}
\tag{20}
$$

$$
= \begin{bmatrix} A_{11} & A_{12} & A_{13} & B_{12} \\ A_{21} & A_{22} & A_{23} & 0 \\ 0 & 0 & A_{33} & 0 \\ A_{41} & A_{42} & A_{43} & \Sigma_{\tilde{u}_\nu^l} \end{bmatrix}
\begin{bmatrix} x_1 \\ x_2 \\ x_3 \\ u_2 \end{bmatrix}
+ \begin{bmatrix} B_{11} & A_{14} & B_{13} \\ \Sigma_{v_\nu} & A_{24} & 0 \\ 0 & 0 & 0 \\ 0 & A_{44} & 0 \end{bmatrix}
\begin{bmatrix} u_1 \\ x_4 \\ u_3 \end{bmatrix},
$$

where we have left out the dependence on t for convenience. This is now a square system.

But it is clear that we can extract a square subsystem in many different ways by reinterpreting states as inputs or vice versa. A suitable choice will certainly depend on the application. In any case we wish to have a unique solution for suitably chosen inputs, thus we cannot allow the pencil to have more columns than rows. On the other hand, if there are more rows than columns, then the set of controls that will lead to continuous solutions is restricted by these extra algebraic equations.

The previous observations suggest that the design of a practical control problem should be done in such a way, that components which are identically zero should be left off already in the model and the reinterpretation of components as states or controls should be done beforehand.

Thus we will assume in the following that the system has been reordered so that it has square matrices $E(t), A(t)$ with $\nu = 0$, $s_0 = 0$, $\tilde{u}_\nu^l = u_\nu^l$ if we transform it to the condensed form of Theorem 3.2. We call such a subsystem *an underlying square subsystem.*

Observe that although all the transformations that we used are unitary transformations which can in principle be carried out in a numerically stable way, the rank decisions are still an ill-conditioned problem and small perturbations can change the picture completely. See the remarks for constant coefficient systems in [5], which hold here, too.

5. Regularization by feedback. For constant coefficient systems regularizability, i.e., the question whether there exist proportional and/or derivative feedbacks such that the closed loop system has a regular pencil, i.e. is solvable for all consistent initial vectors, has been studied by several authors; see, for example, [13, 26, 3, 5]. We now generalize these results to the variable coefficient case. We introduce the following concepts.

DEFINITION 5.1.

(a) *The descriptor system* (1) *is called* regularizable by proportional feedback *if there exists a (proportional state) feedback* $u(t) = F(t)x(t) + w(t)$ *such that the closed*

loop system

$$E(t)\dot{x}(t) = (A(t) + B(t)F(t))x(t) + B(t)w(t), \qquad x(t_0) = x_0, \tag{21}$$

is uniquely solvable for every consistent initial vector x_0 and any given control $w(t)$.

(b) *The descriptor system* (1) *is called* regularizable by derivative feedback *if there exists a (derivative) feedback $u(t) = G(t)\dot{x}(t) + w(t)$ such that the closed loop system*

$$(E(t) + B(t)G(t))\dot{x}(t) = A(t)x(t) + B(t)w(t), \qquad x(t_0) = x_0, \tag{22}$$

is uniquely solvable for all consistent initial vectors x_0 and any given control $w(t)$.

It is clear from the discussion in the previous section that we need $\tilde{u}_\nu^l = u_\nu^r$ in order to obtain regularizability, otherwise we cannot expect a unique solution or we have to reinterpret certain variables. If $\tilde{u}_\nu^l > u_\nu^r$, then we cannot apply arbitrary controls, and if $u_\nu^r > \tilde{u}_\nu^l$, then the solution will not be unique.

As the following theorem shows, this gives a necessary and sufficient condition if the matrices Σ_{d_ν}, Σ_{v_ν} occurring in the condensed form are invertible everywhere in the given interval; i.e., no rank drops occur, not even at isolated points.

THEOREM 5.2. *Consider system* (1) *in the condensed form* (16)*, and assume that the diagonal matrices $\Sigma_{d_\nu}(t)$, $\Sigma_{v_\nu}(t)$, $\Sigma_{\tilde{u}_\nu^l}(t)$ are pointwise nonsingular in the whole interval $[t_0, t_1]$.*

System (1) *can be regularized by proportional state feedback if and only if $\tilde{u}_\nu^l = u_\nu^r$.*

System (1) *can be regularized by derivative feedback if and only if $\tilde{u}_\nu^l = u_\nu^r$.*

Proof. We have already observed that $\tilde{u}_\nu^l = u_\nu^r$ is a necessary condition. In order to show that this is also sufficient observe that in this case the system can be permuted (by exchanging the last two block rows and columns) to the form

$$\begin{array}{c} d_\nu \\ v_\nu \\ \tilde{u}_\nu^l = u_\nu^r \\ s_\nu \end{array} \begin{bmatrix} \Sigma_{d_\nu} & 0 & 0 & E_{13} \\ 0 & 0 & 0 & E_{23} \\ 0 & 0 & 0 & E_{43} \\ 0 & 0 & 0 & E_{33} \end{bmatrix} \begin{bmatrix} \dot{x}_1 \\ x_2 \\ x_4 \\ x_3 \end{bmatrix}$$

$$= \begin{bmatrix} A_{11} & A_{12} & A_{14} & A_{13} \\ A_{21} & A_{22} & A_{24} & A_{23} \\ A_{41} & A_{42} & A_{44} & A_{43} \\ 0 & 0 & 0 & A_{33} \end{bmatrix} \begin{bmatrix} x_1 \\ x_2 \\ x_4 \\ x_3 \end{bmatrix} + \begin{bmatrix} B_{11} & B_{12} & B_{13} \\ \Sigma_{v_\nu} & 0 & 0 \\ 0 & \Sigma_{\tilde{u}_\nu^l} & 0 \\ 0 & 0 & 0 \end{bmatrix} \begin{bmatrix} u_1 \\ u_2 \\ u_3 \end{bmatrix}.$$

Since Σ_{d_ν} and $\Sigma_{\tilde{u}_\nu^l}$ are nonsingular in the whole interval, we can choose the proportional feedback

$$\begin{bmatrix} u_1(t) \\ u_2(t) \\ u_3(t) \end{bmatrix} = \begin{bmatrix} F_{11}(t) & F_{12}(t) & 0 & 0 \\ F_{21}(t) & F_{22}(t) & F_{23}(t) & 0 \\ 0 & 0 & 0 & 0 \end{bmatrix} \begin{bmatrix} x_1(t) \\ x_2(t) \\ x_4(t) \\ x_3(t) \end{bmatrix} + w(t)$$

such that

$$\Sigma_{v_\nu}(t) \begin{bmatrix} F_{11}(t) & F_{12}(t) \end{bmatrix} = \begin{bmatrix} -A_{21}(t) & I - A_{22}(t) \end{bmatrix}$$

and

$$\Sigma_{\tilde{u}_\nu^l}(t) \begin{bmatrix} F_{21}(t) & F_{22}(t) & F_{23}(t) \end{bmatrix} = \begin{bmatrix} -A_{41}(t) & -A_{42}(t) & I - A_{44}(t) \end{bmatrix}.$$

This choice gives a closed loop system

$$\begin{matrix} d_\nu \\ v_\nu \\ \tilde{u}_\nu^l = u_\nu^r \\ s_\nu \end{matrix} \begin{bmatrix} \Sigma_{d_\nu} & 0 & 0 & E_{13} \\ 0 & 0 & 0 & E_{23} \\ 0 & 0 & 0 & E_{43} \\ 0 & 0 & 0 & E_{33} \end{bmatrix} \begin{bmatrix} \dot{x}_1 \\ x_2 \\ x_4 \\ x_3 \end{bmatrix}$$

$$= \begin{bmatrix} A_{11} & A_{12} & A_{14} & A_{13} \\ 0 & I & A_{24} & A_{23} \\ 0 & 0 & I & A_{43} \\ 0 & 0 & 0 & A_{33} \end{bmatrix} \begin{bmatrix} x_1 \\ x_2 \\ x_4 \\ x_3 \end{bmatrix} + \begin{bmatrix} B_{11} & B_{12} & B_{13} \\ \Sigma_{v_\nu} & 0 & 0 \\ 0 & \Sigma_{\tilde{u}_\nu^l} & 0 \\ 0 & 0 & 0 \end{bmatrix} \begin{bmatrix} w_1 \\ w_2 \\ w_3 \end{bmatrix}.$$

Recall that the solutions components x_3 are constrained to be zero. If we remove these equations, then the remaining system has strangeness index $\mu = 0$, i.e., is uniquely solvable for all consistent initial values.

Similarly, in the case of derivative feedback we choose the derivative feedback

$$\begin{bmatrix} u_1(t) \\ u_2(t) \\ u_3(t) \end{bmatrix} = \begin{bmatrix} 0 & F_{12}(t) & 0 & 0 \\ 0 & 0 & F_{13}(t) & 0 \\ 0 & 0 & 0 & 0 \end{bmatrix} \begin{bmatrix} \dot{x}_1(t) \\ x_2(t) \\ x_4(t) \\ x_3(t) \end{bmatrix} + w(t)$$

such that $\Sigma_{v_\nu} F_{12} = -I$ and $\Sigma_{\tilde{u}_\nu^l} F_{23} = -I$.

This choice gives a closed loop system

$$\begin{matrix} d_\nu \\ v_\nu \\ \tilde{u}_\nu^l = u_\nu^r \\ s_\nu \end{matrix} \begin{bmatrix} \Sigma_{d_\nu} & 0 & 0 & E_{13} \\ 0 & I & 0 & E_{23} \\ 0 & 0 & I & E_{43} \\ 0 & 0 & 0 & E_{33} \end{bmatrix} \begin{bmatrix} \dot{x}_1 \\ x_2 \\ x_4 \\ x_3 \end{bmatrix}$$

$$= \begin{bmatrix} A_{11} & A_{12} & A_{14} & A_{13} \\ A_{21} & A_{22} & A_{24} & A_{23} \\ A_{43} & A_{42} & A_{44} & A_{43} \\ 0 & 0 & 0 & A_{33} \end{bmatrix} \begin{bmatrix} x_1 \\ x_2 \\ x_4 \\ x_3 \end{bmatrix} + \begin{bmatrix} B_{11} & B_{12} & B_{13} \\ \Sigma_{v_\nu} & 0 & 0 \\ 0 & \Sigma_{\tilde{u}_\nu^l} & 0 \\ 0 & 0 & 0 \end{bmatrix} \begin{bmatrix} w_1 \\ w_2 \\ w_3 \end{bmatrix},$$

which is as required. □

It is clear that weaker assumptions can be considered in Theorem 5.2 by allowing rank jumps in the matrices Σ_{d_ν}, Σ_{v_ν} at isolated points and using a weak solvability concept. This topic is currently under investigation.

Note further that there is still quite a lot of freedom in the choice of the feedback, and the freedom may be used to improve the robustness of the system as was done for constant coefficient systems in [3, 4, 12]. Unfortunately, so far it is not really clear what robustness means for variable coefficient systems of the type considered.

6. Conclusion. We have shown that under some smoothness assumptions every linear time-varying descriptor system can be transformed to a condensed form which displays free state components which can be interpreted as inputs, fixed controls which can be interpreted as states and form a regularizable subsystem, solution components which are constrained to be zero coming from higher index components that are unchanged by feedback, plus equations which hold trivially. In principle this structure can be obtained from a sequence of smooth singular value decompositions for which numerical methods are available. From a practical point of view, however,

the uncontrollable higher index part and the other removable parts are very sensitive to perturbations which may change the whole system structure. In view of this, modeling or linearization which leads to such components should be avoided.

Acknowledgment. We thank Werner Rath for helpful comments.

REFERENCES

[1] K. Brenan, S. Campbell, and L. Petzold, *Numerical Solution of Initial-Value Problems in Differential Algebraic Equations*, Elsevier–North Holland, New York, 1989.

[2] A. Bunse-Gerstner, R. Byers, V. Mehrmann, and N. K. Nichols, *Numerical computation of an analytic singular value decomposition of a matrix valued function*, Numer. Math., 60 (1991), pp. 1–40.

[3] A. Bunse-Gerstner, V. Mehrmann, and N. Nichols, *Regularization of descriptor systems by derivative and proportional state feedback*, SIAM J. Matrix Anal. Appl., 13 (1992), pp. 46–67.

[4] A. Bunse-Gerstner, V. Mehrmann, and N. Nichols, *Regularization of descriptor systems by output feedback*, IEEE Trans. Automat. Control, 39 (1994), pp. 1742–1748.

[5] R. Byers, T. Geerts, and V. Mehrmann, *Descriptor systems without controllability at infinity*, SIAM J. Control Optim., 35 (1997), to appear.

[6] S. Campbell, *Linearization of DAEs along trajectories*, Z. Angew. Math. Phys., 46 (1995), pp. 70–84.

[7] S. Campbell, N. Nichols, and W. Terrell, *Duality, observability, and controllability for linear time-varying descriptor systems*, Circuits Systems Signal Process., 10 (1991), pp. 455–470.

[8] S. Campbell and W. Terrell, *Observability for linear time-varying descriptor systems*, SIAM J. Matrix Anal. Appl., 12 (1991), pp. 484–496.

[9] S. L. Campbell, *A general form for solvable linear time varying singular systems of differential equations*, SIAM J. Math. Anal., 18 (1987), pp. 1101–1114.

[10] S. L. Campbell, *The numerical solution of higher index linear time varying singular systems of differential equations*, SIAM J. Sci. Statist. Comput., 6 (1988), pp. 334–348.

[11] C. De Boor, *An empty exercise*, SIGNUM Newsletter, 25 (1990), pp. 2–6.

[12] L. Elsner, C. He, and V. Mehrmann, *Completion of a matrix so that the inverse has minimum norm. Application to the regularization of descriptor control problems*, in Linear Algebra for Control Theory, P. Van Dooren and B. Wyman, eds., Springer-Verlag, New York, 1993, pp. 75–86.

[13] L. Fletcher, J. Kautsky, and N. Nichols, *Eigenstructure assignment in descriptor systems*, IEEE Trans. Automat. Control, AC-31 (1986), pp. 1138–1141.

[14] M. Fliess, J. Levine, and P. Rouchon, *Index of a general linear time-varying implicit system*, in European Control Conference 1991, Hermes, Paris, 1991, pp. 769–772.

[15] E. Griepentrog and R. März, *Differential-Algebraic Equations and Their Numerical Treatment*, Teubner-Verlag, Leipzig, 1986.

[16] M. Günther and P. Rentrop, *Multirate Row-Methods and Latency of Electric Circuits*, Tech. report TUM-M9208, Mathematisches Institut, Technische Universität München, Arcisstr. 21, Postfach 202420, D-8000 München 2, 1992.

[17] P. Kunkel and V. Mehrmann, *Smooth factorizations of matrix valued functions and their derivatives*, Numer. Math., 60 (1991), pp. 115–132.

[18] P. Kunkel and V. Mehrmann, *Canonical forms for linear differential-algebraic equations with variable coefficients*, J. Comput. Appl. Math., 56 (1994), pp. 225–251.

[19] P. Kunkel and V. Mehrmann, *Generalized inverses of differential-algebraic operators*, SIAM J. Matrix Anal. Appl., 17 (1996), pp. 426–442.

[20] P. Kunkel and V. Mehrmann, *A new look at pencils of matrix valued functions*, Linear Algebra Appl., 212/213 (1994), pp. 215–248.

[21] P. Kunkel and V. Mehrmann, *Local and Global Invariants of Linear Differential-Algebraic Equations and Their Relation*, Fakultät für Mathematik SPC 95-25, Technical University Chemnitz, D-09107 Chemnitz, 1995, submitted for publication.

[22] P. Kunkel and V. Mehrmann, *A new class of discretization methods for the solution of linear differential-algebraic equations*, SIAM J. Numer. Anal., 33 (1996), pp. 1941–1961.

[23] P. Kunkel, V. Mehrmann, W. Rath, and J. Weickert, *A new software package for linear differential algebraic equations*, SIAM J. Sci. Comput., 18 (1997), to appear.

[24] J. J. Loiseau, K. Özçaldiran, M. Malabre, and N. Karcanias, *Feedback canonical forms of singular systems*, Kybernetika (Prague), 27 (1991), pp. 289–305.
[25] V. Mehrmann and W. Rath, *Numerical methods for the computation of analytic singular value decompositions*, Electron. Trans. Numer. Anal., 1 (1993), pp. 72–88.
[26] K. Özcaldiran and F. Lewis, *On regularizability of singular systems*, IEEE Trans. Automat. Control, AC-35 (1990), pp. 1156–1160.
[27] P. Rabier and W. Rheinboldt, *Classical and generalized solutions of time-dependent linear differential algebraic equations*, Linear Algebra Appl., to appear.
[28] P. Rabier and W. Rheinboldt, *A geometric treatment of implicit differential–algebraic equations*, J. Differential Equations, 109 (1994), pp. 110–146.
[29] W. Rath, *Canonical forms for linear descriptor systems with variable coefficients*, Linear Algebra Appl., to appear.
[30] B. Simeon, C. Führer, and P. Rentrop, *Differential-algebraic equations in vehicle system dynamics*, Surveys Math. Indust., 1 (1991), pp. 1–37.
[31] B. Simeon, F. Grupp, C. Führer, and P. Rentrop, *A nonlinear truck model and its treatment as a multibody system*, Technical report TUM-M9204, Mathematisches Institut, TU München, München, FRG, 1992.
[32] T. Schmidt and M. Hou, *Rollringgetriebe*, Internal report, Sicherheitstechnische Regelungs- und Meßtechnik, Bergische Universität, GH Wuppertal, Wuppertal, FRG, 1992.
[33] P. Van Dooren, *The computation of Kronecker's canonical form of a singular pencil*, Linear Algebra Appl., 27 (1979), pp. 103–141.
[34] K. Wright, *Differential equations for the analytic singular value decomposition of a matrix*, Numer. Math., 3 (1992), pp. 283–295.

SIAM J. CONTROL OPTIM.
Vol. 35, No. 1, pp. 134–159, January 1997

007

ON CONVERGENCE OF ATTAINABILITY SETS FOR CONTROLLED TWO-SCALE STOCHASTIC LINEAR SYSTEMS*

YURI KABANOV† AND SERGEI PERGAMENSHCHIKOV‡

Abstract. A limit of attainability sets is found for a linear two-scale stochastic system for the case when the diffusion coefficient of the fast variable is of order $\varepsilon^{1/2}$. The attainability set is defined as the set of distributions of attainable terminal values of solutions of stochastic differential equations. As a corollary we calculate a limit of the optimal value of the terminal cost in the stochastic Mayer problem.

Key words. controlled stochastic differential equations, two-scale system, singular perturbations, attainability sets, Mayer problem, Hausdorff metric

AMS subject classifications. 93E20, 93C73

PII. S0363012994269685

Introduction. In mathematical modeling of complex systems with processes having two essentially different "velocities," fast variables are usually described by singularly perturbed differential equations, i.e., by equations having a small parameter ε on the left-hand side. In general, there is a hope that the reduced limiting model (when the parameter is equal to zero) is more simple and can be used as an approximation of the original one which may be rather complicated. This idea seems to be fruitful also in the set-up of controlled systems. However, here an additional difficulty arises since the optimal value of the cost function which depends smoothly on $\varepsilon \in]0,1]$ may have a discontinuity at the most interesting point $\varepsilon = 0$.

To overcome this difficulty in the deterministic setting, an approach based on a study of the convergence of the attainability sets in the Hausdorff metric has been developed; see, e.g., recent work [10]. In the linear case it is possible to find a limit of the attainability sets in a rather explicit way which has been done by Dontchev and Veliov [8]; see also the book [7]. Their result is as follows.

Let us consider the controlled system

$$\dot{x}_t = A_1(t)x_t + A_2(t)y_t + B_1(t)u_t, \quad x_0 = 0, \tag{0.1}$$

$$\varepsilon\dot{y}_t = A_3(t)x_t + A_4(t)y_t + B_2(t)u_t, \quad y_0 = 0, \tag{0.2}$$

where ε is a small positive number; u is any measurable function with values in a convex compact subset of $\mathbf{R}^d$; matrix-valued functions A_i, B_i are continuous; and the eigenvalues of $A_4(t)$ have strictly negative real parts.

Let $K_\varepsilon(t)$ be the attainability set of the system (0.1), (0.2), i.e., the set of all end points (x_T, y_T) corresponding to various admissible controls, and let $K_0^x(T)$ be the attainability set of the reduced system

$$\dot{x}_t = A_0(t)x_t + B_0(t)u_t, \quad x_0 = 0,$$

with the coefficients $A_0 := A_1 - A_2A_4^{-1}A_3$, $B_0 := B_1 - A_2A_4^{-1}B_2$.

*Received by the editors June 8, 1994; accepted for publication (in revised form) October 18, 1995.

http://www.siam.org/journals/sicon/35-1/26968.html

†Bilkent University, Bilkent, 06533, Ankara, Turkey, and Central Economics and Mathematics Institute, Krasikova str., 32, Moscow 117433, Russia. Current address: U.F.R. des Sciences, Laboratoire des Mathématiques, 16, Route de Gray, 25030 Besançon Cedex, France (kabanov@vega.univ-fcomte.fr).

‡Tomsk State University, Tomsk 634041, Russia.

Let us define the set $K_0(T) := \{(x,y) : \ x \in K_0^x(T),\ y \in R(T,x)\}$, where $R(T,x) := -A_4^{-1}(T)A_3(T)x + Y$,

$$Y := \int_0^\infty \exp\{A_4(T)s\}B_2(T)U ds = \left\{y : y = \int_0^\infty \exp\{A_4(T)s\}B_2(T)v_s ds,\ v_s \in V_U\right\}.$$

V_U is the set of all U-valued Borel functions. In other words, if we put $F(x,y) = (x, -A_4^{-1}(T)A_3(T)x + y)$, then $K_0(T)$ is the image of $K_0^x(T) \times Y$ under the mapping F.

THEOREM (see [8], [7]). *The sets $K_\varepsilon(T)$ tend to $K_0(T)$ in the Hausdorff metric as $\varepsilon \to 0$.*

Let us consider for the system (0.1), (0.2) the Mayer problem

$$g(x_T, y_T) \to \min,$$

where g is a continuous function. Then the optimal value for the perturbed problem is

$$J_\varepsilon^* = \min_{K_\varepsilon(T)} g(x,y).$$

From the above theorem it follows immediately that

$$\lim_{\varepsilon \to 0} J_\varepsilon^* = \min_{K_0(T)} g(x,y).$$

In the paper [13] the authors extended the theorem on the convergence of the attainability sets to stochastic differential equations of the form

$$dx_t = (A_1(t)x_t + A_2(t)y_t + B_1(t)u_t)dt + dw_t^x, \quad x_0 = 0, \tag{0.3}$$

$$\varepsilon dy_t = (A_3(t)x_t + A_4(t)y_t + B_2(t)u_t)dt + \sigma(\varepsilon)dw_t^y, \quad y_0 = 0, \tag{0.4}$$

where w^x, w^y are independent Wiener processes and $\sigma(\varepsilon) = O(\varepsilon^{1/2+\delta})$, $\delta > 0$. In the stochastic setting it is natural to define the attainability set as the set of distributions of all terminal random variables (x_T, y_T) when u runs through the set of admissible controls. There are several possible choices for the latter. It seems that the most adequate one is to consider all nonanticipating functions of the trajectories as admissible controls. This implies the need to understand the system (0.3), (0.4) in the weak sense; i.e., the Wiener processes are not given in advance and the solution is actually a probability measure $P^{\varepsilon,u}$ in the space of continuous functions $C[0,T]$. Such a solution can be constructed by the Girsanov theorem. In this case the attainability set $\mathcal{K}_\varepsilon(T)$ is a compact convex set in the space of probability measures equipped with the Prohorov metric. In [13] it was shown that $\mathcal{K}_\varepsilon(T) \to \mathcal{K}_0(T)$ in the Hausdorff metric, where $\mathcal{K}_0(T)$ is the set of probability measures μF^{-1} where $\mu = \mu(dx, dy)$ is such that $\mu(dx, \mathbf{R}^n)$ belongs to the attainable set $\mathcal{K}_0^x(T)$ of the reduced system and $\mu(\mathbf{R}^k, dy)$ belongs to the set $\mathbf{P}(Y)$ of probability measures on Y. The reduced system is given by

$$dx_t = (A_0(t)x_t + B_0(t)u_t)dt + dw_t^x, \quad x_0 = 0, \tag{0.5}$$

where, as in the deterministic case, the coefficients A_0 and B_0 can be obtained if we substitute in (0.3) the expression for y_t which is a formal solution of (0.4) with $\varepsilon = 0$.

Notice that the condition $\delta > 0$ provides a limiting degeneracy of the stochastic equation (0.4) (with a fixed control) to an algebraic one.

In the present paper we prove the convergence result for $\sigma(\varepsilon) = \varepsilon^{1/2}$. In this case $\mathcal{K}_0(T)$ is the set of all measures μF^{-1} such that $\mu(dx, \mathbf{R}^n) \in \mathcal{K}_0^x(T)$ and $\mu(x, dy)$ belong to the convex closure of the set of probability distributions of random variables

$$\xi_0 + \int_0^\infty \exp\{A_4(T)s\} B_2(T) v_s ds,$$

where ξ is the stationary Gaussian Markov process (called also Ornstein–Uhlenbeck) with the zero mean and covariance

$$K(s,t) := \Xi \exp\{A_4'(T)(t-s)\}, \quad s \le t,$$
$$\Xi := \int_0^\infty \exp\{A_4(T)s\} \exp\{A_4'(T)s\} ds,$$

v is any measurable process with values in U such that for any t the random variable v_t is measurable with respect to the σ-algebra $\mathcal{F}^\xi_{\ge t} := \sigma\{\xi_s,\ s \ge t\}$, and prime denotes the matrix transpose. As a corollary of the theorem on convergence of the attainability sets we calculate a limit of the optimal value in the Mayer problem $Eg(x_T^{\varepsilon,u}, y_T^{\varepsilon,u}) \to \min$ when ε tends to zero.

In the last few years singularly perturbed controlled stochastic differential equations have been intensively studied by various methods, mainly based on the theory of weak convergence in the functional spaces or the Bellman–Hamilton–Jacobi equation; see monographs [3], [4], [20] and papers [2], [5] (and the collection [17] for early results). However, almost all studies concern models where the controlled fast variable does not affect the terminal cost. Harold Kushner wrote in his book [20, p. 64]:

> It is hard to deal in any general way with the case where the fast system is also controlled. The main difficulty is due to the fact that the 'stationary measures' which are used to average out the fast variable depend on the control which is used in the fast system. This makes it hard to define the 'averaged problem.'... Similar problems occur in the deterministic case, and it is commonly dealt with there by supposing that the choice of control for the fast system does not alter the steady state value of that system, for each value of the fast variable, i.e., that the fast system is asymptotically stable and the control chosen in a class such that the limit point of that fast system does not depend on the control *when x is fixed*. This assumption essentially 'decouples' the fast and slow system. The assumption seems reasonable and yields good results. Unfortunately, it does not seem possible to find a stochastic analog of this approach which works in any generality.

It worth noticing that the result presented here is nontrivial even for a system with only fast variables. In this case it is clear that the limit of the attainability sets shows to what extent optimal controls (acting on the drift of the process) can follow the change in the scale parameter near the point zero.

The structure of the paper is the following. In section 1 we give the formal description of the problem. Section 2 contains some preliminary explanations and the proof of the result for the simplest one-dimensional model with the fast variable only. The proof of Theorem 1.1 is given in sections 3 and 4. Section 5 is devoted to measure-theoretical aspects which may have some independent interest.

1. Formulations of the results. We consider here the linear stochastic controlled system given by

$$dx_t = (A_1(t)x_t + A_2(t)y_t + B_1(t)u_t)dt + dw_t^x, \quad x_0 = 0, \tag{1.1}$$

$$\varepsilon dy_t = (A_3(t)x_t + A_4(t)y_t + B_2(t)u_t)dt + \sqrt{\varepsilon}dw_t^y, \quad y_0 = 0, \tag{1.2}$$

where w^x and w^y are standard independent Wiener processes with values in $\mathbf{R}^k$ and $\mathbf{R}^n,\ 0 \le t \le T < \infty,\ \varepsilon \in]0,1]$.

We shall understand (1.1), (1.2) as a symbolic notation for the stochastic differential equation in a weak sense when a Wiener process $W = (w^x, w^y)$ is not given in advance and u is a feedback control. Actually, in the following rigorous formulation we could avoid the above representation (which is, in fact, a bit ambiguous) altogether.

We consider as a phase space $\mathbf{R}^m = \mathbf{R}^k \times \mathbf{R}^n$. ($\mathbf{R}^k$ corresponds to the slow and $\mathbf{R}^n$ to the fast variables.) The phase space of control will be a compact convex set $U \subseteq \mathbf{R}^d$. In our matrix notations vectors are column vectors.

The path space of the system is the space $C[0,T]$ of continuous functions $W : [0,T] \to \mathbf{R}^m$. Let $\mathcal{C}_T$ be the Borel σ-algebra on $C[0,T]$, $\mathcal{C}_t^o := \sigma\{W_s,\ s \le t\}$, $\mathcal{C}_t := \mathcal{C}_{t+}^o$. Let $\mathcal{P}$ be the predictable σ-algebra in $C[0,T] \times [0,T]$ corresponding to the filtration $\mathbf{C} = (\mathcal{C}_t)$.

The class of admissible controls $\mathcal{U}$ is defined as the set of all predictable processes $u = (u_t)_{t\in[0,T]}$ with values in U.

Let $A_i = A_i(t),\ B_i = B_i(t)$ be matrix-valued continuous functions of dimensions compatible with (1.1), (1.2); i.e., $A_1(t)$ is a $k \times k$ matrix, $A_4(t)$ is $n \times n$, etc.

We introduce the following notation:

$$f_\varepsilon(W,t,u) = \begin{pmatrix} A_1(t) & A_2(t) \\ \varepsilon^{-1}A_3(t) & \varepsilon^{-1/2}A_4(t) \end{pmatrix} W_t + \begin{pmatrix} B_1(t) \\ \varepsilon^{-1}B_2(t) \end{pmatrix} u_t, \tag{1.3}$$

$$D_\varepsilon := \begin{pmatrix} I_k & 0 \\ 0 & \varepsilon^{-1}I_n(t) \end{pmatrix}, \tag{1.4}$$

where $I_k,\ I_n$ are the identity matrices of corresponding dimensions.

Consider on $(C[0,T], \mathcal{C}_T)$ the probability measure P^ε such that with respect to P^ε the coordinate process W is the Wiener process with the correlation matrix $D_\varepsilon D_\varepsilon'$.

For any admissible control u we define the measure $P^{\varepsilon,u} := \rho_T^\varepsilon(u)P^\varepsilon$ with

$$\rho_T^\varepsilon(u) = \exp\left\{\int_0^T f_\varepsilon(W,s,u_s)' dW_s - \frac{1}{2}\int_0^T |f_\varepsilon(W,s,u_s)' D_\varepsilon|^2 ds\right\}. \tag{1.5}$$

It is well known (see [1] or [16]) that $P^{\varepsilon,u}$ is a probability measure. By the Girsanov theorem the process

$$W_t - \int_0^t f_\varepsilon(W,s,u_s)ds$$

with respect to $P^{\varepsilon,u}$ is the Wiener process with the correlation matrix $D_\varepsilon D_\varepsilon'$. Thus, we can write that

$$dW_t = f_\varepsilon(W,t,u_t)dt + D_\varepsilon dB_t, \quad W_0 = 0,$$

where B is the standard Wiener process.

If we denote the first k components of W and B by x and w^x and the remaining n components by y and w^y, the above representation formally coincides with the system (1.1), (1.2) and the control u will be a nonanticipating functional of the phase trajectory. This explains the terminology where $P^{\varepsilon,u}$ is called a weak solution of (1.1), (1.2) and the model itself usually is referred to as the model with the feedback control.

Let $\mathcal{K}_\varepsilon := \{P^{\varepsilon,u} : \ u \in \mathcal{U}\}$, where $\varepsilon > 0$ is fixed. The set $\mathcal{K}_\varepsilon$ is an analog of the "tube" of trajectories for deterministic systems. Correspondingly, the attainability set $\mathcal{K}_\varepsilon(T) := \{P^{\varepsilon,u}W_T^{-1} : \ u \in \mathcal{U}\}$ is the set of all probability measures on $\mathbf{R}^m$ which are the images of elements of $\mathcal{K}_\varepsilon$ under the mapping $W \mapsto W_T$. It was proved in [1] that $\mathcal{K}_\varepsilon$ is a convex set, hence $\mathcal{K}_\varepsilon(T)$ is also convex. In [1] it was also shown that the set $\{\rho_T^\varepsilon(u) : \ u \in \mathcal{U}\}$ of the attainable densities is sequentially compact in the weak topology of $L^1(P^\varepsilon)$. It follows immediately that $\mathcal{K}_\varepsilon$ and $\mathcal{K}_\varepsilon(T)$ are compact subsets of the corresponding spaces of probability measures $\mathbf{P}(C[0,T])$ and $\mathbf{P}(\mathbf{R}^m)$ equipped with the Prohorov metric.

To formulate the convergence result we need the following assumption.

(A) For all t the real parts of the eigenvalues of $A_4(t)$ have strictly negative real parts:

$$\operatorname{Re}\lambda(A_4(t)) \le -2\kappa < 0. \tag{1.6}$$

Let $\mathcal{K}_0^x(T)$ be the attainability set of the stochastic differential equation

$$dx_t = (A_0(t)x_t + B_0(t)u_t)dt + dw_t^x, \quad x_0 = 0, \tag{1.7}$$

where $A_0 := A_1 - A_2A_4^{-1}A_3, \ B_0 := B_1 - A_2A_4^{-1}B_2$.

Let ξ be the (strong) solution of the following stochastic differential equation with constant coefficients on some filtered probability space $(\Omega, \mathcal{F}, \mathbf{F} = (\mathcal{F}_t), P)$:

$$d\xi_t = A_4(T)\xi_t dt + db_t, \quad \xi_0 = \xi^o, \tag{1.8}$$

where b is a standard Wiener process in $\mathbf{R}^n$ and ξ^o is an independent Gaussian random variable with the zero mean and covariance matrix

$$\Xi := \int_0^\infty \exp\{A_4(T)s\}\exp\{A_4^{'}(T)s\}ds. \tag{1.9}$$

In other words, ξ is the stationary Gaussian Markov process with zero mean and covariance function

$$K(s,t) := E\xi_s x i_t^{'} = \Xi\exp\{A_4^{'}(T)(t-s)\}; \tag{1.10}$$

see, e.g., [16].

Let $\mathcal{V}_U$ be the set of all U-valued processes $v = (v_t)_{t\ge 0}$ such that $v_{1/t}$ is a predictable process with respect to the filtration generated by the process $\xi_{1/t}$, $S_Y^o := \{\mathcal{L}(\xi_0 + I(v)) : \ v \in \mathcal{V}_U\}$, where

$$I(v) := \int_0^\infty \exp\{A_4(T)s\}B_2(T)v_s ds. \tag{1.11}$$

Here and in what follows we use the notation $\mathcal{L}(\eta) := P\eta^{-1}$ for the distribution of the random variable η. The set S_Y^o is compact in $\mathbf{P}(\mathbf{R}^n)$; see Lemma 5.5.

Put $S_Y := \overline{\text{conv}}\ S_Y^o$, the convex closure of S_Y^o in $\mathbf{P}(\mathbf{R}^n)$.

Let S be the set of all probability measures $\mu = \mu(dx, dy)$ on $\mathbf{R}^m = \mathbf{R}^k \times \mathbf{R}^n$ such that

(1) $\mu(x, dy) \in S_Y$;

(2) $\mu(dx, \mathbf{R}^n) \in \mathcal{K}_0^x(T)$.

From the Proposition 5.2 it follows that S is compact in $\mathbf{P}(\mathbf{R}^m)$.

Define a linear mapping $F(x, y) := (x, -A_4^{-1}(T)A_3(T)x + y)$ of $\mathbf{R}^m$ into itself. Put $\mathcal{K}_0(T) := \{\mu F^{-1} : \ \mu \in S\}$.

Our main result is the following theorem.

THEOREM 1.1. *The set $\cup_{\varepsilon\in]0,1]}\mathcal{K}_\varepsilon(T)$ is compact, and as $\varepsilon \to 0$, $\mathcal{K}_\varepsilon(T)$ tend to $\mathcal{K}_0(T)$ in the Hausdorff metric in the space of compact subsets of $\mathbf{P}(\mathbf{R}^m)$.*

For the model (1.1), (1.2) we consider now the Mayer problem, which can be rigorously formulated as the problem to determine the minimal value of the functional

$$J_\varepsilon^* := \inf_{u\in\mathcal{U}} E^{\varepsilon,u} g(W_T) = \inf_{\mu\in\mathcal{K}_\varepsilon(T)} \int g(x, y)\mu(dx, dy), \tag{1.12}$$

where g is a function on $\mathbf{R}^m$ which is integrable with respect to the measures μ from $\mathcal{K}_\varepsilon(T)$.

COROLLARY 1.1. *Assume that g is continuous and bounded. Then*

$$\lim_{\varepsilon\to 0} J_\varepsilon^* = \inf_{\mu\in\mathcal{K}_0(T)} \int g(x, y)\mu(dx, dy). \tag{1.13}$$

Remark 1.1. The definition of the set $\mathcal{V}_U$ seems rather complicated. Essentially, $\mathcal{V}_U$ contains measurable processes v such that for any t the random variable v_t is measurable with respect to the σ-algebra $\mathcal{F}_{\geq t}^{\xi} := \sigma\{\xi_s, \ s \geq t\}$. To avoid a discussion of the measurable structures related to a decreasing family of σ-algebras we prefer to consider the processes in reversed time.

Remark 1.2. There is an alternative description of the set S_Y. Let α be a random variable independent of ξ with values in some Polish space and with a nonatomic distribution. Define the set $\mathcal{V}_U^\alpha$ as the set of all U-valued processes $v = (v_t)_{t\geq 0}$ such that $v_{1/t}$ is a predictable process with respect to the filtration generated by the process $\xi_{1/t}$ and the random variable α. Then $S_Y = \{\mathcal{L}(\xi_0 + I(v)) : \ v \in \mathcal{V}_U^\alpha\}$; see section 5.

Remark 1.3. Evidently, Theorem 1.1 can be applied to the more general optimization problem $J^\varepsilon(u) = F(P^{\varepsilon,u}) \to \min$, where F is any continuous function on $\mathbf{P}(\mathbf{R}^m)$.

We also use in our proof another possible model based on a different (and more traditional) interpretation of the equations (1.1), (1.2). To describe this alternative approach we consider the standard Wiener measure P on $(C[0, T], \mathcal{C}_T)$. Let w^x be the notation for the first k coordinates of the function W and w^y be the notation for the remaining n coordinates. Then for any $u \in \mathcal{U}$ we can find the strong solution $X^{\varepsilon,u} = (x^{\varepsilon,u}, y^{\varepsilon,u})$ of (1.1), (1.2). This model is referred to as the model with the open loop controls (since in this case u is a nonanticipating functional of the "noise").

Let $P_X^{\varepsilon,u} := P(X^{\varepsilon,u})^{-1}$ be the distribution in $C[0, T]$ of the process $X^{\varepsilon,u}$. Certainly, the measure $P_X^{\varepsilon,u}$ need not be equal to $P^{\varepsilon,u}$. Let us consider the sets $\tilde{\mathcal{K}}_\varepsilon := \{P_X^{\varepsilon,u} : \ u \in \mathcal{U}\} \subseteq \mathbf{P}(C[0, T])$ and $\tilde{\mathcal{K}}_\varepsilon(T) := \{P(X_T^{\varepsilon,u})^{-1} : \ u \in \mathcal{U}\} \subseteq \mathbf{P}(\mathbf{R}^m)$. We do not know whether the attainability set $\tilde{\mathcal{K}}_\varepsilon(T)$ coincides with the attainability set $\mathcal{K}_\varepsilon(T)$. However, in our paper [13] it has been shown that there are dense embeddings $\tilde{\mathcal{K}}_\varepsilon \subseteq \mathcal{K}_\varepsilon$ and $\tilde{\mathcal{K}}_\varepsilon(T) \subseteq \mathcal{K}_\varepsilon(T)$ in the sense of total variation convergence (thus, in the weak topology) and that the inclusion $\tilde{\mathcal{K}}_\varepsilon \subseteq \mathcal{K}_\varepsilon$ is strict even in the simplest cases.

This fact, certainly, does not exclude the coincidence of $\tilde{\mathcal{K}}_\varepsilon(T)$ and $\mathcal{K}_\varepsilon(T)$. Nevertheless, the result that there is a dense embedding $\tilde{\mathcal{K}}_\varepsilon(T) \subseteq \mathcal{K}_\varepsilon(T)$ is very helpful since it permits us to apply pathwise techniques similar to that of the deterministic theory.

2. Main ideas and the proof of Theorem 1.1 in the simplest case. We recall some basic facts concerning the Hausdorff metric and convergence of compact sets (for details see, e.g., [11]).

Let (X, d) be a metric space and let $\mathbf{K}_X$ be the class of all its nonempty compact subsets. For $A,\ B \in \mathbf{K}_X$ put $l(A, B) := \sup_{z\in A} d(z, B)$. The Hausdorff distance between A and B is defined by the equality

$$d_H(A, B) := l(A, B) \vee l(B, A).$$

If $A_m \in \mathbf{K}_X$, $m \in \mathbf{Z}_+$, and all A_m are contained in some compact set, then $\lim d_H(A_m, A_0) = 0$ if and only if the following two much more tractable conditions are satisfied for any subsequences of indices (n):

(1) For any convergent sequence $z_n \in A_n$ its limit is a point in A_0.

(2) For any point $z \in A_0$ there exists a subsequence $z_{n_k} \in A_{n_k}$ converging to z.

Notice that if A_n are not subsets of some compact set, the above equivalence fails in general. For the subsets of the real line $A_n := [0, 1] \cup \{n\}$, conditions (1) and (2) are satisfied but A_n do not tend to A_0 in the Hausdorff metric.

The strategy of the proof of Theorem 1.1 is the following. In the first stage we show that for any $\mu_\varepsilon \in \mathcal{K}_\varepsilon(T),\ \varepsilon \in]0, 1]$, there exists $\bar{\mu}_\varepsilon \in \mathcal{K}_0(T)$ such that $d(\bar{\mu}_\varepsilon, \mu_\varepsilon) \to 0$ (d here is the Prohorov metric). Since all $\mathcal{K}_\varepsilon(T)$ are compact this implies that $\cup_{\varepsilon\geq 0}\mathcal{K}_\varepsilon(T)$ is compact and all limit points of $\{\mu_\varepsilon\}$ belongs to $\mathcal{K}_0(T)$; i.e., (1) is fulfilled. Since $\tilde{\mathcal{K}}_\varepsilon(T)$ is dense in $\mathcal{K}_\varepsilon(T)$ it is sufficient to consider only the case when $\mu_\varepsilon \in \tilde{\mathcal{K}}_\varepsilon(T)$. Thus, we can argue with terminal random variables $(x_T^{\varepsilon,u}, y_T^{\varepsilon,u})$ with the distributions μ_ε and approximate them in probability (or in L^p) by random variables $(\bar{x}_T^{\varepsilon,u}, \bar{y}_T^{\varepsilon,u})$ with distributions from $\mathcal{K}_0(T)$.

In the second step of the proof we should find for a given measure $\mu \in \mathcal{K}_0(T)$ the sequence of measures μ_n which are elements of $\tilde{\mathcal{K}}_{\varepsilon_n}(T)$ converging to μ. Again we shall argue with suitably chosen random variables with distributions corresponding to the measures for which we are looking.

Since the proof for the general multidimensional two-scale system requires rather long arguments, we clarify main ideas on the example of a one-dimensional model with constant coefficients and containing only the fast variable.

Let us consider the controlled stochastic differential equation

$$\varepsilon dy_t^{\varepsilon,u} = (-\gamma y_t^{\varepsilon,u} + u_t)dt + \varepsilon^{1/2} dw_t^y, \quad y_0 = 0, \tag{2.1}$$

where u is a predictable process which takes values in $U = [0, 1]$. In this case the set $\mathcal{K}_0(T)$ is the convex closure of the set $\{\mathcal{L}(\xi_0 + I(v)),\ v \in \mathcal{V}_U\}$, where

$$I(v) := \int_0^\infty e^{-\gamma s} v_s ds,$$

ξ is an Ornstein–Uhlenbeck process on some probability space $(\Omega, \mathcal{F}, P)$ with correlation function $K(s, t) = (2\gamma)^{-1} e^{-\gamma|t-s|}$, and $\mathcal{V}_U$ is the set of all U-valued processes v such that $v_{1/t}$ is a predictable process with respect to the filtration generated by the process $\xi_{1/t}$. For our purpose it is more convenient to use the alternative description of $\mathcal{K}_0(T)$ as the set $\{\mathcal{L}(\xi_0 + I(v)),\ v \in \mathcal{V}_U^\alpha\}$, where α is a random variable independent of ξ with values in a Polish space and nonatomic distribution and $\mathcal{V}_U^\alpha$ is the set of

all U-valued processes v such that $v_{1/t}$ is a predictable process with respect to the filtration generated by the process $\xi_{1/t}$ and the random variable α. We understand the equation (2.1) in the strong sense. Its solution can be represented in the following way:

$$y_t^{\varepsilon,u} = \varepsilon^{-1} \int_0^t e^{-\gamma(t-s)/\varepsilon} u_s ds + \eta_t^\varepsilon, \tag{2.2}$$

where

$$\eta_t^\varepsilon := \varepsilon^{-1/2} \int_0^t e^{-\gamma(t-s)/\varepsilon} dw_s^y. \tag{2.3}$$

Put $T_\varepsilon := T(1-\varepsilon^{1/2})$. Let us consider on the interval $[T_\varepsilon, T]$ the Gaussian stationary process

$$\tilde{\xi}_t^\varepsilon := (2\gamma)^{-1/2} \exp\{-\gamma(t-T_\varepsilon)/\varepsilon\}\beta + \varepsilon^{-1/2} \int_{T_\varepsilon}^t e^{-\gamma(t-s)/\varepsilon} dw_s^y,$$

where β is a standard normal random variable independent of the Wiener process w^y (to define β we can extend our canonical coordinate probability space). The process $\tilde{\xi}^\varepsilon$ is the solution of the linear equation

$$\varepsilon d\tilde{\xi}_t^\varepsilon = -\gamma \tilde{\xi}_t^\varepsilon dt + \varepsilon^{1/2} dw_t^y, \quad \tilde{\xi}_{T_\varepsilon}^\varepsilon = (2\gamma)^{-1/2}\beta.$$

Let us consider the Ornstein–Uhlenbeck process $\xi_t^\varepsilon = \tilde{\xi}_{T-\varepsilon t}^\varepsilon,\ t \in [0, T/\sqrt{\varepsilon}]$.

Evidently, $\eta_T^\varepsilon - \xi_0^\varepsilon = \eta_T^\varepsilon - \tilde{\xi}_T^\varepsilon \to 0$ in L^2 as $\varepsilon \to 0$.

For $u \in \mathcal{U}$ we define the process $v_s = v_s^\varepsilon := u_{T-\varepsilon s} I_{[0,T/\sqrt{\varepsilon}[}$.

Now we can write that

$$y_T^{\varepsilon,u} = \eta_T^\varepsilon + \int_0^{T/\sqrt{\varepsilon}} e^{-\gamma s} u_{T-\varepsilon s} ds + \int_{T/\sqrt{\varepsilon}}^{T/\varepsilon} e^{-\gamma s} u_{T-\varepsilon s} ds = \bar{y}_T^{\varepsilon,u} + R^\varepsilon(u),$$

where $\bar{y}_T^{\varepsilon,u} = \xi_0^\varepsilon + I(v)$,

$$R^\varepsilon(u) := \int_{T/\sqrt{\varepsilon}}^{T/\varepsilon} e^{-\gamma s} u_{T-\varepsilon s} ds + \eta_T^\varepsilon - \xi_0^\varepsilon.$$

Since $\sup_{u\in\mathcal{U}} |R^\varepsilon(u)| \to 0$ in probability, to accomplish the first step we need to check only that $\mathcal{L}(\xi_0^\varepsilon + I(v)) \in \mathcal{K}_0(T)$. Indeed, let us take for ξ the process ξ^ε defined above. For any $s \le T/\sqrt{\varepsilon}$ the random variable v_s is measurable with respect to the σ-algebra $\mathcal{C}_{T-\varepsilon s}$. But

$$\begin{aligned}\mathcal{C}_{T-\varepsilon s} &= \sigma\{w_r,\ r \le T_\varepsilon\} \vee \sigma\{w_r,\ T_\varepsilon \le r \le s\} \subseteq \sigma\{w_r,\ r \le T_\varepsilon\} \vee \sigma\{\tilde{\xi}_r^\varepsilon,\ T_\varepsilon \le r \le s\} \\ &= \sigma\{w_r,\ r \le T_\varepsilon\} \vee \sigma\{\xi_r^\varepsilon,\ s \le r \le T/\sqrt{\varepsilon}\},\end{aligned}$$

and we see that $v \in \mathcal{V}_U^\alpha$ where the random variable α is defined as the projection mapping of $C[0,T]$ onto $C[0,T_\varepsilon]$. The above considerations show that the limit of any convergent sequence $\mu^n \in \tilde{\mathcal{K}}_{\varepsilon_n}(T)$ is an element of $\mathcal{K}_0(T)$.

Now we introduce the set $\mathcal{V}_U^{\alpha\prime}$ consisting of all processes

$$v_s = \sum_{i=1}^N \varphi_i I_{]s_i, s_{i+1}]}(s) + u^0 I_{]s_{N+1},\infty[}(s), \tag{2.4}$$

where $0 = s_1 < \cdots < s_{N+1},\ u^0 \in U$, and the U-valued random variables φ_i have the form

$$\varphi_i = f_i(\alpha, \xi(r_1^i), \dots, \xi(r_{M_i}^i)), \quad s_{i+1} < r_j^i \le s_N. \tag{2.5}$$

Let $\mathcal{K}_0'(T) := \{\mathcal{L}(\xi_0 + I(v)),\ v \in \mathcal{V}_U^{\alpha'}\}$. It is easy to show that the set $\{I(v),\ v \in \mathcal{V}_U^{\alpha'}\}$ is dense in $\{I(v),\ v \in \mathcal{V}_U\}$ in probability. Thus, $\mathcal{K}'(T)$ is dense in $\mathcal{K}_0(T)$ in $\mathbf{P}(\mathbf{R})$.

Let $\mu \in \mathcal{K}'(T)$. This means that μ is the distribution of a random variable $\chi := \xi_0 + I(v)$ where v is of the form (2.4). The result will be proved if we construct a random variable χ^ε and a control u^ε such that $\mathcal{L}(\chi^\varepsilon) = \mathcal{L}(\chi)$ and $\chi^\varepsilon - y_T^{u^\varepsilon,\varepsilon} \to 0$ in probability. To this aim it is enough to find on the coordinate probability space $(C[0,T], \mathcal{C}, P)$ a stationary Gaussian Markov process ξ^ε with correlation function $K(s,t)$, a standard normal random variable α^ε independent on ξ^ε, and an admissible control $u^\varepsilon \in \mathcal{U}$ such that $\xi_0^\varepsilon - \eta_T^\varepsilon \to 0$ in probability (η_T^ε is defined by (2.3)), and

$$\int_0^\infty e^{-\gamma s} v_s^\varepsilon ds - \varepsilon^{-1} \int_0^T e^{-\gamma(T-s)/\varepsilon} u_s^\varepsilon ds \to 0,$$

where v^ε is the process given by the formula (2.4) if we substitute ξ^ε, φ^ε, and α^ε for ξ, φ, and α. Indeed, in this case the random variable $\chi^\varepsilon := \xi_0^\varepsilon + I(v^\varepsilon)$ meets the required properties.

The process ξ^ε can be constructed in the following way. For sufficiently small ε let $T_\varepsilon^k := T(1 - k\varepsilon^{1/2}),\ k = 1,2,3$. Put

$$\alpha^\varepsilon := (w_{T_\varepsilon^2} - w_{T_\varepsilon^3})/(T_\varepsilon^2 - T_\varepsilon^3)^{1/2},$$

$$\beta^\varepsilon := (2\gamma)^{-1/2}(w_{T_\varepsilon^1} - w_{T_\varepsilon^2})/(T_\varepsilon^1 - T_\varepsilon^2)^{1/2},$$

$$\tilde{\xi}_t^\varepsilon := \exp\{(t - T_\varepsilon^1)/\varepsilon\}\beta^\varepsilon + \varepsilon^{-1/2} \int_{T_\varepsilon^1}^t e^{-\gamma(t-s)/\varepsilon} dw_s, \quad t \ge T_\varepsilon^1.$$

Define the process ξ^ε on $[0, \varepsilon^{-1/2}T]$ by the equality $\xi_t^\varepsilon := \tilde{\xi}_{T-\varepsilon t}^\varepsilon$.

Evidently,

$$\xi_0^\varepsilon - \eta_T^\varepsilon = \exp\{(T - T_\varepsilon^1)/\varepsilon\}\beta^\varepsilon - \varepsilon^{-1/2} \int_0^{T_\varepsilon^1} e^{-\gamma(T-s)/\varepsilon} dw_s \to 0 \text{ in } L^2.$$

For sufficiently small ε we put

$$u^\varepsilon := u^0 I_{[0, t_{N+1}[} + \sum_{i=1}^{N+1} \varphi_i^\varepsilon I_{[t_{i+1}, t_i[},$$

where $t_i := T - \varepsilon s_i,\ i \le N+1$.

The random variables φ_i^ε are $\mathcal{C}_{t_{i+1}}$-measurable. Thus, $u^\varepsilon \in \mathcal{U}$. It follows that

$$\begin{aligned}\int_0^\infty e^{-\gamma s} v_s^\varepsilon ds - \varepsilon^{-1} \int_0^T e^{-\gamma(T-s)/\varepsilon} u_s^\varepsilon ds &= \int_0^\infty e^{-\gamma s} v_s^\varepsilon ds - \int_0^{T/\varepsilon} e^{-\gamma s} u_{T-\varepsilon s}^\varepsilon ds \\ &= \int_{T/\varepsilon}^\infty e^{-\gamma s} v_s^\varepsilon ds \to 0.\end{aligned}$$

The proof of the result for this particular case is finished.

3. Proof of Theorem 1.1. Part 1. We use the notation $\| f \|_t := \sup_{s\le t} |f_s|$ (omitting the subscript $t = T$) and denote by C different constants which do not depend on ε and u.

In the following statements the solution of (1.1), (1.2) (as well as that of (3.1)) is understood in the strong sense as given on the probability space $(C[0,T], \mathcal{C}_T, P)$.

PROPOSITION 3.1. *Let $(x_T^{\varepsilon,u}, y_T^{\varepsilon,u})$ be the solution of* (1.1), (1.2) *corresponding to some $u \in \mathcal{U}$, and let $\bar{x}^u$ be the solution of the reduced equation*

$$d\bar{x}_t^u = (A_0(t)\bar{x}_t^u + B_0(t)u_t)dt + dw_t^x, \quad \bar{x}_0^u = 0. \tag{3.1}$$

Then for any $p \in [1, \infty[$

$$\sup_\varepsilon \sup_{u\in\mathcal{U}} E \| x^{\varepsilon,u} \|^p < \infty, \tag{3.2}$$

$$\lim_{\varepsilon\to 0} \sup_{u\in\mathcal{U}} E \| x^{\varepsilon,u} - \bar{x}^u \|^p = 0, \tag{3.3}$$

$$\sup_\varepsilon \sup_{u\in\mathcal{U}} \sup_{t\le T} E|y_t^{\varepsilon,u}|^p < \infty. \tag{3.4}$$

Proof. Let us introduce for $\varepsilon^{-1}A_4(t)$ the fundamental matrix $\Psi^\varepsilon(t,s)$, which is the solution of the linear matrix equation

$$\frac{\partial \Psi^\varepsilon(t,s)}{\partial t} = \varepsilon^{-1}A_4(t)\Psi^\varepsilon(t,s), \quad \Psi^\varepsilon(s,s) = I_n. \tag{3.5}.$$

Since A_4 is continuous and the eigenvalues satisfy (1.6), there exists a constant L such that

$$|\Psi^\varepsilon(t,s)| \le Le^{-\kappa(t-s)/\varepsilon} \tag{3.6}$$

for all $s \le t \le T$ and $\varepsilon \in]0,1]$; see, e.g., [18]. In particular, from the above bound it follows that for all $t \le T$ and $\varepsilon \in]0,1]$

$$\frac{1}{\varepsilon}\int_0^t |\Psi^\varepsilon(t,s)|ds \le L/\kappa. \tag{3.7}$$

Using the fundamental matrix, the equation (1.2) can be solved with respect to $y = y^{\varepsilon,u}$ and we get the representation

$$y_t^{\varepsilon,u} = \frac{1}{\varepsilon}\int_0^t \Psi^\varepsilon(t,s)[A_3(s)x_s^{\varepsilon,u} + B_2(s)u_s]ds + \eta_t^\varepsilon, \tag{3.8}$$

where

$$\eta_t^\varepsilon := \frac{1}{\sqrt{\varepsilon}}\int_0^t \Psi^\varepsilon(t,s)dw_s^y. \tag{3.9}$$

The process η^ε is the solution of the linear stochastic equation

$$d\eta_t^\varepsilon = \varepsilon^{-1}A_4(t)\eta_t^\varepsilon dt + \varepsilon^{-1/2}dw_t^y, \quad \eta_0^\varepsilon = 0. \tag{3.10}$$

We shall use the following properties of η^ε following, e.g., from Theorem 3.1 in [14]: there exists a constant C_p such that

$$\sup_{t\ge 0} E|\eta_t^\varepsilon|^p \le C_p \tag{3.11}$$

for any $p \in [1, \infty[$ and

$$E \parallel \eta^\varepsilon \parallel^p \leq C_p \varepsilon^{-1/4} \tag{3.12}$$

for any $p \in [4, \infty[$.

Substituting (3.8) in the equation (1.1) written in the integral form we come to the following representation for the slow variable:

$$x_t^{\varepsilon,u} = \int_0^t [A_1(s) x_s^{\varepsilon,u} + B_1(s) u_s] ds$$
$$+ \int_0^t \left\{ A_2(s) \frac{1}{\varepsilon} \int_0^s \Psi^\varepsilon(s,r) [A_3(r) x_r^{\varepsilon,u} + B_2(r) u_r] dr \right\} ds + \zeta_t^\varepsilon + w_t^x, \tag{3.13}$$

where

$$\zeta_t^\varepsilon := \int_0^t A_2(s) \eta_s^\varepsilon ds. \tag{3.14}$$

LEMMA 3.1. *For any $p \in [1, \infty[$ there exists a constant c_p such that for all $\varepsilon \in]0, 1]$ it holds that*

$$E \parallel \zeta^\varepsilon \parallel^p \leq c_p, \tag{3.15}$$

$$\lim_{\varepsilon \to 0} E \parallel \zeta^\varepsilon \parallel^p = 0. \tag{3.16}$$

Proof. Since A_2 is bounded, (3.15) follows immediately from the Jensen inequality and (3.11). To prove (3.16) we consider the approximation of $D := A_2 A_4^{-1}$ by the step functions

$$D^N := \sum_{i=1}^N D_{t_i} I_{]t_{i-1}, t_i]},$$

where $t_i := iT/N$. Using (3.10) we have

$$\zeta_t^\varepsilon = \int_0^t D_s^N A_4(s) \eta_s^\varepsilon ds + \int_0^t (D_s - D_s^N) A_4(s) \eta_s^\varepsilon ds$$
$$= \varepsilon \sum_{i=1}^N D_{t_i} [\eta_{t_i \wedge t}^\varepsilon - \eta_{t_{i-1} \wedge t}^\varepsilon - \varepsilon^{1/2} (w_{t_i \wedge t}^y - w_{t_{i-1} \wedge t}^y)] + \int_0^t (D_s - D_s^N) A_4(s) \eta_s^\varepsilon ds.$$

This implies the bound

$$\parallel \zeta^\varepsilon \parallel \leq 2\varepsilon^{1/2} (\varepsilon^{1/2} \parallel \eta^\varepsilon \parallel + \parallel w^y \parallel) + C \delta_N \int_0^T |\eta_s^\varepsilon| ds, \tag{3.17}$$

where $\delta_N := \parallel D - D^N \parallel \to 0$ as $N \to \infty$ due to continuity of α.

Notice that (3.12) implies that the family of random variables $\{\varepsilon^{1/2} \parallel \eta^\varepsilon \parallel,\ \varepsilon \in]0, 1]\}$ is bounded in L^p (for any finite p). It follows from (3.11) that the family of integrals on the right-hand side of (3.17) is also bounded in L^p. Thus,

$$\limsup_{\varepsilon \to 0} \parallel \zeta^\varepsilon \parallel \leq C \delta_N$$

and (3.16) holds.

From the representation (3.13) and bounds (3.6), (3.15) it is easy to deduce that

$$E \parallel x^{\varepsilon,u} \parallel_t^{2p} \leq C \left(1 + \int_0^t E \parallel x^{\varepsilon,u} \parallel_s^{2p} ds\right),$$

and the standard application of the Gronwall–Bellman lemma gives (3.2).

Put $\bar{\Delta}_t^{x,\varepsilon,u} := x_t^{\varepsilon,u} - \bar{x}_t^u$. The relations (3.1), (3.13) imply that

$$\bar{\Delta}_t^{x,\varepsilon,u} = \int_0^t A_0(s)\bar{\Delta}_t^{x,\varepsilon,u} ds + R_t^{\varepsilon,u}, \tag{3.18}$$

where

$$\begin{aligned} R_t^{\varepsilon,u} &:= \int_0^t A_2(s) \left[\frac{1}{\varepsilon}\int_0^s \Psi^\varepsilon(s,r)A_3(r)x_r^{\varepsilon,u} dr + A_4^{-1}(r)A_3(r)x_r^{\varepsilon,u}\right] ds \\ &+ \int_0^t A_2(s) \left[\frac{1}{\varepsilon}\int_0^s \Psi^\varepsilon(s,r)B_2(r)u_r dr + A_4^{-1}(r)B_2(r)u_r\right] ds + \zeta_t^\varepsilon. \end{aligned} \tag{3.19}$$

It follows from (3.18) that

$$E \parallel \bar{\Delta}^{x,\varepsilon,u} \parallel_t^p \leq C \left(\int_0^t E \parallel \bar{\Delta}^{x,\varepsilon,u} \parallel_s^p ds + E \parallel R^{\varepsilon,u} \parallel^p\right),$$

and by the Gronwall–Bellman lemma we have

$$E \parallel \bar{\Delta}^{x,\varepsilon,u} \parallel_t^p \leq CE \parallel R^{\varepsilon,u} \parallel^p e^{CT}.$$

Thus, to prove (3.3) we need to show that

$$\lim_{\varepsilon\to 0} \sup_{u\in\mathcal{U}} E \parallel R^{\varepsilon,u} \parallel^p = 0.$$

But this relation follows from (3.2), (3.16) and the following statement (see [15, Lemma 3.1] or [13, Lemma 3.2]).

LEMMA 3.2. *For any $\varepsilon \in]0,1]$, $\eta > 0$, and bounded measurable function h the following holds:*

$$\begin{aligned} &\left\| \int_0^\cdot A_2(s) \left[\frac{1}{\varepsilon}\int_0^s \Psi^\varepsilon(s,r)h_r dr + A_2(s)A_4^{-1}(s)h_s\right] ds \right\| \\ &\qquad \leq \parallel h \parallel T(C_1\eta + \varepsilon C_2(\eta)), \end{aligned} \tag{3.20}$$

where C_1, $C_2(\eta)$ depend on A_2 and A_4.

At last, the property (3.4) of uniform boundedness in L^p of values of the fast variables for the fixed time follows from the representation (3.8) and (3.2), (3.7), and (3.11).

PROPOSITION 3.2. *Let $(x^{\varepsilon,u}, y^{\varepsilon,u})$ be the solution of* (1.1), (1.2) *corresponding to some $u \in \mathcal{U}$, and let $\bar{x}^u$ be the solution of the reduced equation* (3.1). *Let the random variable $\bar{y}_T^{\varepsilon,u}$ be defined by*

$$\bar{y}_T^{\varepsilon,u} := -A_4^{-1}(T)A_3(T)\bar{x}_T^u + \int_0^\infty \exp\{A_4(T)r\}B_2(T)v_r^\varepsilon dr + \tilde{\xi}_T^\varepsilon, \tag{3.21}$$

where $v^\varepsilon_r := u_{T-r\varepsilon} I_{[0,T/\sqrt{\varepsilon}]}(r) + u^0 I_{]T/\sqrt{\varepsilon},\infty[}(r)$, u^0 *is an arbitrary point in* U,

$$\tilde{\xi}^\varepsilon_T := \exp\{\varepsilon^{-1}A_4(T)(T-T^\varepsilon\}\beta + \frac{1}{\sqrt{\varepsilon}}\int_{T_\varepsilon}^T \exp\{\varepsilon^{-1}A_4(T)(T-s)\}dw^y_s, \tag{3.22}$$

$T_\varepsilon := (1-\sqrt{\varepsilon})T$, β *is a Gaussian random variable with the zero mean and covariance* Ξ, *and the matrix* Ξ *is defined in* (1.9).

Then for any $p \in [1,\infty[$

$$\lim_{\varepsilon\to 0}\sup_{u\in\mathcal{U}} E|y^{\varepsilon,u}_T - \bar{y}^{\varepsilon,u}_T|^p = 0. \tag{3.23}$$

Proof. Let $\tilde{y}^{\varepsilon,u}$ be the solution of the stochastic differential equation

$$\varepsilon d\tilde{y}^{\varepsilon,u}_t = (A_3(T)\bar{x}^u_T + A_4(T)\tilde{y}^{\varepsilon,u}_t + B_2(T)u_t)dt + \sqrt{\varepsilon}dw^y_t \quad \tilde{y}^{\varepsilon,u}_0 = 0. \tag{3.24}$$

Put

$$\tilde{\Delta}^{y,\varepsilon,u}_t := y^{\varepsilon,u}_t - \tilde{y}^{\varepsilon,u}_t, \quad \widehat{x}^{\varepsilon,u}_t := x^{\varepsilon,u}_t - x^{\varepsilon,u}_T,$$

$\widehat{A}_i(t) := A_i(t) - A_i(T)$, $\quad \widehat{B}_i(t) := B_i(t) - B_i(T)$.

The process $\tilde{\Delta}^{y,\varepsilon,u}$ is the solution of the ordinary differential equation

$$d\tilde{\Delta}^{y,\varepsilon,u}_t = (A_4(T)\tilde{\Delta}^{y,\varepsilon,u}_t + \varphi^{\varepsilon,u}_t)dt, \quad \tilde{\Delta}^{y,\varepsilon,u}_0 = 0,$$

where

$$\varphi^{\varepsilon,u}_t := \widehat{A}_4(t)y^{\varepsilon,u}_t + \widehat{A}_3(t)x^{\varepsilon,u}_t + A_3(T)\widehat{x}^{\varepsilon,u}_t + A_3(T)\bar{\Delta}^{x,\varepsilon,u}_T + \widehat{B}_2(t)u_t.$$

Thus,

$$\tilde{\Delta}^{y,\varepsilon,u}_T = \frac{1}{\varepsilon}\int_0^T \exp\{\varepsilon^{-1}A_4(T)(T-s)\}\varphi^{\varepsilon,u}_s ds. \tag{3.25}$$

By virtue of (1.6) for all $t \geq 0$ we have that

$$|\exp\{\varepsilon^{-1}A_4(T)t\}| \leq Ce^{-2\kappa t/\varepsilon}. \tag{3.26}$$

Taking into account (3.2), (3.4) and the boundedness of U, we get from (3.25) that the L^p-norm of $\tilde{\Delta}^{y,\varepsilon,u}_T$ is bounded by

$$C\frac{1}{\varepsilon}\int_0^T e^{-2\kappa(T-s)/\varepsilon}(|\widehat{A}_4(s)| + |\widehat{A}_3(s)| + f^\varepsilon_s + \bar{g}^\varepsilon + |\widehat{B}_2(s)|)ds, \tag{3.27}$$

where

$$f^\varepsilon_s := \sup_{u\in\mathcal{U}}(E|x^{\varepsilon,u}_s - x^{\varepsilon,u}_T|^p)^{1/p}, \quad \bar{g}^\varepsilon := \sup_{u\in\mathcal{U}}(E|\bar{\Delta}^{x,\varepsilon,u}_T|^p)^{1/p}.$$

Let $\bar{f}_s$ be the function similar to f^ε_s but defined for $\bar{x}^u$. It follows from (3.3) that for any $\delta > 0$ we have $f^\varepsilon_s \leq \bar{f}_s + \delta$ for all sufficiently small ε. But it is clear from the equation (3.1) that $\lim_{s\to T}\bar{f}_s = 0$. Taking into account the above remarks we check easily that the expression (3.27) tends to zero as $\varepsilon \to 0$ and, hence,

$$\lim_{\varepsilon\to 0}\sup_{u\in\mathcal{U}} E|y^{\varepsilon,u}_T - \tilde{y}^{\varepsilon,u}_T|^p = 0. \tag{3.28}$$

Now we show that

$$\lim_{\varepsilon\to 0}\sup_{u\in\mathcal{U}} E|\bar{y}_T^{\varepsilon,u}-\tilde{y}_T^{\varepsilon,u}|^p=0. \tag{3.29}$$

Indeed,

$$\bar{y}_T^{\varepsilon,u}-\tilde{y}_T^{\varepsilon,u}=\left(-A_4^{-1}(T)-\frac{1}{\varepsilon}\int_0^T \exp\{\varepsilon^{-1}A_4(T)(T-s)\}ds\right)A_3(T)\bar{x}_T^u$$
$$+\int_{T/\varepsilon}^{\infty}\exp\{A_4(T)r\}B_2(T)u_r^0dr-\int_{T/\sqrt{\varepsilon}}^{T/\varepsilon}\exp\{A_4(T)r\}B_2(T)u_{T-\varepsilon r}dr$$
$$+\exp\{\varepsilon^{-1/2}A_4(T)T\}\beta-\frac{1}{\sqrt{\varepsilon}}\int_0^{T_\varepsilon}\exp\{\varepsilon^{-1}A_4(T)(T-s)\}dw_s^y.$$

Evidently, L^p-norms of all terms on the right-hand side of this identity tend to zero and the convergence of the first one is uniform in $u\in\mathcal{U}$ by virtue of (3.2) and (3.3). Thus, (3.29) holds. The relations (3.28), (3.29) imply (3.23).

Proposition 3.2 is proved.

Assume that sequence $\mathcal{L}(x_T^{\varepsilon_n,u_n},y_T^{\varepsilon_n,u_n})$ converges in $\mathbf{P}(\mathbf{R}^m)$ to some μ. Choose in the representation (3.22) the random variable β independent of W. It follows from Propositions 3.1, 3.2 that the sequence $\mathcal{L}(\bar{x}_T^{u_n},\bar{y}_T^{\varepsilon_n,u_n})$ converges to the same limit. Let us introduce the modified controls $\hat{u}_n=u_nI_{[0,T_{\varepsilon_n}]}+u^0I_{]T_{\varepsilon_n},T]}$, where u^0 is a fixed point from U. Since $\bar{x}_T^{u_n}-\bar{x}_T^{\hat{u}_n}$ tends to zero in probability, the sequence $\mathcal{L}(\bar{x}_T^{\hat{u}_n},\bar{y}_T^{\varepsilon_n,u_n})$ converges to μ and we need to check only that $\mathcal{L}(\bar{x}_T^{\hat{u}_n},\bar{y}_T^{\varepsilon_n,u_n})\in\mathcal{K}_0(T)$. To show this notice that $\bar{x}_T^{\hat{u}_n}$ is a function of the natural projection

$$i^{\varepsilon_n}:\{w_t^x,w_t^y,\ t\in[0,T]\}\mapsto(\{w_t^x,\ t\in[0,T]\},\{w_t^y,\ t\in[0,T_{\varepsilon_n}]\}).$$

As in section 2 it can be shown that the regular conditional distribution of the random variable $\xi_0^{\varepsilon_n}+I(v^{\varepsilon_n})$ for a fixed value i^{ε_n} belongs to S. Since S is a convex closed set and $\bar{x}_T^{\hat{u}_n}$ is a measurable function on i^{ε_n}, it follows from Lemma 5.6 that the regular conditional distribution of $\xi_0^{\varepsilon_n}+I(v^{\varepsilon_n})$ for a fixed value $\bar{x}_T^{\hat{u}_n}$ also belongs to S, implying the result.

4. Proof of Theorem 1.1. Part 2. Now we must show that for any measure $\mu F^{-1}\in\mathcal{K}_0(T)$ there exists a sequence $\mu_n\in\mathcal{K}_{\varepsilon_n}(T)$ which converges to μF^{-1} in $\mathbf{P}(\mathbf{R}^n)$. It is sufficient to find such a sequence for an arbitrary μF^{-1} from the set $\tilde{\mathcal{K}}_0(T)$ which is dense in $\mathcal{K}_0(T)$ in the total variation topology. The latter property holds since the attainability set $\tilde{\mathcal{K}}_0^x$ corresponding to the strong solutions of (2.1) is dense in $\mathcal{K}_0^x$ in the total variation topology. Thus, there are dense embeddings $\tilde{\mathcal{K}}_0\subseteq\mathcal{K}_0$ and $\tilde{\mathcal{K}}_0(T)\subseteq\mathcal{K}_0(T)$.

Let us fix $\delta>0$ and a measure $\mu=m(x,dy)\nu(dx)$ such that $\mu F^{-1}\mathcal{K}_0(T)$. By definition $\nu=\mathcal{L}(\bar{x}_T^u)$, where $\bar{x}^u$ is a solution of the reduced equation (2.1) corresponding to some admissible control u. Let $\nu_h:=\mathcal{L}(\bar{x}_{T-h}^u),\ \mu_h(dx,dy):=m(x,dy)\nu_h(dx),\ h\in[0,T]$. Then there exists $h_0>0$ such that

$$d(\mu F^{-1},\mu_hF^{-1})\le\delta \tag{4.1}$$

for all $h\in]0,h_0]$.

To prove (4.1) we use the following.

LEMMA 4.1. *Let $\bar{x}^u$ be the solution of* (3.1). *Then*

$$\lim_{s\to 0}\sup_{u\in\mathcal{U}} \mathrm{Var}(\mathcal{L}(\bar{x}^u_{T-s}) - \mathcal{L}(\bar{x}^u_T)) = 0. \tag{4.2}$$

Proof. For any $u \in \mathcal{U}$ let $u^r := uI_{[0,T-r]} + u^0 I_{]T-r,T]}$, where u^0 is an arbitrary point in $\mathcal{U}$. It follows from the bound for the total variation distance in terms of the Hellinger process h_t (see [12, Theorems 2.2 and 5.1]) that

$$\mathrm{Var}(\mathcal{L}(\bar{x}^u) - \mathcal{L}(\bar{x}^{u^r})) \le Cr^{1/2}. \tag{4.3}$$

(Notice that in the considered situation the Hellinger process for the pair $(\mathcal{L}(\bar{x}^u), \mathcal{L}(\bar{x}^{u^r}))$ has the form

$$h_t = \int_0^t I_{[r,T]}(\tau)|B_0(\tau)(\widehat{u}_\tau - u^0)|^2 d\tau,$$

where $\widehat{u}_s$ takes values in U.)

Fix $\gamma > 0$ and $r > 0$ such that $Cr^{1/2} \le \gamma$. For any $s \in [0, r]$ we have

$$\mathcal{L}(\bar{x}^{u^r}_{T-s}) = \mathcal{L}(\bar{x}^u_{T-r}) * \mathcal{N}(a_s, K_s),$$

where $*$ denotes the convolution, $\mathcal{N}(a_s, K_s)$ is the nondegenerate Gaussian distribution with the mean

$$a_s := \int_{T-r}^{T-s} B_0(\tau)u^0 d\tau$$

and covariance

$$K_s := \int_{T-r}^{T-s} \Phi_0(T-s,\tau)\Phi_0'(T-s,\tau)d\tau,$$

and $\Phi_0(T-s,\tau)$ is the fundamental matrix corresponding to $A_0(t)$. In particular,

$$\mathcal{L}(\bar{x}^{u^r}_T) = \mathcal{L}(\bar{x}^u_{T-r}) * \mathcal{N}(a_0, K_0).$$

The well-known inequality

$$\mathrm{Var}(F * G - F * \tilde{G}) \le \mathrm{Var}(G - \tilde{G})$$

implies that

$$\mathrm{Var}(\mathcal{L}(\bar{x}^{u^r}_{T-s}) - \mathcal{L}(\bar{x}^{u^r}_T)) \le \mathrm{Var}(\mathcal{N}(a_s, K_s) - \mathcal{N}(a_0, K_0)),$$

where the right-hand side tends to zero as $s \to 0$.

Thus, for sufficiently small s we have

$$\sup_{u\in\mathcal{U}} \mathrm{Var}(\mathcal{L}(\bar{x}^{u^r}_{T-s}) - \mathcal{L}(\bar{x}^{u^r}_T)) \le \gamma. \tag{4.4}$$

It follows from (4.3) and (4.4) that

$$\sup_{u\in\mathcal{U}} \mathrm{Var}(\mathcal{L}(\bar{x}^u_{T-s}) - \mathcal{L}(\bar{x}^u_T)) \le 3\gamma$$

and the lemma is proved.

Since

$$\mathrm{Var}(\mu F^{-1} - \mu_h F^{-1}) = \mathrm{Var}(\mu - \mu_h) = \mathrm{Var}(\nu - \nu_h) \to 0$$

by virtue of the above lemma, the relation (4.1) holds.

Furthermore, there exists $h_1 > 0$

$$\sup_{\varepsilon} \sup_{z \in \mathcal{U}_h(u)} d(\mathcal{L}(x^{\varepsilon,z}_{T-h}, y^{\varepsilon,z}_T), \mathcal{L}(x^{\varepsilon,z}_T, y^{\varepsilon,z}_T)) \le \delta, \tag{4.5}$$

where $\mathcal{U}_h(u)$ is the set consisting of all $z \in \mathcal{U}$ such that

$$zI_{[0,T-h]} = uI_{[0,T-h]}. \tag{4.6}$$

The relation (4.5) is an evident corollary of Proposition 3.1 and the following.

LEMMA 4.2. *Let $(\xi^{(i)}_{\iota,h})$, $\iota \in I(h)$, $h \in [0,T]$, $i = 1,\ 2$, be two families of random variables with values in $\mathbf{R}^m$ such that*

$$\sup_h \sup_{\iota \in I(h)} E|\xi^{(i)}_{\iota,h}|^p < \infty, \quad i = 1,\ 2,$$

$$\lim_{h \to 0} \sup_{\iota \in I(h)} E|\xi^{(1)}_{\iota,h} - \xi^{(2)}_{\iota,h}|^p = 0$$

for some $p > 0$. Then for any bounded continuous function f on $\mathbf{R}^m$

$$\lim_{h \to 0} \sup_{\iota \in I(h)} |Ef(\xi^{(1)}_{\iota,h}) - f(\xi^{(2)}_{\iota,h})| = 0.$$

The proof of Lemma 4.2 is easy and is omitted.

Lemma 4.2 implies also the existence of $h_2 > 0$ such that

$$\sup_{\iota} d(\mathcal{L}(\bar{x}^u_{T-h}, -A_4(T)A_3(T)\bar{x}^u_{T-h} + \eta_\iota), \mathcal{L}(\bar{x}^u_{T-h}, -A_4(T)A_3(T)\bar{x}^u_T + \eta_\iota)) \le \delta, \tag{4.7}$$

where the family (η_ι) consists of all random variables with distribution from S_Y.

Let us consider some $h \le h_0 \wedge h_1 \wedge h_2$. The desired result will be proved if we find for any sufficiently small ε an admissible control $z = z^\varepsilon$ satisfying (4.6) such that

$$d(\mathcal{L}(x^{\varepsilon,z}_{T-h}, y^{\varepsilon,z}_T), \mu_h F^{-1}) \le 2\delta. \tag{4.8}$$

Indeed, it follows from (4.1), (4.5), and (4.8) that

$$d(\mathcal{L}(x^{\varepsilon,z}_T, y^{\varepsilon,z}_T), \mu_h F^{-1}) \le 4\delta,$$

and this means that any point in $\mathcal{K}_0(T)$ can be approximated by points from $\mathcal{K}_\varepsilon(T)$.

Let $(\Omega, \mathcal{F}, P)$ be a probability space with a countably generated σ-algebra. Assume that on this space we have independent random elements $\zeta,\ \alpha,\ \xi$, where ζ has the distribution ν_h, i.e., the same distribution as $\bar{x}^u_{T-h}$; α has the standard normal distribution; ξ is a stationary Gaussian Markov process with zero mean and covariance function given by (1.8), (1.9). Let us consider the set $\mathcal{V}^\alpha_U$ of all U-valued processes which are predictable with respect to the filtration generated by $\xi_{1/t}$ and α (we denote by $\mathcal{P}$ the corresponding predictable σ-algebra in $\Omega \times \mathbf{R}_+$).

LEMMA 4.3. *There is a function $v : \Omega \times \mathbf{R}_+ \times \mathbf{R}^m \to U$ which is measurable with respect to $\mathcal{P} \otimes \mathcal{B}(\mathbf{R}^m)$ such that $v(.,x) \in \mathcal{V}^\alpha$ for all $x \in \mathbf{R}^m$ and $\mathcal{L}(\xi_0 + I(v(.,x)))$ is equal to $\mu(x, dy)$ for ν_h almost all $x \in \mathbf{R}^m$.*

Proof. Evidently, $v \mapsto \mathcal{L}(\xi_0 + I(v))$ is a continuous, hence measurable, mapping from the space $\mathbf{V} := L^1(\Omega \times \mathbf{R}_+, \mathcal{P}, \rho)^d$ into $\mathbf{P}(\mathbf{R}^n)$, where $\rho(d\omega, dt) = e^{-2\kappa t} P(d\omega)dt$. Thus, the multivalued mapping

$$\Gamma : x \mapsto \{v \in \mathbf{V} : \ v(\omega, t) \in U \ \rho \text{ a.e.}, \ \mathcal{L}(\xi_0 + I(v)) = \mu(x, .)\}$$

has a measurable graph. Hence, it admits a measurable selector $x \mapsto V(x)$. Notice that $V(x)$ as an element of $\mathbf{V}$ is a class of ρ-equivalent functions. To choose from $V(x)$ a representative in a measurable way we proceed as follows. Let (v^i) be a sequence of elements from $\mathcal{V}_U^\alpha$ which is dense in $\mathcal{V}_U^\alpha \cap \mathbf{V}$, $j(x, l) := \min\{i : \| v(x) - v^i \| \leq 1/l\}$. Then $v^{j(l)} = v^{j(x,l)}(\omega, t)$ is a $\mathcal{P} \otimes \mathcal{B}(\mathbf{R}^m)$-measurable function with values in U. The sequence $v^{j(x,l)}$ converges to $V(x)$ in $\mathbf{V}$. Since U is bounded, the sequence $v^{j(l)}$ converges to V in $L^1(\Omega \times \mathbf{R}_+ \times \mathbf{R}^m, \mathcal{P} \otimes \mathcal{B}(\mathbf{R}^m), \rho \times \nu_h)^d$. Hence, there exists a subsequence which converges $\rho \times \nu_h$ a.e. to some $\mathcal{P} \otimes \mathcal{B}(\mathbf{R}^m)$-measurable function $v = v(\omega, t, x)$. For ν_h almost all x we have the inclusion $v(., x) \in V(x)$ implying that $\mathcal{L}(\xi_0 + I(v(., x))) = \mu(x, dy)$ for such x.

It follows from the above lemma that the measure μ_h is the distribution of the random variable $(\zeta, \xi_0 + I(v(., \zeta)))$, i.e.,

$$\mu_h = \mathcal{L}(\zeta, \xi_0 + I(v(., \zeta))). \tag{4.9}$$

Generalizing the arguments of section 2 we introduce a set $\mathcal{V}_U^{(\alpha,\zeta)\prime}$ consisting of all functions

$$v(s, x) = \sum_{i=1}^{N} \varphi_i(x) I_{]s_i, s_{i+1}]}(s) + u^0 I_{]s_{N+1}, \infty[}(s), \tag{4.10}$$

where $0 = s_1 < \cdots < s_{N+1}$, $u^0 \in U$, and $\varphi_i(x)$ have the form

$$\varphi_i(x) = f_i(\alpha, \xi(r_1^i), \ldots, \xi(r_{M_i}^i), x), \quad s_{i+1} < r_j^i \leq s_N, \tag{4.11}$$

and the functions f_i are measurable with respect to their arguments and take values in U.

Assume that the representation (4.9) holds with $v \in \mathcal{V}_U^{(\alpha,\zeta)\prime}$. There is a freedom in the choice of ζ, α, and ξ which we use in the following constructions.

Put $T_\varepsilon^k := T(1 - k\varepsilon^{1/2})$, $k = 1, 2, 3$, $\zeta := \bar{x}_{T-h}^u$.

Define

$$\alpha^\varepsilon := (w_{T_\varepsilon^2}^{y,1} - w_{T_\varepsilon^3}^{y,1})/(T_\varepsilon^2 - T_\varepsilon^3)^{1/2},$$

where $w^{y,1}$ is the first component of the vector process w^y,

$$\beta^\varepsilon := \Xi^{1/2}(w_{T_\varepsilon^1}^y - w_{T_\varepsilon^2}^y)/(T_\varepsilon^1 - T_\varepsilon^2)^{1/2}.$$

Let us consider on $[T_\varepsilon^1, T]$ the linear stochastic differential equation

$$\varepsilon d\tilde{\xi}_t^\varepsilon = A_4(T)\tilde{\xi}_t^\varepsilon dt + \varepsilon^{1/2} dw_t^y, \quad \tilde{\xi}_{T_\varepsilon^1}^\varepsilon = \beta^\varepsilon.$$

Put $\xi_t^\varepsilon := \tilde{\xi}_{T-\varepsilon t}^\varepsilon$, $t \in [0, \varepsilon^{-1/2}T]$. For sufficiently small ε we define the admissible control

$$z^\varepsilon := uI_{[0,t_{N+1}[} + \sum_{i=1}^{N+1} \varphi_i^\varepsilon(\bar{x}_{T-h}^u) I_{[t_{i+1}, t_i[},$$

where $t_i := T - \varepsilon s_i$, $i \leq N + 1$, and φ_i^ε is constructed in accordance with (4.11).

It follows from Propositions 3.1 and 3.2 that

$$(x^{\varepsilon,z^\varepsilon}_{T-h}, y^{\varepsilon,z^\varepsilon}_T) - (\bar{x}^u_{T-h}, -A_4(T)A_3(T)\bar{x}^u_T + \xi^\varepsilon_0 + I(v(.,\bar{x}^u_{T-h}))) \to 0$$

in probability as $\varepsilon \to 0$. Thus,

$$d(\mathcal{L}(x^{\varepsilon,z^\varepsilon}_{T-h}, y^{\varepsilon,z^\varepsilon}_T), \mathcal{L}(\bar{x}^u_{T-h}, -A_4(T)A_3(T)\bar{x}^u_T + \xi^\varepsilon_0 + I(v(.,\bar{x}^u_{T-h})))) \le \delta \tag{4.12}$$

for all sufficiently small ε. Taking into account (4.7) we get from here the desired inequality (4.8).

Part 2 of Theorem 1.1 is proved now for the case when μ_h is given by (4.9) with $v \in \mathcal{V}_U^{(\alpha,\zeta)\prime}$. Since the set $\{I(v) : \ v \in \mathcal{V}_U^{(\alpha,\zeta}\prime\}$ is dense in probability in the set $\{I(v) : \ v \in \mathcal{V}_U^{\alpha,\zeta}\}$, the result holds for the general case as well.

5. On a compactness of some subsets in the space of probability measures.

5.1. Notations and preliminaries. Let X be a Polish space with the Borel σ-algebra $\mathcal{X}$ and $\mathbf{P}(X)$ be a space of all probability measures on X with the topology of weak convergence. It is well known that $\mathbf{P}(X)$ equipped by the Prohorov metric is again a Polish space. The relative compactness of a subset $A \subseteq \mathbf{P}(X)$ is equivalent to its tightness. The last means that for any $\varepsilon > 0$ there exists a compact set $K \subseteq X$ such that $m(K) \ge 1 - \varepsilon$ for all $m \in A$.

We shall use the notation $m(f) = \int_X f(x)m(dx)$. We denote by $\mathcal{L}(\xi)$ the distribution of a random variable ξ.

Let $(X, \mathcal{X})$ and $(Y, \mathcal{Y})$ be two Polish spaces. We denote by $\mathcal{M}(X, Y)$ the set of stochastic kernels from $(X, \mathcal{X})$ to $(Y, \mathcal{Y})$ that is mappings $\mu : X \times \mathcal{Y} \to ([0,1], \mathcal{B}[0,1])$ such that $x \mapsto \mu(x, \Gamma)$ is $\mathcal{X}$-measurable for any $\Gamma \in \mathcal{Y}$ and $\mu(x, .) \in \mathbf{P}(Y)$ for any $x \in X$.

It is easy to check that the mapping $\mu : X \times \mathcal{Y} \to ([0,1], \mathcal{B}[0,1])$ is in $\mathcal{M}(X, Y)$ if and only if one of the following equivalent conditions is satisfied:

(1) The mapping $x \mapsto \mu(x, .)$ is $\mathcal{X}$-measurable (i.e., $\mu(x, .)$ is a $\mathbf{P}(Y)$-valued random variable).

(2) For any $f \in C_b(Y)$ (the set of all bounded continuous functions on Y) the mapping $x \mapsto \mu(x, f)$ is $\mathcal{X}$-measurable (i.e., $\mu(x, f)$ is a real-valued random variable).

The Skorohod Representation Theorem. *Let Y be a Polish space and $m_n \in \mathbf{P}(Y)$ be a sequence converging in $\mathbf{P}(Y)$ to some m. Then on the probability space $([0,1], \mathcal{B}[0,1], dx)$ there exist Y-valued random variables $\tilde{\xi}_n$ and $\tilde{\xi}$ such that $\mathcal{L}(\tilde{\xi}_n) = m_n$, $\mathcal{L}(\tilde{\xi}) = m$, and $\tilde{\xi}_n \to \tilde{\xi}$ pointwise.*

The Measurable Isomorphism Theorem. *Let $(X, \mathcal{X}$ be an uncountable Polish space. Then there is a one-to-one mapping $i : X \to [0,1]$ such that $i(\Gamma) \in \mathcal{B}[0,1]$ for any $\Gamma \in \mathcal{X}$ and $i^{-1}(A) \in \mathcal{X}$ for any $A \in \mathcal{B}[0,1]$.*

Another useful result is that any Polish space X is homeomorphic to a G_δ-subset of the Hilbert cube $[0,1]^{\mathbf{N}}$. For further information see, e.g., [6], [9].

5.2. For $\mu \in \mathcal{M}(X, Y)$, $m \in \mathbf{P}(X)$, and $\Gamma \in \mathcal{Y}$, the integral $\int_X \mu(x, \Gamma)m(dx)$ defines a probability measure on $(Y, \mathcal{Y})$ which we shall denote by $\int_X \mu(x, .)m(dx)$.

Lemma 5.1. *Let $(X, \mathcal{X})$ be a Polish space with nonatomic measure ν on it, let S be a compact set in $\mathbf{P}(Y)$, and let $\mathcal{M})$ be the set consisting of all stochastic kernels μ from $(X, \mathcal{X}$ to $(Y, \mathcal{Y})$ such that $\mu(x, .) \in S$ for all $x \in X$. Then the set*

$$K = \left\{ m \in \mathbf{P}(Y) : \ m(.) = \int_X \mu(x, .)\nu(dx), \ \mu \in \mathcal{M} \right\}$$

is a convex compact subset in $\mathbf{P}(Y)$ coinciding with $\overline{\text{conv}}S$.

Proof. By virtue of the measurable isomorphism theorem we can consider only the case when $(X, \mathcal{X}) = ([0,1], \mathcal{B}[0,1])$. Assume at first that $\nu(dx) = dx$, i.e., ν is the Lebesgue measure. Convexity of $\mathcal{M}$ is clear: if measures $m_i(.) = \int_X \mu_i(x,.)dx,\ i = 1,\ 2$, belong to K, $\alpha > 0,\ \beta > 0,\ \alpha + \beta = 1$, then the measure $\alpha m_1(.) + \beta m_2(.) = \int_X \mu(x,.)dx$ with

$$\mu(x,.) = I_{[0,\alpha]}(x)\mu_1(\alpha^{-1}x,.) + I_{]1-\beta,1]}(x)m_2(\beta^{-1}(x-1+\beta),.)$$

also belonging to K. The tightness of K follows easily from the tightness of S. To prove that K is closed, let us consider the sequence $m_n(.) = \int \mu_n(x,.)dx \in K$ converging to some $m(.)$ in $\mathbf{P}(Y)$. Notice that elements of $\mathcal{M}$ are random variables with values in the compact subset S of a Polish space. Thus, the set of distributions of these random variables $\{\mathcal{L}(\mu) :\ \mu \in \mathcal{M}\}$ is relatively compact in $\mathbf{P}(\mathbf{P}(Y))$. Taking, if necessary, a subsequence we can assume that $\mathcal{L}(\mu_n)$ tend to some $\mathcal{L}$ in $\mathbf{P}(\mathbf{P}(Y))$. By the Skorohod representation theorem on the probability space $([0,1], \mathcal{B}[0,1], dx)$ there exist S-valued random variables $\tilde{\mu}_n$ and $\tilde{\mu}$ such that $\tilde{\mu}_n(x,.) \to \tilde{\mu}(x,.)$ for all x when $n \to \infty$ and $\mathcal{L}(\tilde{\mu}) = m,\ \mathcal{L}(\tilde{\mu}_n) = \mathcal{L}(\mu_n)$ for all n.

The last equality means that for any $f \in C_b(Y)$ the distribution of the random variable $\tilde{\mu}_n(f)$ coincides with the distribution of $\mu_n(f)$. It follows that for any $f \in C_b(Y)$

$$m(f) = \lim_{n\to\infty} m_n(f) = \lim_{n\to\infty} \int \mu_n(x,f)dx = \lim_{n\to\infty} \int \tilde{\mu}_n(x,f)dx = \int \tilde{\mu}(x,f)dx.$$

Thus, $m(.) = \int \tilde{\mu}(x,.)dx \in K$.

The general case when ν is any nonatomic measure on $[0,1], \mathcal{B}[0,1]$ is easily reduced to the considered one by the quantile transformation. Indeed, let $F(t) := \nu([0,t],\ C(t) := \inf\{s :\ F(s) > t\}$. Then we have the identities

$$\int \mu(x,.)dx = \int \mu(F(x),.)\nu(dx), \qquad \int \mu(x,.)\nu(dx) = \int \mu(C(x),.)dx$$

which show that K does not depend on the measure ν.

Evidently, $S \subseteq K$. Hence, $\overline{\text{conv}}\, S \subseteq K$. Let $m_0(.) = \int \mu(t,.)dt$ be a point in K which does not belong to $\overline{\text{conv}}\, S$. By the separation theorem a convex compact set and a point outside it can be strictly separated by a continuous linear functional. This means that there exists $f \in C_b(Y)$ such that $\inf_{m \in \overline{\text{conv}}\, S} m(f) < m_0(f)$. It follows that $\int \mu(t,f)dt < m_0(f)$ in contradiction with the assumption that $m_0 \in K$.

Remark 5.1. If ν has atoms, then we can assert only that K is a subset of $\overline{\text{conv}}\, S$, even when S is compact.

5.3. Convergence of measure-valued martingales.

PROPOSITION 5.1. *Let $(\Omega, \mathcal{F}, P)$ be a probability space with an increasing family of σ-algebras $(\mathcal{F}_n)$ such that $\mathcal{F} = \sigma\{\mathcal{F}_n,\ n \in \mathbf{N}\}$. Let $\mu_n(\omega,.)$ be a stochastic kernel from $(\Omega, \mathcal{F}_n)$ to $(Y, \mathcal{Y})$ such that for any $f \in C_b(Y)$ the sequence $(\mu_n(f), \mathcal{F}_n)$ is a martingale. Assume that for almost all ω the sequence $\mu_n(\omega,.)$ is tight. Then for almost all ω there exists a limit $\mu(.)$ of $\mu_n(\omega,.)$ in $\mathbf{P}(Y)$ and $E(\mu(f) \mid \mathcal{F}_n) = \mu_n(f)$ for all $f \in C_b(Y)$ and $n \in \mathbf{N}$.*

Proof. To clarify ideas we start from the case when $Y = \mathbf{R}$. Let $M_n(\omega, y) = \mu_n(\omega,]-\infty, y])$ be the distribution function of $\mu_n(\omega,.)$. Evidently, $(M_n(y), \mathcal{F}_n)$ is a bounded martingale for all $y \in \mathbf{R}$ and by the Doob theorem it converges almost surely (a.s.) to $M^0(y)$. There is a set Ω_1 with $P(\Omega_1) = 1$ such that for all $\omega \in \Omega_1$

and all rationals r we have convergence of $M_n(\omega, r)$ to $M^0(\omega, r)$. Put $M(\omega, y) = \inf\{M^0(\omega, r) :\ r \in \mathbf{Q},\ r > y\}$ for $\omega \in \Omega_1$. Let $M(\omega, .)$ be equal to any distribution function outside Ω_1. The assumption on tightness implies that $M(\omega, .)$ is a probability distribution function and for any $\omega \in \Omega_1$ we have that $M_n(\omega, y)$ tends to $M(\omega, y)$ at any point y where the function $M(\omega, .)$ is continuous.

As any Polish space is homeomorphic to a G_δ-subset of $H = [0,1]^{\mathbf{N}}$ we can assume in general case that Y is the intersection of open subsets G_n in H. The closure $\bar{Y}$ of Y is a compact subset of H. Thus, $C_b(\bar{Y})$ is separable. Let A be a countable dense subset of $C_b(\bar{Y})$ closed under finite sums and multiplication by rationals. For any $f \in A$ the sequence $\mu_n(\omega, f)$ converges to some $\mu_f(\omega)$ for all ω from a set Ω_f with $P(\Omega_f) = 1$. It is possible to find a set Ω_1 with $P(\Omega_1) = 1$ such that for all $\omega \in \Omega_1,\ f,\ g \in A$, and rational a and b

$$\mu_{af+bg}(\omega) = a\mu_f(\omega) + b\mu_g(\omega).$$

Evidently,

$$\mid \mu_f(\omega) - \mu_g(\omega) \mid \leq \parallel f - g \parallel,\ \omega \in \Omega_1,$$

where $\parallel . \parallel$ is a uniform norm in $C_b(\bar{Y})$, and the function $f \mapsto \mu_f(\omega)$ can be extended uniquely to the continuous positive linear functional on $C_b(\bar{Y})$ which by the Riesz theorem has the form $\mu_f(\omega) = \mu(\omega, f)$ for some measure $\mu(\omega, .)$ on $\bar{Y}$. For $\omega \in \Omega_1$ we put $\mu(\omega, .)$ equal to any fixed probability measure on Y. We show that μ is the kernel we are seeking. Notice that $\mu(\omega, Y) = 1$. Fix $\omega \in \Omega_1$. By the assumption there exists a subsequence $\mu_{n'}(\omega, .)$ which converges in $\mathbf{P}(Y)$ to a measure $\mu'(\omega, .)$ on Y. We can extend $\mu_{n'}(\omega, .)$ and $\mu'(\omega, .)$ to $\bar{Y}$ in a trivial way. Then for $f \in A$ we have

$$\int_{\bar{Y}} f(y)\mu'(\omega, dy) = \int_Y f(y)\mu'(\omega, dy) = \lim_{n\to\infty} \int_Y f(y)\mu_{n'}(\omega, dy)$$
$$= \lim_{n\to\infty} \int_{\bar{Y}} f(y)\mu_{n'}(\omega, dy) = \int_{\bar{Y}} f(y)\mu(\omega, dy).$$

It follows that the probability measures $\mu'(\omega, .)$ and $\mu(\omega, .)$ coincide, and, since any convergent subsequence has the same limit, the whole sequence $\mu_n(\omega, .)$ converges in $\mathbf{P}(Y)$ to $\mu_n(\omega, .)$.

The result is proved.

5.4. Let X and Y be Polish spaces. Any measure $m \in \mathbf{P}(X \times Y)$ can be desintegrated, that is, can be represented as $m(dx, dy) = \mu(x, dy)\nu(dx)$, where ν is the image of m under the projection mapping $X \times Y$ onto X and μ is an element of $\mathcal{M}(X, Y)$ (regular conditional probability) defined ν a.s. uniquely.

LEMMA 5.2. *Let S_Y be a convex compact subset in $\mathbf{P}(Y)$, and let S be the set of all $m \in \mathbf{P}([0,1] \times Y)$ such that $m(dx, dy) = \mu(x, dy)dx$ with $\mu(x, .) \in S_Y$ for all $t \in [0,1]$. Then S is a convex compact set.*

Proof. The problem is to prove that S is closed. Let us consider for any $\Delta = [a, b] \subseteq [0,1],\ b > a$, the set

$$K_\Delta = \left\{ m \in \mathbf{P}(Y) :\ m(.) = \frac{1}{b-a} \int_\Delta \mu(x, .)dx,\ \mu(x, .) \in S_Y \text{ for all } x \in \Delta \right\},$$

which is, by Lemma 5.1, a convex compact set in $\mathbf{P}(Y)$. Let L be the set of all $m \in \mathbf{P}([0,1] \times Y)$ such that the image of m under the projection mapping $X \times Y$

onto X is the Lebesgue measure (this means that $m(dx,dy)=\mu(x,dy)dx$ without any restriction on μ). Evidently, L is a closed convex set in $\mathbf{P}([0,1]\times Y)$.

Define the continuous affine mapping $f_\Delta : L \to \mathbf{P}(Y)$ by the formula $f_\Delta : m \mapsto m_\Delta$ where $m_\Delta(\Gamma) = m(\Delta\times\Gamma)/(b-a)$. The result will be proved if we show that $S = \cap_\Delta f_\Delta^{-1}(K_\Delta)$. The inclusion $S \subseteq \cap_\Delta f_\Delta^{-1}(K_\Delta)$ is evident. To prove the opposite inclusion let us consider the measure m from L which belongs to $\cap_\Delta f_\Delta^{-1}(K_\Delta)$. Let us define the dyadic σ-algebras $\mathcal{F}_l = \sigma\{\Delta_{k,l},\ k=1,\dots,\ 2^l\}$, where $\Delta_{0,l}=[0,2^{-l}]$, $\Delta_{k,l}=](k-1)2^{-l},k2^{-l}]$, $k\ge 1$. Using Lemma 5.1 it is easy to show that for any l there exists a stochastic kernel μ_l such that $\mu_l(x,.)\in S_Y$ for all $t\in[0,1]$ and

$$m(A\times .) = \int_A \mu_l(x,.)dx$$

for all $A\in\mathcal{F}_l$. Put

$$m_l(t,.) = \sum_{k=1}^{2^l} I_{\Delta_{k,l}}(t) m_{l,k}(.)$$

where

$$m_{l,k}(.) = 2^l \int_{\Delta_l} \mu_l(x,.)dx \in S$$

according to Lemma 5.1. By Proposition 5.1 on convergence of measure-valued martingales, the sequence $\mu_l(x,.)$ tends to $\mu(x,.)$ in $\mathbf{P}(Y)$ for almost all x and

$$\int_A \mu_l(x,.)dx = \int_A \mu(x,.)dx$$

for all $A\in\mathcal{F}_l$. Thus, we find a stochastic kernel μ such that $\mu(x,.)\in S_Y$ for all $x\in[0,1]$ and $m(A\times\Gamma)=\int_A\mu(x,\Gamma)dx$ for all $A\in\mathcal{B}_l,\ l\in\mathbf{N}$, and $\Gamma\in\mathcal{Y}$. It follows that $m(dx,dy)=\mu(x,dy)dt$. Hence, $m\in S$ and the lemma is proved.

5.5.

LEMMA 5.3. *Let $(X,\mathcal{X})$ be any uncountable Polish space with a probability measure ν on it. Then there exists an increasing family of σ-algebras $(\mathcal{X}_l)$, $l\in\mathbf{N}$, such that*

(1) *$\mathcal{X}_l$ is generated by a finite partition of X to the sets $A_{k,l}$, $k=1,\dots,\ r_l$;*

(2) *$\mathcal{X}=\sigma\{\mathcal{X}_l,\ l\in\mathbf{N}\}$;*

(3) *$\nu(\partial A_{k,l})=0$ for any k and l (∂A denotes the boundary of A).*

Proof. Since a Polish space is homeomorphic to G_δ-subsets of $H=[0,1]^{\mathbf{N}}$, we can assume without loss of generality that X is a Borel subset of H. Moreover, it is sufficient to construct the family $(\mathcal{X}_l)$ for the space H (then the σ-algebras $\mathcal{X}_l\cap X=\{A\cap X,\ X\in\mathcal{X}_l\}$ will have the desired properties for X). Let $\varepsilon\in[0,1/2[$. Let us define the partitions of the interval $[0,1]$ by points $a^\varepsilon_{k2^{-l}}$, $k=0,\ \dots,\ 2^l$, in the following recurrent way. Let $a^\varepsilon_0=0$, $a^\varepsilon_1=1$, $a^\varepsilon_{2^{-l}}=2^{-1}+\varepsilon$. Starting from the lth partition we define for k even the point $a^\varepsilon_{k2^{-l-1}}=(a^\varepsilon_{k2^{-l}}+a^\varepsilon_{(k+1)2^{-l}})/2$; i.e., we construct the ordinary dyadic partitions on both intervals $[0,2^{-1}+\varepsilon]$ and $]2^{-1}+\varepsilon,1]$.

Evidently, diameters of the partitions tend to zero as $l\to\infty$.

Put

$$\Delta^\varepsilon_{1,l}=[0,a^\varepsilon_{2^{-l}}],\ \Delta^\varepsilon_{k,l}=]a^\varepsilon_{(k-1)2^{-l}},a^\varepsilon_{k2^{-l}}],\ k=1,\dots,\ 2^l,$$
$$\Gamma^\varepsilon=\{a^\varepsilon_{k2^{-l}},\ k=1,\ \dots,\ 2^l,\ l\in\mathbf{N}\}.$$

Let $\Delta^{\varepsilon}_{k_1,\dots,k_l,l} = \{x : x_1 \in \Delta^{\varepsilon}_{k_l,l}, \dots, x_l \in \Delta^{\varepsilon}_{k_1,1}\}$, $\mathcal{X}^{\varepsilon}_l = \sigma\{\Delta^{\varepsilon}_{k_1,\dots k_l,l},\ k_i \le 2^l\}$. Notice that the set N_d of superscripts $\varepsilon \in [0, 1/2[$ such that Γ^{ε} are disjoint is uncountable (this follows from the observation that $\Gamma^{\varepsilon} \cap \Gamma^{\eta} = \oslash$ if $\mathbf{Q}\varepsilon + \mathbf{Q} \ne \mathbf{Q}\eta + \mathbf{Q}$ and there are uncountably many different sets $\mathbf{Q}\varepsilon + \mathbf{Q}$). Let's consider the countable subset N_p of N_d containing all superscripts ε such that at least one of the probabilities $\nu(x :\ x_k \in \Gamma^{\varepsilon}),\ k \in \mathbf{N}$, is positive. Thus, $N_d \setminus N_p$ is uncountable. It is clear that for any $\varepsilon \in N_d \setminus N_p$ the sequence of σ-algebras $\mathcal{X}^{\varepsilon}_l$ has the needed properties.

5.6. The following assertion is a generalization of Lemma 5.2.

PROPOSITION 5.2. *Let S_X be a compact subset in $\mathbf{P}(X)$, and let S_Y be a convex compact subset in $\mathbf{P}(Y)$. Assume that all elements of S_X are nonatomic. Let S be the set of all $m \in \mathbf{P}(X \times Y)$ such that $m(dx, dy) = \mu(x, dy)\nu(dx)$ with $\mu(x, .) \in S_Y$ for all x and $\nu(.) \in S_X$. Then S is a compact set.*

Proof. Since the relative compactness is evident, we need to show only that S is closed. Let us consider the sequence $m_n \in S$ with $m_n(dx, dy) = \mu_n(x, dy)\nu_n(dx)$ which tends in $\mathbf{P}(X \times Y)$ to $m(dx, dy) = \mu(x, dy)\nu(dx)$. As ν_n tends to ν in $\mathbf{P}(X)$ and S_X is a compact, $\nu \in S$.

To prove that $m \in S$ for all x, we construct a sequence of stochastic kernels $\tilde{\mu}_l$ such that $\tilde{\mu}_l(x, .) \in S_Y$ for any x, $\tilde{\mu}_l(x, .)$ converges ν-a.s. to some $\tilde{\mu}(x, .)$, and $\tilde{\mu}(x, dy)\nu(dx) = \mu(x, dy)\nu(dx)$.

Let us consider the σ-algebras $\mathcal{X}_l = \sigma\{A_{k,l},\ k = 1, \dots,\ r_l\}$, $l \in \mathbf{N}$, defined in Lemma 5.3. Since $\nu(\partial A_{k,l}) = 0$, the sequence of measures $m_n(A_{k,l} \times .)$ converges in $\mathbf{P}(Y)$ to the measure $m(A_{k,l} \times .)$ for any set $A_{k,l}$. From Lemma 5.1 it follows that for any $l \in \mathbf{N}$ there exists a stochastic kernel μ_l such that $\mu_l(t, .) \in S_Y$ for all $t \in [0, 1]$ and

$$m(A \times .) = \int_A \mu_l(x, .)\nu(dx)$$

for all $A \in \mathcal{X}_l$. Let

$$\tilde{\mu}_l(x, .) = \sum_{k=1}^{2^l} I_{A_{k,l}}(x) m_{l,k}(.),$$

where

$$m_{l,k}(.) = \frac{1}{\nu(A_{k,l})} \int_{A_{k,l}} \mu_l(x, .)\nu(dx) \in S_Y$$

according to Lemma 5.1 (if $\nu(A_{k,l}) = 0$ we can put $m_{l,k}(.)$ to be equal to any point of S_Y). By Proposition 5.1 on the convergence of measure-valued martingales the sequence $\tilde{\mu}_l(x, .)$ tends to $\tilde{\mu}(x,)$ in $\mathbf{P}(Y)$ for almost all x and

$$\int_A \tilde{\mu}_l(x, .)\nu(dx) = \int_A \tilde{\mu}(x, .)\nu(dx)$$

for all $A \in \mathcal{X}_l$. Thus, we found a stochastic kernel μ such that $\tilde{\mu}(x, .) \in S_Y$ for all $x \in [0, 1]$ and $m(A \times \Gamma) = \int_A \tilde{\mu}(x, \Gamma)\nu(dx)$ for all $A \in \mathcal{X}_l$, $l \in \mathbf{N}$, and $\Gamma \in \mathcal{Y}$. It follows that $m(dx, dy) = \tilde{\mu}(x, dy)\nu(dx)$. Hence, $m \in S$.

Remark 5.2. Walter Schachermayer suggested the following simpler proof of the above result without the assumption that measures from S_X are nonatomic. At first, notice that $S_Y = \cup_{j=1}^{n} \Gamma_j$, where $\Gamma_j := \{\mu :\ \mu(f_j) \le \beta_j\}$, $f_j \in C_b(Y)$, $\beta_j \in \mathbf{R}$.

Indeed, from the Hahn–Banach theorem it follows that S_Y is an intersection of sets of this type. Their complements form an open covering of the open set $\mathbf{P}(Y)\setminus S_Y$. Since a Polish space is Lindelöf it contains a countable covering $\bar{\Gamma}_j,\ j\in\mathbf{N}$. Assume now that for the limiting measure $m(dx,dy)=\mu(x,dy)\nu(dx)$ there exists a set of positive ν-measure where $\mu(x,.)\notin S_Y$. The above representation for S_Y implies that there exists a set $B=\{x:\ \mu(x,f)>\beta\}$ with $\nu(B)>0$. Let $g_k\in C_b(X)$ be a sequence converging in $L^1(\nu)$ to I_B. Since $\mu_n(x,.)\in S_Y$ we have that $\mu_n(x,f)\le\beta$. Thus,

$$\lim_{k\to\infty}\lim_{n\to\infty}\int\int g_k(x)f(y)m_n(dx,dy)=\lim_{k\to\infty}\lim_{n\to\infty}\int g_k(x)\mu_n(x,f)\nu_n(dx)$$
$$\le\lim_{k\to\infty}\beta\int g_k(x)\nu(dx)=\beta\nu(B).$$

From the other side,

$$\lim_{k\to\infty}\lim_{n\to\infty}\int\int g_k(x)f(y)m_n(dx,dy)=\lim_{k\to\infty}\lim_{n\to\infty}\int\int g_k(x)f(y)m(dx,dy)$$
$$=\lim_{k\to\infty}\int g_k(x)\mu(dx,f)\nu(dx)=\int_B\mu(dx,f)\nu(dx)>\beta\nu(B),$$

and we get a contradiction to the assumption that $\mu(x,.)$ does not belong to S_Y ν-a.s.

5.7. Now we consider the following problem.

Let $(\Omega,\mathcal{F},P)$ be a probability space, $\mathcal{P}$ be a σ-algebra in the product $\Omega\times\mathbf{R}_+$ such that $\mathcal{P}\subseteq\mathcal{F}\otimes\mathcal{B}(\mathbf{R}_+)$, Γ is a measurable set-valued mapping from $(\mathbf{R}_+,\mathcal{B}(\mathbf{R}_+))$ to $\mathbf{R}^q$. Measurability means that the graph $\operatorname{Gr}\Gamma=\{(t,x):\ x\in\Gamma(t)\}$ is a $\mathcal{B}(\mathbf{R}_+)\otimes\mathcal{B}(\mathbf{R}^q)$-measurable set. We shall assume that $\Gamma(t)$ are closed sets and there exists a function $r\in L^1(\mathbf{R}_+,dt)$ such that $|\Gamma(t)|\le r_t$ for all t. Let $\mathcal{V}$ be a set of all $\mathcal{P}$-measurable functions f on $\Omega\times\mathbf{R}_+$ such that $f(\omega,t)\in\Gamma(t)$. Define the set K in $\mathbf{P}(\mathbf{R}^q)$ as

$$K:=\left\{\mathcal{L}(\phi):\ \phi=\int_0^\infty f(t)dt,\ f\in\mathcal{V}\right\}.$$

The question is if K is a compact set. We give here only a partial answer to this question imposing some specific assumption on the structure of the σ-algebra $\mathcal{P}$.

Let $w=(w_t)$ be a d-dimensional Wiener process on $(\Omega,\mathcal{F},P)$, $\mathcal{F}_t^{o,w}=\sigma\{w_s,\ s\le t\}$, $\mathcal{F}_t^w=\mathcal{F}_{t+}^{o,w}\vee\mathcal{N}$, where $\mathcal{N}$ is a family of all sets from $\mathcal{F}$ of zero probability. In other words, $\mathbf{F}^w=(\mathcal{F}_t^w)$ is the minimal filtration generated by the Wiener process and satisfying the usual assumptions.

LEMMA 5.4. *Assume that $\mathcal{P}$ is the predictable σ-algebra generated by $\mathbf{F}^w$ and $\Gamma(t)$ is a convex set for all t. Then K is a compact set.*

Proof. Since random variables ϕ are bounded by some constant, K is relatively compact and it remains to show that K is closed.

Let us consider the sequence $f^n\in\mathcal{V}$ such that the corresponding sequence of distribution $\mathcal{L}(\phi^n)$ converges in $\mathbf{P}(\mathbf{R}^q)$. Define the random processes

$$\phi_t^n=\int_0^t f^n(\omega,s)ds.$$

Using the criteria of relative compactness in $\mathbf{P}(C^{q+d}(\mathbf{R}_+))$ (the space $C^{q+d}(\mathbf{R}_+)$ is equipped with the metric $\sum_j 2^{-j}\parallel x\parallel_j(1+\parallel x\parallel_j)^{-1}$), we can assume without loss

of generality that the sequence $\mathcal{L}((\phi^n, w))$ converges to some $\mathcal{L}$ in $\mathbf{P}(C^{q+d}(\mathbf{R}_+))$. The Skorohod theorem asserts that on some probability space $(\tilde{\Omega}, \tilde{\mathcal{F}}, \tilde{P})$ (actually, on the standard unit interval) there are processes $(\tilde{\phi}^n, \tilde{w}^n),\ n \in \mathbf{N}$, and $(\tilde{\phi}, \tilde{w})$ such that $\mathcal{L}(\tilde{\phi}^n, \tilde{w}^n) = \mathcal{L}(\phi^n, w),\ \mathcal{L}(\tilde{\phi}, \tilde{w}) = \mathcal{L}$, and $(\tilde{\phi}^n, \tilde{w}^n)$ converges to $(\tilde{\phi}, \tilde{w})$ in $C^{q+d}(\mathbf{R}_+)$ pointwise.

It is easy to show that the following properties hold:

(1) The process $\tilde{\phi}^n$ is adapted with respect to $(\tilde{\mathcal{F}}_t^n)$, where $\tilde{\mathcal{F}}_t^n := \sigma\{\tilde{w}_s^n,\ s \le t\}$ and

$$\tilde{\phi}_t^n(\tilde{\omega}) = \int_0^t \tilde{f}^n(\tilde{\omega}, s)ds \tag{5.1}$$

with $\tilde{\mathcal{P}}^n$-measurable $\tilde{f}^n$ such that $\tilde{f}^n(\tilde{\omega}, s) \in \Gamma(s)$ for $(\tilde{\omega}, s)$ (where $\tilde{\mathcal{P}}^n$ is the predictable σ-algebra generated by $(\tilde{\mathcal{F}}_t^n)$).

(2) The process $\tilde{\phi}$ is adapted with respect to $(\tilde{\mathcal{F}}_t)$, where $\tilde{\mathcal{F}}_t := \sigma\{\tilde{w}_s,\ s \le t\}$ and

$$\tilde{\phi}_t(\tilde{\omega}) = \int_0^t \tilde{f}(\tilde{\omega}, s)ds \tag{5.2}$$

with $\tilde{\mathcal{P}}$-measurable $\tilde{f}$ such that $\tilde{f}(\tilde{\omega}, s) \in \Gamma(s)$ for $(\tilde{\omega}, s)$ (where $\tilde{\mathcal{P}}$ is the predictable σ-algebra generated by the minimal filtration with the usual assumptions for $\tilde{w}$).

Let us prove that $\tilde{\phi}^n$ is adapted with respect to $(\tilde{\mathcal{F}}_t^n)$. Fix $t \in \mathbf{R}_+$ and define the Wiener process $\widehat{w}_s^n = \tilde{w}_{s+t}^n - \tilde{w}_t^n,\ s \in \mathbf{R}_+$, which is independent of $\tilde{\mathcal{F}}_t^n$. It is sufficient to show that $\tilde{E}(\tilde{\phi}_t^n \mid \tilde{\mathcal{F}}_t^n) = \tilde{\phi}_t^n$ ($\tilde{P}$-a.s.) or, equivalently, that

$$\tilde{E}\tilde{E}(\tilde{\phi}_t^n \mid \tilde{\mathcal{F}}_t^n)h(\tilde{w}^n)g(\widehat{w}^n) = \tilde{E}\tilde{\phi}_t h(\tilde{w}^n)g(\widehat{w}^n)$$

for any bounded continuous functions $h : C^d[0,t] \to \mathbf{R}$ and $g : C^d(\mathbf{R}_+) \to \mathbf{R}$ (the argument of h, in fact, is the restriction of $\tilde{w}^n$ to $[0,t]$). Since $h(\tilde{w}^n)$ is $\tilde{\mathcal{F}}_t^n$-measurable, it follows from properties of the conditional expectations that the above equality holds if and only if

$$\tilde{E}\tilde{\phi}_t^n h(\tilde{w}^n)\tilde{E}g(\widehat{w}^n) = \tilde{E}\tilde{\phi}_t^n h(\tilde{w}^n)g(\widehat{w}^n). \tag{5.3}$$

But $\mathcal{L}(\tilde{\phi}^n, \tilde{w}^n) = \mathcal{L}(\phi^n, w)$, and the last identity is equivalent to the following one:

$$E\phi_t^n h(w)Eg(w) = E\phi_t^n h(w)g(w'),$$

where $w'_s = w_{s+t} - w_t,\ s \in \mathbf{R}_+$, which holds because ϕ^n is adapted with respect to $(\mathcal{F}_t^n)$.

Taking a limit in (5.3) we get that

$$\tilde{E}\tilde{\phi}_t h(\tilde{w})\tilde{E}g(\tilde{w}) = \tilde{E}\tilde{\phi}_t h(\tilde{w})g(\widehat{w}),$$

where $\widehat{w}_s = \tilde{w}_{s+t} - \tilde{w}_t,\ s \in \mathbf{R}_+$. As above, this means that $\tilde{\phi}_t = \tilde{E}(\tilde{\phi}_t \mid \tilde{\mathcal{F}}_t)$; i.e., $\tilde{\phi}$ is adapted with respect to $(\tilde{\mathcal{F}}_t)$. The representation (5.1) follows from the definition of ϕ^n and coincidence of $\mathcal{L}(\tilde{\phi}^n, \tilde{w}^n)$ and $\mathcal{L}(\phi^n, w)$. To obtain the representation (5.2) we notice that by the Komloś theorem [19] for the bounded sequence $\tilde{f}^n$, there exists a subsequence (n_j) such that $(\tilde{f}^{n_1} + \cdots + \tilde{f}^{n_k})/k$ converge to some function $\tilde{f}^0$ for almost all $(\tilde{\omega}, t)$. It follows that

$$\tilde{\phi}_t(\tilde{\omega}) = \int_0^t \tilde{f}^0(\tilde{\omega}, s)ds. \tag{5.4}$$

The convexity assumption implies that $\tilde{f}^0(\tilde{\omega},s)\in\Gamma(s)$ for almost all $(\tilde{\omega},s)$, and we can assume without loss of generality that $\tilde{f}^0(\tilde{\omega},s)\in\Gamma(s)$ for all $(\tilde{\omega},s)$. This means that the trajectories of $\tilde{\phi}$ are absolutely continuous functions. Let

$$\tilde{f}'(\tilde{\omega},s)=\limsup_{m\to\infty}\sum_{i=2}^{m}I_{\Delta_i}(s)2^m(\tilde{\phi}_{t_{i-1}}(\tilde{\omega})-\tilde{\phi}_{t_{i-2}}(\tilde{\omega})),$$

where $t_i=i2^{-m}$, $\Delta_i=t_i-t_{i-1}$. Clearly, $\tilde{f}'(\tilde{\omega},s)$ is a $\tilde{\mathcal{P}}$-measurable function, and for all $\tilde{\omega}$ and almost all s it coincides with $\tilde{f}^0(\tilde{\omega},s)\in\Gamma(s)$. Thus, the following function gives the representation (5.2) with the required properties:

$$\tilde{f}(\tilde{\omega},s)=\tilde{f}'(\tilde{\omega},s)I_A+x(s)I_{\bar{A}},$$

where $A=\{(\tilde{\omega},s):\ \tilde{f}'(\tilde{\omega},s)\in\Gamma(s)\}$, $x(s)$ is any Borel function such that $x(s)\in\Gamma(s)$.

Properties (1) and (2) imply the result. Indeed, it follows from (2) and Lemma 2.1 in [13] that there exists a predictable function $a(x,s):C^d(\mathbf{R}_+)\times\mathbf{R}_+\to\mathbf{R}^q$ such that $\tilde{f}(\tilde{\omega},s)=a(\tilde{w}(\tilde{\omega}),s)$. Evidently, we can modify $a(x,s)$ in such a way that $a(x,s)\in\Gamma(s)$ for all (x,s). Let us define on the original probability space $(\Omega,\mathcal{F},P)$ the process

$$\phi_t(\omega)=\int_0^t f(\omega,s)ds$$

with $f(\omega,s)=a(w(\omega),s)$. Since $f\in\mathcal{V}$ and $\mathcal{L}(\phi)=\mathcal{L}(\tilde{\phi})=\mathcal{L}$ it follows that the limit of $\mathcal{L}(\phi^n)$ belongs to K and the lemma is proved.

5.8. Now we apply the previous result to our specific setting.

LEMMA 5.5. *The set* $S_Y^0:=\{\mathcal{L}(\xi_0+I(v):\ v\in\mathcal{V}_U)\}$ *is compact in* $\mathbf{P}(\mathbf{R}^n)$.

Proof. Reversing the time and taking into account the notations of the previous subsection we can reduce the problem to the question of whether the set

$$K:=\left\{\mathcal{L}(\phi):\ \phi=\int_0^\infty f(t)dt,\ f\in\mathcal{V}\right\}$$

is compact. Here $\Gamma(t)=-s^{-2}\exp\{A_4(T)/s\}B_2(T)U$ and the σ-algebra $\mathcal{P}$ is generated by the time reverse of the Ornstein–Uhlenbeck process $\xi_{1/t}$, or, equivalently, by the process $\eta_t:=t\xi_{1/t}$. The process η (as well as ξ) is defined in the present context only up to the distribution. For example, we can take as η the process defined by the stochastic differential equations

$$d\eta_t=t^{-2}(tI-A)\eta_t dt+dw_t,\quad \eta_0=0, \tag{5.5}$$

where I is the unit matrix and w is the Wiener process. This representation can be deduced from the differential equation for the Ornstein–Uhlenbeck process by the Ito formula. But from equation (5.5) it follows that $\mathcal{F}_t^{o,w}=\sigma\{\eta_s,s\le t\}$ and the needed result is a corollary of Lemma 5.4.

5.9. Let η_i be random variables with values in Polish spaces $(X_i,\mathcal{X}_i)$, $i=1,\ 2,\ 3$, let ν_i be the distribution of η_i, and let $\mu_{ij}(x_j,dx_i)$ be the regular conditional distribution of η_i given η_j.

LEMMA 5.6. *Let* $\eta_3=f(\eta_2)$ *for some measurable function* $f:X_2\to X_3$, *and let* S_1 *be a compact convex set in* $\mathbf{P}(X_1)$. *Assume that* $\mu_{12}(x_2,dx_1)\in S_1$ *for all* x_2. *Then* $\mu_{13}(x_3,dx_1)\in S_1$ *for* ν_3*-almost all* x_3.

Proof. The assertion follows from the relation

$$\mu_{13}(x_3, dx_1) = \int_{X_2} \mu_{12}(x_2, dx_1)\mu_{23}(x_3, dx_2) \quad (\nu_3\text{-a.e.})$$

and Remark 5.1.

Acknowledgment. The authors express their thanks to Walter Schachermayer for helpful discussions of functional-analytic aspects of the problem.

REFERENCES

[1] V. E. Beneš, *Existence of optimal control laws*, SIAM J. Control, 9 (1971), pp. 446–475.
[2] A. Bensoussan, *On some singular perturbation problems arising in stochastic control*, Stochastic Anal. Appl., 2 (1984), pp. 13–53.
[3] A. Bensoussan, *Perturbation Methods in Optimal Control*, J. Wiley/Gauthier Villars, New York, 1988.
[4] A. Bensoussan, *Optimal Control of Partially Observed Systems*, Cambridge University Press, Cambridge, UK, 1992.
[5] A. Bensoussan and G. L. Blankenship, *Singular perturbations in stochastic control*, in Singular Perturbations and Asymptotic Analysis in Control Systems, Lecture Notes in Control and Inform. Sci. 90, P. Kokotovič, A. Bensoussan, and G. Blankenship, eds. Springer-Verlag, Berlin, 1987.
[6] C. Dellacherie and P.-A. Meyer, *Probabilities and Potentials*, North-Holland, Amsterdam, 1978.
[7] A. L. Dontchev. *Perturbations, Approximations and Sensitivity Analysis of Optimal Control Systems*, Lecture Notes in Control and Inform. Sci. 52, Springer-Verlag, Berlin, 1983.
[8] A. L. Dontchev and V. M. Veliov, *Singular perturbation in Mayer's optimization problem for linear systems*, SIAM J. Control Optim., 21 (1983), pp. 566–581.
[9] S. N. Ethier and T. G. Kurtz, *Markov Processes: Characterization and Convergence*, J. Wiley, New York, 1986.
[10] V. Gaitsgory, *Suboptimization of singularly perturbed control systems*, SIAM J. Control Optim., 30 (1992), pp. 1228–1249.
[11] W. Hildenbrandt, *Core and Equilibria of a Large Economy*, Princeton University Press, Princeton, NJ, 1974.
[12] Yu. M. Kabanov, R. Sh. Liptser, and A. N. Shiryaev, *On the variation distance for the probability measures defined on a filtered space*, Probability Theory Related Fields, 71 (1986), pp. 19–35.
[13] Yu. M. Kabanov and S. M. Pergamenshchikov, *Optimal control of singularly perturbed linear stochastic systems*, Stochastics Stochastics Rep., 36 (1991), pp. 109–135.
[14] Yu. M. Kabanov, S. M. Pergamenshchikov, and J. M. Stoyanov, *Asymptotic expansions for singularly perturbed stochastic differential equations*, in New Trends in Probability and Statistics, Vol. 1., V. V. Sazonov and T. L. Shervashidze, eds., Coronet Books, Philadelphia, PA, 1991, pp. 413–435.
[15] Yu. M. Kabanov and W. Runggaldier, *On control of two-scale stochastic systems with linear dynamics in the fast variables*, Math. Control Systems Signals, to appear.
[16] I. Karatzas and S. E. Shreve, *Brownian Motion and Stochastic Calculus*, Springer-Verlag, New York, 1988.
[17] P. V. Kokotovič and H. K. Khalil, eds., *Singular Perturbations in Systems and Control*, IEEE Press, New York, 1986.
[18] P. V. Kokotovič, H. K. Khalil, and J. O'Reilly, *Singular Perturbation Methods in Control: Analysis and Design*, Academic Press, New York, 1986.
[19] J. Komlóś, *A generalization of a problem of Steinhaus*, Acta Math. Sci. (Hung.), 18 (1967), pp. 217–229.
[20] H. J. Kushner, *Weak Convergence Methods and Singularly Perturbed Stochastic Control and Filtering Problems*, Birkhäuser, Boston, 1990.

SIAM J. CONTROL OPTIM.
Vol. 35, No. 1, pp. 160–177, January 1997

008

ANALYSIS OF ROBUST $\mathbf{H}_2$ PERFORMANCE USING MULTIPLIER THEORY*

ERIC FERON†

Abstract. In this paper, the problem of determining the worst-case $\mathbf{H}_2$ performance of a control system subject to linear time-invariant uncertainties is considered. A set of upper bounds on the performance is derived, based on the theory of stability multipliers and the solution of an original optimal control problem. The numerical issues raised by the resulting computational problems are discussed; in particular, newly developed interior-point convex optimization methods, combined with linear matrix inequalities, apply very well to the fast and accurate solution of these problems. The new results compare favorably with prior ones. The method can be extended to other types of perturbations.

Key words. $\mathbf{H}_2$ performance, stability multipliers, convex optimization, linear matrix inequalities

AMS subject classifications. 93B40, 93D05, 93D09, 93D10, 93D25

PII. S0363012994266504

1. Introduction.

Among all performance indices known to control engineering, the $\mathbf{H}_2$ performance index holds a special place for historical and practical reasons. The historical reasons are that minimizing the $\mathbf{H}_2$ norm of a linear control system via feedback, better known as the LQG problem, is among the first optimal control problems to have been solved analytically. (For an extensive presentation and bibliography, see [1].) The practical reason is that this problem can be solved using reliable and fast computational procedures [2, 20, 39].

It is, however, well known that the performance of the LQG-optimal controller can be very sensitive to perturbations on the nominal system [11]. In view of this fact, devising analysis and synthesis tools that will respectively evaluate and minimize worst-case $\mathbf{H}_2$ norms of control systems is especially relevant.

In this paper, we consider the following specific problem: given a linear control system perturbed by linear time-invariant (LTI) perturbations, what is its worst-case $\mathbf{H}_2$ norm? This question has remained open until recently when some attempts have been made at its solution. Packard and Doyle [26] and Bernstein and Haddad [4, 5, 6] are among the first to consider the problem of robust $\mathbf{H}_2$ performance in the face of dynamic and parametric uncertainty. Stoorvogel [37, 38], Petersen, Rotea, and McFarlane [30, 31] find bounds on the worst-case $\mathbf{H}_2$ norm of a system subject to norm-bounded, noncausal, possibly nonlinear, and time-varying uncertainties. Peres, Geromel, and Souza [28, 29] find upper bounds on the $\mathbf{H}_2$ norm of linear time-varying and uncertain LTI systems based on quadratic Lyapunov functions. The book and papers by Boyd, El Ghaoui, Feron, and Balakrishnan [9, 14, 7, 13] show that the computation of all these bounds on $\mathbf{H}_2$ performance can be reduced to convex optimization problems involving linear matrix inequalities, which can be solved via efficient convex optimization techniques. In [9, 13], attempts are made to refine the upper bounds

*Received by the editors April 21, 1994; accepted for publication (in revised form) October 24, 1995. This work was supported by the National Science Foundation grant ECS-9409715, Army Research Office grant ARO DAAL03-92-G0115 (Center for Intelligent Control Systems), and the Charles Stark Draper Career Development Chair at MIT.

http://www.siam.org/journals/sicon/35-1/26650.html

†Department of Aeronautics and Astronautics, Massachusetts Institute of Technology, Cambridge, MA 02139 (feron@mit.edu).

on $\mathbf{H}_2$ performance when dealing with particular classes of perturbations such as static nonlinearities and parametric uncertainties, using Lur'e Lyapunov functions and causal multipliers. Other attempts at obtaining reliable upper bounds on robust $\mathbf{H}_2$ performance include the recent paper by Paganini, Doyle, and D'Andrea [27].

In this paper, we propose to extend the results presented in [9, 13] by using noncausal multipliers to evaluate the worst-case $\mathbf{H}_2$ norm of linear systems perturbed by LTI perturbations. Using noncausal multipliers is a well-known technique to determine the stability of uncertain systems (see [10, 40] and references therein) and has proved to yield effective computational procedures [36, 34]. We believe this paper is the first attempt to use them to determine robust $\mathbf{H}_2$-performance of linear systems subject to linear perturbations. It is organized as follows.

The first part is devoted to a few definitions and notations. In particular, we recall the notions of boundedness, positivity, and passivity of operators.

In the second part, we formulate the robust $\mathbf{H}_2$ analysis problem and sketch our line of attack to get upper bounds on worst-case $\mathbf{H}_2$ performance. We present a new upper bound on robust $\mathbf{H}_2$ performance, based on the use of certain dynamic Lagrange multipliers.

In the third part of this paper, we present a way to compute the upper bound on robust performance using convex optimization and linear matrix inequalities. In particular, we exhibit convenient linear families of finite-dimensional multipliers to perform this computation.

In the fourth part, we discuss the obtained results: in particular, we study conditions for the obtained upper bound to be finite. We also study special cases and show they correspond to results having already appeared in the literature. A numerical example that illustrates the usefulness of dynamic multipliers to determine accurate upper bounds on robust performance is provided.

2. Notation. In this paper, $\mathbf{R}$ (resp., $\mathbf{C}$) denotes the set of real (resp., complex) numbers. $\mathbf{R}_+$ denotes the set of nonnegative real numbers. $\mathbf{R}^{n\times p}$ (resp., $\mathbf{C}^{n\times p}$) is the vector space of $n \times p$ real (resp., complex) matrices. $\mathbf{R}^{n\times 1}$ ($\mathbf{C}^{n\times 1}$) is abbreviated $\mathbf{R}^n$ ($\mathbf{C}^n$). For the random variable x, $\mathbf{E}\, x$ denotes the expected value of x. For any matrix X, X^T denotes its transpose and X^* denotes its complex conjugate transposed. The identity matrix is noted I. If X is invertible, then its inverse is noted X^{-1}. When X is not invertible, its Moore–Penrose inverse is noted $X^\dagger$. When $X \in \mathbf{R}^{n\times n}$, $\mathbf{Tr}\, X$ denotes the trace of X. A square matrix X is said stable if all its eigenvalues lie in the open left complex half-plane. Given a set of matrices $X_1 \in \mathbf{R}^{n_1\times p_1}, \ldots, X_N \in \mathbf{R}^{n_N\times p_N}$ and defining $n = \sum_{i=1}^N n_i$, $p = \sum_{i=1}^N p_i$, $\mathbf{diag}(X_1, \ldots, X_N)$ denotes the $n \times p$ matrix

$$\begin{bmatrix} X_1 & 0 & \cdots & 0 \\ 0 & X_2 & \ddots & \vdots \\ \vdots & \ddots & \ddots & 0 \\ 0 & \cdots & 0 & X_N \end{bmatrix}.$$

Note that the X_is need not be square. From time to time, when no ambiguity is possible, $\mathbf{diag}(X_1, \ldots, X_N)$ is noted $\mathbf{diag}_{i=1}^L(X_i)$, or $\mathbf{diag}_i(X_i)$.

For any two matrices X and $Y \in \mathbf{R}^{n\times n}$, the inequality $X \le Y$ means that X and Y are symmetric and that the difference $Y - X$ is positive. The inequality $X < Y$ means that X and Y are symmetric and that the difference $Y - X$ is positive definite.

$\mathbf{L}_2(\mathbf{R})$ denotes the Hilbert space of functions h mapping $\mathbf{R}$ into $\mathbf{R}^{n\times p}$ which satisfy

$$\int_{-\infty}^{\infty} \mathbf{Tr}\, h(t)^T h(t) dt < \infty.$$

It is equipped with the standard scalar product

$$\forall g, h \in \mathbf{L}_2(\mathbf{R}), \quad \langle g, h\rangle \triangleq \int_{-\infty}^{\infty} \mathbf{Tr}\, g(t)^T h(t) dt,$$

and for all $h \in \mathbf{L}_2(\mathbf{R})$, the Euclidean norm $\langle h, h\rangle^{1/2}$ of h is denoted $\|h\|_2$. $\mathbf{L}_2(\mathbf{R}_+)$ denotes the subspace of $\mathbf{L}_2(\mathbf{R})$ made of the functions h satisfying $h(t) = 0$ when $t < 0$. $\mathbf{L}_{2e}$ denotes the space of functions h mapping $\mathbf{R}$ into $\mathbf{R}^{n\times p}$ satisfying $h(t) = 0$ for $t < 0$ and

$$\forall t \geq 0, \quad \int_{-\infty}^{t} \mathbf{Tr}\, h(t)^T h(t) dt < \infty.$$

Following the usage of Francis [16], we suppress the dependence of these spaces on the integers n and p.

For a given operator Δ and a function p mapping $\mathbf{R}$ into $\mathbf{R}^n$, $(\Delta p)(t)$ denotes the value taken by the image function Δp at time t. An operator Δ is said to be *causal* if for any function p and any time t, $(\Delta p)(t)$ depends only on the past values of p up to time t. It is said *anticausal* if $(\Delta p)(t)$ depends only on the future values of p from time t. In any other case, Δ is said *noncausal*.

Given a set of operators $\Delta_1, \ldots, \Delta_L$, $\mathbf{diag}(\Delta_1, \ldots \Delta_L)$ stands for the operator which maps the function taking the value $\begin{bmatrix} u_1(t)^T & \ldots & u_L(t)^T \end{bmatrix}^T$ at time t to the function taking the value $\begin{bmatrix} (\Delta_1 u_1)(t)^T & \ldots & (\Delta_L u_L)(t)^T \end{bmatrix}^T$ at time t.

Let H map $\mathbf{L}_2(\mathbf{R})$ into $\mathbf{L}_2(\mathbf{R})$ and be linear. The *adjoint* of H, denoted H^*, is the unique linear operator satisfying

$$\langle x, Hy\rangle = \langle H^* x, y\rangle \quad \forall x, y \in \mathbf{L}_2(\mathbf{R}).$$

Defining s as the usual Laplace variable, the transfer function of H is denoted $H(s)$ whenever it exists.

Let us now introduce the notions of finite gain, positivity, and passivity that we will use throughout this paper.

DEFINITION 2.1 (see [32]). *An operator F mapping $\mathbf{L}_2(\mathbf{R})$ into $\mathbf{L}_2(\mathbf{R})$ has finite gain (or, equivalently, is bounded) if there exists a positive δ such that for any $u \in \mathbf{L}_2$*

$$\|Fu\|_2 \leq \delta \|u\|_2 .$$

The smallest such δ is called the gain of F and denoted $\|F\|_\infty$.

DEFINITION 2.2 (see [10]). *A linear operator G mapping $\mathbf{L}_2(\mathbf{R})$ into $\mathbf{L}_2(\mathbf{R})$ is said to be positive if for any $u \in \mathbf{L}_2(\mathbf{R})$*

$$\langle Gu, u\rangle \geq 0.$$

G is said to be strictly positive if there exists $\delta > 0$ such that for any $u \in \mathbf{L}_2(\mathbf{R})$

$$\langle Gu, u\rangle \geq \delta \|u\|_2^2 .$$

DEFINITION 2.3 (see [10]). *A linear, causal operator G mapping $\mathbf{L}_{2e}$ into $\mathbf{L}_{2e}$ is said to be passive if for any $u \in \mathbf{L}_{2e}$ and $T \geq 0$*

$$\int_0^T (Gu)^T u \, dt \geq 0.$$

3. Problem statement and line of attack. Consider the system

$$\begin{aligned} \frac{d}{dt}x(t) &= Ax(t) + B_p p(t) + B_w w(t), \quad x(0) = x_0, \\ q(t) &= C_q x(t) + D_{qp} p(t), \\ z(t) &= C_z x(t), \\ p(t) &= (\Delta q)(t), \end{aligned} \tag{1}$$

where $x : \mathbf{R} \to \mathbf{R}^n$, $p : \mathbf{R} \to \mathbf{R}^{n_p}$, $q : \mathbf{R} \to \mathbf{R}^{n_p}$, $z : \mathbf{R} \to \mathbf{R}^{n_z}$, and $w : \mathbf{R} \to \mathbf{R}^{n_w}$ and all quantities are equal to 0 for $t < 0$. Assume that the matrix A is stable. Δ is a perturbation that satisfies the following set of assumptions:

$$\begin{aligned} &\Delta = \mathbf{diag}(\Delta_1, \ldots, \Delta_{n_p}), \\ &\forall u \in \mathbf{L}_2(\mathbf{R}), \quad (\Delta_i u)(t) = \int_0^\infty \delta_i(\tau) u(t-\tau) \, d\tau, \\ &\int_0^\infty |\delta_i(\tau)| \, d\tau < \infty, \\ &\Delta_i \text{ is passive}, \quad i = 1, \ldots, n_p. \end{aligned} \tag{2}$$

In the literature, the passivity assumption on Δ is often replaced by a finite-gain assumption [12]. Standard loop-transformations allow us to move almost freely from one framework to the other (see [10, p. 215] and the end of this paper for a more detailed discussion).

Much of the existing literature is devoted to studying the robust stability of the system (1) against the uncertainty Δ in the following sense: Assume $w = 0$ and any initial condition x_0; then the signals x, p, q, and z belong to $\mathbf{L}_2(\mathbf{R}_+)$.

In this paper, we assume the system (1) to be robustly stable, and we are interested in evaluating its worst-case $\mathbf{H}_2$ performance against the uncertainty Δ: Let

$$H_\Delta(s) = H_{zw}(s) + H_{zp}(s)\Delta(s)(I - H_{qp}(s)\Delta(s))^{-1} H_{qw}(s),$$

where

$$\begin{aligned} H_{zw}(s) &= C_z(sI - A)^{-1} B_w, \\ H_{zp}(s) &= C_z(sI - A)^{-1} B_p, \\ H_{qp}(s) &= C_q(sI - A)^{-1} B_p + D_{qp}, \\ H_{qw}(s) &= C_q(sI - A)^{-1} B_w. \end{aligned} \tag{3}$$

The $\mathbf{H}_2$ norm of the system (1) is defined as

$$\|H_\Delta\|_2 = \left(\frac{1}{2\pi} \int_{-\infty}^{\infty} \mathbf{Tr} \, H_\Delta(j\omega)^* H_\Delta(j\omega) \, d\omega \right)^{1/2}.$$

Equivalently, using Parseval's theorem, $\|H_\Delta\|_2$ may also be expressed as $\|h_\Delta\|_2$, where h_Δ is the impulse matrix of H_Δ. In the subsequent developments of this paper, it will also be very convenient to express it as

$$\|H_\Delta\|_2 = \left(\mathbf{E}\,\|z\|_2^2\right)^{1/2}, \tag{4}$$

where z is the output of the system (1) with the following assumptions: the input w is identically 0 and the initial condition x_0 is equal to $B_w u$, where u is a random variable satisfying $\mathbf{E}\,uu^T = I$. (The expectation appearing in (4) is therefore to be taken with respect to u.)

The robust $\mathbf{H}_2$ analysis problem is to compute the worst-case $\mathbf{H}_2$ norm of the system (1) over all possible values of Δ that satisfy (2). This computation is in general quite a complicated problem. Thus, we propose to replace it by the computation of upper bounds on robust $\mathbf{H}_2$ norm, using a technique similar to the classical technique of Lagrange multipliers: Consider any family $\mathcal{M}$ of operators M mapping $\mathbf{L}_2(\mathbf{R})$ into $\mathbf{L}_2(\mathbf{R})$ such that the operator $M^*\Delta$ is positive for any Δ satisfying (2). The following lemma gives us an *upper bound* on the worst-case $\mathbf{H}_2$ norm of the system (1).

LEMMA 3.1. *We have the inequality*

$$\max_{\Delta} \mathbf{E}\,\|z\|_2^2 \le \min_{M\in\mathcal{M}} \mathbf{E} \max_{\tilde{p}\in\mathbf{L}_2(\mathbf{R}_+)} \|\tilde{z}\|_2^2 + 2\langle \tilde{p}, M\tilde{q}\rangle, \tag{5}$$

where $\tilde{p}$, $\tilde{q}$, and $\tilde{z}$ are the inputs and outputs of the system

$$\begin{aligned} \frac{d}{dt}\tilde{x}(t) &= A\tilde{x}(t) + B_p\tilde{p}(t), \;\; \tilde{x}(0) = B_w u, \\ \tilde{q}(t) &= C_q\tilde{x}(t) + D_{qp}\tilde{p}(t), \\ \tilde{z}(t) &= C_z\tilde{x}(t), \end{aligned} \tag{6}$$

where all variables belong to $\mathbf{L}_2(\mathbf{R}_+)$, and u is a random variable satisfying $\mathbf{E}\,uu^T = I$.

Proof. Consider the system (1). For any $M \in \mathcal{M}$, we have $\langle p, Mq\rangle = \langle \Delta q, Mq\rangle = \langle M^*\Delta q, q\rangle \ge 0$, since $M^*\Delta$ is positive. Therefore, for any Δ satisfying (2), any initial condition x_0, and any $M \in \mathcal{M}$, we have $\|z\|_2^2 \le \|z\|_2^2 + 2\langle p, Mq\rangle$. Since $p \in \mathbf{L}_2(\mathbf{R}_+)$, we furthermore have

$$\|z\|_2^2 + 2\langle p, Mq\rangle \le \max_{\tilde{p}\in\mathbf{L}_2(\mathbf{R}_+)} \|\tilde{z}\|_2^2 + 2\langle \tilde{p}, M\tilde{q}\rangle,$$

where the right-hand side of the inequality may be infinite. Taking expected values (with respect to the random variable u) on both sides of these inequalities, we conclude that

$$\mathbf{E}\,\|z\|_2^2 \le \mathbf{E} \max_{\tilde{p}\in\mathbf{L}_2(\mathbf{R}_+)} \|\tilde{z}\|_2^2 + 2\langle \tilde{p}, M\tilde{q}\rangle.$$

This ends the proof of our lemma. □

Note that the multiplier M can indeed be seen as a *Lagrange multiplier*. Such an approach is not unlike the one encountered in the papers by Yakubovich [43, 15, 44] and Megretsky [21, 22, 23], where it is named the $\mathcal{S}$-procedure. In the remainder of this paper, we will show that a suitable choice of the family of multipliers $\mathcal{M}$ allows the right-hand side of (5) to be easily computed.

4. Upper bound computation via linear families of finite-dimensional multipliers.

4.1. Linear families of finite-dimensional operators. Following an idea arising in [8, 33, 9], we consider finite-dimensional, noncausal operators $M \in \mathcal{M}$, where

$$\mathcal{M} = \left\{ \begin{array}{l} M = \mathbf{diag}(M_1, \ldots, M_{n_p}), \\ M_i(s) = m_{i0} + \displaystyle\sum_{j=1}^{N} \frac{m_{ij}}{(s+1)^j} + \frac{m_{ij}}{(-s+1)^j}, \\ M_i(j\omega) \geq 0 \ \ \forall \omega \in \mathbf{R}, \\ m_{i,j} \in \mathbf{R}, \ \ 1 \leq i \leq n_p, \ \ 0 \leq j \leq N \end{array} \right\}.$$

(We refer the reader to [16] for a complete discussion of the representation of noncausal operators via transfer functions with unstable poles.) Thus, the set $\mathcal{M}$ is parameterized by the real numbers m_{ij}, $i = 1, \ldots, n_p$, $j = 0, \ldots, N$. Each transfer function $M_i(s)$ may alternatively be written as $M_i(s) = C_{M_i}(sI - A_{M_i})^{-1}B_{M_i} + D_{M_i}$, where

$$\begin{aligned} A_{Mi} &= \left[\begin{array}{cc} A_{\mathrm{ca}i} & 0 \\ 0 & -A_{\mathrm{ac}i} \end{array} \right], \quad B_{Mi} = \left[\begin{array}{c} B_{\mathrm{ca}i} \\ -B_{\mathrm{ac}i} \end{array} \right], \\ C_{Mi} &= \left[\begin{array}{cc} C_{\mathrm{ca}i} & C_{\mathrm{ac}i} \end{array} \right], \qquad D_{Mi} = m_{i0}, \end{aligned} \tag{7}$$

and

$$\begin{aligned} A_{\mathrm{ca}i} &= \left[\begin{array}{ccccc} -1 & 1 & 0 & \cdots & 0 \\ 0 & -1 & 1 & \ddots & \vdots \\ \vdots & \ddots & \ddots & \ddots & 0 \\ \vdots & & \ddots & -1 & 1 \\ 0 & \cdots & \cdots & 0 & -1 \end{array} \right], \quad A_{\mathrm{ac}i} = A_{\mathrm{ca}i}, \ \ A_{\mathrm{ca}i} \in \mathbf{R}^{N \times N}, \\ B_{\mathrm{ca}i} &= \left[\begin{array}{cccc} 0 & \cdots & 0 & 1 \end{array} \right]^T, \qquad B_{\mathrm{ac}i} = B_{\mathrm{ca}i}, \\ C_{\mathrm{ca}i} &= \left[\begin{array}{ccc} m_{iN} & \cdots & m_{i1} \end{array} \right], \qquad C_{\mathrm{ac}i} = C_{\mathrm{ca}i}, \quad i = 1, \ldots, n_p. \end{aligned} \tag{8}$$

Likewise, the transfer function $M(s)$ may also be written as $M(s) = C_M(sI - A_M)^{-1}B_M + D_M$, with

$$\begin{aligned} A_M &= \left[\begin{array}{cc} A_{\mathrm{ca}} & 0 \\ 0 & -A_{\mathrm{ac}} \end{array} \right], \quad B_M = \left[\begin{array}{c} B_{\mathrm{ca}} \\ -B_{\mathrm{ac}} \end{array} \right], \\ C_M &= \left[\begin{array}{cc} C_{\mathrm{ca}} & C_{\mathrm{ac}} \end{array} \right], \qquad D_M = \mathbf{diag}_i(m_{i0}), \end{aligned} \tag{9}$$

and

$$\begin{aligned} A_{\mathrm{ca}} &= \mathbf{diag}_i(A_{\mathrm{ca}i}), \quad A_{\mathrm{ac}} = \mathbf{diag}_i(A_{\mathrm{ac}i}), \\ B_{\mathrm{ca}} &= \mathbf{diag}_i(B_{\mathrm{ca}i}), \quad B_{\mathrm{ac}} = \mathbf{diag}_i(B_{\mathrm{ac}i}), \\ C_{\mathrm{ca}} &= \mathbf{diag}_i(C_{\mathrm{ca}i}), \quad C_{\mathrm{ac}} = \mathbf{diag}_i(C_{\mathrm{ac}i}), \\ & i = 1, \ldots, n_p. \end{aligned} \tag{10}$$

To check that $\mathcal{M}$ is indeed an admissible set, we must check that for any $M \in \mathcal{M}$ and any Δ satisfying (2), $M^*\Delta$ is positive. Since $M^*\Delta$ is a diagonal operator, we just need to check that $M_i^*\Delta_i$ is positive for $i = 1, \ldots, n_p$. From [10, p. 174], passivity of Δ_i is equivalent to the inequality

$$\Delta_i(j\omega)^* + \Delta_i(j\omega) \geq 0 \ \ \forall \omega \in \mathbf{R}. \tag{11}$$

By hypothesis, $M_i(j\omega)$ is real and nonnegative. Thus, the inequality (11) implies

$$\Delta_i(j\omega)^* M_i(j\omega) + M_i(j\omega)\Delta_i(j\omega) \geq 0. \tag{12}$$

Thus, positivity of $M_i^*\Delta_i$ holds, since by Parseval's theorem, we have

$$\begin{aligned}&\langle u, M^*\Delta_i u\rangle \\ &\quad = \frac{1}{2\pi}\int_{-\infty}^{\infty} u(j\omega)^*(\Delta_i(j\omega)^* M_i(j\omega) + M_i(j\omega)\Delta_i(j\omega))u(j\omega)d\omega \quad \forall u \in \mathbf{L}_2(\mathbf{R}).\end{aligned}$$

The inequality $M_i(j\omega) + M_i(j\omega)^* \geq 0$ for all $\omega \in \mathbf{R}$ can be expressed in a convenient form via a straightforward application of Theorems 3 and 4 of Willems [41], subsequently corrected in [42].

LEMMA 4.1. *The inequality*

$$M_i(j\omega) + M_i(j\omega)^* \geq 0 \ \ \forall \omega \in \mathbf{R}$$

is satisfied if and only if there exists a symmetric matrix P_i satisfying

$$\begin{bmatrix} A_{M_i}^T P_i + P_i A_{Mi} & P_i B_{Mi} - C_{M_i}^T \\ B_{M_i}^T P_i - C_{Mi} & -(D_{Mi} + D_{M_i}^T) \end{bmatrix} \leq 0. \tag{13}$$

Note that this lemma requires controllability of (A_{M_i}, B_{M_i}) to hold and that this assumption is indeed satisfied.

When $\tilde{q} \in \mathbf{L}_2(\mathbf{R}_+)$, a simple state-space representation of $M\tilde{q}$ can be given that will be useful in the subsequent developments in this paper.

LEMMA 4.2. *For any $\tilde{q} \in \mathbf{L}_2(\mathbf{R}_+)$, we can write $M\tilde{q}$ as the output of the system*

$$\begin{aligned}\frac{d}{dt}x_M &= A_M x_M + B_M \tilde{q}, \qquad x_M(0) = \begin{bmatrix} 0 \ x_{\mathrm{ac}0}^T \end{bmatrix}^T, \\ M\tilde{q} &= C_M x_M + D_M \tilde{q},\end{aligned} \tag{14}$$

where $x_{\mathrm{ac}0}$ is the unique initial condition such that $\lim_{t\to\infty} x_M = 0$, given by

$$x_{\mathrm{ac}0} = \int_0^{\infty} e^{A_{\mathrm{ac}}\tau} B_{\mathrm{ac}} \tilde{q}(\tau) d\tau. \tag{15}$$

A proof of this lemma may be found in Appendix A.

4.2. Upper bound computation. Having identified an appropriate family of multipliers $\mathcal{M}$, we can now describe a numerical implementation of the upper bound on worst-case $\mathbf{H}_2$ norm given in Lemma 3.1.

We first proceed to compute $\mathbf{E}\max_{\tilde{p}\in\mathbf{L}_2(\mathbf{R}_+)} \|\tilde{z}\|^2 + \langle \tilde{p}, M\tilde{q}\rangle$ for a given operator M, where $\tilde{p}$, $\tilde{q}$, and $\tilde{z}$ satisfy (6) and $\tilde{x}(0) = B_w u$, where $\mathbf{E}\, uu^T = I$. Introduce the augmented system

$$
\begin{aligned}
&\frac{d}{dt}\bar{x}(t) = A_{MH}\bar{x}(t) + B_{MH}\tilde{p}(t), \qquad \bar{x}(0) = \begin{bmatrix} \tilde{x}(0)^T & 0 & x_{\mathrm{ac}0}^T \end{bmatrix}^T,\\
&(M\tilde{q})(t) = C_{MH}\bar{x}(t) + D_{MH}\tilde{p}(t),\\
&\tilde{z}(t) = C_{MHz}\bar{x}(t),
\end{aligned} \tag{16}
$$

where $A_{MH},\ B_{MH},\ C_{MH},\ D_{MH}$, and C_{MHz} are given by

$$
\begin{aligned}
&A_{MH} = \begin{bmatrix} A & 0 \\ B_M C_q & A_M \end{bmatrix}, \quad B_{MH} = \begin{bmatrix} B_p \\ B_M D_{qp} \end{bmatrix},\\
&C_{MH} = \begin{bmatrix} D_M C_q & C_M \end{bmatrix}, \quad D_{MH} = D_M D_{qp},\\
&C_{MHz} = \begin{bmatrix} C_z & 0 & 0 \end{bmatrix},
\end{aligned}
$$

and $x_{\mathrm{ac}0}$ is given by (15). From Lemma 4.2, computing $\max_{\tilde{p}\in\mathbf{L}_2(\mathbf{R}_+)} \|\tilde{z}\|_2^2 + 2\langle \tilde{p}, M\tilde{q}\rangle$ is equivalent to computing

$$
\max_{\tilde{p}\in\mathbf{L}_2(\mathbf{R}_+)} \int_0^\infty \begin{bmatrix} \bar{x}(t) \\ \tilde{p}(t) \end{bmatrix}^T \begin{bmatrix} C_{MHz}^T C_{MHz} & C_{MH}^T \\ C_{MH} & D_{MH} + D_{MH}^T \end{bmatrix} \begin{bmatrix} \bar{x}(t) \\ \tilde{p}(t) \end{bmatrix} dt, \tag{17}
$$

where $\tilde{p}$ and $\bar{x}$ satisfy (16). If the initial condition $\bar{x}(0)$ was constant, the solution to this quadratic optimal control problem could be obtained by standard methods such as the ones described in [41]. Unfortunately, this is not the case, because the noncausal multiplier M is involved, which makes $\bar{x}(0)$ depend on $\tilde{p}$ through the relation (15). In fact, assuming that (A_{MH}, B_{MH}) is controllable (we will make this technical assumption from now on), $x_{\mathrm{ac}0}$ spans all of $\mathbf{R}^{Nn_p}$ as $\tilde{p}$ spans all of $\mathbf{L}_2(\mathbf{R}_+)$. Therefore, in order to compute (17) subject to the constraints (16) and (15), we propose the following two-step strategy:

(i) Fix $x_{\mathrm{ac}0}$. Compute (17) subject to the constraints (16) and

$$
\int_0^\infty e^{A_{\mathrm{ac}}\tau} B_{\mathrm{ac}} \tilde{q}(\tau) d\tau = x_{\mathrm{ac}0}.
$$

(ii) Maximize the resulting solution over $x_{\mathrm{ac}0} \in \mathbf{R}^{Nn_p}$.

From Lemma 4.2, step (i) is equivalent to computing (17) subject to the constraints (16) and $\lim_{t\to\infty} \bar{x}(t) = 0$. This is a well-known problem whose solution is given by Willems or Yakubovich, for example.

LEMMA 4.3 (see [41, Theorem 3]; [44, Theorem 3]). *Assume that (A_{MH}, B_{MH}) is controllable. The value of* (17) *subject to the constraints* (16) *and $\lim_{t\to\infty} \bar{x}(t) = 0$ is finite if and only if there exists a symmetric matrix P satisfying the matrix inequality*

$$
\begin{bmatrix} A_{MH}^T P + P A_{MH} + C_{MHz}^T C_{MHz} & P B_{MH} + C_{MH}^T \\ B_{MH}^T P + C_{MH} & D_{MH} + D_{MH}^T \end{bmatrix} \le 0. \tag{18}
$$

It is then given by $\bar{x}(0)^T P^- \bar{x}(0)$, where P^- is the smallest (in the sense of the partial ordering of symmetric matrices) among all matrices P satisfying (18).

A proof of this lemma may be found in Appendix B.

In particular, we see that P^- is independent from the initial condition $\bar{x}(0)$. Therefore, step (ii) is simply done by maximizing

$$
\begin{bmatrix} x_0^T & 0 & x_{\mathrm{ac}0}^T \end{bmatrix} P^- \begin{bmatrix} x_0^T & 0 & x_{\mathrm{ac}0}^T \end{bmatrix}^T
$$

over $x_{\mathrm{ac}0}$. Partitioning P^- as

$$P^- = \begin{bmatrix} P_{11}^- & P_{12}^- & P_{13}^- \\ P_{12}^{-T} & P_{22}^- & P_{23}^- \\ P_{13}^{-T} & P_{23}^{-T} & P_{33}^- \end{bmatrix},$$

this problem is equivalent to maximizing

$$\phi(x_{\mathrm{ac}0}) = x_0^T P_{11}^- x_0 + 2x_0^T P_{13}^- x_{\mathrm{ac}0} + x_{\mathrm{ac}0}^T P_{33}^- x_{\mathrm{ac}0}.$$

The function ϕ is quadratic in $x_{\mathrm{ac}0}$. Thus, it has a maximum if and only if $P_{33}^- \leq 0$ and it has a stationary point $x_{\mathrm{ac}0}$, solution to the equation $P_{13}^{-T} x_0 = -P_{33}^- x_{\mathrm{ac}0}$. In this case, the maximum value of ϕ is given by

$$x_0^T (P_{11}^- - P_{13}^- P_{33}^{-\dagger} P_{13}^{-T}) x_0. \tag{19}$$

Assume now that $x_0 = B_w u$, where u is a random variable satisfying $\mathbf{E}\, uu^T = I$. Then

$$\mathbf{E} \max_{\tilde{p} \in \mathbf{L}_2(\mathbf{R}_+)} \|\tilde{z}\|_2^2 + 2\langle \tilde{p}, M\tilde{q} \rangle \tag{20}$$

is finite if and only if

$$\begin{aligned} &P_{33}^- \leq 0, \\ &\forall u \in \mathbf{R}^{n_w}, \ \ \exists x_{\mathrm{ac}0} \text{ such that } P_{13}^{-T} B_w u = -P_{33}^- x_{\mathrm{ac}0}. \end{aligned} \tag{21}$$

Then, from (19), the value of (20) is

$$\begin{aligned} &\mathbf{E}\, u^T B_w^T (P_{11}^- - P_{13}^- P_{33}^{-\dagger} P_{13}^{-T}) B_w u \\ &\qquad = \mathbf{Tr}\, B_w^T (P_{11}^- - P_{13}^- P_{33}^{-\dagger} P_{13}^{-T}) B_w. \end{aligned} \tag{22}$$

It is possible to write the second condition in (21) in a more compact manner by remarking that it is equivalent to the requirement that $P_{13}^{-T} B_w$ lie in the range of P_{33}^- or, equivalently, in the nullspace of $I - P_{33}^- P_{33}^{-\dagger} = I - P_{33}^{-\dagger} P_{33}^-$. Thus, this condition may also be written $(I - P_{33}^{-\dagger} P_{33}^-) P_{13}^{-T} B_w = 0$.

Introducing the symmetric matrix $\Gamma \in \mathbf{R}^{n_w \times n_w}$ as a slack variable, we can also write the value of (22) together with the condition (21) as the minimum value of $\mathbf{Tr}\, \Gamma$ subject to the conditions

$$\begin{aligned} &B_w^T (P_{11}^- - P_{13}^- P_{33}^{-\dagger} P_{13}^{-T}) B_w \leq \Gamma, \\ &P_{33}^- \leq 0, \\ &(I - P_{33}^{-\dagger} P_{33}^-) P_{13}^{-T} B_w = 0. \end{aligned} \tag{23}$$

Using Schur complements (see [9, p. 28] for details), it is also the minimum value of $\mathbf{Tr}\, \Gamma$ subject to the single constraint

$$\begin{bmatrix} B_w^T P_{11}^- B_w - \Gamma & B_w^T P_{13}^- \\ P_{13}^{-T} B_w & P_{33}^- \end{bmatrix} \leq 0.$$

We now remark that for a general symmetric matrix P partitioned as

$$P = \begin{bmatrix} P_{11} & P_{12} & P_{13} \\ P_{12}^T & P_{22} & P_{23} \\ P_{13}^T & P_{23}^T & P_{33} \end{bmatrix},$$

the matrix

$$X(P) = \begin{bmatrix} B_w^T P_{11} B_w & B_w^T P_{13} \\ P_{13}^T B_w & P_{33} \end{bmatrix}$$

varies monotonically with P, meaning that if $P_1 \le P_2$, then $X(P_1) \le X(P_2)$. Thus, given P^- as defined in Lemma 4.3, we compute the value of $\mathbf{E} \max_{\tilde{p} \in \mathbf{L}_2(\mathbf{R}_+)} \|\tilde{z}\|_2^2 + 2\langle \tilde{p}, M\tilde{q} \rangle$ by minimizing $\mathbf{Tr}\,\Gamma$ over the variables P and Γ subject to the matrix constraints (18) and

$$\begin{bmatrix} B_w^T P B_w - \Gamma & B_w^T P_{13} \\ P_{13}^T B_w & P_{33} \end{bmatrix} \le 0. \tag{24}$$

Thus the value of $\min_{M \in \mathcal{M}} \mathbf{E} \max_{\tilde{p} \in \mathbf{L}_2} \|\tilde{z}\|_2^2 + 2\langle \tilde{p}, M\tilde{q} \rangle$ is obtained by minimizing $\mathbf{Tr}\,\Gamma$ over the variables P, Γ, and $M \in \mathcal{M}$ subject to the matrix constraints (18) and (24). Remarking that $M \in \mathcal{M}$ if and only if the inequality (13) holds, we can now summarize the computation of the upper bound on robust $\mathbf{H}_2$ performance in the following theorem.

THEOREM 4.4. *Consider system* (6). *The quantity*

$$\min_{M \in \mathcal{M}} \mathbf{E} \max_{\tilde{p} \in \mathbf{L}_2(\mathbf{R}_+)} \|\tilde{z}\|_2^2 + 2\langle \tilde{p}, M\tilde{q} \rangle,$$

where $\tilde{z}$, $\tilde{p}$, *and* $\tilde{q}$ *satisfy* (6), *is computed as the minimum of* $\mathbf{Tr}\,\Gamma$ *over the variables* Γ, P, $P_1, \ldots, P_{n_p}$, m_{ij}, $i = 1, \ldots, n_p$, $j = 0, \ldots, N$, *satisfying the constraints*

$$\begin{bmatrix} A_{Mi}^T P_i + P_i A_{Mi} & P_i B_{Mi} - C_{Mi}^T \\ B_{Mi}^T P_i - C_{Mi} & -(D_{Mi} + D_{Mi}^T) \end{bmatrix} \le 0, \quad i = 1, \ldots, n_p, \tag{25}$$

$$\begin{bmatrix} A_{MH}^T P + P A_{MH} + C_{MHz}^T C_{MHz} & P B_{MH} + C_{MH}^T \\ B_{MH}^T P + C_{MH} & D_{MH} + D_{MH}^T \end{bmatrix} \le 0, \tag{26}$$

and

$$\begin{bmatrix} B_w^T P_{11} B_w - \Gamma & B_w^T P_{13} \\ P_{13}^T B_w & P_{33} \end{bmatrix} \le 0, \tag{27}$$

where

$$P = \begin{bmatrix} P_{11} & P_{12} & P_{13} \\ P_{12}^T & P_{22} & P_{23} \\ P_{13}^T & P_{23}^T & P_{33} \end{bmatrix} \tag{28}$$

has been partitioned conformally with the dimensions of A, A_{ca}, *and* A_{ac} (*where* A_{ca} *and* A_{ac} *are given by* (10)).

5. Discussion. In this section, we discuss the main result of this paper and compare it with some previous approaches.

5.1. Computational issues. We see that Theorem 4.4 gives us *effective means* to compute the upper bound on the worst-case $\mathbf{H}_2$ norm of the system (1): indeed, we have to minimize the linear objective $\mathbf{Tr}\,\Gamma$ over the variables Γ, P, $P_1, \ldots, P_{n_0}$, m_{ij}, $i = 1, \ldots, n_p$, $j = 0, \ldots, N$, which appear *linearly* in the matrix constraints (25)–(27). In particular, new interior-point convex optimization algorithms will solve this problem very efficiently [25, 9]. Note also that the size of the optimization problem grows with the dimension N of the family of multipliers used. Note finally that the solution of the optimization problem in Theorem 4.4 via interior-point methods requires that all soft inequality signs of the form $\leq$ appearing in the constraints (25)–(27) be replaced by strict inequality signs. This does not present significant problems in most practical cases. (For a detailed discussion, we refer the reader to [9, section 2.5]).

5.2. When the obtained bound is finite. Theorem 4.4 provides an upper bound for the worst-case $\mathbf{H}_2$ norm of the system (1). However, it does not guarantee that this upper bounds is finite. Thus, it is interesting to examine cases for which this bound is guaranteed to be finite.

One such case arises when there exists $M \in \mathcal{M}$ such that $-MH_{qp}$ is strictly positive, where H_{qp} is the operator whose transfer function is given in (3). Then there exists a positive δ such that

$$\forall p \in \mathbf{L}_2(\mathbf{R}), \quad \langle p, MH_{qp}p\rangle \leq -\delta\langle p, p\rangle.$$

Define the impulse matrices

$$h_{zw}(t) = \begin{cases} C_z e^{At} B_w \text{ for } t \geq 0, \\ 0 \text{ otherwise} \end{cases}$$

and

$$h_{qw}(t) = \begin{cases} C_q e^{At} B_w \text{ for } t \geq 0, \\ 0 \text{ otherwise.} \end{cases}$$

Then, for any $\lambda > 0$, any initial condition $\tilde{x}(0) = B_w u$ and any $\tilde{p} \in \mathbf{L}_2(\mathbf{R}_+)$ in the system (6), we have

$$\begin{aligned} &\|\tilde{z}\|_2^2 + 2\lambda\langle\tilde{p}, M\tilde{q}\rangle \\ &= \|h_{zw}u + H_{zp}\tilde{p}\|_2^2 + 2\lambda\langle\tilde{p}, MH_{qp}\tilde{p} + Mh_{qw}u\rangle \\ &\leq \|h_{zw}u\|_2^2 + 2(\|h_{zw}u\|_2 \|H_{zp}\|_\infty + \lambda \|h_{qw}u\|_2 \|M\|_\infty) \|\tilde{p}\|_2 + (\|H_{zp}\|_\infty^2 - 2\lambda\delta) \|\tilde{p}\|_2^2 . \end{aligned}$$

Choosing $\lambda = (\|H_{zp}\|_\infty^2 + 1)/2\delta$, we therefore have

$$\begin{aligned} &\|\tilde{z}\|_2^2 + 2\lambda\langle\tilde{p}, M\tilde{q}\rangle \\ &\quad \leq \|h_{zw}u\|_2^2 + 2(\|h_{zw}u\|_2 \|H_{zp}\|_\infty + \lambda \|h_{qw}u\|_2 \|M\|_\infty) \|\tilde{p}\|_2 - \|\tilde{p}\|_2^2 \\ &\quad \leq \|h_{zw}u\|_2^2 + (\|h_{zw}u\|_2 \|H_{zp}\|_\infty + \lambda \|h_{qw}u\|_2 \|M\|_\infty)^2 \\ &\quad \leq \|h_{zw}u\|_2^2 + 2(\|h_{zw}u\|_2^2 \|H_{zp}\|_\infty^2 + \lambda^2 \|h_{qw}u\|_2^2 \|M\|_\infty^2). \end{aligned}$$

(Note that the quantities $\|M\|_\infty$ and $\|H_{zp}\|_\infty$ are well defined, since the corresponding transfer functions have no poles on the imaginary axis.) Taking expected values with respect to the random variable u, we have

$$\begin{aligned}&\mathbf{E}\max_{\tilde{p}\in\mathbf{L}_2(\mathbf{R}_+)} \|\tilde{z}\|_2^2 + 2\lambda\langle\tilde{p}, M\tilde{q}\rangle \\ &\qquad \le \|h_{zw}\|_2^2 + 2(\|h_{zw}\|_2^2 \|H_{zp}\|_\infty^2 + \lambda^2 \|h_{qw}\|_2^2 \|M\|_\infty^2).\end{aligned}$$

Noting that $M \in \mathcal{M}$ implies $\lambda M \in \mathcal{M}$, we conclude our upper bound is finite.

It is interesting to remark that the strict positivity of $-MH_{qp}$ is one of the conditions used in the classical theory of stability multipliers to prove stability of the system (1) as described in [10, p. 203]. Thus, whenever stability of the system (1) can be proven via stability multipliers, then we can provide finite bounds on its worst-case $\mathbf{H}_2$ performance. Note that numerical methods involving linear matrix inequalities to prove robust stability of the system (1) using linear families of finite-dimensional multipliers may be found in [3, 35].

5.3. Special cases and comparison with earlier results. In this section, we investigate what happens when considering special cases of the system (1) and of the multiplier M.

Consider first the case when the system (1) is perfectly known, that is, $B_p = 0, C_q = 0, D_{qp} = 0$. The optimization problem in Theorem 4.4 is solved by choosing $m_{ij} = 0$ for all i and j; $P_i = 0$ for all i;

$$P = \begin{bmatrix} P_{11} & 0 & 0 \\ 0 & 0 & 0 \\ 0 & 0 & 0 \end{bmatrix},$$

where P_{11} satisfies $A^T P_{11} + P_{11} A + C_z^T C_z = 0$; and $\Gamma = B_w^T P B_w$. P_{11} is the *observability Gramian*, and the obtained bound is then exact.

Second, consider the case when $N = 0$ and $n_p = 1$, that is, when the multiplier $M = m_{10} = m$ is simply a nonnegative scalar. Then, applying Theorem 4.4 leads to the computation of

$$\min_{m\ge 0} \mathbf{E} \max_{p\in\mathbf{L}_2(\mathbf{R}_+)} \|\tilde{z}\|^2 + 2m\langle\tilde{p}, \tilde{q}\rangle \tag{29}$$

via convex programming and linear matrix inequalities.

This special case in which scalar, memoryless multipliers are used has already appeared in the literature—for example, in the papers by Stoorvogel [37, 38], although in a different format: in these papers, the problem under consideration is to compute the worst-case $\mathbf{H}_2$ norm of the system

$$\begin{aligned}\frac{d}{dt}x(t) &= \hat{A}x(t) + \hat{B}_p\hat{p}(t) + B_w w(t), \quad x(0) = x_0, \\ \hat{q}(t) &= \hat{C}_q x(t) + \hat{D}_{qp}\hat{p}(t), \\ z(t) &= C_z x(t),\end{aligned} \tag{30}$$

when $\hat{p}$ and $\hat{q} \in \mathbf{L}_2(\mathbf{R}_+)$ are subject to the constraint

$$\|\hat{p}\|_2^2 \le \|\hat{q}\|_2^2. \tag{31}$$

Stoorvogel obtains an upper bound on the worst-case $\mathbf{H}_2$ norm for this system by relaxing the constraint (31) and by computing

$$\min_{m\geq 0} \mathbf{E} \max_{\hat{p}\in \mathbf{L}_2(\mathbf{R}_+)} \|z\|_2^2 + 2m(\|\hat{q}\|_2^2 - \|\hat{p}\|_2^2), \tag{32}$$

where z, $\hat{p}$, and $\hat{q}$ satisfy (30), subject to the boundary conditions $x(0) = B_w u$, $w(t) = 0$, and u is a random variable satisfying $\mathbf{E}\, uu^T = I$. This formulation can be cast in our framework the following way: Introduce the scattering variables $q = (\hat{p}+\hat{q})/2$ and $p = -(\hat{p}-\hat{q})/2$. Then, following the same reasoning as in [10, p. 215], the system (30) and the constraint (31) can also be written as

$$\begin{aligned} \frac{d}{dt}x(t) &= Ax(t) + B_p p(t) + B_w w(t), \quad x(0) = x_0, \\ q(t) &= C_q x(t) + D_{qp} p(t), \\ z(t) &= C_z x(t), \end{aligned} \tag{33}$$

and

$$\langle p, q\rangle \geq 0, \tag{34}$$

where A, B_p, C_q, and D_{qp} are determined from $\hat{A}$, $\hat{B}_p$, $\hat{C}_q$, and $\hat{D}_{qp}$ via the relations

$$\begin{aligned} A &= \hat{A} + \hat{B}_p(I - \hat{D}_{qp})^{-1}\hat{C}_q, \quad & B_p &= -2\hat{B}_p(I - \hat{D}_{qp})^{-1}, \\ C_q &= (I - \hat{D}_{qp})^{-1}\hat{C}_q, & D_{qp} &= -(I + \hat{D}_{qp})(I - \hat{D}_{qp})^{-1}. \end{aligned} \tag{35}$$

(These relations are valid if and only if $I - \hat{D}_{qp}$ is invertible.) The upper bound (32) can then be written as

$$\min_{m\geq 0} \mathbf{E} \max_{p\in \mathbf{L}_2(\mathbf{R}_+)} \|z\|_2^2 + 2m\langle p, q\rangle, \tag{36}$$

where z, p, and q satisfy (33), subject to the boundary conditions $x(0) = B_w u$, $w(t) = 0$, and u is a random variable satisfying $\mathbf{E}\, uu^T = I$. But then the quantity (36) is the same as the special case (29). Note that the only difference between Stoorvogel's system as given by (33) and (34) and the system (1) is that the uncertainty relationship $p(t) = (\Delta q)(t)$, where Δ is a passive, LTI operator, has been replaced by the uncertainty relationship $\langle p, q\rangle \geq 0$. The latter relationship describes a larger class of uncertainties than the former relationship, since whenever $p = \Delta q$ for Δ LTI and passive, then $\langle p, q\rangle \geq 0$. Similar comments may be made about the results presented in [30, 9, 13, 19].

6. Example. In this section, we present an example to illustrate the developed method, and compare it with earlier results. We consider the system (1) with

$$A = \begin{bmatrix} -0.1 & -0.1 & 0 & 0 & 0 \\ 0.1 & 0 & 0 & 0 & 0 \\ 0 & 0 & -0.2 & -0.2 & 0 \\ 0 & 0 & 0.1 & 0 & 0 \\ 0 & 0 & 0 & 0 & -2 \end{bmatrix}, \quad B_p = \begin{bmatrix} 0.1 \\ 0 \\ 0.1 \\ 0 \\ 0.001 \end{bmatrix}, \quad B_w = \begin{bmatrix} 0 \\ 0 \\ 0 \\ 0 \\ 1 \end{bmatrix},$$

$$C_q = \begin{bmatrix} 11.98 & 0 & 0.01 & 0 & 20 \end{bmatrix}, \qquad D_{qp} = -12,$$

$$C_z = \begin{bmatrix} 0 & 0 & 0 & 2 & 0 \end{bmatrix},$$

TABLE 1
Upper bound on square of $\mathbf{H}_2$ *performance as a function of multiplier order.*

N	0	1	2	3	4	5	6
Upper bound	341.17	137.87	99.03	83.59	75.81	71.39	68.69

and the perturbation Δ is any passive, LTI, and single-input, single-output system. It is easy to check (via a Nyquist plot, for example) that the transfer function $-H(s) = -C_q(sI - A)^{-1}B_p - D_{qp}$ has positive dissipation such that by application of the passivity theorem [10], the system (1) is stable. Using the software described in [18, 24, 17], we have recorded in Table 1 the obtained bounds on the square of its worst-case $\mathbf{H}_2$ norm as a function of N. ($2N$ is the order of the noncausal multiplier which is used.) Thus, $N = 0$ corresponds to the use of simple constant-gain, memoryless multipliers. As can be seen, the use of dynamic, noncausal multipliers improves the estimate on the square of the worst-case $\mathbf{H}_2$ norm by a factor of 5. Note also that the best upper bound converges to a steady state value quite fast with the size of the multiplier. This result is indeed obtained at the expense of increased computations.

7. Conclusion and extensions. In this paper, we have considered the problem of determining an upper bound for the worst-case $\mathbf{H}_2$ norm of linear systems subject to LTI uncertainties, by extending the theory of stability multipliers to handle $\mathbf{H}_2$ performance.

We have shown this bound appears as the solution of a convex optimization problem involving linear matrix inequalities. Thus there exist algorithms that will compute it fast and accurately.

We have shown this bound is always sharper than the ones devised earlier for larger classes of uncertainties. One example shows that this improvement can be significant.

This work can be extended in many directions. For example, the theory of stability multipliers has proved to be effective not only on LTI perturbations but also on other classes of uncertainties, including memoryless, sector-bounded, and monotonic nonlinearities and constant, unknown linear gains (parametric uncertainties). Thus, this paper could be easily extended to these cases (with the restriction that $\mathbf{H}_2$ norms of nonlinear systems require careful definition). The set of allowable multipliers $\mathcal{M}$ would then be different.

Appendix A: Proof of Lemma 4.2. Using inverse Fourier transforms, it is easy to show that for any $\tilde{q} \in \mathbf{L}_2(\mathbf{R})$, we have

$$(37)\quad (M\tilde{q})(t) = \int_{-\infty}^{t} C_{\mathrm{ca}} e^{A_{\mathrm{ca}}(t-\tau)} B_{\mathrm{ca}} \tilde{q}(\tau) d\tau + D_M \tilde{q}(t) + \int_{t}^{\infty} C_{\mathrm{ac}} e^{A_{\mathrm{ac}}(\tau-t)} B_{\mathrm{ac}} \tilde{q}(\tau) d\tau.$$

Thus, $(M\tilde{q})(t)$ is the sum of three parts: the first integral accounts for the causal part of H, the midterm represents a possible feedthrough term, and the second integral accounts for the anticausal part of M. (This is the reason for using the subscripts "ca" and "ac," which stand for "causal" and "anticausal," respectively.)

When $\tilde{q} \in \mathbf{L}_2(\mathbf{R}_+)$, the first integral in (37) is easily computed as $r_{\mathrm{ca}}(t)$, the output of the system

$$\begin{aligned} \frac{d}{dt} x_{\mathrm{ca}}(t) &= A_{\mathrm{ca}} x_{\mathrm{ca}}(t) + B_{\mathrm{ca}} \tilde{q}(t), \quad x_{\mathrm{ca}}(0) = 0, \\ r_{\mathrm{ca}}(t) &= C_{\mathrm{ca}} x_{\mathrm{ca}}(t). \end{aligned}$$

The second integral can be transformed the following way:

$$\begin{aligned}\int_t^\infty & C_{\rm ac}e^{A_{\rm ac}(\tau-t)}B_{\rm ac}\tilde q(\tau)d\tau \\ &= \int_0^\infty C_{\rm ac}e^{A_{\rm ac}(\tau-t)}B_{\rm ac}\tilde q(\tau)d\tau - \int_0^t C_{\rm ac}e^{A_{\rm ac}(\tau-t)}B_{\rm ac}\tilde q(\tau)d\tau \\ &= C_{\rm ac}\left[e^{-A_{\rm ac}t}\int_0^\infty e^{A_{\rm ac}\tau}B_{\rm ac}\tilde q(\tau)d\tau - \int_0^t e^{-A_{\rm ac}(t-\tau)}B_{\rm ac}\tilde q(\tau)d\tau\right].\end{aligned}$$

Therefore, defining

$$x_{\rm ac0} = \int_0^\infty e^{A_{\rm ac}\tau}B_{\rm ac}\tilde q(\tau)d\tau$$

($x_{\rm ac0}$ is always defined and finite, since $\tilde q \in \mathbf{L}_2(\mathbf{R}_+)$ and $A_{\rm ac}$ is stable), the second integral term in (37) can be written as $r_{\rm ac}(t)$, the output of the system

$$\begin{aligned}\frac{d}{dt}x_{\rm ac}(t) &= -A_{\rm ac}x_{\rm ac}(t) - B_{\rm ac}\tilde q(t), \quad x_{\rm ac}(0) = x_{\rm ac0}, \\ r_{\rm ac}(t) &= C_{\rm ac}x_{\rm ac}(t).\end{aligned}$$

Thus we obtain the expression (14) for $(M\tilde q)(t)$. The fact that $\lim_{t\to\infty} x_{\rm ca}(t) = 0$ is a direct consequence of the fact that $A_{\rm ca}$ is stable; the fact that $\lim_{t\to\infty} x_{\rm ac}(t) = 0$ follows from the identity

$$x_{\rm ac}(t) = \int_t^\infty e^{A_{\rm ac}(\tau-t)}B_{\rm ac}\tilde q(\tau)d\tau.$$

Conversely, consider the system (14), and let $x_{\rm ac0}$ be such that $\lim_{t\to\infty} x_{\rm ac}(t) = 0$. Then

$$x_{\rm ac}(t) = e^{-A_{\rm ac}t}x_{\rm ac0} - \int_0^t e^{-A_{\rm ac}(t-\tau)}\tilde q(\tau)d\tau.$$

If $\lim_{t\to\infty} x_{\rm ac}(t) = 0$, then $\lim_{t\to\infty} e^{A_{\rm ac}t}x_{\rm ac}(t) = 0$. Therefore, from the above equality, we must have

$$\lim_{t\to\infty}\int_0^t e^{A_{\rm ac}\tau}\tilde q(\tau)d\tau = x_{\rm ac0},$$

which proves that $x_{\rm ac0}$ is indeed unique. □

Appendix B: Proof of Lemma 4.3. From the literature on linear quadratic control, it is well known that the value of (17) subject to the constraint $\lim_{t\to\infty}\bar x = 0$ is a quadratic function of $\bar x(0)$ whenever it is finite. See for example [1, p. 21]. Denote this quadratic form as $\bar x(0)^T P^-\bar x(0)$. We first prove that P^- must satisfy the inequality (18). This is done by first noting that for any input $\tilde p$ and corresponding trajectory $\bar x$, the dissipation inequality

$$\frac{d}{dt}(\bar x(t)^T P^-\bar x(t)) \le -w(\bar x(t),\tilde p(t)) \tag{38}$$

must be satisfied, where

$$w(\bar x(t),\tilde p(t)) = \begin{bmatrix}\bar x(t)\\ \tilde p(t)\end{bmatrix}^T \begin{bmatrix} C_{MHz}^T C_{MHz} & C_{MH}^T \\ C_{MH} & D_{MH}+D_{MH}^T\end{bmatrix}\begin{bmatrix}\bar x(t)\\ \tilde p(t)\end{bmatrix}.$$

Indeed, for any initial condition $\bar{x}(t)$ and any input $\tilde{p}_1$ defined over the time interval $[t \;\; t+dt]$, standard dynamic programming arguments yield the inequality

$$\int_t^{t+dt} w(\bar{x},\tilde{p}_1)dt + \max_{\tilde{p}\in\mathbf{L}_2}\int_{t+dt}^{\infty} w(\bar{x},\tilde{p})dt \le \max_{\tilde{p}\in\mathbf{L}_2}\int_t^{\infty} w(\bar{x},\tilde{p})dt.$$

Letting dt tend toward 0 yields the equivalent differential form

$$w(\bar{x}(t),\tilde{p}(t)) \le -\frac{d}{dt}\max_{\tilde{p}\in\mathbf{L}_2}\int_t^{\infty} w(\bar{x},\tilde{p})dt,$$

which yields (38). Noting that this quadratic inequality must hold for any $\bar{x}$ and $\tilde{p}$ and that

$$\frac{d}{dt}\bar{x}(t)^T P^- \bar{x}(t) = 2\bar{x}(t)^T P^-(A_{MH}\bar{x}(t) + B_{MH}\tilde{p}(t))$$

yields the equivalent inequality (18).

Conversely, consider any feasible solution P to the matrix inequality (18). Then, the inequality (38) is automatically satisfied. Integrating this inequality from zero to infinity with any $\tilde{p} \in \mathbf{L}_2$ such that $\lim_{t\to\infty}\bar{x}(t) = 0$ implies

$$\bar{x}(0)^T P\bar{x}(0) \ge \int_0^{\infty} w(\bar{x},\tilde{p})dt.$$

Thus, whenever P^- exists, we have

P^- satisfies (18) and
$\bar{x}(0)^T P^- \bar{x}(0) \le \bar{x}(0)^T P\bar{x}(0) \;\; \forall(\bar{x}(0), P)$,
where P satisfies (18).

Thus, P^- must be the smallest feasible solution to the inequality (18). □

Acknowledgments. The author would like to thank V. (Ragu) Balakrishnan and A. Megretski for numerous and fruitful discussions.

REFERENCES

[1] B. Anderson and J. B. Moore, *Optimal Control: Linear Quadratic Methods*, Prentice-Hall, Englewood Cliffs, NJ, 1990.

[2] W. F. Arnold and A. J. Laub, *Generalized eigenproblem algorithms and software for algebraic Riccati equations*, Proc. IEEE, 72 (1984), pp. 1746–1754.

[3] V. Balakrishnan, Y. Huang, A. Packard, and J. Doyle, *Linear matrix inequalities in analysis with multipliers*, in Proc. American Control Conf., Baltimore, MD, June 1994, pp. 1228–1232.

[4] D. Bernstein and W. Haddad, *LQG control with an $\mathbf{H}_\infty$ performance constraint: A Riccati equation approach*, in Proc. American Control Conf., Atlanta, GA, June 1988, pp. 796–802.

[5] D. Bernstein and W. Haddad, *LQG control with an $\mathbf{H}_\infty$ performance bound: A Riccati equation approach*, IEEE Trans. Automat. Control, AC-34 (1989), pp. 293–305.

[6] D. Bernstein and W. Haddad, *Robust stability and performance analysis for linear dynamic systems*, IEEE Trans. Automat. Control, AC-34 (1989), pp. 751–758.

[7] S. Boyd, V. Balakrishnan, E. Feron, and L. El Ghaoui, *Control system analysis and synthesis via linear matrix inequalities*, in Proc. American Control Conf., vol. 2, San Francisco, CA, June 1993, pp. 2147–2154.

[8] S. Boyd and C. Barratt, *Linear Controller Design: Limits of Performance*, Prentice-Hall, Englewood Cliffs, NJ, 1991.

[9] S. Boyd, L. El Ghaoui, E. Feron, and V. Balakrishnan, *Linear Matrix Inequalities in System and Control Theory*, SIAM Stud. Appl. Math. 15, SIAM, Philadelphia, 1994.
[10] C. A. Desoer and M. Vidyasagar, *Feedback Systems: Input-Output Properties*, Academic Press, New York, 1975.
[11] J. Doyle, *Guaranteed margins for LQG regulators*, IEEE Trans. Automat. Control, AC-23 (1978), pp. 756–757.
[12] J. Doyle, *Analysis of feedback systems with structured uncertainties*, IEE Proc., 129-D (1982), pp. 242–250.
[13] E. Feron, *Linear Matrix Inequalities for the Problem of Absolute Stability of Control Systems*, Ph.D. thesis, Stanford University, Stanford, CA, Oct. 1993.
[14] E. Feron, V. Balakrishnan, S. Boyd, and L. El Ghaoui, *Numerical methods for* $\mathbf{H}_2$ *related problems*, in Proc. American Control Conf., vol. 4, Chicago, June 1992, pp. 2921–2922.
[15] A. L. Fradkov and V. A. Yakubovich, *The* $\mathcal{S}$*-procedure and duality relations in nonconvex problems of quadratic programming*, Vestnik Leningrad Univ. Math., 6 (1979), pp. 101–109 (in Russian, 1973).
[16] B. A. Francis, *A Course in* $\mathbf{H}_\infty$ *Control Theory*, Lecture Notes in Control and Inform. Sci. 88, Springer-Verlag, New York, 1987.
[17] P. Gahinet, A. Nemirovskii, A. J. Laub, and M. Chilali, *The LMI Control Toolbox*, The MathWorks Inc., 1995.
[18] P. Gahinet and A. S. Nemirovskii, *General-purpose LMI solvers with benchmarks*, in Proc. IEEE Conf. on Decision and Control, San Antonio, TX, Dec. 1993, pp. 3162–3165.
[19] S. R. Hall and J. P. How, *Mixed* $\mathbf{H}_2/\mu$ *performance bounds using dissipation theory*, in Proc. IEEE Conf. on Decision and Control, San Antonio, Texas, Dec. 1993.
[20] A. J. Laub, *A Schur method for solving algebraic Riccati equations*, IEEE Trans. Automat. Control, AC-24 (1979), pp. 913–921.
[21] A. Megretsky, $\mathcal{S}$*–procedure in optimal non-stochastic filtering*, Technical report TRITA/MAT-92-0015, Department of Mathematics, Royal Institute of Technology, S-100 44 Stockholm, Sweden, Mar. 1992.
[22] A. Megretsky, *Power distribution approach in robust control*, Technical report TRITA/MAT-92-0027, Department of Mathematics, Royal Institute of Technology, S-100 44 Stockholm, Sweden, 1992.
[23] A. Megretsky, *Necessary and sufficient conditions of stability: A multiloop generalization of the circle criterion*, IEEE Trans. Automat. Control, AC-38 (1993), pp. 753–756.
[24] A. S. Nemirovskii and P. Gahinet, *The projective method for solving linear matrix inequalities*, in Proc. American Control Conf., Baltimore, MD, June 1994, pp. 840–844.
[25] Y. Nesterov and A. Nemirovsky, *Interior-Point Polynomial Methods in Convex Programming*, SIAM Stud. Appl. Math. 13, SIAM, Philadelphia, 1994.
[26] A. Packard and J. C. Doyle, *Robust control with an* $\mathbf{H}_2$ *performance objective*, in Proc. American Control Conf., Minneapolis, MN, June 1987, pp. 2141–2146.
[27] F. Paganini, R. D'Andrea, and J. Doyle, *Behavioral approach to robustness analysis*, in Proc. American Control Conf., Baltimore, MD, June 1994, pp. 2782–2786.
[28] P. L. D. Peres and J. C. Geromel, $\mathbf{H}_2$ *control for discrete-time systems, optimality and robustness*, Automatica, 29 (1993), pp. 225–228.
[29] P. L. D. Peres, S. R. Souza, and J. C. Geromel, *Optimal* $\mathbf{H}_2$ *control for uncertain systems*, in Proc. American Control Conf., Chicago, June 1992.
[30] I. R. Petersen and D. C. McFarlane, *Optimal guaranteed cost control of uncertain linear systems*, in Proc. American Control Conf., Chicago, June 1992.
[31] I. R. Petersen, D. C. McFarlane, and M. A. Rotea, *Optimal guaranteed cost control of discrete-time uncertain systems*, in 12th IFAC World Congress, Sydney, Australia, July 1993, pp. 407–410.
[32] F. Riesz and B. Sz.-Nagy, *Functional Analysis*, Dover, New York, 1990.
[33] M. G. Safonov and R. Y. Chiang, *Real/complex* K_m*-synthesis without curve fitting*, in Digital and Numeric Techniques and Their Applications in Control Systems, Part 2, Control and Dyn. Syst. 56, C. T. Leondes, ed., Academic Press, New York, 1993, pp. 303–324.
[34] M. G. Safonov and P. H. Lee, *A multiplier method for computing real multivariable stability margin*, in IFAC World Congress, Sydney, Australia, July 1993.
[35] M. G. Safonov, J. Ly, and R. Y. Chiang, *On computation of multivariable stability margin using generalized Popov multiplier—LMI approach*, in Proc. American Control Conf., Baltimore, MD, June 1994.
[36] M. G. Safonov and G. Wyetzner, *Computer-aided stability criterion renders Popov criterion obsolete*, IEEE Trans. Automat. Control, AC-32 (1987), pp. 1128–1131.

[37] A. A. Stoorvogel, *The robust $\mathbf{H}_2$ problem: A worst case design*, in Conf. on Decision and Control, Brighton, England, Dec. 1991, pp. 194–199.
[38] A. A. Stoorvogel, *The robust $\mathbf{H}_2$ problem: A worst case design*, IEEE Trans. Automat. Control, 38 (1993), pp. 1358–1370.
[39] P. Van Dooren, *A generalized eigenvalue approach for solving Riccati equations*, SIAM J. Sci. Statist. Comput., 2 (1981), pp. 121–135.
[40] J. C. Willems, *The Analysis of Feedback Systems*, Research Monographs 62, MIT Press, Cambridge, MA, 1969.
[41] J. C. Willems, *Least squares stationary optimal control and the algebraic Riccati equation*, IEEE Trans. Automat. Control, AC-16 (1971), pp. 621–634.
[42] J. C. Willems, *On the existence of a nonpositive solution to the Riccati equation*, IEEE Trans. Automat. Control, AC-19 (1974).
[43] V. A. Yakubovich, *The $\mathcal{S}$-procedure in non-linear control theory*, Vestnik Leningrad Univ. Math., 4 (1977), pp. 73–93 (in Russian, 1971).
[44] V. A. Yakubovich, *Nonconvex optimization problem: The infinite-horizon linear-quadratic control problem with quadratic constraints*, Systems Control Lett., 19 (1992), pp. 13–22.

SIAM J. CONTROL OPTIM.
Vol. 35, No. 1, pp. 178–193, January 1997

009

A NEW NONSMOOTH EQUATIONS APPROACH TO NONLINEAR COMPLEMENTARITY PROBLEMS*

HOUYUAN JIANG† AND LIQUN QI†

Abstract. Based on Fischer's function, a new nonsmooth equations approach is presented for solving nonlinear complementarity problems. Under some suitable assumptions, a local and Q-quadratic convergence result is established for the generalized Newton method applied to the system of nonsmooth equations, which is a reformulation of nonlinear complementarity problems. To globalize the generalized Newton method, a hybrid method combining the generalized Newton method with the steepest descent method is proposed. Global and Q-quadratic convergence is established for this hybrid method. Some numerical results are also reported.

Key words. nonlinear complementarity problems, nonsmooth equations, semismoothness, uniform P-functions

AMS subject classifications. 90C33, 65K10

PII. S0363012994276494

1. Introduction.

Consider the following nonlinear complementarity problem (denoted the NCP):

$$\text{find } x \in \mathcal{R}^n \text{ such that } x \geq 0, \quad F(x) \geq 0, \quad \text{and } x^T F(x) = 0, \tag{1}$$

where $F : \mathcal{R}^n \to \mathcal{R}^n$ is continuously differentiable and the superscript T denotes the transpose operator.

The NCP has been used as a general framework for quadratic programming, linear complementarity problems, mathematical programming and some equilibrium problems. Designing algorithms for solving the NCP became extremely popular in the last decade, although it has a relatively long history. Many different approaches such as fixed point, homotopy, projection, Newton, smooth or nonsmooth optimization, smooth or nonsmooth equations, and many other methods have appeared. For an extensive survey of the NCP, we refer the reader to [34] and references therein; see also [13].

Josephy [17] presented a generalized Newton method for solving variational inequalities which contain the NCP as a special case. The Josephy–Newton method was shown to be convergent locally as well as Q-quadratically. However, this method does not converge globally. Therefore, the design of good global methods for solving the NCP becomes a challenging endeavor.

Based on the above motivation, Pang [33] reformulated the NCP as a system of nonsmooth equations. This makes it possible to extend the classical damped Newton method for solving smooth equations to the nonsmooth case. Harker and Xiao [14] also converted the NCP into a system of nonsmooth equations in a different way. Pang and Gabriel [35] further combined the nonsmooth equations reformulation with sequential quadratic programming, resulting in a so-called NE/SQP method. They established that the NE/SQP method converges both globally and locally Q-quadratically. The

*Received by the editors November 2, 1994; accepted for publication (in revised form) November 8, 1995. This research was supported by the Australian Research Council.
http://www.siam.org/journals/sicon/35-1/27649.html
†School of Mathematics, University of New South Wales, Sydney, NSW 2052, Australia (jiang@solution.maths.unsw.edu.au, qi@solution.maths.unsw.edu.au).

idea of reformulating the NCP as a system of equations is not new. Actually, Mangasarian [25] reformulated the NCP as a system of smooth equations (see also [41]). But the singularity of the system confines the application of Mangasarian's method. Kanzow [19] further established many ways for reformulating the NCP. Chen and Mangasarian [4] proposed smoothing methods where a class of smooth functions approximate certain nonsmooth functions arising in the reformulations of the NCP. The further study of those methods has been done by Chen and Harker [3].

Another approach is to reformulate the NCP as a smooth unconstrained minimization problem. Mangasarian and Solodov [26] introduced a smooth function in such a way that any global minimizer of the unconstrained minimization problem with this function as the objective function is a solution of the NCP. Yamashita and Fukushima [43] proved that any stationary point of the unconstrained minimization problem proposed by Mangasarian and Solodov is a solution of the NCP if F is continuously differentiable and strongly monotone in $\mathcal{R}^n$. This shows that any method for solving unconstrained minimization problems is applicable for the NCP in this special case. Kanzow [18] gave some more approaches to characterize the NCP as unconstrained minimization problems. Geiger and Kanzow [12] proved that one of these approaches has a good property as Yamashita and Fukushima observed for the Mangasarian–Solodov function, but with a weaker assumption. Reformulating the NCP as a smooth constrained minimization problem is also a direction to follow. This approach is due to Fukushima's [11] and Auchmuty's [1] remarkable work; see [11, 1] for more details. Recently, Moré [30] formulated the NCP as a bound-constrained nonlinear least squares problem which may be classified as part of the smooth constrained minimization reformulation approach. For nonsmooth constrained minimization reformulations of the NCP, we refer the reader to [27].

Other methods for solving the NCP include homotopy or continuation methods. The idea is also based on the equivalence between the NCP and a system of equations; see [2, 20, 21, 22, 42].

The discussion above shows that the approaches of reformulating the NCP into smooth or nonsmooth equations, smooth or nonsmooth minimization problems are very promising. It generates a new future for the designing of computational methods for solving the NCP.

In this paper, we shall propose a new nonsmooth equations–based method for the NCP. This work is more related to nonsmooth equations and smooth unconstrained minimization based methods. We first transform the NCP into a system of nonsmooth equations in the next section by employing a function introduced by Fischer [8]. In section 3, we establish locally Q-quadratic convergence of the generalized Newton method for solving the NCP under some suitable conditions. Unlike other nonsmooth equations–based methods, we solve a system of linear equations instead of a mixed complementarity problem or a quadratic programming problem for each inner iteration. To globalize the generalized Newton method, we propose in section 4 a hybrid method by combining the generalized Newton method with a minimization technique. This hybrid method enjoys not only global convergence but also locally Q-quadratic convergence under some assumptions. In section 5, some numerical results are reported for both the generalized Newton and the hybrid methods.

2. The nonsmooth equations reformulation. Let $\phi : R^2 \to R$ be defined by

$$\phi(a, b) = \sqrt{a^2 + b^2} - a - b. \tag{2}$$

Fischer [8] first introduced this function to reformulate the Karush–Kuhn–Tucker (KKT) optimality conditions of nonlinear programming problems as systems of nonsmooth equations. Kanzow [18, 19] used this same function to reformulate linear and nonlinear complementarity problems as smooth nonlinear programs or systems of smooth equations. Jiang [16] studied sensitivity properties of nonsmooth variational inequalities by employing nonsmooth analysis to the function ϕ.

One basic property of this function is that

$$\phi(a,b)=0 \Longleftrightarrow a\geq 0, \quad b\geq 0, \quad ab=0. \tag{3}$$

From the above characterization, (1) can be recast as a system of nonsmooth equations defined by

$$H(x)=\begin{pmatrix} H_1(x) \\ \vdots \\ H_n(x) \end{pmatrix}=\begin{pmatrix} \phi(x_1,F_1(x)) \\ \vdots \\ \phi(x_n,F_n(x)) \end{pmatrix}=0 \tag{4}$$

in the sense that x solves (1) if and only if x solves (4).

Note that H is locally Lipschitz on $\mathcal{R}^n$ and Fréchet differentiable only on the set Ω, where

$$\Omega=\{x\in\mathcal{R}^n | x_i^2+(F_i(x))^2>0, i=1,...,n\}.$$

Let $G:\mathcal{R}^n\to\mathcal{R}^n$ be locally Lipschitz on $\mathcal{R}^n$. Then Clarke's generalized Jacobian of G at x, denoted by $\partial G(x)$, can be defined as the convex hull of the set

$$\{\lim_{x^k\to x} \bigtriangledown G(x^k) \mid G \text{ is differentiable at } x^k\in\mathcal{R}^n\}.$$

As an analogue to Fischer's analysis [8], we study Clarke's generalized Jacobian of H at all points on $\mathcal{R}^n$. For $x\in\Omega$, we have

$$\bigtriangledown H(x)=\mathrm{diag}(\gamma_i(x))\bigtriangledown F(x)+\mathrm{diag}(\mu_i(x)),$$

where $\mathrm{diag}(\alpha_i)$ denotes a diagonal matrix with diagonal elements $\alpha_1,\alpha_2,...,\alpha_n$, and

$$\gamma_i(x)=F_i(x)((F_i(x))^2+x_i^2)^{-1/2}-1,$$

$$\mu_i(x)=x_i((F_i(x))^2+x_i^2)^{-1/2}-1.$$

Clearly, for $i=1,...,n$,

$$(\gamma_i(x)+1)^2+(\mu_i(x)+1)^2=1. \tag{5}$$

It also follows that

$$\mu_i(x)=-1, \quad \gamma_i(x)=0 \Longleftrightarrow F_i(x)>0, \quad x_i=0,$$

$$\mu_i(x)=0, \quad \gamma_i(x)=-1 \Longleftrightarrow F_i(x)=0, \quad x_i>0.$$

For $x\notin\Omega$, by the definition of Clarke's generalized Jacobian and some simple calculations, any $V\in\partial H(x)$ can be represented as follows

$$V=\mathrm{diag}(\gamma_i)\bigtriangledown F(x)+\mathrm{diag}(\mu_i), \tag{6}$$

where

$$(\gamma_i + 1)^2 + (\mu_i + 1)^2 \leq 1, \quad i = 1, 2, ..., n. \tag{7}$$

Remark. The reverse of the above assertion is not true in general. However, one can find a matrix $V \in \partial H(x)$ for any $x \in \mathcal{R}^n$. For any given $x \in \mathcal{R}^n$, let $\mathcal{I} = \{i | x_i = F_i(x) = 0\}$ and u be a vector of $\mathcal{R}^n$ with $u_i \neq 0$ for $i \in \mathcal{I}$. Define

$$\gamma_i = \begin{cases} F_i(x)((F_i(x))^2 + x_i^2)^{-1/2} - 1, & i \notin \mathcal{I}, \\ \nabla F_i(x)^T u((\nabla F_i(x)^T u)^2 + u_i^2)^{-1/2} - 1, & i \in \mathcal{I}, \end{cases}$$

and

$$\mu_i = \begin{cases} x_i((F_i(x))^2 + x_i^2)^{-1/2} - 1, & i \notin \mathcal{I}, \\ u_i((\nabla F_i(x)^T u)^2 + u_i^2)^{-1/2} - 1, & i \in \mathcal{I}. \end{cases}$$

It is not hard to prove that $\mathrm{diag}(\gamma_i)\nabla F(x) + \mathrm{diag}(\mu_i)$ is a Clarke generalized Jacobian of H at x. This technique is also used by De Luca, Facchinei, and Kanzow [5].

3. The algorithm and its local convergence. Since Robinson's and Pang's pioneering works [40], [32], numerical methods for solving nonsmooth equations have been developed quite extensively. In particular, the generalized Newton method was proved to be convergent locally and superlinearly under the key assumption that the considered function is semismooth at solution points; see Qi and Sun [38] and Qi [37] for more details. Kojima and Shindo [23], Kummer [24], Pang [33], Pang and Qi [36], Ralph [39], and many more studied nonsmooth equations from different perspectives. Our goal here is to establish a local and superlinear convergence result when the generalized Newton method is applied to (4).

We now present our generalized Newton method for solving the system (4).

ALGORITHM A.

Step 1. Choose initial point $x^0 \in \mathcal{R}^n$ and Let $k = 0$.

Step 2. Choose $V^k \in \partial H(x^k)$ and solve the following Newton equations for the direction $d^k \in \mathcal{R}^n$:

$$H(x^k) + V^k d^k = 0. \tag{8}$$

Step 3. Set $x^{k+1} = x^k + d^k$. If x^{k+1} solves $H(x) = 0$, stop. Otherwise, let $k := k + 1$ and go to Step 2.

Note that d^k is not unique in (8) if V^k is not nonsingular. Moreover, it is not known if (8) is solvable or not. We shall provide conditions to ensure the solvability of (8).

Recall that a function $G : \mathcal{R}^n \to \mathcal{R}^n$ is said to be monotone if

$$(y - x)^T(G(y) - G(x)) \geq 0 \quad \text{for all } x, y \in \mathcal{R}^n.$$

Furthermore, G is called a uniform P-function [29] if there is $\alpha > 0$ such that

$$\max_{1 \leq i \leq n} (y_i - x_i)(G_i(y) - G_i(x)) \geq \alpha ||y - x||^2 \text{ for all } x, y \in \mathcal{R}^n.$$

Both monotone functions and uniform P-functions are very broad and have very fine properties. It is known [31] that a smooth function is monotone on $\mathcal{R}^n$ if and only

if its Jacobian function is positive semidefinite on $\mathcal{R}^n$. The following lemma shows a similar relation to the above characterization for uniform P-functions.

LEMMA 3.1. *Suppose that $G : \mathcal{R}^n \to \mathcal{R}^n$ is a smooth uniform P-function with modulus $\alpha > 0$. Then for any $x \in \mathcal{R}^n$ and $0 \neq d \in \mathcal{R}^n$*

$$\max_{1\leq i\leq n} \; d_i \bigtriangledown G_i(x)^T d \geq \alpha ||d||^2.$$

Proof. Let $y = x + td$ for $t > 0$. Then the result follows from the mean value theorem. □

PROPOSITION 3.2. *Suppose $F : \mathcal{R}^n \to \mathcal{R}^n$ is continuously differentiable. If F is a uniform P-function, then V is nonsingular for any $V \in \partial H(x)$ and $x \in \mathcal{R}^n$.*

Proof. Let $V \in \partial H(x)$ for $x \in \mathcal{R}^n$. Let d solve the equations $Vd = 0$. By (6), there exist constants $\gamma_1, \ldots, \gamma_n, \mu_1, \ldots, \mu_n$ satisfying (7) such that

$$\mathrm{diag}(\gamma_i) \bigtriangledown F(x)d + \mathrm{diag}(\mu_i)d = 0,$$

namely,

$$\begin{cases} \mu_1 d_1 + \gamma_1 \bigtriangledown F_1(x)d = 0, \\ \qquad\qquad \vdots \\ \mu_n d_n + \gamma_n \bigtriangledown F_n(x)d = 0. \end{cases}$$

Multiplying the ith equation by d_i, we have

$$\begin{cases} \mu_1 d_1^2 + d_1\gamma_1 \bigtriangledown F_1(x)d = 0, \\ \qquad\qquad \vdots \\ \mu_n d_n^2 + d_n\gamma_n \bigtriangledown F_n(x)d = 0. \end{cases}$$

By (7), if $\gamma_i = 0$ then $d_i = 0$. Consequently,

$$\max_{1\leq i\leq n} d_i \bigtriangledown F_i(x)d \leq \max_{1\leq i\leq n, \gamma_i\neq 0} \{0, -\mu_i/\gamma_i d_i^2\} \leq 0.$$

But the uniform P-function property of F implies that $d = 0$. Thus the nonsingularity of V follows. □

By a similar argument with some minor modifications, one can prove the following result.

PROPOSITION 3.3. *Suppose $F : \mathcal{R}^n \to \mathcal{R}^n$ is continuously differentiable. If $\bigtriangledown F(x)$ is positive definite for all $x \in \mathcal{R}^n$, then V is nonsingular for any $V \in \partial H(x)$ and $x \in \mathcal{R}^n$.*

Remarks. (1) The positive definiteness assumption of Proposition 3.3 cannot be weakened to the condition that F is strictly monotone on $\mathcal{R}^n$. This is shown by the following example. Let $F(x) = x^3$. Then F is strictly monotone on $\mathcal{R}$ and $H(x) = \sqrt{x^2 + x^6} - x - x^3$. An easy calculation shows that $\partial H(0) = [-2, 0]$. Therefore, $0 \in \partial H(0)$ is singular. (2) Clearly, for the nonsingularity of the generalized Jacobian of H at a given point, say x in Propositions 3.2 and 3.3, it is sufficient to impose the uniform P-function property or the positive definiteness of the Jacobian of the function F only in an open neighborhood of x.

It is known [37] that the semismooth condition imposed on the function at the solution point plays a key role in establishing the superlinear convergence of the generalized Newton method for nonsmooth equations. We now verify this property

for H. Denote $y \to_d x$ if $y \to x$, $y \neq x$ and $(y-x)/||y-x|| \to d/||d||$ for some $d \in \mathcal{R}^n \setminus \{0\}$. A function $G : \mathcal{R}^n \to \mathcal{R}^n$ is said to be semismooth at $x \in \mathcal{R}^n$ if G is Lipschitz continuous in an open neighborhood of x and the following limit exists:

$$\lim_{V \in \partial G(y), y \to_d x} Vd.$$

Furthermore, the function G is said to be semismooth in a given domain if it is semismooth at each point of the domain. A semismooth function has many fine properties. We introduce some of them in the following lemma taken from [37, 38].

LEMMA 3.4. *If G is semismooth at x, then it is directionally differentiable at x; i.e., $G'(x,d)$ exists for any $d \in \mathcal{R}^n$ and*

$$\lim_{\|d\| \to 0} \frac{G(x+d) - G(x) - G'(x,d)}{\| d \|} = 0, \tag{9}$$

$$\lim_{V \in \partial G(x+d), \|d\| \to 0} \frac{Vd - G'(x,d)}{\| d \|} = 0. \tag{10}$$

Qi and Sun [38] also introduced 1-order semismoothness. A function G is called 1-order semismooth at x if G is semismooth at x and for any $d \in \mathcal{R}^n$:

$$G(x+d) - G(x) - G'(x,d) = O(\| d \|^2). \tag{11}$$

We are now ready to establish the semismooth property for H.

PROPOSITION 3.5. *Suppose that $F : \mathcal{R}^n \to \mathcal{R}^n$ is twice continuously differentiable. Then H defined by* (4) *is* 1*-order semismooth on $\mathcal{R}^n$.*

Proof. Clearly, we only need to prove that H is 1-order semismooth at points on which H is not differentiable, i.e., $x \notin \Omega$. Note that H is 1-order semismooth at a point x if and only if H_i is 1-order semismooth at x for all i. Without loss of generality, we prove only H_1 is 1-order semismooth at the point x such that $x_1 = 0$ and $F_1(x) = 0$. A quick calculation yields for any $d \in \mathcal{R}^n$

$$H_1'(x,d) = \sqrt{d_1^2 + (\bigtriangledown F_1(x)^T d)^2} - d_1 - \bigtriangledown F_1(x)^T d.$$

Similarly, if $d_1^2 + (\bigtriangledown F_1(x)^T d)^2 > 0$, then

$$\begin{aligned}
&H_1(x+d) - H_1(x) - H_1'(x,d) \\
&= \sqrt{d_1^2 + (\bigtriangledown F_1(x)^T d + O(||d||^2))^2} - \sqrt{d_1^2 + (\bigtriangledown F_1(x)^T d)^2} + O(||d||^2) \\
&= \frac{O(||d||^2)(2 \bigtriangledown F_1(x)^T d + O(||d||^2))}{\sqrt{d_1^2 + (\bigtriangledown F_1(x)^T d)^2 + O(||d||^2)} + \sqrt{d_1^2 + (\bigtriangledown F_1(x)^T d)^2}} + O(||d||^2) \\
&= O(||d||^2),
\end{aligned}$$

which shows the 1-order semismoothness of H_1 at x. If $d_1^2 + (\bigtriangledown F_1(x)^T d)^2 = 0$, then $d_1 = 0 = \bigtriangledown F_1(x)^T d$. Hence the above estimate still holds:

$$H_1(x+d) - H_1(x) - H_1'(x,d) = O(||d||^2).$$

The desired result follows. ∎

THEOREM 3.6. *Let $F : \mathcal{R}^n \to \mathcal{R}^n$ be twice continuously differentiable. Suppose that F is a uniform P-function on $\mathcal{R}^n$. Suppose that x^* is a solution of* (1). *Then the sequence $\{x^k\}$ generated by Algorithm* A *is Q-quadratically convergent to x^* if x^0 is sufficiently to x^*.*

Proof. The result follows from the application of Theorem 3.2 of [38]. □

Remark. The above convergence result still holds if we only assume that any $V \in \partial H(x^*)$ is nonsingular where x^* is the solution point. The latter is guaranteed by a condition presented in the remark after Proposition 3.3.

4. A globally convergent algorithm. It is well known that the Newton method for solving a system of smooth equations does not necessarily converge if the initial point is chosen arbitrarily. Traditionally, the Gauss–Newton method is a remedy to achieve both global and local superlinear convergence; namely, a globalization convergence strategy is applied to the merit function defined by the 2-norm operator. Unfortunately, this method cannot be applied directly to the nonsmooth case, although some special Gauss–Newton methods do work; see [33, 14]. These methods solve either a mixed linear complementarity problem or a quadratic programming problem at each inner iteration. Therefore, it is highly desirable to solve some simpler problems rather than mixed linear complementarity problems or quadratic programming problems in some cases. This is our goal in this section.

Define a merit function $\theta : \mathcal{R}^n \to \mathcal{R}$ by

$$\theta(x) = \frac{1}{2}\sum_{i=1}^{n} \phi(x_i, F_i(x))^2 = \frac{1}{2}||H(x)||^2.$$

Geiger and Kanzow [12] used this function to reformulate nonlinear complementarity problems. They showed that solving (1) is equivalent to finding stationary points of the unconstrained optimization problem

$$\min_{x\in\mathcal{R}^n} \theta(x), \tag{12}$$

whenever the continuously differentiable function F is monotone and (1) is solvable.

Note that θ is continuously differentiable on the whole space $\mathcal{R}^n$. This suggests that it is possible to invoke any global convergence algorithm for solving (12). Clearly, the Newton method combined with a line search or a trust region strategy is one option. However, the singularity of the second-order derivative of θ at the solution point makes it doubtful to obtain superlinear convergence. Moreover, much computational work is needed to get the second-order derivative of θ. For this reason, using the Newton method combined with a line search or a trust region strategy is not recommended for solving (12). It is known that the steepest descent method can guarantee the global convergence of (12). But one of the disadvantages of this method is that it usually performs poorly, particularly when the solution point is close.

Based on the above discussion, we propose a hybrid method which combines the steepest descent method for solving (12) and the generalized Newton method for solving (4). The idea is simple. We first check if the Newton step gives a sufficient decrease to θ. If it does, another point is obtained. Otherwise, the steepest descent method is applied to the minimization (12). It also provides a new iteration point. Moreover, when the iterated point is close enough to the solution point of (1), the algorithm never requires any steepest descent step. Then the superlinear convergence result follows directly, provided that the generalized Newton method possesses this property locally.

Define a function $\Phi : \mathcal{R}^2 \to \mathcal{R}$,

$$\Phi(a,b) = \frac{1}{2}(\phi(a,b))^2.$$

Geiger and Kanzow [12] have established the following lemma.

LEMMA 4.1.

(1) Φ *is continuously differentiable for all* $a, b \in \mathcal{R}$, *in particular* $\bigtriangledown\Phi(0,0) = (0,0)^T$.

(2) $\bigtriangledown_a\Phi(a,b) \bigtriangledown_b \Phi(a,b) \geq 0$ *for* $a, b \in \mathcal{R}$.

(3) $\bigtriangledown_a\Phi(a,b) \bigtriangledown_b \Phi(a,b) = 0$ *implies* $\Phi(a,b) = 0$.

Remark. The inverse of (3) in Lemma 4.1 is also true by a simple argument.

We are now ready to present the hybrid algorithm.

ALGORITHM B.

Step 1. Choose initial point $x^0 \in \mathcal{R}^n$, parameters $\sigma \in (0, 1/2), \rho \in (0,1)$, and $\beta \in (0,1)$. Let $k = 0$.

Step 2. Choose $V^k \in \partial H(x^k)$ and solve the following Newton equations for the direction $d^k \in \mathcal{R}^n$:

$$H(x^k) + V^k d^k = 0. \tag{13}$$

Step 3. If (13) is solvable and $\theta(x^k + d^k) \leq \beta\theta(x^k)$, go to Step 5. Otherwise, go to Step 4.

Step 4. Let $\bar{d^k} = - \bigtriangledown \theta(x^k)$. Find a minimum nonnegative integer, say m, such that

$$\theta(x^k + \rho^m \bar{d^k}) \leq \theta(x^k) + \sigma\rho^m \bigtriangledown \theta(x^k)^T \bar{d^k}. \tag{14}$$

Let $d^k = \rho^m \bar{d^k}$, and go to Step 5.

Step 5. Set $x^{k+1} = x^k + d^k$. If x^{k+1} solves $H(x) = 0$, stop. Otherwise, let $k := k + 1$ and go to Step 2.

To prove the existence of a cluster point of the sequence $\{x^k\}$ generated by Algorithm B, we need the following result. It is a generalization of Theorem 3.2 of Geiger and Kanzow [12]. The technique of the proof is also due to [12].

PROPOSITION 4.2. *Suppose that* $F : \mathcal{R}^n \to \mathcal{R}^n$ *is a continuous and uniform P-function. Let* $L(x^0) = \{x \in \mathcal{R}^n | \theta(x) \leq \theta(x^0)\}$, *where* $x^0 \in \mathcal{R}^n$. *Then* $L(x^0)$ *is compact.*

Proof. Clearly, $L(x^0)$ is closed. Suppose that $L(x^0)$ is not compact for a given $x^0 \in \mathcal{R}^n$; namely, there exists a sequence of $\{x^k\} \subseteq L(x^0)$ such that $\lim_{k\to\infty} ||x^k|| = \infty$. Let

$$I = \{1 \leq i \leq n \mid \{x_i^k\} \text{ is unbounded}\}.$$

Clearly, $I \neq \emptyset$ due to our assumption. Define a new sequence $\{y^k\}$ associated with $\{x^k\}$ by

$$y_i^k = \begin{cases} 0 & \text{if } i \in I, \\ x_i^k & \text{otherwise.} \end{cases}$$

Since F is a uniform P-function, there exists $\alpha > 0$ such that

$$\begin{aligned}
\alpha \textstyle\sum_{i\in I}(x_i^k)^2 &= \alpha||x^k - y^k||^2 \\
&\leq \max_{1\leq i\leq n}(x_i^k - y_i^k)(F_i(x^k) - F_i(y^k)) \\
&= \max\{0, \max_{i\in I} x_i^k(F_i(x^k) - F_i(y^k))\} \\
&\leq \sqrt{\textstyle\sum_{i\in I}(x_i^k)^2}\, \max_{i\in I}|F_i(x^k) - F_i(y^k)| \\
&\leq \sqrt{\textstyle\sum_{i\in I}(x_i^k)^2}\, \textstyle\sum_{i\in I}|F_i(x^k) - F_i(y^k)|.
\end{aligned}$$

Consequently, by the definition of I there exists a subset K such that for $k \in K$

$$\alpha\sqrt{\sum_{i\in I}(x_i^k)^2} \leq \sum_{i\in I}|F_i(x^k) - F_i(y^k)|.$$

From the continuity of F, the boundedness of $\{y^k\}$, the definition of I, and the fact that I contains only a finite number of elements, the above inequality implies that there exists $i_0 \in I$ such that $\{|F_{i_0}(x^k)|, k \in K\}$ is unbounded. Since $i_0 \in I$, it follows from the definition of I that $\{x_{i_0}^k\}$ is also unbounded. By Lemma 3.1 of [12], $\{\phi(x_{i_0}^k, F_{i_0}(x^k)), k \in K\}$ is unbounded. Thus, $\{\theta(x^k)\}$ is unbounded, which is a contradiction. □

It is well known that some local minimum points of an unconstrained optimization problem are not global minimum points in general. This shows that some cluster points generated by the above algorithm are not solutions of (12) in general. Fortunately, under some assumptions, the local minimum points of (12) coincide with its global minimum points. We present this in the following proposition.

PROPOSITION 4.3. *Suppose that $F : \mathcal{R}^n \to \mathcal{R}^n$ is continuously differentiable, that F is a uniform P-function. Then x^* is a global minimum point of* (12) *if and only if x^* is a stationary point of* (12)*, i.e., $\bigtriangledown\theta(x^*) = 0$; moreover, such a point x^* must be a solution of* (1)*.*

Proof. The "only if" part is obvious. Now we suppose x^* is a stationary point of (12), namely, $\bigtriangledown\theta(x^*) = 0$. Some simple calculations show that

$$\bigtriangledown\theta(x) = \begin{pmatrix} \bigtriangledown_a\Phi(x_1, F_1(x)) \\ \vdots \\ \bigtriangledown_a\Phi(x_n, F_n(x)) \end{pmatrix} + \bigtriangledown F(x)^T \begin{pmatrix} \bigtriangledown_b\Phi(x_1, F_1(x)) \\ \vdots \\ \bigtriangledown_b\Phi(x_n, F_n(x)) \end{pmatrix}.$$

Denote $\Phi_a = (\bigtriangledown_a\Phi(x_1^*, F_1(x^*)), \ldots, \bigtriangledown_a\Phi(x_n^*, F_n(x^*)))^T$, and $\Phi_b = (\bigtriangledown_b\Phi(x_1^*, F_1(x^*)), \ldots, \bigtriangledown_b\Phi(x_n^*, F_n(x^*)))^T$. Hence,

$$0 = \bigtriangledown\theta(x^*) = \begin{pmatrix} \bigtriangledown_a\Phi(x_1^*, F_1(x^*)) + (\bigtriangledown F(x^*)^T\Phi_b)_1 \\ \vdots \\ \bigtriangledown_a\Phi(x_n^*, F_n(x^*)) + (\bigtriangledown F(x^*)^T\Phi_b)_n \end{pmatrix},$$

where $(\nabla F(x^*)^T\Phi_b)_i$ denotes the ith element of the column vector $\nabla F(x^*)^T\Phi_b$. Multiplying the ith equation above by $\bigtriangledown_b\Phi(x_i^*, F_i(x^*))$, we have for $i = 1, 2, \ldots, n$,

$$\bigtriangledown_b\Phi(x_i^*, F_i(x^*))\, \bigtriangledown_a \Phi(x_i^*, F_i(x^*)) + \bigtriangledown_b\Phi(x_i^*, F_i(x^*))(\bigtriangledown F(x^*)^T\Phi_b)_i = 0.$$

This implies

$$\max_{1\le i\le n} \bigtriangledown_b\Phi(x_i^*, F_i(x^*))(\bigtriangledown F(x^*)^T\Phi_b)_i = -\min_{1\le i\le n} \bigtriangledown_b\Phi(x_i^*, F_i(x^*))\bigtriangledown_a\Phi(x_i^*, F_i(x^*)) \le 0,$$

where the inequality is due to (2) of Lemma 4.1. By the uniform P-function property of F, both $\nabla F(x^*)$ and $\nabla F(x^*)^T$ are P-matrices. It follows from the definition of P-matrix that $\Phi_b = 0$, which shows that x^* is a global solution of (12) by (3) of Lemma 4.1. This completes the proof. □

It is clear that if Algorithm B stops after a finite number of steps, then a solution to (1) is obtained. In what follows, we assume that Algorithm B generates a sequence $\{x^k\}$. We now establish global convergence for Algorithm B.

THEOREM 4.4. *Suppose that $F : \mathcal{R}^n \to \mathcal{R}^n$ is twice continuously differentiable, and $\{x^k\}$ is a sequence generated by Algorithm* B. *Then*

(1) *$\{x^k\}$ converges to the unique solution x^* of* (1) *if F is a uniform P-function on $\mathcal{R}^n$. Furthermore, $\{x^k\}$ converges quadratically to x^*.*

(2) *x^* is a solution of* (1) *if F is monotone on $\mathcal{R}^n$ and if x^* is an accumulation point of $\{x^k\}$. Moreover, $\{x^k\}$ converges quadratically to x^* if $\bigtriangledown F(x^*)$ is positive definite.*

Proof. (1) Suppose that there are an infinite number of Newton steps, say $k \in K$ in Algorithm B. Then $\lim_{k\to\infty, k\in K} ||H(x^k)|| = 0$. Since F is a uniform P-function on $\mathcal{R}^n$, Proposition 4.2 shows that $\{x^k, k \in K\}$ is bounded. Assume that x^{**} is an accumulation point of $\{x^k, k \in K\}$. Then x^{**} is a solution to (1). But x^{**} should be the unique solution x^* of (1) by the uniform P-function property of F.

Suppose that there are only a finite number of Newton steps taken in Algorithm B. Then Algorithm B eventually reduces to the steepest descent algorithm for solving the unconstrained optimization problem (12). Proposition 4.2 shows that $\{x^k\}$ is bounded. Assuming that x^{**} is an accumulation point of $\{x^k\}$, we have that x^{**} is a stationary point of (12) by the optimization literature of [10]. By Proposition 4.3, x^{**} is a solution of (1). Clearly, it follows that $x^* = x^{**}$ by the uniform P-function property of F.

Thus we have proved that x^k converges to x^*. Next we show that Step 3 of Algorithm B is always successful eventually. Suppose x^k is sufficiently close to the solution point x^*. Then the generalized Newton method gives

$$||H(x^k)|| = ||V^k d^k||, \tag{15}$$

$$||x^k + d^k - x^*|| = o(||x^k - x^*||) = o(||d^k||). \tag{16}$$

By the semismoothness of H, (15), and (16), we have

$$\begin{aligned} ||H(x^k + d^k)|| &= ||H(x^*) + \bar{V}(x^k + d^k - x^*) + o(||x^k + d^k - x^*||)|| \\ &= ||\bar{V}(x^k + d^k - x^*) + o(||x^k + d^k - x^*||)||, \end{aligned}$$

where $\bar{V} \in \partial H(x^k + d^k)$. By Proposition 3.2, any generalized Jacobian of H at x^* is nonsingular. Moreover, there exists a constant $c > 0$ such that for any $V \in \partial H(x)$, $d \in \mathcal{R}^n$, and x sufficiently close to x^* we have

$$||Vd||/||d|| \ge c.$$

Consequently,

$$\frac{||H(x^k + d^k)||}{||H(x^k)||} = \frac{||\bar{V}(x^k + d^k - x^*) + o(||x^k + d^k - x^*||)||}{||V^k d^k||} = o(1).$$

TABLE 1
Number of iterations for Example 1.

n	8	16	32	64	128	256
Algorithm A	4	4	4	4	4	4
Algorithm B	4	4	4	4	4	4
Steepest descent method	37	31	30	32	32	31

The above argument means that Algorithm B always implements the Newton step after a finite number of iterations. The remaining results follow from Theorem 3.6.

(2) The proof is analogous to that of (1). We omit the detail. □

Remarks. (1) Algorithm B does not use the second-order derivative of the merit function but still achieves the second-order convergence results. (2) From a numerical point of view, the steepest descent method is not a good option. One may propose other algorithms for solving the NCP by combining the generalized Newton method and other global methods of unconstrained optimization.

5. Numerical results. In this section, we present some numerical experiments for three algorithms, i.e., the generalized Newton method (Algorithm A), the hybrid algorithm (Algorithm B), and the steepest descent method to $\theta(x)$. For Algorithm A we used $\theta(x) < \epsilon$ as the stopping criteria. Algorithm B stops if either $\theta(x) < \epsilon$ or $|| \bigtriangledown \theta(x)|| < \epsilon_1$. When F is not a uniform P-function, the stationary point of θ may not be a solution of the NCP. In this case, Algorithm B may converge to a stationary point of θ but which is not a solution of the NCP. Thus, the second stopping rule $|| \bigtriangledown \theta(x)|| < \epsilon_1$ can be necessary for some cases.

Throughout the computational experiments, the parameters used in the algorithms were set as $\epsilon = 10^{-6}$, $\epsilon_1 = 10^{-5}$, $\sigma = 10^{-4}$, $\rho = 0.5$, $\beta = 0.95$. All computational results were undertaken on a DEC 5000 workstation by MATLAB.

Example 1. This example was used by Geiger and Kanzow [12]. Let $F(x) = Mx + q$, where

$$M = \begin{pmatrix} 4 & -1 & 0 & \cdots & 0 \\ -1 & 4 & -1 & \cdots & 0 \\ 0 & -1 & 4 & \cdots & 0 \\ \vdots & \vdots & \vdots & \vdots & \vdots \\ 0 & 0 & 0 & \cdots & -1 \\ 0 & 0 & 0 & \cdots & 4 \end{pmatrix}, \quad q = (-1, \ldots, -1)^T.$$

Since F is strongly monotone on $\mathcal{R}^n$, the corresponding NCP has a unique solution. Table 1 lists the results for this example with initial point $x = (0, \ldots, 0)$ for different dimensions n.

Example 2. This example is also taken from Geiger and Kanzow [12]. Define $F(x) = Mx + q$, where

$$M = \text{diag}(1/n, 2/n, \ldots, 1), \qquad q = (-1, \ldots, -1)^T.$$

Again F is strongly monotone on $\mathcal{R}^n$. The corresponding strong monotonicity modulus depends on the dimension n and approaches zero when n tends to infinity. Table 2 gives the results for this example with starting point $x = (0, \ldots, 0)$ for different dimensions n.

From Tables 1 and 2, one observes that both the generalized Newton and the hybrid methods work quite well for Examples 1 and 2, respectively. However, the

TABLE 2
Number of iterations for Example 2.

n	8	16	32	64	128	256
Algorithm A	4	4	4	4	4	4
Algorithm B	4	4	4	4	4	4
Steepest descent method	384	1586	6465	$> 10^4$	$> 10^4$	$> 10^4$

TABLE 3
Number of iterations for Example 3.

Starting point	Algorithm A	Algorithm B
(0, 0, 0, 0)	13*	20*
(1, 1, 1, 1)	8*	8
(100, 100, 100, 100)	11*	7*
(10^5, 10^5, 10^5, 10^5)	11*	7*
(-10^5, -10^5, -10^5, -10^5)	11*	7*
(1234, 2345, 3456, 4567)	17*	53*

steepest descent method demonstrates very poor performance, especially for Example 2, in which the problem becomes ill conditioned as n increases. For this reason, we choose not to test the steepest descent method for the remaining examples.

Example 3. This example was used by Pang and Gabriel [35], Mangasarian and Solodov [26], and Kanzow [19] with four variables. Let

$$\begin{aligned} F_1(x) &= 3x_1^2 + 2x_1x_2 + 2x_2^2 + x_3 + 3x_4 - 6, \\ F_2(x) &= 2x_1^2 + x_1 + x_2^2 + 10x_3 + 2x_4 - 2, \\ F_3(x) &= 3x_1^2 + x_1x_2 + 2x_2^2 + 2x_3 + 9x_4 - 9, \\ F_4(x) &= x_1^2 + 3x_2^2 + 2x_3 + 3x_4 - 3. \end{aligned}$$

This example has one degenerate solution $(\frac{\sqrt{6}}{2}, 0, 0, \frac{1}{2})$ and one nondegenerate solution $(1, 0, 3, 0)$. This result is summarized in Table 3 using different starting points. The asterisk (*) denotes that the limit point generated by the algorithms is the degenerate solution, otherwise it is the nondegenerate solution.

Example 4. This example was tested by Kanzow [19] with five variables defined by

$$F_i(x) = 2\exp\left(\sum_{i=1}^{5}(x_i - i + 2)^2\right)(x_i - i + 2), \qquad 1 \le i \le 5.$$

This example has one degenerate solution, namely, $(0, 0, 1, 2, 3)$. The results are listed in Table 4 using different starting points. The asterisk denotes that the algorithm does not converge due to the inability of the computer to deal with very large number overflow.

Example 5. This example is a modification of Mathiesen [28] tested in [19]. Let

$$\begin{aligned} F_1(x) &= -x_2 + x_3 + x_4, \\ F_2(x) &= x_1 - (4.5x_3 + 2.7x_4)/(x_2 + 1), \\ F_3(x) &= 5 - x_1 - (0.5x_3 + 0.3x_4)/(x_3 + 1), \\ F_4(x) &= 3 - x_1. \end{aligned}$$

TABLE 4
Number of iterations for Example 4.

Starting point	Algorithm A	Algorithm B
(0, 0, 0, 0, 0)	20	20
(1, 1, 1, 1, 1)	18	*
(−1, −1, −1, −1, −1)	37	37
(2, 2, 2, 2, 2)	85	*
(−2, −2, −2, −2, −2)	61	61
(10, 10, 10, 10, 10)	*	*
(3, 2, 1, 2, 3)	1	1
(1, 0, 1, 3, 5)	20	6

TABLE 5
Number of iterations for Example 5.

Starting point	Algorithm A	Algorithm B
(1, 1, 1, 1)	3	3
(2, 2, 2, 2)	4	6
(−2 , −2, −2, −2)	4	4
(10, 10, 10, 10)	5	5
(− 10, −10, −10, −10)	6	*
(1234, 2345, 3456, 4567)	6	6

This example has infinitely many solutions $(\lambda, 0, 0, 0)$, where $\lambda \in [0, 3]$. For $\lambda = 1, 3$, the solutions are degenerate, and for $\lambda \in (0, 3)$ nondegenerate. The test results for Example 5 are listed in Table 5 using different starting points. For all starting points, the limit points generated by both the generalized Newton method and the hybrid method are the degenerate solutions if the algorithms converge. The asterisk denotes the algorithm fails due to discontinuity of F when $x_2 = -1$ or $x_3 = -1$.

Example 6. This example is problem 35 of Hock and Schittkowski [15] and was tested by Geiger and Kanzow [12]. The problem is defined by

$$\begin{aligned} \min \quad & f(x) = 9 - 8x_1 - 6x_2 - 4x_3 + 2x_1^2 + 2x_2^2 + x_3^2 + 2x_1x_2 + 2x_1x_3 \\ \text{s.t.} \quad & 3 - x_1 - x_2 - x_3 \geq 0, \\ & 0 \leq x_i, \quad i = 1, 2, 3. \end{aligned}$$

The original problem is a convex programming problem. Its KKT optimality conditions lead to a monotone complementarity problem with four variables. This example has one optimal solution $x = (4/3, 7/9, 4/9)$. The test results for Example 6 are listed in Table 6 using different initial starting points.

We next test some economic equilibrium problems with larger sizes, but with only standard starting points.

Example 7. This is a problem arising in a spatial equilibrium model with dimension 42; see [4, 5, 35] for more details. The numerical results for Example 7 are listed in Table 7 using the starting point $(0, \ldots, 0)$.

Example 8. This is a 50-variable traffic equilibrium problem with elastic demand; see [4, 5, 35] for more details. The numerical results for Example 8 are listed in Table 7 using the standard starting point defined below. $x_1, x_2, x_3, x_{10}, x_{11}, x_{20}, x_{21}, x_{22}, x_{29}, x_{30}, x_{40}, x_{45}$ are all ones; $x_{39}, x_{42}, x_{43}, x_{46}$ are equal to 7; $x_{41}, x_{47}, x_{48}, x_{50}$ are equal to 6; x_{44} and x_{49} are equal to 10; and all other elements are zeros.

TABLE 6
Number of iterations for Example 6.

Starting point	Algorithm A	Algorithm B
(0, 0, 0, 0)	4	4
(1, 1, 1, 1)	5	93
(−1, −1, −1, −1)	4	4
(10, 10, 10, 10)	5	5
(−10, −10, −10, −10)	4	4
(100, 100, 100, 100)	5	5
(−100, −100, −100, −100)	4	4
$(10^5, 10^5, 10^5, 10^5)$	5	5

TABLE 7
Number of iterations for Examples 7 *and* 8.

Problem	Starting point	Algorithm A	Algorithm B
Spatial eq.	$(0, \ldots, 0)$	14	14
Traffic eq.	as in Example 8	11	28

6. Conclusions. In this paper, we have presented an equivalent reformulation of the NCP as a system of nonsmooth equations. The generalized Newton method applied to the system of nonsmooth equations has been shown to be locally and Q-quadratically convergent. The hybrid method, which is a combination of the generalized Newton method and a minimization technique for solving the NCP, enjoys both global convergence and local Q-quadratic convergence under some conditions. From a numerical point of view, the steepest descent direction method used in the hybrid method is not a good option. However, the numerical results reported still show that the approaches presented in this paper are promising. Therefore, for global methods, numerical results should appear better than those in this paper if a better unconstrained minimization method is used combining with the generalized Newton method.

Very recently, De Luca, Facchinei, and Kanzow [5] proposed a global convergent algorithm for solving the NCP. They use the same local approach as our generalized Newton method. However, their global approach is quite different from ours. It appears that their global strategy is better than ours from the robustness point of view. In particular, it was pointed out [5] that the solution d^k of (8) is always a descent direction of the merit function θ if x^k is not a solution of the NCP. Regarding the nonsingularity of the generalized Jacobians of H, De Luca, Facchinei, and Kanzow [5] provided more general conditions which are weaker than uniform P-functions. As for the equivalence between the stationary point set of the merit function θ and the solution set of the NCP, the condition presented in Proposition 4.3 can be weakened substantially; see [5, 6, 7, 9] for more details. Tseng [42] proved that the R_0-function property, which is weaker than the uniform P-function property, of F is a sufficient condition for ensuring the compactness of the level set in Proposition 4.2. For some further refinements and more discussions, we refer the reader to [5, 9]. All those improvements show that the combination of nonsmooth equation and unconstrained optimization approaches for the solution of the NCP is very encouraging.

Acknowledgments. The authors are grateful to F. Facchinei, A. Fischer, C. Kanzow, and two anonymous referees for their helpful and detailed comments.

REFERENCES

[1] G. Auchmuty, *Variational principles for variational inequalities*, Numer. Funct. Anal. Optim., 10 (1989), pp. 863–874.
[2] B. Chen and P. T. Harker, *A continuation method for monotone variational inequalities*, Math. Programming, 69 (1995), pp. 237–253.
[3] B. Chen and P. T. Harker, *A Class of Smooth Approximations to Nonlinear Complementarity Problems*, Department of Management and Systems, Washington State University, Pullman, 1995, preprint.
[4] C. Chen and O. L. Mangasarian, *A class of smoothing functions for nonlinear and mixed complementarity problems*, Comput. Optim. Appl., 5 (1996), pp. 97–138.
[5] T. De Luca, F. Facchinei, and C. Kanzow, *A semismooth equation approach to the solution of nonlinear complementarity problems*, Math. Programming, to appear.
[6] F. Facchinei and J. Soares, *A new merit function for nonlinear complementarity problems and a related algorithm*, SIAM J. Optim., (1997), to appear.
[7] F. Facchinei and J. Soares, *Testing a new class of algorithms for nonlinear complementarity problems*, in F. Giannessi, ed., Variational Inequalities and Network Equilibrium Problems, Plenum Press, New York, 1994.
[8] A. Fischer, *A special Newton-type optimization method*, Optimization, 24 (1992), pp. 269–284.
[9] A. Fischer, *An NCP-function and its use for the solution of complementarity problems*, in Recent Advances in Nonsmooth Optimization, D. Du, L, Qi, and R. Womersley, eds., World Scientific Publishers, River Edge, NJ, 1995, pp. 88–105.
[10] R. Fletcher, *Practical Methods of Optimization*, 2nd ed., John Wiley, New York, 1987.
[11] M. Fukushima, *Equivalent differentiable optimization problems and descent methods for asymmetric variational inequality problems*, Math. Programming, 58 (1992), pp. 99–110.
[12] C. Geiger and C. Kanzow, *On the resolution of monotone complementarity problems*, Comput. Optim. Appl., 5 (1996), pp. 155–173.
[13] P. T. Harker and J.-S. Pang, *Finite-dimensional variational inequality and nonlinear complementarity problem: A survey of theory, algorithms and applications*, Math. Programming, 48 (1990), pp. 161–220.
[14] P. T. Harker and B. Xiao, *Newton method for the nonlinear complementarity problem: A B-differentiable equation approach*, Math. Programming (Series B), 48 (1990), pp. 339–357.
[15] W. Hock and K. Schittkowski, *Test Examples for Nonlinear Programming Codes*, Lecture Notes in Econom. and Math. Systems 187, Springer-Verlag, Berlin, 1981.
[16] H. Jiang, *Local properties of solutions of nonsmooth variational inequalities*, Optimization, 33 (1995), pp. 119–132.
[17] N. H. Josephy, *Newton's Method for Generalized Equations*, Technical summary report 1965, Mathematics Research Center, University of Wisconsin, Madison, WI, 1979.
[18] C. Kanzow, *Nonlinear complementarity as unconstrained optimization*, J. Optim. Theory Appl., 88 (1996), pp. 139–155.
[19] C. Kanzow, *Some equation-based methods for the nonlinear complementarity problem*, Optim. Methods Software, 3 (1994), pp. 327–340.
[20] M. Kojima, N. Megiddo, and S. Mizuno, *A general framework of continuation methods for complementarity problems*, Math. Oper. Res., 18 (1993), pp. 945–963.
[21] M. Kojima, N. Megiddo, and T. Noma, *Homotopy continuation methods for nonlinear complementarity problems*, Math. Oper. Res., 16 (1991), pp. 754–774.
[22] M. Kojima, T. Noma, and A. Yoshise, *Global convergence in interior-point algorithms*, Math. Programming, 65 (1994), pp. 43–72.
[23] M. Kojima and S. Shindo, *Extension of Newton and Quasi-Newton methods to systems of PC^1 equations*, J. Oper. Res. Soc. Japan, 29 (1986), pp. 352–374.
[24] B. Kummer, *Newton's method for nondifferentiable functions*, in Mathematical Research: Advances in Mathematical Optimization, J. Guddat, et al., eds., Akademie Verlag, Berlin, 1988, pp. 114–125.
[25] O. L. Mangasarian, *Equivalence of the complementarity problem to a system of nonlinear equations*, SIAM J. Appl. Math., 31 (1976), pp. 89–92.
[26] O. L. Mangasarian and M. V. Solodov, *Nonlinear complementarity as unconstrained and constrained minimization*, Math. Programming (Series B), 62 (1993), pp. 277–297.
[27] P. Marcotte and J. P. Dussault, *A note on a globally convergent Newton method for solving variational inequalities*, Oper. Res. Lett., 6 (1987), pp. 35–42.
[28] L. Mathiesen, *An algorithm based on a sequence of linear complementarity problems applied to a Walrasian equilibrium model: An example*, Math. Programming, 37 (1987), pp. 1–18.

[29] J. J. Moré, *Classes of function and feasibility conditions in nonlinear complementarity problems*, Math. Programming, 6 (1974), pp. 327–338.
[30] J. J. Moré, *Global Methods for Nonlinear Complementarity Problems*, Math. Oper. Res., 21 (1996), pp. 589–614.
[31] J. M. Ortega and W. C. Rheinboldt, *Iterative Solution of Nonlinear Equations in Several Variables*, Academic Press, New York, 1970.
[32] J.-S. Pang, *Newton's methods for B-differentiable equations*, Math. Oper. Res., 15 (1990), pp. 311–341.
[33] J.-S. Pang, *A B-differentiable equation based, globally and locally quadratically convergent algorithm for nonlinear programs, complementarity, and variational inequality problems*, Math. Programming, 51 (1991), pp. 101–131.
[34] J.-S. Pang, *Complementarity problems*, in Handbook of Global Optimization, R. Horst and P. Pardalos, eds., Kluwer Academic Publishers, Boston, 1994, pp. 271–338.
[35] J.-S. Pang and S. A. Gabriel, *NE/SQP: A robust algorithm for the nonlinear complementarity problem*, Math. Programming, 60 (1993), pp. 295–337.
[36] J.-S. Pang and L. Qi, *Nonsmooth equations: Motivation and algorithms*, SIAM J. Optim., 3 (1993), pp. 443–465.
[37] L. Qi, *Convergence analysis of some algorithms for solving nonsmooth equations*, Math. Oper. Res., 18 (1993), pp. 227–244.
[38] L. Qi and J. Sun, *A nonsmooth version of Newton's method*, Math. Programming, 58 (1993), pp. 353–368.
[39] D. Ralph, *Global convergence of damped Newton's method for nonsmooth equations, via the path search*, Math. Oper. Res., 19 (1994), pp. 352–389.
[40] S. Robinson, *Newton's method for a class of nonsmooth functions*, Set-Valued Anal., 2 (1994), pp. 291–305.
[41] P. K. Subramanian, *Gauss-Newton methods for the complementarity problem*, J. Optim. Theory Appl., 77 (1993), pp. 467–482.
[42] P. Tseng, *An infeasible path-following method for monotone complementarity problems*, SIAM J. Optim., 7 (1997), to appear.
[43] N. Yamashita and M. Fukushima, *On stationary points of the implicit Lagrangian for nonlinear complementarity problems*, J. Optim. Theory Appl., 84 (1995), pp. 653–663.

SIAM J. CONTROL OPTIM.
Vol. 35, No. 1, pp. 194–204, January 1997

010

FAMILIES OF SOLUTIONS OF MATRIX RICCATI DIFFERENTIAL EQUATIONS*

M. PAVON† AND D. D'ALESSANDRO‡

Abstract. The J. C. Willems–Coppel–Shayman geometric characterization of solutions of the algebraic Riccati equation (ARE) is extended to *asymmetric* Riccati *differential* equations with *time-varying coefficients*. The coefficients do not need to satisfy any *definiteness*, *periodicity*, or *system-theoretic condition*. More precisely, given any two solutions $X_1(t)$ and $X_2(t)$ of such equation on a given interval $[t_0, t_1]$, we show how to construct a family of solutions of the same equation of the form $X(t) = (I - \pi(t))X_1(t) + \pi(t)X_2(t)$, where π is a suitable matrix-valued function. Even when specialized to the case of X_1 and X_2 equilibrium solutions of a symmetric equation with constant coefficients, our results considerably extend the classical ones, as no further assumption is made on the pair X_1, X_2 and on the coefficient matrices.

Key words. asymmetric Riccati differential equation, families of solutions, geometric characterization, invariant subspaces, projection-preserving differential equation

AMS subject classifications. 34A05, 34A26, 34A34, 15A24, 49K15

PII. S0363012994276330

1. Families of solutions of the RDE.

Consider the asymmetric Riccati differential equation (RDE)

$$\dot{X} = AX + XB + XPX + Q, \tag{1.1}$$

where X is $m \times n$ and A, B, P, Q are continuous, matrix-valued functions with real entries on $[t_0, t_1]$ of dimension $m \times m, n \times n, n \times m$, and $m \times n$, respectively. As is well known, the symmetric version of (1.1), i.e., when $n = m$, $B = A^T$, $P = P^T$, $Q = Q^T$, plays a central role in many fields of applied mathematics, including optimal control and estimation, and has therefore been intensively studied. General Riccati equations such as (1.1) arise in the theory of differential games [3], in state-space solutions to H^∞ problems [10], in polynomial factorization [5], in problems of feedback control [1], and in the singular perturbation of boundary value problems [4]; see the introductions of [17, 3, 12] for further information. A further example is provided by equation (1.7) below, which is asymmetric even when (1.1) is symmetric.

All through this paper, X_1 and X_2 denote two fixed but arbitrary solutions of (1.1) on the time interval $[t_0, t_1]$. Moreover, let $\Delta_{12} := X_2 - X_1$. There exists a one-to-one correspondence between solutions of (1.1) and solutions of the homogeneous Riccati equation

$$\dot{\Delta} = A_{X_1}\Delta + \Delta B_{X_1} + \Delta P \Delta, \tag{1.2}$$

where $A_{X_1} := A + X_1 P$ and $B_{X_1} := B + PX_1$, given by $X \leftrightarrow \Delta = X - X_1$. Thus, all results below concerning solutions of (1.1) may also be viewed as results concerning

*Received by the editors October 31, 1994; accepted for publication (in revised form) November 10, 1995.

http://www.siam.org/journals/sicon/35-1/27633.html

†Dipartimento di Elettronica e Informatica, Università di Padova, via Gradenigo 6/A, and LADSEB-CNR, 35131 Padova, Italy (pavon@ladseb.pd.cnr.it). The research of this author at LADSEB-CNR was supported in part by CNR through a CNR-DFG Bilateral Project, and the research at the Institute of Theoretical Dynamics, University of California, Davis, was supported by Office of Naval Research grant USN-N00014-89-J-3135.

‡Dipartimento di Elettronica e Informatica, Università di Padova, via Gradenigo 6/A, 35131 Padova, Italy (daless@maya.dei.unipd.it).

solutions of (1.2), where the roles of X_1 and X_2 are played by the zero solution and Δ_{12}, respectively.

Jan Willems' classification of solutions of the ARE [23] was used in [14] to classify all output-induced minimal stochastic realizations of a given process. In [2, Theorem 8.3], this classification was extended to the nonstationary case. Its implications for the RDE, however, were not pursued there. Jan Willems' original derivation of the geometric parametrization of solutions of the ARE relied on first establishing a similarity relation involving two "extreme" closed-loop matrices [23, Lemma 8]. The latter result was generalized to the symmetric, nonsingular (i.e., Δ_{12} invertible), time-varying situation in [18, Theorem 5.5]. It can indeed be extended to our very general setting, and its consequences are far reaching.

LEMMA 1.1. *Let X be any solution of* (1.1) *on* $[t_0, t_1]$ *and let* $\Delta_i := X - X_i$, $i = 1, 2$. *Let* $\phi(\cdot,\cdot)$ *and* $\psi(\cdot,\cdot)$ *be the transition matrices corresponding to* $A_X := A + XP$ *and* $-B_X := -(B + PX)$, *respectively. Let* $\phi_i(\cdot,\cdot)$ *and* $\psi_i(\cdot,\cdot)$, $i = 1, 2$, *be the transition matrices corresponding to* $A_{X_i} := A + X_iP$ *and* $-B_{X_i} := -(B + PX_i)$, *respectively. Then, for* $i = 1, 2$ *and for all* s *and* t *in* $[t_0, t_1]$, *we have*

$$\Delta_i(t)\psi_i(t,s) = \phi(t,s)\Delta_i(s), \tag{1.3}$$

$$\Delta_i(t)\psi(t,s) = \phi_i(t,s)\Delta_i(s). \tag{1.4}$$

Proof. Notice that Δ_i satisfies

$$\begin{aligned}\dot{\Delta}_i &= A_{X_i}\Delta_i + \Delta_i B_{X_i} + \Delta_i P\Delta_i \\ &= A_X\Delta_i + \Delta_i B_{X_i} = A_{X_i}\Delta_i + \Delta_i B_X.\end{aligned} \tag{1.5}$$

From (1.5) it follows that

$$\begin{aligned}\frac{\partial(\Delta_i(t)\psi_i(t,s))}{\partial t} &= \dot{\Delta}_i(t)\psi_i(t,s) + \Delta_i(t)\frac{\partial\psi_i(t,s)}{\partial t} \\ = \dot{\Delta}_i(t)\psi_i(t,s) - \Delta_i(t)B_{X_i}\psi_i(t,s) &= A_X(t)\Delta_i(t)\psi_i(t,s).\end{aligned}$$

Hence, both sides of (1.3) satisfy

$$\frac{\partial W(t,s)}{\partial t} = A_X(t)W(t,s).$$

Since they coincide for $s = t$, they coincide everywhere. Exchanging the roles of X and X_i, we get (1.4) from (1.3). □

COROLLARY 1.2. $\Delta_i(t), i = 1, 2$, *has constant rank on* $[t_0, t_1]$.

Proof. By (1.3), $\Delta_i(t) = \phi(t, t_0)\Delta_i(t_0)\psi_i(t_0, t)$. □

COROLLARY 1.3. *Let X be any solution of* (1.1) *on* $[t_0, t_1]$, *and let* $i = 1, 2$. *Suppose that* $\ker\Delta_{12}(t_0) \subseteq \ker\Delta_i(t_0)$. *Then* $\ker\Delta_{12}(t) \subseteq \ker\Delta_i(t)$ *for all* $t \in [t_0, t_1]$.

Proof. Let $x \in R^n$ be such that $\Delta_{12}(t)x = 0$. By (1.3), we get $\Delta_{12}(t_0)\psi_i(t_0, t)x = 0$. Thus, $\psi_i(t_0, t)x$ is in the kernel of $\Delta_{12}(t_0)$. By hypothesis, $\Delta_i(t_0)\psi_i(t_0, t)x = 0$. The latter implies $\phi(t, t_0)\Delta_i(t_0)\psi_i(t_0, t)x = 0$. Using equation (1.3) again, we get $\Delta_i(t)x = 0$. □

Obviously, the above result holds true if t_0 is replaced by any other time s in $[t_0, t_1]$. Let us agree that all through the paper $\pi(t)$ denotes an $m \times m$ matrix function on $[t_0, t_1]$.

THEOREM 1.4. *The matrix function* $X(t) = (I - \pi(t))X_1(t) + \pi(t)X_2(t)$ *is a solution of* (1.1) *on* $[t_0, t_1]$ *if and only if* $\pi(t)$ *is a* C^1 *function satisfying*

$$\dot{\pi}\Delta_{12} = [A_{X_1}\pi - \pi A_{X_1} - \pi\Delta_{12}P(I - \pi)]\Delta_{12}. \tag{1.6}$$

Conversely, let $X(t)$ *be a solution of* (1.1) *on* $[t_0, t_1]$ *with* $\ker \Delta_{12}(t_0) \subseteq \ker \Delta_1(t_0)$ *where* $\Delta_1 = X - X_1$. *Then there exists a* C^1 *function* $\pi(t)$ *satisfying* (1.6) *such that* $X(t) = (I - \pi(t))X_1(t) + \pi(t)X_2(t)$. *Moreover, if* $\mathrm{Rank}\Delta_{12}(t_0) = m$, (1.6) *may be replaced by the* auxiliary Riccati differential equation (ARDE)

$$\dot{\pi} = A_{X_1}\pi - \pi A_{X_1} - \pi\Delta_{12}P(I - \pi), \tag{1.7}$$

and there is a one-to-one correspondence between solutions of (1.1) *and solutions of* (1.7).

Proof. Let $\mathcal{R}(X) := AX + XB + XPX + Q$. If $X(t) = (I - \pi(t))X_1(t) + \pi(t)X_2(t)$, we get

$$\begin{aligned}
\mathcal{R}(X) &= A[(I-\pi)X_1 + \pi X_2] + [(I-\pi)X_1 + \pi X_2]B \\
&\quad + [(I-\pi)X_1 + \pi X_2]P[(I-\pi)X_1 + \pi X_2] + Q \\
&= (I-\pi)\mathcal{R}(X_1) + \pi\mathcal{R}(X_2) - (I-\pi)AX_1 - (I-\pi)X_1PX_1 - \pi AX_2 \\
&\quad - \pi X_2PX_2 + A(I-\pi)X_1 + A\pi X_2 + [(I-\pi)X_1 + \pi X_2]P[(I-\pi)X_1 + \pi X_2] \\
&= (I-\pi)\mathcal{R}(X_1) + \pi\mathcal{R}(X_2) - \pi A\Delta_{12} + A\pi\Delta_{12} \\
&\quad - (I-\pi)X_1P\pi X_1 - \pi X_2P(I-\pi)X_2 + \pi X_2P(I-\pi)X_1 + (I-\pi)X_1P\pi X_2 \\
&= (I-\pi)\mathcal{R}(X_1) + \pi\mathcal{R}(X_2) + [-\pi A + A\pi + (I-\pi)X_1P\pi - \pi X_2P(I-\pi)]\Delta_{12} \\
&= (I-\pi)\mathcal{R}(X_1) + \pi\mathcal{R}(X_2) \\
&\quad + [-\pi A + A\pi + (I-\pi)X_1P\pi - \pi\Delta_{12}P(I-\pi) - \pi X_1P(I-\pi)]\Delta_{12} \\
&= (I-\pi)\mathcal{R}(X_1) + \pi\mathcal{R}(X_2) + [A_{X_1}\pi - \pi A_{X_1} - \pi\Delta_{12}P(I-\pi)]\Delta_{12}.
\end{aligned}$$

If π is of class C^1, it then follows that X is a solution of (1.1) if and only if (1.6) holds. Conversely, suppose that X is a solution of (1.1) on $[t_0, t_1]$ such that $\ker \Delta_{12}(t_0) \subseteq \ker \Delta_1(t_0)$. By Corollary 1.3, the inclusion $\ker \Delta_{12}(t) \subseteq \ker \Delta_i(t)$ holds for all $t \in [t_0, t_1]$. Then there exist $m \times m$-valued matrix functions $Z(t)$ such that

$$\Delta_1(t) = Z(t)\Delta_{12}(t) \tag{1.8}$$

for all $t \in [t_0, t_1]$. Notice that (1.8) already implies that $X(t) = (I - Z(t))X_1(t) + Z(t)X_2(t)$. Thus, the proof of the converse will be complete if we can show that among the functions Z satisfying (1.8) there is at least one $\tilde{Z}$ of class C^1. In that case, we can take $\pi = \tilde{Z}$. To this end, notice that, in view of Lemma 1.1, any function Z satisfying (1.8) also satisfies

$$\Delta_1(t)\psi_2(t, t_0) = Z(t)\phi_1(t, t_0)\Delta_{12}(t_0). \tag{1.9}$$

This leads us to introduce the function $\tilde{Z}$ defined by

$$\tilde{Z}(t) = \Delta_1(t)\psi_2(t, t_0)\Delta_{12}^{\#}(t_0)\phi_1(t_0, t),$$

where $\Delta_{12}^{\#}(t_0)$ denotes the Moore–Penrose pseudoinverse of $\Delta_{12}(t_0)$. The function $\tilde{Z}$ is clearly continuously differentiable. We show next that indeed $\tilde{Z}$ satisfies $\Delta_1(t) = \tilde{Z}(t)\Delta_{12}(t)$. Observe that the latter is equivalent to

$$\Delta_1(t)[I - \psi_2(t, t_0)\Delta_{12}^{\#}(t_0)\phi_1(t_0, t)\Delta_{12}(t)] = 0. \tag{1.10}$$

Now let Z be any function satisfying (1.8). Using (1.9) in (1.10), we see that the latter is equivalent to

$$Z(t)\phi_1(t, t_0)\Delta_{12}(t_0)\psi_2(t_0, t)[I - \psi_2(t, t_0)\Delta_{12}^{\#}(t_0)\phi_1(t_0, t)\Delta_{12}(t)] = 0.$$

Obtaining $\Delta_{12}(t)$ from Lemma 1.1 and using properties of transition matrices, it can be verified that the latter equation is in turn equivalent to

$$Z(t)\phi_1(t,t_0)\Delta_{12}(t_0)[I-\Delta_{12}^{\#}(t_0)\Delta_{12}(t_0)]\psi_2(t_0,t)=0. \tag{1.11}$$

Because of $\Delta(I-\Delta^{\#}\Delta)=0$, the preceding identity (1.11) is valid. Finally, suppose that $\mathrm{Rank}\Delta_{12}(t_0)=m$. By Corollary 1.2 the same is true for $\Delta_{12}(t)$, $t\in[t_0,t_1]$. The one-to-one map between the solution sets of (1.1) and (1.7) is then given by $\pi(t):=[X(t)-X_1(t)]\Delta_{12}^{-R}(t)$, where Δ_{12}^{-R} denotes any right inverse of Δ_{12}. □

Remark 1.5. Obviously, in Theorem 1.4 (and in the following), we could have considered combinations of the form $X(t)=X_1(t)(I-\sigma(t))+X_2(t)\sigma(t)$. The assumption for the converse part would then read $\ker\Delta_{12}(t_0)^T\subseteq\ker\Delta_1(t_0)^T$. Equation (1.6) would be replaced by the equation

$$\Delta_{12}\dot{\sigma}=\Delta_{12}[\sigma B_{X_1}-B_{X_1}\sigma-(I-\sigma)P\Delta_{12}\sigma].$$

Remark 1.6. Notice that if π_1 and π_2 are two C^1 functions generating the same solution X of (1.1) on $[t_0,t_1]$, i.e., $X(t)=X_1(t)+\pi_1(t)\Delta_{12}(t)=X_1(t)+\pi_2(t)\Delta_{12}(t)$, then necessarily $[\pi_1(t)-\pi_2(t)]\Delta_{12}(t)=0$ for all t in $[t_0,t_1]$. If $\Delta_{12}(t_0)$ admits a right inverse, then $\pi_1=\pi_2$.

At first sight, the correspondence between solutions of (1.1) and solutions of (1.6) or (1.7) established by Theorem 1.4 appears rather disappointing. Indeed, in the best case, we still have to deal with an asymmetric Riccati equation, the ARDE, with the only apparent advantage that π, A_{X_1}, and $\Delta_{12}P$ are all square $m\times m$-dimensional. Notice that solutions X_1 and X_2 of (1.1) correspond to the equilibrium solutions zero and identity, respectively, of (1.6) and (1.7). Nevertheless, the power of this connection will shortly be apparent. Indeed, (1.6) and (1.7) lend themselves naturally to a geometric characterization of a subclass of their solutions; see Theorems 2.3 and 2.5 below.

We conclude this section with a result relating different ϕ transition matrices. This result, which will not be needed in what follows, appears to be of interest for nonstationary stochastic realization [2]. Indeed, it extends a result for feedback matrices corresponding to different solutions of the symmetric ARE that was applied to stationary stochastic realization in [11, Lemma 4.1].

PROPOSITION 1.7. *Let X be any solution of* (1.1) *on* $[t_0,t_1]$. *If* $X(t)-X_1(t)=\Delta_1(t)=\pi(t)\Delta_{12}(t)$ *on* $[t_0,t_1]$ *for some function* π, *we have, in the notation of Lemma* 1.1,

$$\{\phi(t,s)\pi(s)-\pi(t)\phi_2(t,s)\}\Delta_{12}(s)=0, \tag{1.12}$$

$$\{\phi(t,s)(I-\pi(s))-(I-\pi(t))\phi_1(t,s)\}\Delta_{12}(s)=0. \tag{1.13}$$

If π *is projection valued, it follows that*

$$(I-\pi(t))\phi(t,s)\pi(s)\Delta_{12}(s)=0, \tag{1.14}$$

$$\pi(t)\phi(t,s)(I-\pi(s))\Delta_{12}(s)=0. \tag{1.15}$$

If, moreover, π *is* C^1, *the latter gives*

$$(I-\pi)(\dot{\pi}-A_X\pi)\Delta_{12}=0, \tag{1.16}$$

$$\pi(\dot{\pi}+A_X(I-\pi))\Delta_{12}=0. \tag{1.17}$$

Proof. Employing (1.3) twice, once for X and once for X_2, we get

$$\phi(t,s)\pi(s)\Delta_{12}(s) = \phi(t,s)\Delta_1(s) = \Delta_1(t)\psi_1(t,s)$$
$$= \pi(t)\Delta_{12}(t)\psi_1(t,s) = \pi(t)\phi_2(t,s)\Delta_{12}(s),$$

which is (1.12). Similarly, (1.13) is established. If π is a C^1, projection-valued function, differentiating (1.14) and the equation $\pi(t) = \pi(t)^2$ with respect to t, we get $[(I-\pi(t))A_X - \dot{\pi}(t)]\phi(t,s)\pi(s)\Delta_{12}(s) = 0$ and $\dot{\pi}(t)\pi(t) = (I-\pi(t))\dot{\pi}(t)$, respectively. Evaluating the first at $s = t$ and then using the second, we get (1.16). Similarly, we get (1.17) from (1.15). □

2. Geometric results. The first step in establishing a geometric characterization of certain families of solutions of (1.1) consists of rewriting (1.6) and (1.7). Simply rearranging terms, we get that these equations are equivalent to

$$[\dot{\pi} - (I-\pi)A_{X_1}\pi + \pi A_{X_2}(I-\pi)]\Delta_{12} = 0, \tag{2.1}$$
$$\dot{\pi} - (I-\pi)A_{X_1}\pi + \pi A_{X_2}(I-\pi) = 0, \tag{2.2}$$

where $A_{X_2} := A + X_2P = A_{X_1} + \Delta_{12}P$.

LEMMA 2.1. *If π is a projection for all times, i.e., $\pi(t) = \pi(t)^2$ for t in $[t_0,t_1]$, then it satisfies* (2.1) *if and only if it satisfies the system of equations*

$$(I-\pi)[\dot{\pi} - A_{X_1}\pi]\Delta_{12} = 0, \tag{2.3}$$
$$\pi[(I-\dot{\pi}) - A_{X_2}(I-\pi)]\Delta_{12} = 0. \tag{2.4}$$

Proof. Multiplying (2.1) on the left first by $(I-\pi)$ and then by π, we get (2.3) and (2.4), respectively. Conversely, obtaining $\pi\dot{\pi}\Delta_{12}$ from (2.4) and plugging it into (2.3), we get (2.1). □

Remark 2.2. Equations (2.3), (2.4) can be obtained from (1.16), (1.17), observing that

$$(I-\pi)A_{X_1} = (I-\pi)A_X,$$
$$\pi A_{X_2} = \pi A_X.$$

Equations (2.1), (2.2), (2.3), (2.4) enjoy a certain symmetry. Indeed, they are invariant under the permutation $\pi \leftrightarrow (I-\pi)$, $X_1 \leftrightarrow X_2$. Lemma 2.1 above singles out a subclass of solutions of (2.1) and, by Theorem 1.4, of (1.1). This subclass may also be described as the solutions on $[t_0,t_1]$ of the following *implicit system*:

$$[0, I-\pi, \pi]\dot{\pi}\Delta_{12} = [\pi - \pi^2, (I-\pi)A_{X_1}\pi\Delta_{12}, -\pi A_{X_2}(I-\pi)\Delta_{12}]. \tag{2.5}$$

The following result provides a geometric characterization of the *projection-valued* solutions of (2.1). The question of *existence* of such solutions will be addressed in Theorem 2.7 below.

THEOREM 2.3. *Let $X(t) = (I-\pi(t))X_1(t) + \pi(t)X_2(t)$ be a solution of* (1.1) *on $[t_0,t_1]$. Let $M(t) := \pi(t)\Delta_{12}(t)R^n$ and $N(t) := (I-\pi(t))\Delta_{12}(t)R^n$. Then, for s and t in $[t_0,t_1]$, we have*

$$M(t) = \phi_1(t,s)M(s), \tag{2.6}$$
$$N(t) = \phi_2(t,s)N(s). \tag{2.7}$$

Moreover, we also have

$$M(t) = \phi(t,s)M(s), \tag{2.8}$$

$$N(t) = \phi(t,s)N(s). \tag{2.9}$$

Conversely, Let $\{M(t)\}$ *and* $\{N(t)\}$, $t \in [t_0, t_1]$, *be two families of subspaces of* R^m *providing a direct sum decomposition of* $\Delta_{12}(t)R^n$. *Let* π *be a* C^1 *function such that* $\pi(t)x = x\ \forall x \in M(t)$ *and* $\pi(t)y = 0\ \forall y \in N(t)$. *If* (2.6), (2.7) *hold for all* s *and* t *in* $[t_0, t_1]$, *then* $X(t) = (I - \pi(t))X_1(t) + \pi(t)X_2(t)$ *is a solution of the* RDE (1.1) *on* $[t_0, t_1]$.

Proof. By Lemma 1.1, $\Delta_1(t) = \phi_1(t,s)\Delta_1(s)\psi(s,t)$. Replacing $\Delta_1(t)$ with $\pi(t)\Delta_{12}(t)$, we get $\pi(t)\Delta_{12}(t)R^n = \phi_1(t,s)\pi(s)\Delta_{12}(s)R^n$, namely (2.6) holds. Formula (2.7) is proven similarly. Lemma 1.1 also gives $\Delta_1(t) = \phi(t,s)\Delta_1(s)\psi_1(s,t)$. The same argument as above then gives (2.8). Similarly, (2.9) is established. To prove the converse, notice that (2.6), (2.7) imply

$$[I - \pi(t)]\phi_1(t,s)\pi(s)\Delta_{12}(s) = 0, \tag{2.10}$$

$$\pi(t)\phi_2(t,s)[I - \pi(s)]\Delta_{12}(s) = 0. \tag{2.11}$$

Evaluating the derivatives of (2.10) and (2.11) with respect to t on the diagonal $t = s$, we get (2.3) and (2.4). The latter imply that (1.6) holds, and consequently X is a solution of (1.1). □

Remark 2.4. Notice that the first half of the theorem holds for *any* solution X of (1.1) of the form $X(t) = (I - \pi(t))X_1(t) + \pi(t)X_2(t)$, namely even when π is not projection valued. In that case, however, the spaces $M(t)$ and $N(t)$ do not need to form a direct sum. Observing once more that $X - X_1 = \pi\Delta_{12}$ and $X_2 - X = (I - \pi)\Delta_{12}$, we also see that the spaces $M(t)$ and $N(t)$ are *uniquely determined* by the solution X and do not depend on the particular projection π used in the definition.

In the important case where Δ_{12} has full row rank, Theorem 2.3 reads as follows.

THEOREM 2.5. *Assume that* $\Delta_{12}(t_0)$ *has full row rank. Let* $X(t) = (I - \pi(t))X_1(t) + \pi(t)X_2(t)$ *be a solution of* (1.1) *on* $[t_0, t_1]$. *Let* $M(t)$ *and* $N(t)$ *denote the range of* $\pi(t)$ *and the range of* $(I - \pi(t))$, *respectively. Then, for* s *and* t *in* $[t_0, t_1]$, *relations* (2.6), (2.7), (2.8) *and* (2.9) *hold true. Conversely, let* $\pi(\cdot)$ *be a* C^1, *projection-valued function on* $[t_0, t_1]$, *and let* $M(t)$ *and* $N(t)$ *denote the range of* $\pi(t)$ *and the range of* $(I - \pi(t))$, *respectively. If the propagation relations* (2.6) *and* (2.7) *hold for all* s *and* t *in* $[t_0, t_1]$, *then* $X(t) = (I - \pi(t))X_1(t) + \pi(t)X_2(t)$ *is a solution of the* RDE (1.1) *on the same time interval.*

Theorems 2.3 and 2.5 provide the desired geometric characterization of a subclass of solutions of the ARDE (2.2) and, consequently, of the RDE (1.1). Notice that, in the case $m = n$, Remark 2.4 gives that the first half of Theorem 2.5 applies to *any* solution of (1.1) on $[t_0, t_1]$. Indeed, in this case, $\Delta_{12}(t)$ is invertible at all times, and $\ker \Delta_{12}(t) \subseteq \ker \Delta_1(t)$ is trivially satisfied. Hence, any solution X of (1.1) can be expressed as $X(t) = (I - \pi(t))X_1(t) + \pi(t)X_2(t)$. For the purpose of immediate comparison, we state below Jan Willems' classical result; cf. also [6, 19, 20, 21, 13] (the latter should also be compared with Theorems 3.3 and 4.2 below).

THEOREM 2.6. *In equation* (1.1), *let* $n = m$, $B = A^T$, $P = P^T$, $Q = Q^T$. *Suppose moreover that* P *is negative semidefinite and that the pair* (A, P) *is reachable. Let* X_- *and* X_+ *denote two symmetric equilibrium solutions of* (1.1) *such that the corresponding closed-loop matrices* $A_- := A + X_-P$ *and* $A_+ := A + X_+P$ *have all their eigenvalues in the closed right and left half-planes, respectively. Suppose that*

$\Delta := X_+ - X_-$ is positive definite. Then X is another symmetric equilibrium solution of (1.1) *if and only if it can be expressed as*

$$X = (I - \pi)X_- + \pi X_+$$

where π projects onto an A_--invariant subspace and $I - \pi$ projects onto an A_+-invariant subspace.

We now turn to the question of *existence of projection-valued solutions* of (2.2) ((1.7)) (equivalently, of solutions of the implicit system (2.5) if $\Delta_{12}(t_0)$ has full row rank). The following remarkable result basically says that (2.2) is a *projection-preserving differential equation.*

THEOREM 2.7. *Let π be a solution of* (2.2) *on $[t_0, t_1]$. Suppose that $\pi(t_0)$ is a projection. Then $\pi(t)$ is a projection for all t in $[t_0, t_1]$.*

Proof. Let us rewrite (2.2) as

$$\dot{\pi} = A_{X_1}\pi - \pi A_{X_2} + \pi\Delta_{12}P\pi.$$

Then

$$\begin{aligned}\frac{d\pi^2}{dt} = \dot{\pi}\pi + \pi\dot{\pi} &= [A_{X_1}\pi - \pi A_{X_2} + \pi\Delta_{12}P\pi]\pi + \pi[A_{X_1}\pi - \pi A_{X_2} + \pi\Delta_{12}P\pi] \\ &= A_{X_1}\pi^2 - \pi A_{X_2}\pi + \pi\Delta_{12}P\pi^2 + \pi A_{X_1}\pi - \pi^2 A_{X_2} + \pi^2\Delta_{12}P\pi.\end{aligned}$$

Hence,

$$\frac{d(\pi^2 - \pi)}{dt} = A_{X_1}(\pi^2-\pi)-(\pi^2-\pi)A_{X_2}-(\pi^2-\pi)\Delta_{12}P(\pi^2-\pi)+\pi^2\Delta_{12}P\pi^2-\pi\Delta_{12}P\pi.$$

Adding and subtracting the quantity $\pi^2\Delta_{12}P\pi$ in the right-hand side and rearranging terms, we finally get

$$\frac{d(\pi^2 - \pi)}{dt} = (A_{X_1}+\pi^2\Delta_{12}P)(\pi^2-\pi)-(\pi^2-\pi)(A_{X_2}-\Delta_{12}P\pi)-(\pi^2-\pi)\Delta_{12}P(\pi^2-\pi).$$

Let $F_1 := A_{X_1} + \pi^2\Delta_{12}P$ and $F_2 := A_{X_2} - \Delta_{12}P\pi$. It follows that, if $\pi(t)$ is a solution of (2.2), then, on the same time interval, $\pi^2 - \pi$ is a solution of the homogeneous Riccati equation

$$\dot{X} = F_1X - XF_2 - X\Delta_{12}PX, \tag{2.12}$$

and F_1 and F_2 are there bounded. Since $\pi^2(t_0) - \pi(t_0) = 0$, by uniqueness of the solution of equation (2.12) starting at zero, it follows that $\pi^2(t) - \pi(t) = 0$ on all of $[t_0, t_1]$. □

The above proof actually establishes an amplification of Theorem 2.7. We record it below because it is of interest on its own.

PROPOSITION 2.8. *Let A_1 and A_2 be $m \times m$ continuous matrix functions on $[t_0, t_1]$. Let Y be an $m \times m$ matrix function solving the homogeneous Riccati equation*

$$\dot{Y} = A_1Y - YA_2 + Y(A_2 - A_1)Y \tag{2.13}$$

on $[t_0, t_1]$. If there exists a time $\bar{t} \in [t_0, t_1]$ such that $Y(\bar{t}) = Y(\bar{t})^2$, then $Y(t) = Y(t)^2$ on all of $[t_0, t_1]$.

Remark 2.9. Notice that $Y_1 \equiv 0$ and $Y_2 \equiv I$ are two equilibrium solutions of (2.13). Also notice that the corresponding closed-loop matrices are $A_1 + 0(A_2 - A_1) = A_1$

and $A_1 + I(A_2 - A_1) = A_2$. Now let Y be as in the proposition above—namely, a projection-valued solution of (2.13)—and let $M(t)$ and $N(t)$ be the range spaces of $Y(t)$ and $I - Y(t)$, respectively. Then, by Theorem 2.5, the propagation properties (2.6) and (2.7) hold true, where ϕ_1 and ϕ_2 are the transition matrices corresponding to A_1 and A_2, respectively. Finally, if A_1 and A_2 are constant and $Y = Y^2$ is an equilibrium solution of (2.13), Y projects onto a subspace invariant for A_1 along a subspace invariant for A_2.

Remark 2.10. The geometric results of this section provide a procedure to produce new solutions of (1.1). For instance, in the full-rank case, let π_0 be any projection, and let M_0 and N_0 be the ranges of π_0 and $I-\pi_0$, respectively. Define $M(t) := \phi_1(t, t_0)M_0$ and $N(t) := \phi_2(t, t_0)N_0$. Let $\bar{t}$ be the largest time such that for $t_0 \le t < \bar{t}$, $M(t)$ and $N(t)$ give a direct sum decomposition of R^m (by continuity, $\bar{t} > 0$). Let $\pi(t)$ be the projection such that $M(t)$ and $N(t)$ are the ranges of $\pi(t)$ and $I - \pi(t)$, respectively. Then π solves (2.2) and $X = (I-\pi)X_1 + \pi X_2$ solves (1.1) on $[t_0, \bar{t})$. Using an explicit expression for π in terms of bases for its range and the range of $I - \pi$, it is easily seen that $\pi(t)$ becomes unbounded as t tends to $\bar{t}$. If $\Delta_{12}(t_0)$ has full row rank, it follows that the corresponding solution $X(\cdot)$ has a finite escape time (see, e.g., [16, 7, 8]) at $t = \bar{t}$.

We conclude the section with an example that illustrates Remark 2.10 as well as Proposition 2.8 and Remark 2.9.

Example 2.11. Consider equation (2.13) with $m = 2$ and

$$A_1 = \begin{pmatrix} 1 & 0 \\ 0 & 0 \end{pmatrix}, \quad A_2 = \begin{pmatrix} 0 & 1 \\ 0 & 0 \end{pmatrix}.$$

Choose as reference solutions $Y_1 = 0$ and $Y_2 = I$, and let π_0 be given by

$$\pi_0 = \begin{pmatrix} 1 & 0 \\ 0 & 0 \end{pmatrix}.$$

Clearly π_0 is a projection, in fact an orthogonal projection. We have that $M_0 = \binom{R}{0}$ and $N_0 = \binom{0}{R}$. Next notice that the transition matrices $\phi_1(t, s)$ and $\phi_2(t, s)$ are given here by

$$\phi_1(t,s) = e^{A_1(t-s)} = \begin{pmatrix} e^{t-s} & 0 \\ 0 & 1 \end{pmatrix}, \quad \phi_2(t,s) = e^{A_2(t-s)} = \begin{pmatrix} 1 & t-s \\ 0 & 1 \end{pmatrix}.$$

Hence, $M(t) = \phi_1(t, t_0)M_0$ is the span of the vector $\binom{1}{0}$ and $N(t) = \phi_2(t, t_0)N_0$ is the span of the vector $\binom{t-t_0}{1}$. Notice that $M(t)$ and $N(t)$ provide a direct sum decomposition of R^2 for all $t \ge t_0$. The projection $\pi(t)$ with range $M(t)$ and kernel $N(t)$ is given by

$$\pi(t) = \begin{pmatrix} 1 & t_0 - t \\ 0 & 0 \end{pmatrix}.$$

The corresponding solution of (2.13) is $Y(t) = 0 + \pi(t)(I - 0) = \pi(t)$, namely $\pi(t)$ itself. This is no surprise. Since $Y(t_0) = 0 + \pi_0(I - 0) = \pi_0$ is a projection, Theorem 2.7 implies that $Y(t)$ has to be a projection for all t. Notice that $\pi(t)$ is unbounded as t tends to infinity. This is possible because $\pi(t)$ for $t > 0$ is not an *orthogonal* projection, although π_0 is orthogonal.

3. Geometric results: The case where X_1 and X_2 are equilibrium solutions. All through this section we assume that X_1 and X_2 are equilibrium solutions of (1.1). The coefficients A, B, P, and Q may still be time varying.

PROPOSITION 3.1. *Let X be an equilibrium solution of* (1.1) *and let $\Delta_i = X - X_i$, $i = 1, 2$. Then for all t in $[t_0, t_1]$,*

$$A_X(t)\Delta_i = -\Delta_i B_{X_i}(t), \tag{3.1}$$

$$A_{X_i}\Delta_i = -\Delta_i B_X(t). \tag{3.2}$$

It follows that if $(\xi(t), \lambda(t))$ is an eigenvector-eigenvalue pair for $B_{X_i}(t)$ so that $B_{X_i}(t)\xi(t) = \lambda(t)\xi(t)$, then either $\Delta_i\xi(t) = 0$ or $(\Delta_i\xi(t), -\lambda(t))$ is an eigenvector-eigenvalue pair for $A_X(t)$. Similarly, it follows for $A_{X_i}(t)$ and $B_X(t)$. If Δ_i admits a right inverse, we get the relations

$$A_X(t) = -\Delta_i B_{X_i}(t)\Delta_i^{-R},$$
$$A_{X_i}(t) = -\Delta_i B_X(t)\Delta_i^{-R}.$$

In particular, if $m = n$ and Δ_{12} is invertible, we have

$$A_{X_2}(t) = -\Delta_{12} B_{X_1}(t)\Delta_{12}^{-1}. \tag{3.3}$$

Proof. Relations (3.1) and (3.2) are a consequence of (1.5). □

Once more, we compare (3.3) with the corresponding classical result. In the notation of Theorem 2.6, let $X_1 = X_-$ and $X_2 = X_+$. Then (3.3) reads $A_+ = -\Delta A_-^T \Delta^{-1}$ which is precisely [23, Lemma 8]. Let us now assume that the coefficients of (1.1) are constant. Theorems 1.4 and 2.3 yield the following result.

THEOREM 3.2. *Let $X = (I - \pi)X_1 + \pi X_2$ be an equilibrium solution of the RDE* (1.1). *Let $M := \pi\Delta_{12}R^n$ and $N := (I - \pi)\Delta_{12}R^n$. Then M is an invariant subspace for A_{X_1} and N is an invariant subspace for A_{X_2}. Moreover, M and N are both invariant for A_X. Conversely, let M and N be two subspaces of R^m providing a direct sum decomposition of $\Delta_{12}R^n$. Let π be an $m \times m$ matrix such that $\pi x = x$ for any x in M and $\pi y = 0$ for any y in N. If M is an invariant subspace for A_{X_1} and N is an invariant subspace for A_{X_2}, then $X = (I - \pi)X_1 + \pi X_2$ is an equilibrium solution of the RDE* (1.1).

Once more, we state independently the result in the case when Δ_{12} has full row rank.

THEOREM 3.3. *Suppose Δ_{12} has full row rank and let X be an equilibrium solution of* (1.1). *Assume that* $\ker \Delta_{12} \subseteq \ker \Delta_1$. *Then there exists an $m \times m$ matrix π such that $X = (I - \pi)X_1 + \pi X_2$. Moreover, the range M of π is invariant for A_{X_1} and for A_X, and the range N of $I - \pi$ is invariant for A_{X_2} and for A_X. Conversely, if π is any oblique projection onto a subspace invariant for A_{X_1} along a subspace invariant for A_{X_2}, then $X = (I - \pi)X_1 + \pi X_2$ satisfies* (1.1).

In order to compare this result with Theorem 2.6, notice that if $m = n$ and Δ_{12} has full rank, the condition $\ker \Delta_{12} \subseteq \ker \Delta_1$ is always satisfied. The additional assumptions of Theorem 2.6 permit us to conclude that if $X = (I - \pi)X_1 + \pi X_2$ is an equilibrium solution of (1.1), π is always a projection.

4. The symmetric Riccati equation. We finally consider the symmetric case where $n = m$, $B = A^T$, $P = P^T$, $Q = Q^T$ but return to the nonequilibrium situation. Equation (1.1) is now

$$\dot{X} = AX + XA^T + XPX + Q. \tag{4.1}$$

We also assume that the two reference solutions X_1 and X_2 take values in the symmetric matrices. Hence, $\Delta_{12}(t)$ is also symmetric at all times. It is then natural to restrict our attention to *symmetric* solutions of (4.1).

LEMMA 4.1. $\phi_2(t,s)\Delta_{12}(s) = \Delta_{12}(t)\phi_1(s,t)^T$.

Proof. By Lemma 1.1, $\phi_2(t,s)\Delta_{12}(s) = \Delta_{12}(t)\psi_1(t,s)$. The conclusion now follows observing that $B_{X_1} = A_{X_1}^T$ implies that $\psi_1(t,s) = \phi_1(s,t)^T$. □

For the sake of simplicity, we only give the main result in the case where Δ_{12} is nonsingular.

THEOREM 4.2. *Let X_1 and X_2 be any two symmetric solutions of* (4.1) *on $[t_0,t_1]$ such that $\Delta_{12}(t_0)$ is invertible. Let $X(t) = (I-\pi(t))X_1(t)+\pi(t)X_2(t)$ be a symmetric solution of* (4.1) *on $[t_0,t_1]$. Let $M(t)$ and $N(t)$ denote the range of $\pi(t)$ and the range of $(I-\pi(t))$, respectively. Then for s and t in $[t_0,t_1]$ the following relations hold true:*

$$\pi(t)\Delta_{12}(t) = \Delta_{12}(t)\pi(t)^T, \tag{4.2}$$

$$(I-\pi(t))\phi_1(t,s)\pi(s) = 0. \tag{4.3}$$

Conversely, let π be a C^1, projection-valued function satisfying for all s and t in $[t_0,t_1]$ (4.2), (4.3). *Then $X(t) = (I-\pi(t))X_1(t)+\pi(t)X_2(t)$ is also a symmetric solution of the RDE* (4.1) *on $[t_0,t_1]$.*

Proof. Let $X(t) = (I-\pi(t))X_1(t)+\pi(t)X_2(t) = X_1(t)+\pi(t)\Delta_{12}(t)$ be a symmetric solution of (4.1) on $[t_0,t_1]$. The symmetry of $X(t)$ implies that (4.2) must hold. Let $M(t)$ and $N(t)$ denote the range of $\pi(t)$ and of $I-\pi(t)$, respectively. By Theorem 2.5, we have $M(t) = \phi_1(t,s)M(s)$ from which (4.3) follows. Conversely, suppose that (4.2) and (4.3) are verified. From (4.3) we get $\phi_1(t,s)M(s) \subseteq M(t)$. Exchanging the roles of s and t, we see that equality, i.e., equation (2.6), must hold. Now, multiplying equation (4.3) (with s and t exchanged) by $\Delta_{12}(s)^{-1}$ on the left and by $\Delta_{12}(t)$ on the right we get

$$\Delta_{12}(s)^{-1}(I-\pi(s))\phi_1(s,t)\pi(t)\Delta_{12}(t) = 0. \tag{4.4}$$

Transposing (4.4) and using (4.2) twice, we get

$$\pi(t)\Delta_{12}(t)\phi_1(s,t)^T\Delta_{12}(s)^{-1}(I-\pi(s)) = 0.$$

The latter equation, together with Lemma 4.1, now gives equation (2.7). The conclusion now follows from Theorem 2.5. □

5. Closing comments. As is well known, the Riccati differential equation may be viewed as the description in local coordinates of the restriction to a subset of the Lagrangian Grassmannian manifold $\mathcal{L}$ of a vector field on $\mathcal{L}$; see [15, 19]. Our results may then be readily interpreted in that setting. In fact, some may be also directly derived in that setting; see [8], where the case of $l \geq 2$ reference solutions $X_1, X_2, \ldots, X_l$, is also considered (see also [17, Theorem 4]). Similar results may also be derived in the discrete-time setting [9]. Alternative representation formulas for solutions of (1.1) have been proposed in [22] and references therein.

The classification of the solutions of the ARE via invariant subspaces of the Hamiltonian matrix has the disadvantage, when compared with the J. C. Willems classification, that the invariant subspaces need to be J-neutral and complementary to the subspace Span $\binom{0}{I}$; see [13, pp. 67–68]. In [19] it was observed that the disadvantages of Jan Willems method are that, contrary to the Hamiltonian matrix method, it does not lead naturally to a concept of solution at infinity (phenomenon of

the finite escape time; see, e.g., [16, 7]) and it does not have an obvious generalization to the nonsymmetric Riccati equation. Whereas the first disadvantage persists, we observe that this paper has completely removed the second.

Acknowledgment. This paper has considerably profited from the detailed comments of an anonymous reviewer who went so far as to produce a more elegant proof of Theorem 1.4. His or her help is gratefully acknowledged.

REFERENCES

[1] B. D. O. Anderson, *The testing of optimality of linear systems*, Internat. J. Control, 4 (1966), pp. 29–40.
[2] F. Badawi, A. Lindquist, and M. Pavon, *On the Mayne-Fraser smoothing formula and stochastic realization theory for nonstationary linear stochastic systems*, in Proc. 18th IEEE CDC Conf., Fort Lauderdale, FL, Dec. 1979, pp. 505–510A.
[3] T. Başar, *Generalized Riccati equations in dynamic games*, in The Riccati Equation, S. Bittanti, A. Laub, and J. C. Willems, eds., Springer-Verlag, New York, 1989, pp. 293–333.
[4] K. W. Chang, *Singular perturbation of a general boundary value problem*, SIAM J. Math. Anal., 3 (1972), pp. 520–526.
[5] D. J. Clements and B. D. O. Anderson, *Polynomial factorization via the Riccati equation*, SIAM J. Appl. Math., 31 (1976), pp. 179–205.
[6] W. A. Coppel, *Matrix quadratic equations*, Bull. Austral. Math. Soc., 10 (1974), pp. 377–401.
[7] P. E. Crouch and M. Pavon, *On the existence of solutions of the Riccati differential equation*, Systems Control Lett., 9 (1987), pp. 203–206.
[8] D. D'Alessandro, *Invariant manifolds and projective combinations of solutions of the Riccati differential equation*, Linear Algebra Appl., to appear.
[9] D. D'Alessandro, *A superposition theorem for solutions of the Riccati difference equation*, J. Math. Systems Estim. Control, to appear.
[10] J. Doyle, K. Glover, P. Khargonekar, and B. Francis, *State-space solutions to standard H_2 and H_∞ control problems*, IEEE Trans. Automat. Control, 34 (1989), pp. 831–847.
[11] L. Finesso and G. Picci, *A characterization of minimal square spectral factors*, IEEE Trans. Automat. Control, 27 (1982), pp. 122–127.
[12] G. Freiling and G. Jank, *Nonsymmetric matrix Riccati equations*, Z. Anal. Anwendungen, 14 (1995), pp. 259-284.
[13] V. Kučera, *Algebraic Riccati equation: Hermitian and definite solutions*, in The Riccati Equation, S. Bittanti, A. Laub, and J. C. Willems, eds., Springer-Verlag, New York, 1989, pp. 53–88.
[14] A. Lindquist and G. Picci, *On the stochastic realization problem*, SIAM J. Control Optim., 17 (1979), pp. 365–389.
[15] C. F. Martin, *Grassmannian manifolds, Riccati equations and feedback invariants of linear systems*, in Geometrical Methods for the Theory of Linear Systems, C. I. Byrnes and C. F. Martin, eds., Reidel, Dordrecht, the Netherlands, 1980.
[16] C. F. Martin, *Finite escape time for Riccati differential equations*, Systems Control Letters, 1 (1981), pp. 127–131.
[17] J. Medanic, *Geometric properties and invariant manifolds of the Riccati equation*, IEEE Trans. Automat. Control, 27 (1982), pp. 670–677.
[18] M. Pavon, *Optimal interpolation for linear stochastic systems*, SIAM J. Control Optim., 22 (1984), pp. 618-629.
[19] M. A. Shayman, *Geometry of the algebraic Riccati equation*, Parts I and II, SIAM J. Control Optim., 21 (1983), pp. 375–394 and pp. 395–409.
[20] M. A. Shayman, *On the phase portrait of the matrix Riccati equation arising from the periodic control problem*, SIAM J. Control Optim., 23 (1985), pp. 717–751.
[21] M. A. Shayman, *A geometric view of the matrix Riccati equation*, in The Riccati Equation, S. Bittanti, A. Laub, and J. C. Willems, eds., Springer-Verlag, New York, 1989.
[22] M. Sorine and P. Winternitz, *Superposition laws for solutions of differential matrix Riccati equations arising in control theory*, IEEE Trans. Automat. Control, 30 (1985), pp. 266–272.
[23] J. C. Willems, *Least squares stationary optimal control and the algebraic Riccati equation*, IEEE Trans. Automat. Control, 16 (1971), pp. 621–634.

SIAM J. CONTROL OPTIM.
Vol. 35, No. 1, pp. 205–227, January 1997

011

SUFFICIENT OPTIMALITY CONDITIONS FOR OPTIMAL CONTROL SUBJECT TO STATE CONSTRAINTS*

K. MALANOWSKI†

Abstract. Strong second-order sufficient optimality conditions (SSC) are derived for optimal control problems of systems described by nonlinear ODEs subject to mixed control-state and pure state constraints. The obtained SSC are expressed in terms of a modified Legendre–Clebsch condition and the associated Riccati equation. The role of SSC in stability analysis of solutions to parametric optimal control problems is briefly discussed.

Key words. optimal control, nonlinear ordinary differential equations, control and state constraints, second-order sufficient optimality conditions, Riccati equation, parametric optimal control, stability analysis

AMS subject classifications. 49K15, 49K30, 49K40

PII. S0363012994267637

1. Introduction.

The concept of *second-order sufficient optimality conditions* in mathematical programming has been developed for many years. The well-known condition in finite-dimensional programs requires that the Hessian of Lagrangian is positive definite on the critical cone.

However, it was shown by Maurer and Zowe [18] that, in general, this condition is no longer sufficient in infinite-dimensional situations. Instead, Maurer and Zowe derived a stronger coercivity condition (cf. Theorem 5.6 in [18]), which provides some "margin of freedom" and ensures sufficiency of optimality. This concept has been extended by Maurer in [16] to optimization problems with the so-called *two-norm discrepancy* typical for nonlinear optimal control.

Recently, in Dontchev and Hager [1] and in Dontchev et al. [2] this approach has been further developed using three different norms and applied to control constrained optimal control problems.

Along with the second-order sufficient optimality conditions, the concept of *strong second-order sufficient optimality conditions* (SSC) has been introduced and analyzed. For finite-dimensional mathematical programs, SSC require that the Hessian of Lagrangian is positive definite on the affine hull of the critical cone (cf. [26]). It turnes out that SSC play an important role in stability analysis of solutions to parametric programming problems (cf. [8, 9, 25, 26]).

Also in optimal control problems, SSC require that the Hessian of Lagrangian is coercive on the subspace orthogonal to the gradients of some active constraints. Certainly, it is desirable that this subspace is narrow, so in its construction we would like to include as many constraints as possible.

Attempts have been made to express the coercivity condition in the form similar to that known in calculus of variations, which involves the so-called *strengthened Legendre–Clebsch condition* and *strengthened Jacobi condition* (cf., e.g., [4]). This last condition can be expressed in terms of the existence of a bounded solution to a

*Received by the editors May 13, 1994; accepted for publication (in revised form) November 14, 1995. This research was supported by grant 3 P403 002 05 from the Polish State Committee of Scientific Research (Komitet Badań Naukowych).

http://www.siam.org/journals/sicon/35-1/26763.html

†Systems Research Institute, Polish Academy of Sciences, ul. Newelska 6, 01–447 Warsaw, Poland (kmalan@ibspan.waw.pl).

certain *Riccati equation* (cf. [24]). The difficulties with extending these conditions to optimal control problems are connected to the presence of inequality-type constraints and the associated lack of smoothness of the solutions.

The first result in this direction was obtained by Maurer in [16], where only equality-type constraints (the state equation) were used in the construction of the relevant subspace.

Later, some attempts have been made to weaken this condition by including active inequality constraints (cf. [17, 20, 22, 23, 28, 29, 30]). In the last paper by Zeidan [30] she considers both necessary and sufficient second-order optimality conditions for optimal control problems subject to mixed control-state constraints and gives a weak version of SSC.

The main contributions of the present paper is the *derivation of weak SSC for optimal control problems, where, along with the mixed, pure state constraints of first order also are present and used in the construction of the relevant subspace.* That can be viewed as a direct generalization of Zeidan's result.

It should be stressed that the presence of pure state-space constraints complicates the analysis very much since it necessitates the additional discussion of regularity of the solutions and Lagrange multipliers, which is not needed if these constraints are void.

The organization of the paper is as follows: In section 2 we derive SSC for cone-constrained optimization problems in Banach spaces, using the same formalism as in [12]. This result is only a slight reformulation of that due to Dontchev et al. [2].

In sections 3 and 4 the abstract SSC are used to obtain SSC for nonlinear optimal control problems subject to both mixed control-state and pure state constraints. In these conditions active constraints of both types are taken into account.

One of the important applications of SSC is stability analysis of solutions to parametric optimization problems (cf. [1, 2, 3, 8, 9, 10, 11, 12, 13, 14, 25, 26]). In section 5 the possibility of application of the derived SSC to stability analysis of the solutions to parametric optimal control problems is briefly discussed. It turns out that SSC involving active state constraints are too weak to repeat the proof of stability of the solutions given in [13].

Some notation used: $X, Y, Z, \ldots$ denote Banach spaces and $\widehat{X}, \widehat{Y}, \widehat{Z}, \ldots$, Hilbert spaces. Asterisks denote dual spaces. $L(X, Y)$ is the Banach space of linear continuous operators from X into Y.

The norms in Banach and Hilbert spaces are denoted by $[\![\cdot]\!]$ and $\|\cdot\|$, respectively, with a subscript refering to the space.

For $f : X \times Y \mapsto Z$, $D_x f(x, y), D_y f(x, y), D^2_{xy} f(x, y), \ldots$ denote the respective Fréchet derivatives in the corresponding arguments.

$\mathbb{R}^n$ is the n-dimensional Euclidean space with the inner product denoted by $\langle x, y\rangle$ and the norm $|x| = \langle x, x\rangle^{\frac{1}{2}}$.

$L^2(0, T; \mathbb{R}^n)$ is the Hilbert space of square integrable vector functions, with the inner product

$$(x, y) = \int_0^T \langle x(t), y(t)\rangle dt \quad \text{and the norm } \|x\|_2 = (x, x)^{\frac{1}{2}}.$$

$L^\infty(0, T; \mathbb{R}^n)$ is the Banach space of essentially bounded vector functions with the norm

$$\|x\|_\infty = \max_i \operatorname*{ess\,sup}_{t\in[0,T]} |x^i(t)|.$$

$C^0(0,T;\mathbb{R}^n)$ and $C^1(0,T;\mathbb{R}^n)$ are the spaces of continuous and continuously differentiable vector functions, respectively, equipped with the usual norms.

$$W^{1,p}(0,T;\mathbb{R}^n) = \{x \in L^p(0,T;\mathbb{R}^n) \mid \dot{x} \in L^p(0,T;\mathbb{R}^n)\}, \qquad p = 2, \infty,$$

denote Sobolev spaces of absolutely continuous functions with the norms $\|x\|_{1,2} = \{|x(0)|^2 + \|\dot{x}\|_2^2\}^{\frac{1}{2}}$ and $\|x\|_{1,\infty} = \max\{|x(0)|, \|\dot{x}\|_\infty\}$, respectively.

2. SSC for abstract optimization problems. In this section we discuss SSC for cone-constrained optimization problems with two-norm discrepancy in Banach spaces.

Let Z and Y be two Banach spaces, the space of arguments and constraints, respectively. Moreover, two Hilbert spaces $\widehat{Z}$ and $\widehat{Y}$ are given such that $Z \subset \widehat{Z}$ and $Y \subset \widehat{Y}$ with dense and continuous embeddings. The duality pairing between $\widehat{Z}^*$ and $\widehat{Z}$ or Z^* and Z is denoted by $(\cdot,\cdot)_{\widehat{Z}}$. We put $\widehat{Y}^* = \widehat{Y}$, and by $(\cdot,\cdot)_{\widehat{Y}}$ we denote the inner product in $\widehat{Y}$, extended by continuity to $Y^* \times Y$. We denote

$$X = Z \times Y, \qquad \widehat{X} = \widehat{Z} \times \widehat{Y}.$$

In Y there is given a closed convex cone K with vertex at the origin, which induces a partial order in Y. By $\widehat{K}$ we denote the closure of K in $\widehat{Y}$, and by

$$K^+ = \{\lambda \in Y^* \mid (\lambda, y)_{\widehat{Y}} \geq 0 \quad \text{for all } y \in K\}, \tag{2.1}$$

the positive polar cone to K.

We consider the following optimization problem:

$$\min_{z \in Z} F(z) \qquad \text{subject to} \qquad \varphi(z) \in K. \tag{P}$$

We assume the following.

(I.1) The functions $F : Z \mapsto \mathbb{R}^1$ and $\varphi : Z \mapsto Y$ are twice Fréchet differentiable, and the following compatibility conditions are satisfied:

$$\begin{array}{lll} D_z F(z) \in \widehat{Z}^*, & D^2_{zz} F(z) \in L(\widehat{Z}, \widehat{Z}^*), & \\ D_z \varphi(z) \in L(\widehat{Z}, \widehat{Y}), & D_z \varphi^*(z)\lambda \in \widehat{Z}^*, & D^2_{zz} \varphi^*(z)\lambda \in L(\widehat{Z}, \widehat{Z}^*) \end{array} \tag{2.2}$$

for all $z \in Z$ and $\lambda \in Y$.
Moreover

$$\begin{aligned} &\lim \|D^2_{zz} F(z_1) - D^2_{zz} F(z_2)\|_{\widehat{Z} \mapsto \widehat{Z}^*} = 0, \\ &\lim \|D_z \varphi(z_1) - D_z \varphi(z_2)\|_{\widehat{Z} \mapsto \widehat{Y}} = 0, \\ &\lim \|D^2_{zz} \varphi^*(z_1)\lambda_1 - D^2_{zz} \varphi^*(z_2)\lambda_2\|_{\widehat{Z} \mapsto \widehat{Z}^*} = 0 \end{aligned} \tag{2.3}$$

for $[\![z_1 - z_2]\!]_Z \to 0$ and $[\![\lambda_1 - \lambda_2]\!]_Y \to 0$.

Let $z_0 \in Z$ be a given point feasible for (P). We will find sufficient conditions under which z_0 is a locally unique minimizer of (P).

Let us start with the constraint qualifications, which will be formulated in the same way as in [12]. To this end, for any $y \in K$ we define the following subspace of Y:

$$M_y = (K + [y]) \cap (-K + [y]), \tag{2.4}$$

and denote $M_0 := M_{y_0}$, where $y_0 = \varphi(z_0)$ and $[y]$ is the one-dimensional subspace generated by the element y. We assume that

(I.2) there exists a subspace $M \subset M_0$, closed in Y-topology, and a linear continuous mapping

$$\Pi \in L(Y,Y) \cup L(\widehat{Y},\widehat{Y}), \qquad \Pi : Y \mapsto M$$

such that

$$D_z\varphi(z_0)Z + \Pi Y = Y, \qquad D_z\varphi(z_0)\widehat{Z} + \Pi\widehat{Y} = \widehat{Y}. \tag{2.5}$$

Moreover, there exists a neighborhood $\mathcal{Y}_0 \subset Y$ of $y_0 = \varphi(z_0)$ such that

$$M \subset M_y \quad \text{for all } y \in \mathcal{Y}_0 \cap K. \tag{2.6}$$

If we define

$$S_0 \in L(X,Y) \cap L(\widehat{X},\widehat{Y}), \qquad S_0 \begin{pmatrix} z \\ y \end{pmatrix} := D_z\varphi(z_0)z + \Pi y, \tag{2.7}$$

then condition (2.5) amounts to that $S_0 \in L(X,Y) \cup L(\widehat{X},\widehat{Y})$ is surjective.

Let us introduce the following Lagrangian associated with (P):

$$\mathcal{L} : Z \times Y^* \mapsto \mathbb{R}^1, \qquad \mathcal{L}(z,\lambda) = F(z) - \varphi^*(z)\lambda. \tag{2.8}$$

We assume that

(I.3) there exists a Lagrange multiplier $\lambda_0 \in Y^*$ associated with z_0 such that the following Kuhn–Tucker conditions hold:

$$\begin{aligned} &D_z\mathcal{L}(z_0,\lambda_0) := D_zF(z_0) - D_z\varphi^*(z_0)\lambda_0 = 0, \\ &(\lambda_0, \varphi(z_0))_{\widehat{Y}} = 0, \qquad \lambda_0 \in K^+. \end{aligned} \tag{2.9}$$

It follows from (I.2) (cf. Lemma 3.1 and Theorem 4.8 in [12]) that λ_0 is defined uniquely and it belongs to $\widehat{Y}$. We assume that λ_0 is more regular. Namely,

(I.4) $\lambda_0 \in Y$.

By (I.2) and (I.4) we have $D^2_{zz}\mathcal{L}(z_0,\lambda_0) \in L(\widehat{Z},\widehat{Z}^*)$ and

$$|(D^2_{ZZ}\mathcal{L}(z_0,\lambda_0)x,z)_{\widehat{Z}}| \le c\|x\|_{\widehat{Z}}\|z\|_{\widehat{Z}} \quad \text{for all } x,z \in \widehat{Z}. \tag{2.10}$$

In a way similar to (2.4), for $\lambda \in K^+$ we define

$$N_\lambda = (K^+ + [\lambda]) \cap (-K^+ + [\lambda]), \tag{2.11}$$

denote $N_0 := N_{\lambda_0}$, and assume that

(I.5) there exist a subspace $N \subset N_0 \subset Y^*$ and a constant $\gamma > 0$ such that

$$\begin{aligned} &(D^2_{zz}\mathcal{L}(z_0,\lambda_0)z,z)_{\widehat{Z}} \ge \gamma\|z\|^2_{\widehat{Z}} \\ &\text{for all } z \in E_0 := \{z \in \widehat{Z} \mid D_z\varphi(z_0)z \in \widehat{N}^\perp\}, \end{aligned} \tag{2.12}$$

where $\widehat{N}^\perp$ is the closure in $\widehat{Y}$ of the subspace $\{y \in Y \mid (\lambda,y)_{\widehat{Y}} = 0 \quad \text{for all } \lambda \in N\}$. Moreover, there exists $\sigma > 0$ such that

$$\Lambda_0 \cap (K^+ + N) \subset K^+, \tag{2.13}$$

where $\Lambda_0 = \{\lambda \in Y \mid [\![\lambda - \lambda_0]\!]_Y \le \sigma\}$. We also require that

$$[\![y^\perp]\!]_Y \le c[\![y]\!]_Y \qquad \text{for all } y \in Y, \tag{2.14}$$

where $y^{\perp}$ denotes the projection of y onto $\widehat{N}^{\perp}$ orthogonal in $\widehat{Y}$. Note that it follows from (2.9) that

$$M \subset \widehat{N}^{\perp}. \tag{2.15}$$

Remark 2.1. Condition (I.5) is a counterpart of the SSC in finite-dimensional mathematical programs (cf., e.g., [8, 9, 26]), since we require that coercivity is satisfied on a whole subspace. We are interested in this type of condition, having in mind further applications in stability analysis. A counterpart of the *second-order sufficient optimality condition* for finite-dimensional mathematical programs (cf., e.g., [3]) is used in [2], where in our terminology it is assumed that coercivity condition (2.12) is satisfied on the more narrow set

$$\{z \in Z \mid \varphi(z_0) + D_z\varphi(z_0)z \in \widehat{N}^{\perp} \cap K,\ \|z - z_0\|_Z \le \beta \quad \text{for some } \beta > 0\}.$$

In this context we should mention another similar but not identical sufficient optimality condition, which was introduced in pioneering papers by Maurer and Zowe [18] and Maurer [16] and later used in [27]. In [16] it is assumed that the coercivity condition (2.12) is satisfied on the set

$$\{z \in Z \mid D_z\varphi(z_0)z \in K + [\varphi(z_0)],\ (D_z\varphi^*(z_0)\lambda_0, z)_{\widehat{Z}} \le \beta\|z\|_{\widehat{Z}} \quad \text{for some } \beta > 0\}.$$

Remark 2.2. A condition similar to (2.12) was introduced in [11]. However, in that case only the one-norm situation was considered, whereas (2.12) constitutes the essence of the so called *two-norm discrepancy* for (P). Namely, the problem is well defined and differentiable in the stronger topology of space Y, whereas the coercivity condition (2.12) is satisfied only in the weaker norm $\widehat{Y}$, in which problem (P) is not differentiable.

In assumption (I.5) the additional condition (I.4) on regularity of the Lagrange multiplier is crucial. It allows us to formulate stability condition (2.13) for a neighborhood Λ_0 in the stronger topology of Y, rather than in the topology of the dual space Y^*, as was done in [11].

The following theorem is actually a weakened version of Theorem 1 in [2]. For the sake of completeness, the proof of the theorem based on that in [2] is given in Appendix A. Later, some elements of this proof will be used in the proof of Proposition 4.5.

THEOREM 2.3. *If assumptions* (I.1)–(I.5) *hold, then there exist constants* $\rho_1 > 0$ *and* $\gamma_1 > 0$ *such that*

$$\begin{aligned} &F(z) \ge F(z_0) + \gamma_1\|z - z_0\|^2_{\widehat{Z}} \\ &\text{for all feasible } z \in \mathcal{O}_0^{\rho_1} := \{z \in Z \mid \|z - z_0\|_Z < \rho_1\}; \end{aligned} \tag{2.16}$$

i.e., z_0 is a locally unique minimizer of (P).

To illustrate the nature of the abstract assumptions (I.1)–(I.5), let us consider the following simple example:

$$\min_{z \in L^\infty(0,1)} F(z) := \int_0^1 f(z(t))dt \quad \text{subject to } \varphi(z(t)) \ge 0 \text{ for all } t \in [0,1], \tag{E}$$

where $f(\cdot)$ and $\varphi(\cdot)$ are twice continuously differentiable.

To put (E) in the framework of (P) we define $Z = Y = L^\infty(0,1)$, $\widehat{Z} = \widehat{Y} = L^2(0,1)$, and $K = \{y \in L^\infty(0,1) \mid y(t) \ge 0 \text{ for all } t \in [0,1]\}$. It is easy to see that conditions (I.1) are satisfied.

Let $z_0 \in L^\infty(0,1)$ be a feasible element. For $\alpha \geq 0$ introduce the set $\Omega_\alpha = \{t \in [0,1] \mid z_0(t) < \alpha\}$ and define the set M in (I.2) as

$$M = \{y \in L^\infty(0,1) \mid y(t) = 0 \text{ for all } t \in [0,1] \setminus \Omega_{\bar\alpha}\},$$

where $\bar\alpha > 0$ is an arbitrary number.

It is easy to see that condition (I.2) is satisfied if

$$D_z\varphi(z_0(t)) \neq 0 \quad \text{for all } t \in \Omega_{\bar\alpha}. \tag{2.17}$$

Let us introduce the following Lagrangian:

$$\mathcal{L}(z,\lambda) = \int_0^1 \ell(z(t),\lambda(t))dt = \int_0^1 [f(z(t)) - \lambda(t)\varphi(z(t))]\,dt.$$

Conditions (I.3) and (I.4) amounts to the existence of a Lagrange multiplier $\lambda_0 \in L^\infty(0,1)$ such that the Kuhn–Tucker conditions are satisfied:

$$\begin{aligned} &D_z\mathcal{L}(z_0,\lambda_0) = 0,\\ &\lambda_0(t)z_0(t) = 0, \qquad \lambda_0(0) \geq 0 \quad \text{for all } t \in [0,1]. \end{aligned}$$

By (2.17) the Lagrange multiplier λ_0 is defined uniquely.

For $\sigma \geq 0$ introduce the set $\Xi_\sigma = \{t \in [0,1] \mid \lambda_0(t) > \sigma\}$, and define the set N in (I.5) as

$$N = \{\lambda \in L^\infty(0,1) \mid \lambda(t) = 0 \text{ for all } t \in [0,1] \setminus \Xi_{\bar\sigma}\},$$

where $\bar\sigma > 0$ is an arbitrary number.

Condition (I.5) is satisfied if $\gamma > 0$ exists such that

$$(D_{zz}\mathcal{L}(z_0,\lambda_0)z,z) \geq \gamma \text{ for all } z \in \{z \in L^2(0,1) \mid (\lambda,z) = 0 \text{ for all } \lambda \in N\}, \quad \text{i.e.,}$$

$$D_{zz}\ell(z_0(t),\lambda_0(t)) \geq \gamma > 0 \quad \text{for all } t \in [0,1] \setminus \Xi_{\bar\sigma}. \tag{2.18}$$

It should be stressed that, except some technicalities, the nature of assumptions (I.2) and (I.5) for optimal control problems is the same as (2.17) and (2.18).

3. Optimal control problems: Constraint qualifications. In this section we consider a class of optimal control problems for nonlinear ordinary differential equations and discuss conditions under which constraint qualifications (I.2) of the abstract problem (P) are satisfied. The problems considered are quite general. They are subject to mixed initial-terminal constraints as well as to pointwise inequality constraints, both mixed control-state and pure state. Hence they include problems considered in [1, 2, 12] (pure control constraints), in [15, 17, 22, 30] (mixed constraints), and in [11, 13, 23] (state constraints).

Our model optimal control problem is the following:

(O) Find $(x_0,u_0) \in W^{1,\infty}(0,T;\mathbb{R}^n) \times L^\infty(0,T;\mathbb{R}^m)$ such that

$$F(x_0,u_0) = \min\left\{F(x,u) := \int_0^T f^0(x(t),u(t))dt + g(x(T))\right\} \tag{3.1}$$

subject to

$$\dot x(t) - f(x(t),u(t)) = 0 \qquad \text{for a.a. } t \in [0,T], \tag{3.2}$$

$$\xi(x(0),x(T)) = 0, \tag{3.3}$$

$$\theta(x(t),u(t)) \leq 0 \qquad \text{for a.a. } t \in [0,T], \tag{3.4}$$

$$\vartheta(x(t)) \leq 0 \qquad \text{for all } t \in [0,T], \tag{3.5}$$

where $\xi : \mathbb{R}^n \times \mathbb{R}^n \mapsto \mathbb{R}^d$, $\theta : \mathbb{R}^n \times \mathbb{R}^m \mapsto \mathbb{R}^k$, $\vartheta : \mathbb{R}^n \mapsto \mathbb{R}^l$.

Denote by $I = \{1, ..., k\}$ and $J = \{1, ..., l\}$ the sets of indices of constraints.

It is assumed that

(II.1) functions $f^0(\cdot,\cdot), g(\cdot), f(\cdot,\cdot), \xi(\cdot,\cdot), \theta(\cdot,\cdot), \vartheta(\cdot)$ and $D_x\vartheta(\cdot)$ are twice Fréchet differentiable in all arguments, and the respective derivatives are locally Lipschitz continuous.

We assume that for a fixed feasible point (x_0, u_0), at which we will find sufficient optimality conditions, the following regularity condition is satisfied:

(II.2) u_0 is a piecewise continuous function with finite number of discontinuity points t_k.

In order to represent (O) in form (P) we put

$$Z = W^{1,\infty}(0,T;\mathbb{R}^n) \times L^\infty(0,T;\mathbb{R}^m), \quad \widehat{Z} = W^{1,2}(0,T;\mathbb{R}^n) \times L^2(0,T;\mathbb{R}^m).$$

For $z = (x,u)$ belonging to Z or $\widehat{Z}$ we define the norms

$$[\![z]\!]_Z = \max\{\|x\|_{1,\infty}, \|u\|_\infty\} \text{ and } \|z\|_{\widehat{Z}} = \{\|x\|_{1,2}^2 + \|u\|_2^2\}^{\frac{1}{2}}.$$

Moreover we put

$$F(z) = F(x,u), \qquad \varphi(z) = (\dot{x} - f(x,u), \xi(x(0),x(T)), -\theta(x,u), -\vartheta(x)).$$

Hence it is natural to choose

$$Y = L^\infty(0,T;\mathbb{R}^n) \times \mathbb{R}^d \times L^\infty(0,T;\mathbb{R}^k) \times W^{1,\infty}(0,T;\mathbb{R}^l),$$
$$\widehat{Y} = L^2(0,T;\mathbb{R}^n) \times \mathbb{R}^d \times L^2(0,T;\mathbb{R}^k) \times W^{1,2}(0,T;\mathbb{R}^l),$$

with

$$K = K_1 \times K_2 \times K_3 \times K_4, \tag{3.6}$$

where

$$K_1 = \{0\}, \qquad K_2 = \{0\},$$
$$K_3 = \{u \in L^\infty(0,T;\mathbb{R}^k) \mid u^i(t) \geq 0, \quad i = 1,\dots,k, \quad \text{for a.a. } t \in [0,T]\},$$
$$K_4 = \{x \in W^{1,\infty}(0,T;\mathbb{R}^l) \mid x^j(t) \geq 0, \quad j = 1,\dots,l, \quad \text{for all } t \in [0,T]\}.$$

The cone $\widehat{K}$ is defined as in (3.6) but with Y substituted by $\widehat{Y}$.

For $y = (p,q,r,s)$ belonging to Y or $\widehat{Y}$ we define the norms

$$[\![y]\!]_Y = \max\{\|p\|_\infty, |q|, \|r\|_\infty, \|s\|_{1,\infty}\} \text{ and } \|y\|_{\widehat{Y}} = \{\|p\|_2^2 + |q|^2 + \|r\|_2^2 + \|s\|_{1,2}^2\}^{\frac{1}{2}}.$$

As in section 2, we denote $X = Z \times Y$ and $\widehat{X} = \widehat{Z} \times \widehat{Y}$.

Note that by (II.1) conditions (I.1) are satisfied. Now we will formulate conditions under which (I.2) holds.

To simplify notation we put

$$\begin{aligned}
&A(t) := D_x f(x_0(t), u_0(t)), && B(t) := D_u f(x_0(t), u_0(t)),\\
&\Xi_0 := D_{x(0)}\xi(x_0(0), x_0(T)), && \Xi_T := D_{x(T)}\xi(x(0), x_0(T)),\\
&\Theta_x(t) := D_x\theta(x_0(t), u_0(t)), && \Theta_u(t) := D_u\theta(x_0(t), u_0(t)),\\
&\Upsilon(t) := D_x\vartheta(x_0(t)).
\end{aligned} \tag{3.7}$$

Moreover denote by

$$I^a(t) = \{i \in I \mid \theta^i(x_0(t), u_0(t)) = 0\}, \qquad J^a(t) = \{j \in J \mid \vartheta^j(x_0(t)) = 0\}$$

the sets of indices of active constraints.

For $\alpha \geq 0$ we introduce the sets

$$\begin{aligned} \Psi^i_\alpha &:= \left\{t \in [0,T] \mid \theta^i(x_0(t), u_0(t)) < -\alpha\right\}, \quad i = 1, ..., k, \\ \Omega^j_\alpha &:= \left\{t \in [0,T] \mid \vartheta^j(x_0(t)) < -\alpha\right\}, \quad j = 1, ..., l, \end{aligned} \tag{3.8}$$

and define the following functions

$$\psi^i_\alpha(t) = \begin{cases} \theta^i(x_0(t), u_0(t)) + \alpha & \text{if} \quad t \in \Psi^i_\alpha, \\ 0 & \text{if} \quad t \notin \Psi^i_\alpha \end{cases} \tag{3.9}$$

and

$$\omega^j_\alpha(t) = \begin{cases} \vartheta^j(x_0(t)) + \alpha & \text{if} \quad t \in \Omega^j_\alpha, \\ 0 & \text{if} \quad t \notin \Omega^j_\alpha. \end{cases} \tag{3.10}$$

We introduce the $(k \times k)$ and $(l \times\ l)$ diagonal matrices

$$U_\alpha(t) = \operatorname{diag} \psi^i_\alpha(t), \qquad T_\alpha(t) = \operatorname{diag} \omega^j_\alpha(t) \tag{3.11}$$

and define $(k+l) \times (m+k+l)$ matrices

$$V_\alpha(t) = \begin{bmatrix} \Theta_u(t) & U_\alpha(t) & 0 \\ \Upsilon(t)B(t) & 0 & T_\alpha(t) \end{bmatrix}. \tag{3.12}$$

Let us choose any $\alpha \geq 0$ and define the subspace $M^\alpha \subset Y$ by

$$M^\alpha = M^\alpha_1 \times M^\alpha_2 \times M^\alpha_3 \times M^\alpha_4, \tag{3.13}$$

where

$$\begin{aligned} &M^\alpha_1 = 0, \qquad M^\alpha_2 = 0, \\ &M^\alpha_3 = \left\{u \in L^\infty(0,T;\mathbb{R}^k) \mid u^i(t) = 0 \text{ for all } t \in [0,T] \setminus \Psi^i_\alpha, \quad i = 1, ..., k\right\}, \\ &M^\alpha_4 = \left\{x \in W^{1,\infty}(0,T;\mathbb{R}^l) \mid x^j(t) = 0 \text{ for all } t \in [0,T] \setminus \Omega^j_\alpha, \quad j = 1, ..., l\right\}. \end{aligned}$$

It is easy to see that for any $\alpha > 0$, $M^\alpha \subset M_0$, where M_0 is defined in (2.4). Moreover, if we choose as $\mathcal{Y}^\alpha_0 \subset Y$ the open ball of radius $\alpha/2$ about $y_0 = \varphi(z_0)$, then (2.6) is satisfied.

Define the mapping $\Pi^\alpha := (\Pi^\alpha_1, \Pi^\alpha_2, \Pi^\alpha_3, \Pi^\alpha_4) : Y \mapsto M^\alpha$ by

$$\begin{aligned} &\Pi^\alpha_1 = 0, && \Pi^\alpha_2 = 0, \\ &(\Pi^\alpha_3 y)(t) = U_\alpha(t) y(t), && (\Pi^\alpha_4 y)(t) = T_\alpha(t) y(t). \end{aligned} \tag{3.14}$$

Condition (I.2) will be fully satisfied if we are able to show that there exists $\alpha > 0$ such that for $\Pi := \Pi^\alpha$ surjectivity conditions (2.5) hold. We assume that

(II.3) there exists $\beta > 0$ such that

$$|V_0(t) V_0(t)^* \chi| \geq \beta |\chi| \tag{3.15}$$

for all $\chi \in \mathbb{R}^{k+l}$ and all $t \in [0,T]$.

(II.4) For any $\widetilde{q} \in \mathbb{R}^{d+l}$ an element $(\eta, \upsilon) \in W^{1,\infty}(0,T;\mathbb{R}^{n+l}) \times L^{\infty}(0,T;\mathbb{R}^{m+k+l})$ exists such that the following equations hold (the complete controllability condition is satisfied):

$$\begin{aligned} &\dot{\eta}(t) - \widetilde{A}(t)\eta(t) - \widetilde{B}(t)\upsilon(t) = 0, \\ &\widetilde{P}_0\eta(0) + \widetilde{P}_T\eta(T) = \widetilde{q}, \end{aligned} \tag{3.16}$$

where

$$\widetilde{A}(t) = \begin{bmatrix} \widetilde{A}_{11}(t) & \widetilde{A}_{12}(t) \\ \widetilde{A}_{21}(t) & \widetilde{A}_{22}(t) \end{bmatrix},$$

$$\widetilde{A}_{11}(t) = A(t) - B(t) \begin{bmatrix} \Theta_u(t)^* & B(t)^*\Upsilon(t)^* \end{bmatrix} [V_0(t)V_0(t)^*]^{-1} \times \begin{bmatrix} \Theta_x(t) \\ \Upsilon(t)A(t) + \dot{\Upsilon}(t) \end{bmatrix},$$

$$\widetilde{A}_{12}(t) = -B(t) \begin{bmatrix} \Theta_u(t)^* & B(t)^*\Upsilon(t)^* \end{bmatrix} [V_0(t)V_0(t)^*]^{-1} \begin{bmatrix} 0 \\ \dot{T}_0(t) \end{bmatrix},$$

$$\widetilde{A}_{21}(t) = \begin{bmatrix} 0 & T_0(t)^* \end{bmatrix} [V_0(t)V_0(t)^*]^{-1} \begin{bmatrix} \Theta_x(t) \\ \Upsilon(t)A(t) + \dot{\Upsilon}(t) \end{bmatrix},$$

$$\widetilde{A}_{22}(t) = \begin{bmatrix} 0 & T_0(t)^* \end{bmatrix} [V_0(t)V_0(t)^*]^{-1} \begin{bmatrix} 0 \\ \dot{T}_0(t) \end{bmatrix},$$

$$\widetilde{B}(t) = \begin{bmatrix} B(t) & 0 & 0 \\ 0 & 0 & I \end{bmatrix} - \begin{bmatrix} B(t)\Theta_u(t)^* & B(t)B(t)^*\Upsilon(t)^* \\ 0 & T_0(t)^* \end{bmatrix} \times [V_0(t)V_0(t)^*]^{-1} V_0(t),$$

$$\widetilde{P}_0 = \begin{bmatrix} \Xi_0 & 0 \\ \Upsilon(0) & T_0(0) \end{bmatrix}, \qquad \widetilde{P}_T = \begin{bmatrix} \Xi_T & 0 \\ 0 & 0 \end{bmatrix}.$$

Remark 3.1. Condition (II.3) means that all gradients $D_u\theta^i(x_0(t), u_0(t))$, $i \in I^a(t)$, of the active mixed constraints and all gradients $D_x\vartheta^j(x_0(t))$, $j \in J^a(t)$, of the active state constraints, transformed into the space $\mathbb{R}^m$ by means of the mapping $B^*(t) : \mathbb{R}^n \mapsto \mathbb{R}^m$, are jointly linearly independent, uniformly on $[0,T]$. This condition implies that we restrict ourselves to the *first-order* state constraints (cf. [6]).

Remark 3.2. By a standard argument we find that system (3.16) is completely controllable if and only if the following rank condition is satisfied:

$$\operatorname{rank}\left\{ [\widetilde{P}_0 + \widetilde{P}_T\Phi(T)][\widetilde{P}_0 + \widetilde{P}_T\Phi(T)]^* + \int_0^T G(t)G(t)^* dt \right\} = d + l, \tag{3.17}$$

where Φ is the solution of the homogeneous matrix equation

$$\dot{\Phi}(t) = \widetilde{A}(t)\Phi(t), \qquad \Phi(0) = I,$$

and $G(t) = \widetilde{P}_T\Phi(T)\Phi(t)^{-1}\widetilde{B}(t)$.

LEMMA 3.3. *If assumptions* (II.3) *and* (II.4) *are satisfied, then there exists* $\alpha > 0$ *such that the mapping* $\Pi := \Pi_\alpha$ *defined by* (3.14) *satisfies conditions* (I.2).

Proof. Condition (2.6) is satisfied by construction of Π_α, so it remains to check (2.5). We will prove the first equation in (2.5). The proof of the second one is the same.

We have to show that there exist $\alpha > 0$ such that, for any

$$\begin{aligned} &p \in L^\infty(0,T;\mathbb{R}^n), && q \in \mathbb{R}^d, \\ &r \in L^\infty(0,T;\mathbb{R}^k), && s \in W^{1,\infty}(0,T;\mathbb{R}^l), \end{aligned}$$

there exists a solution

$$\begin{aligned} &\sigma \in L^\infty(0,T;\mathbb{R}^m), && \zeta \in W^{1,\infty}(0,T;\mathbb{R}^n), \\ &\rho \in L^\infty(0,T;\mathbb{R}^k), && \tau \in W^{1,\infty}(0,T;\mathbb{R}^l) \end{aligned}$$

of the following system of equations:

$$\dot\zeta(t) - A(t)\zeta(t) - B(t)\sigma(t) = p(t), \tag{3.18}$$

$$\Xi_0\zeta(0) + \Xi_T\zeta(T) = q, \tag{3.19}$$

$$\Theta_x(t)\zeta(t) + \Theta_u(t)\sigma(t) + U_\alpha(t)\rho(t) = r(t), \tag{3.20}$$

$$\Upsilon(t)\zeta(t) + T_\alpha(t)\tau(t) = s(t). \tag{3.21}$$

First we will show that (3.18)–(3.21) have a solution for $\alpha = 0$. From the proof it will follow that it also has a solution for a sufficiently small $\alpha > 0$.

Note that (3.21) is equivalent to

$$\Upsilon(0)\zeta(0) + T_0\tau(0) = s(0), \tag{3.22}$$

$$\dot\Upsilon(t)\zeta(t) + \Upsilon(t)\dot\zeta(t) + \dot T_0(t)\tau(t) + T_0(t)\dot\tau(t) = \dot s(t). \tag{3.23}$$

Multiplying (3.18) by $\Upsilon(t)$ and using (3.23) we obtain

$$\begin{aligned} &\Upsilon(t)B(t)\sigma(t) + T_0(t)\dot\tau(t) \\ &\qquad = (\dot s(t) - \Upsilon(t)p(t)) - (\Upsilon(t)A(t) + \dot\Upsilon(t))\zeta(t) - \dot T_0(t)\tau(t). \end{aligned} \tag{3.24}$$

Combining (3.20) and (3.24) yields

$$\begin{aligned} &V_0 \begin{bmatrix} \sigma(t) \\ \rho(t) \\ \dot\tau(t) \end{bmatrix} := \begin{bmatrix} \Theta_u(t)\sigma(t) + U_0(t)\rho(t) \\ \Upsilon(t)B(t)\sigma(t) + T_0(t)\dot\tau(t) \end{bmatrix} \\ &= \begin{bmatrix} r(t) - \Theta_x(t)\zeta(t) \\ (\dot s(t) - \Upsilon(t)p(t)) - (\Upsilon(t)A(t) + \dot\Upsilon(t))\zeta(t) - \dot T_0(t)\tau(t) \end{bmatrix}. \end{aligned} \tag{3.25}$$

By (II.3), any solution of (3.25) can be expressed in the form

$$\begin{aligned} &\begin{bmatrix} \sigma(t) \\ \rho(t) \\ \dot\tau(t) \end{bmatrix} = V_0(t)^*[V_0(t)V_0(t)^*]^{-1} \\ &\times \begin{bmatrix} r(t) - \Theta_x(t)\zeta(t) \\ (\dot s(t) - \Upsilon(t)p(t)) - (\Upsilon(t)A(t) + \dot\Upsilon(t))\zeta(t) - \dot T_0(t)\tau(t) \end{bmatrix} \\ &+ \big(I - V_0(t)^*[V_0(t)V_0(t)^*]^{-1}V_0(t)\big)\,\upsilon(t), \end{aligned} \tag{3.26}$$

where $\upsilon(t) \in \mathbb{R}^{m+k+l}$ is an arbitrary vector.

Combining (3.18) with the first and the third row of (3.26) we obtain

$$\begin{bmatrix} \dot\zeta(t) \\ \dot\tau(t) \end{bmatrix} - \begin{bmatrix} \widetilde A_{11}(t) & \widetilde A_{12}(t) \\ \widetilde A_{21}(t) & \widetilde A_{22}(t) \end{bmatrix} \begin{bmatrix} \zeta(t) \\ \tau(t) \end{bmatrix} - \begin{bmatrix} \widetilde B_1(t) \\ \widetilde B_2(t) \end{bmatrix} \upsilon(t) = \begin{bmatrix} \widetilde C_1(t) \\ \widetilde C_2(t) \end{bmatrix}, \tag{3.27}$$

where $\widetilde A_{ij}$ and $\widetilde B(t) = \left[\begin{smallmatrix} \widetilde B_1(t) \\ \widetilde B_2(t) \end{smallmatrix}\right]$ are given in (3.16) and

$$\begin{bmatrix} \widetilde C_1(t) \\ \widetilde C_2(t) \end{bmatrix} = \begin{bmatrix} B(t)\Theta_u(t)^* & B(t)B(t)^*\Upsilon(t)^* \\ 0 & T_0(t)^* \end{bmatrix} \times [V_0(t)V_0(t)^*]^{-1} \begin{bmatrix} r(t) \\ \dot s(t) - \Upsilon(t)p(t) \end{bmatrix} + \begin{bmatrix} p(t) \\ 0 \end{bmatrix}.$$

On the other hand from (3.19) and (3.22) we obtain

$$\widetilde P_0 \begin{bmatrix} \zeta(0) \\ \tau(0) \end{bmatrix} + \widetilde P_T \begin{bmatrix} \zeta(T) \\ \tau(T) \end{bmatrix} = \begin{bmatrix} q \\ s(0) \end{bmatrix}. \tag{3.28}$$

It is easy to see that for any (p, q, r, s) there exists a control function υ as well as functions ζ and τ such that (3.27) along with (3.28) is satisfied if and only if (II.4) holds. Having (ζ, τ, υ) we find σ and ρ from the first two rows of (3.26).

Thus, we have proved that (3.18)–(3.21) are satisfied for $\alpha = 0$. Now we will briefly show that these equations are also satisfied for sufficiently small $\alpha > 0$.

By (II.2), (II.3), and slight modification of Lemma 7.2 in [12], we find that for $\alpha > 0$ sufficiently small

$$|V_\alpha(t)V_\alpha(t)^*\chi| \ge \frac{1}{2}\beta|\chi|$$

for all $\chi \in \mathbb{R}^{k+l}$ and all $t \in [0, T]$.

Hence we can repeat the above argument with 0 substituted by $\alpha > 0$ to obtain (3.27) and (3.28), where all data are functions of α. The matrix on the left-hand side of (3.17) becomes a continuous function of α, so shrinking α, if necessary, we find that the rank condition is satisfied. This completes the proof of the lemma. □

Remark 3.4. Note that Lemma 3.3 assures that conditions (II.3) and (II.4), originally assumed at $\alpha = 0$, actually are satisfied with some "margin of freedom" $\alpha > 0$. This result was proven using the assumption (II.2) of piecewise continuity of u_0. It allows us to verify the abstract assumption (I.2), and in section 5 it will play a crucial role in stability analysis.

4. SSC for optimal control problems. This section is devoted to deriving SSC.

Let us introduce the following Lagrangian associated with (O):

$$\begin{aligned} &\mathcal{L} : W^{1,\infty}(0,T;\mathbb{R}^n) \times L^\infty(0,T;\mathbb{R}^m) \times (L^\infty(0,T;\mathbb{R}^n))^* \times \mathbb{R}^d \\ &\qquad \times \left(L^\infty(0,T;\mathbb{R}^k)\right)^* \times \left(W^{1,\infty}(0,T;\mathbb{R}^l)\right)^* \mapsto \mathbb{R}^1, \\ &\mathcal{L}(x,u,q,\rho,\kappa,\mu) = F(x,u) + (q, \dot x - f(x,u)) + \langle \rho, \xi(x(0),x(T))\rangle \\ &\qquad\qquad + (\kappa, \theta(x,u)) + \langle \mu(0), \vartheta(x(0))\rangle + (\dot\mu, D_x\vartheta(x)f(x,u)). \end{aligned} \tag{4.1}$$

Note that the Lagrangian is in the so-called *Pontryagin form* with absolutely continuous adjoint function q (cf. section 7 in [6] as well as [5] and [19]). The state constraints are considered in $W^{1,\infty}(0,T;\mathbb{R}^l)$, where the general form of a linear functional is given by $\langle \mu(0), y(0)\rangle + (\dot\mu, \dot y)$. Accordingly, the terms of Lagrangian (4.1) corresponding to the state constraints (3.5) are obtained as follows:

$$\begin{aligned}\langle \mu(0), \vartheta(x(0))\rangle + \left(\dot\mu_0, \frac{d}{dt}\vartheta(x)\right) &= \langle \mu(0), \vartheta(x(0))\rangle + (\dot\mu_0, D_x\vartheta(x)\dot x)\\ &= \langle \mu(0), \vartheta(x(0))\rangle + (\dot\mu_0, D_x\vartheta(x)f(x,u)).\end{aligned}$$

As in (I.3) we assume that

(II.5) there exist Lagrange multipliers $(q_0, \rho_0, \kappa_0, \mu_0) \in Y^*$ associated with (x_0, u_0) such that the following Kuhn–Tucker conditions are satisfied:

$$D_x\mathcal{L}(x_0, u_0, q_0, \kappa_0, \rho_0, \mu_0) = 0, \tag{4.2}$$

$$D_u\mathcal{L}(x_0, u_0, q_0, \kappa_0, \rho_0, \mu_0) = 0, \tag{4.3}$$

$$(\kappa_0, \theta(x_0, u_0)) = 0, \quad \kappa_0 \in K_3^+, \tag{4.4}$$

$$\langle \mu_0(0), \vartheta(x_0(0))\rangle + (\dot\mu_0, D_x\vartheta(x_0)f(x_0, u_0)) = 0, \quad \mu_0 \in K_4^+. \tag{4.5}$$

By Lemma 3.3 the multipliers $(q_0, \rho_0, \kappa_0, \mu_0)$ are defined uniquely and belong to $\widehat Y$.

Recall (cf. [21]) that

$$\begin{aligned}\widehat K_4^+ = \{\mu \in W^{1,2}(0,T:\mathbb{R}^n) \mid \mu^j(t) \ge 0, \dot\mu^j(t) \text{ is nonincreasing}\\ \text{and } 0 \le \dot\mu^j(t) \le \mu^j(t), \quad j \in J\}.\end{aligned} \tag{4.6}$$

Let us introduce the following augmented Hamiltonian:

$$\begin{aligned}\mathcal{H}(t) = f^0(x_0(t), u_0(t)) - \langle q_0(t), f(x_0(t), u_0(t))\rangle\\ +\langle \kappa_0(t), \theta(x_0(t), u_0(t))\rangle + \langle \dot\mu_0(t), D_x\vartheta(x_0(t))f(x_0(t), u_0(t))\rangle.\end{aligned} \tag{4.7}$$

Condition (4.2) takes on the form of the adjoint equation

$$\dot q_0(t) - D_x\mathcal{H}(t) = 0,$$

along with the boundary conditions

$$-q_0(0) + D_{x(0)}\xi(x_0(0), x_0(T))^*\rho_0 + D_x\vartheta(x_0(0))^*\mu_0(0) = 0,$$

$$q_0(T) + D_{x(T)}\xi(x_0(0), x_0(T))^*\rho_0 + D_x g(x_0(T)) = 0,$$

while (4.3) can be expressed as

$$D_u\mathcal{H}(t) = 0 \qquad \text{for a.a. } t \in [0,T].$$

The Hessian of Lagrangian (4.1), evaluated at the reference point, can be expressed in terms of Hamiltonian (4.7) as follows:

$$\begin{aligned}&\left((y,v), \begin{pmatrix} D^2_{xx}\mathcal{L}_0 & D^2_{xu}\mathcal{L}_0 \\ D^2_{ux}\mathcal{L}_0 & D^2_{uu}\mathcal{L}_0 \end{pmatrix}(y,v)\right)\\ &\quad = \int_0^T \begin{bmatrix} y(t) \\ v(t) \end{bmatrix}^* \begin{bmatrix} D^2_{xx}\mathcal{H}(t) & D^2_{xu}\mathcal{H}(t) \\ D^2_{ux}\mathcal{H}(t) & D^2_{uu}\mathcal{H}(t) \end{bmatrix} \begin{bmatrix} y(t) \\ v(t) \end{bmatrix} dt\\ &\qquad + \begin{bmatrix} y(0) \\ y(T) \end{bmatrix}^* \begin{bmatrix} \mathcal{R}_{11} & \mathcal{R}_{12} \\ \mathcal{R}_{21} & \mathcal{R}_{22} \end{bmatrix} \begin{bmatrix} y(0) \\ y(T) \end{bmatrix},\end{aligned} \tag{4.8}$$

where $\mathcal{L}_0 := \mathcal{L}(x_0, u_0, q_0, \rho_0, \kappa_0, \mu_0)$ and

$$\mathcal{R}_{11} = D^2_{x(0)x(0)}\xi(x_0(0), x_0(T))\rho_0 + D^2_{xx}\vartheta(x_0(0))^*\mu_0(0),$$
$$\mathcal{R}_{12} = \mathcal{R}^*_{21} = D^2_{x(0)x(T)}\xi(x_0(0), x_0(T))\rho_0,$$
$$\mathcal{R}_{22} = D^2_{x(T)x(T)}\xi(x_0(0), x_0(T))\rho_0 + D^2_{xx}g(x_0(T)).$$

We have to show that (4.8) is positive definite on the subspace defined in (I.5). We are going to construct this subspace. To this end we need several technical results. Let us start with the discussion of the regularity of primal and dual optimal variables. Using the same argument as in [5] we obtain the following regularity of the multipliers.

LEMMA 4.1. *If* (II.1)–(II.5) *are satisfied, then* $\dot q_0, \kappa_0$, *and* $\dot\mu_0$ *are continuous and uniformly bounded on all subintervals* $(t_k, t_{k+1}) \subset [0, T]$.

Assume that the following modified Legendre–Clebsch condition is satisfied:

$$\langle u, D^2_{uu}\mathcal{H}(t)u\rangle \geq \gamma|u|^2, \qquad \gamma > 0, \tag{4.9}$$

for all

$$u \in \{u \in \mathbb{R}^m \quad |\langle D_u\theta^i(x_0(t), u_0(t), h_0), u\rangle = 0 \qquad \text{for all } i \in I^0(t),$$
$$\langle D_x\vartheta^j(x_0(t), h_0)D_uf(x_0(t), u_0(t), h_0), u\rangle = 0 \quad \text{for all } j \in J^0(t)\}$$

and for all $t \in [0, T]$.

Using the same argument as in [5] (cf. also [13]) we obtain the following regularity result.

LEMMA 4.2. *If* (II.1)–(II.5) *and* (4.9) *are satisfied, then* $(u_0, \dot q_0, \kappa_0, \dot\mu_0)$ *are uniformly Lipschitz continuous functions on all subintervals* $(t_k, t_{k+1}) \subset [0, T]$.

Note that strict complementarity condition for state constraints at some t amounts to $-\ddot\mu_0(t) > 0$. In order to construct the subspace N needed in (I.5) we introduce the sets of indices $I^+_\alpha(t)$ and $J^+_\alpha(t)$ of those constraints active at t for which strict complementarity condition is satisfied with the margin $\alpha \geq 0$:

$$I^+_\alpha(t) = \{i \in I^a(t) \mid \kappa^i_0(t) > \alpha\}, \quad J^+_\alpha(t) = \{j \in J^a(t) \mid -\ddot\mu^j_0(t) > \alpha\}. \tag{4.10}$$

Let us define the following subspaces of $\mathbb{R}^m$:

$$\begin{aligned}
&\mathcal{D}_\alpha(t) = \{u \in \mathbb{R}^m \mid \langle D_u\theta^i(x_0(t), u_0(t)), u\rangle = 0 \quad \text{for } i \in I^+_\alpha(t)\},\\
&\mathcal{E}_\alpha(t) = \{u \in \mathbb{R}^m \mid \langle D_x\vartheta^j(x_0(t))D_uf(x_0(t), u_0(t)), u\rangle = 0\\
&\qquad\qquad \text{for } j \in J^+_\alpha(t)\},\\
&\mathcal{G}_\alpha(t) = \mathcal{D}_\alpha(t) \cap \mathcal{E}_\alpha(t).
\end{aligned} \tag{4.11}$$

Now we are in a position to introduce subspace N needed in (I.5). We put

$$\begin{aligned}
&N = N_1 \times N_2 \times N_3 \times N_4,\\
&N_1 = L^\infty(0, T; \mathbb{R}^n), \qquad N_2 = \mathbb{R}^d,\\
&N_3 = \{\theta \in L^\infty(0, T; \mathbb{R}^k) \mid \theta^i(t) = 0 \ \text{ for } i \notin I^+_\alpha(t)\},\\
&N_4 = \{\mu \in W^{2,\infty}(t_k, t_{k+1}; \mathbb{R}^l) \mid \ddot\mu^j(t) = 0 \ \text{ for } j \notin J^+_\alpha(t)\},
\end{aligned} \tag{4.12}$$

where $\alpha > 0$, and $W^{2,\infty}(t_k, t_{k+1}; \mathbb{R}^l)$ is the space of absolutely continuous functions that are of class $W^{2,\infty}$ on each subinterval (t_k, t_{k+1}), supplied with the norm

$$\|x\|_{2,\infty} = \max\{|x(0)|, |\dot x(0)|, |\dot x(t_k)|, \|\ddot x\|_\infty\}.$$

The subspace N defined in (4.12) does not fully correspond to that defined in (I.5), since N_4 is a closed subspace in $W^{2,\infty}(t_k, t_{k+1}; \mathbb{R}^l)$ rather than in $W^{1,\infty}(0, T; \mathbb{R}^l)$. Accordingly, Λ_0 in (2.13) is the ball of radius $\alpha/2$ in the space

$$\widetilde{Y} = L^\infty(0, T; \mathbb{R}^n) \times \mathbb{R}^d \times L^\infty(0, T; \mathbb{R}^k) \times W^{2,\infty}(t_k, t_{k+1}; \mathbb{R}^l).$$

Note that for $j \in J_0^+(t)$ we have

$$\frac{d}{dt}\vartheta^j(x_0(t)) = D_x\vartheta^j(x_0(t))\dot{x}_0(t) = D_x\vartheta^j(x_0(t))f(x_0(t), u_0(t)) = 0. \tag{4.13}$$

Hence we can use the mixed constraints (4.13), rather than the pure state constraints (3.5), to define the subspace E_0 in (2.12). Namely E_0 is defined as the set of all pairs $(y, v) \in W^{1,2}(0, T; \mathbb{R}^n) \times L^2(0, T; \mathbb{R}^m)$ such that

$$\dot{y}(t) - A(t)y(t) - B(t)v(t) = 0, \tag{4.14}$$

$$\Xi_0 y(0) + \Xi_T y(T) = 0, \tag{4.15}$$

$$\Theta_x^\alpha(t)y(t) + \Theta_u^\alpha(t)v(t) = 0, \tag{4.16}$$

$$C^\alpha(t)y(t) + \Upsilon^\alpha(t)B(t)v(t) = 0, \tag{4.17}$$

where $\Theta_x^\alpha(t)$ (respectively, $\Theta_u^\alpha(t)$) is the matrix whose rows are the functions $D_x\theta^i(x_0(t), u_0(t))$ (respectively, $D_u\theta^i(x_0(t), u_0(t))$) for $i \in I_\alpha^+(t)$. The rows of matrices $C^\alpha(t)$ and $\Upsilon^\alpha(t)$ are given by

$$C^j(t) = D_{xx}^2\vartheta^j(x_0(t))f(x_0(t), u_0(t)) + D_x\vartheta^j(x_0(t))D_x f(x_0(t), u_0(t))$$

and $D_x\vartheta^j(x_0(t))$, respectively, for $j \in J_\alpha^+(t)$.

We have to find conditions under which the quadratic form (4.8) is positive definite on the subspace of pairs satisfying (4.14)–(4.17).

We will need a coercivity condition stronger than (4.9), satisfied with some "margin of freedom." Namely, we assume that

(II.6) there exists a constant $\bar{\alpha} > 0$ such that

$$\langle u, D_{uu}^2\mathcal{H}(t)u\rangle \geq \gamma|u|^2 \quad \text{for all } u \in \mathcal{G}_{\bar{\alpha}}(t) \text{ and all } t \in [0, T]. \tag{4.18}$$

Remark 4.3. We are not going to discuss here conditions under which (4.18) satisfied for $\bar{\alpha} = 0$ is satisfied also for some $\bar{\alpha} > 0$. The reader can find such a discussion in section 4 of [2].

For any $\alpha > 0$, let us define the matrix

$$\mathcal{K}^\alpha(t) = \begin{bmatrix} D_{uu}^2\mathcal{H}(t) & \Theta_u^\alpha(t)^* & B(t)^*\Upsilon^\alpha(t)^* \\ \Theta_u^\alpha(t) & 0 & 0 \\ \Upsilon^\alpha(t)B(t) & 0 & 0 \end{bmatrix}. \tag{4.19}$$

By (II.3) matrix $[\ \Theta_u^\alpha(t)^* \quad B(t)^*\Upsilon^\alpha(t)^*\ ]$ has the full row rank, whereas by (II.6) matrix $D_{uu}^2\mathcal{H}(t)$ is positive definite on the subspace generated by the columns of $[\ \Theta_u^\alpha(t)^* \quad B(t)^*\Upsilon^\alpha(t)^*\ ]$ for any $\alpha \leq \bar{\alpha}$. Hence for any $\alpha \leq \bar{\alpha}$, matrix $\mathcal{K}^\alpha(t)$ is non-singular (cf., e.g., Lemma 3.2 in [5]).

Let us introduce the following *matrix Riccati equation* for a symmetric $(n \times n)$-matrix function Q (cf. [15]):

$$\begin{aligned}\dot{Q}(t) = R^{\bar\alpha}(Q(t)) := &-Q(t)A(t) - A(t)^*Q(t) - D^2_{xx}\mathcal{H}(t)\\ &+\left\{\begin{bmatrix} D^2_{ux}\mathcal{H}(t)\\ \Theta^{\bar\alpha}_x(t)\\ C^{\bar\alpha}(t)\end{bmatrix}^* + Q(t)\begin{bmatrix} B(t)^*\\ 0\\ 0\end{bmatrix}^*\right\}\\ &\times \mathcal{K}^{\bar\alpha}(t)^{-1}\left\{\begin{bmatrix} B(t)^*\\ 0\\ 0\end{bmatrix}Q(t) + \begin{bmatrix} D^2_{ux}\mathcal{H}(t)\\ \Theta^{\bar\alpha}_x(t)\\ C^{\bar\alpha}(t)\end{bmatrix}\right\}.\end{aligned} \tag{4.20}$$

The following assumption is crucial for obtaining coercivity of (4.8).

(II.7) The Riccati equation (4.20) has a solution Q bounded on $[0,T]$, which satisfies the following boundary condition:

$$\begin{aligned}&\begin{bmatrix} y(0)\\ y(T)\end{bmatrix}^*\begin{bmatrix} \mathcal{R}_{11}+Q(0) & \mathcal{R}_{12}\\ \mathcal{R}_{21} & \mathcal{R}_{22}-Q(T)\end{bmatrix}\begin{bmatrix} y(0)\\ y(T)\end{bmatrix} > 0\\ &\text{for all } (y(0),y(T)) \text{ such that } \Xi_0 y(0) + \Xi_T y(T) = 0.\end{aligned} \tag{4.21}$$

LEMMA 4.4. *If* (II.1)–(II.7) *are satisfied, then there exists* $\bar\gamma > 0$ *such that*

$$\left((y,v), \begin{pmatrix} D^2_{xx}\mathcal{L}_0 & D^2_{xu}\mathcal{L}_0\\ D^2_{ux}\mathcal{L}_0 & D^2_{uu}\mathcal{L}_0\end{pmatrix}(y,v)\right) \ge \bar\gamma(\|y\|^2_2 + \|v\|^2_2) \tag{4.22}$$

for all (y,v) *satisfying* (4.14)–(4.17) *with* $\alpha = \bar\alpha$.

Proof. We will follow the idea of the proof of Lemma 4.2 in [15].

Let $Q : [0,T] \mapsto \mathbb{R}^{n\times n}$ be any Lipschitz continuous symmetric matrix function. By a direct computation, using integration by parts, we find that for any pair (y,v) satisfying (4.14) we have

$$\begin{aligned}&\int_0^T \begin{bmatrix} y(t)\\ v(t)\end{bmatrix}^*\begin{bmatrix} D^2_{xx}\mathcal{H}(t) & D^2_{xu}\mathcal{H}(t)\\ D^2_{ux}\mathcal{H}(t) & D^2_{uu}\mathcal{H}(t)\end{bmatrix}\begin{bmatrix} y(t)\\ v(t)\end{bmatrix} dt\\ &= \int_0^T \begin{bmatrix} y(t)\\ v(t)\end{bmatrix}^*\\ &\times \begin{bmatrix} \dot{Q}(t) + Q(t)A(t) + A(t)^*Q(t) + D^2_{xx}\mathcal{H}(t) & D^2_{xu}\mathcal{H}(t) + Q(t)B(t)\\ D^2_{ux}\mathcal{H}(t) + B(t)^*Q(t) & D^2_{uu}\mathcal{H}(t)\end{bmatrix}\\ &\times \begin{bmatrix} y(t)\\ v(t)\end{bmatrix} dt + y(0)^*Q(0)y(0) - y(T)^*Q(T)y(T).\end{aligned}$$

Hence, using (4.8) we obtain

$$
\begin{aligned}
&\left((y,v),\begin{pmatrix} D^2_{xx}\mathcal{L}_0 & D^2_{xu}\mathcal{L}_0 \\ D^2_{ux}\mathcal{L}_0 & D^2_{uu}\mathcal{L}_0 \end{pmatrix}(y,v)\right) \\
&= \int_0^T \begin{bmatrix} y(t) \\ v(t) \end{bmatrix}^* \\
&\times \begin{bmatrix} \dot{Q}(t)+Q(t)A(t)+A(t)^*Q(t)+D^2_{xx}\mathcal{H}(t) & D^2_{xu}\mathcal{H}(t)+Q(t)B(t) \\ D^2_{ux}\mathcal{H}(t)+B(t)^*Q(t) & D^2_{uu}\mathcal{H}(t) \end{bmatrix} \\
&\times \begin{bmatrix} y(t) \\ v(t) \end{bmatrix} dt + \begin{bmatrix} y(0) \\ y(T) \end{bmatrix}^* \begin{bmatrix} \mathcal{R}_{11}+Q(0) & \mathcal{R}_{12} \\ \mathcal{R}_{21} & \mathcal{R}_{22}-Q(T) \end{bmatrix} \begin{bmatrix} y(0) \\ y(T) \end{bmatrix},
\end{aligned}
\tag{4.23}
$$

and if Q is such that (4.21) is satisfied, then to show (4.22) it is enough to prove that the integral term in (4.23) is coercive for (y,v) satisfying (4.14)–(4.17).

To this end, let us introduce slack variables $w \in \mathbb{R}^{i(t)}$ and $z \in \mathbb{R}^{j(t)}$, where $i(t) = \operatorname{card} I^+_{\bar\alpha}(t), j(t) = \operatorname{card} J^+_{\bar\alpha}(t)$. Define the following quadratic form, which is the augmented integrant in (4.23):

$$
\begin{aligned}
&\begin{bmatrix} y \\ v \\ w \\ z \end{bmatrix}^* \mathcal{P}^{\bar\alpha}(t) \begin{bmatrix} y \\ v \\ w \\ z \end{bmatrix} \\
&= \begin{bmatrix} y \\ v \end{bmatrix}^* \begin{bmatrix} \dot{Q}(t)+Q(t)A(t)+A(t)^*Q(t)+D^2_{xx}\mathcal{H}(t) & D^2_{xu}\mathcal{H}(t)+Q(t)B(t) \\ D^2_{ux}\mathcal{H}(t)+B(t)^*Q(t) & D^2_{uu}\mathcal{H}(t) \end{bmatrix} \\
&\times \begin{bmatrix} y \\ v \end{bmatrix} + 2w^*[\Theta^{\bar\alpha}_x(t)y + \Theta^{\bar\alpha}_u(t)v] + 2z^*[C^{\bar\alpha}(t)y + \Upsilon^{\bar\alpha}(t)B(t)v)],
\end{aligned}
\tag{4.24}
$$

where $\mathcal{P}^{\bar\alpha}(t)$ is $(n+m+i(t)+j(t))$-dimensional square matrix given by

$$
\mathcal{P}^{\bar\alpha}(t) = \begin{bmatrix} \mathcal{M}(t) & \mathcal{N}^{\bar\alpha}(t)^* \\ \mathcal{N}^{\bar\alpha}(t) & \mathcal{K}^{\bar\alpha}(t) \end{bmatrix},
$$

with

$$
\mathcal{M}(t) = \dot{Q}(t) + Q(t)A(t) + A(t)^*Q(t) + D^2_{xx}\mathcal{H}(t),
$$

$$
\mathcal{N}^{\bar\alpha}(t) = \begin{bmatrix} B(t)^* \\ 0 \\ 0 \end{bmatrix} Q(t) + \begin{bmatrix} D^2_{ux}\mathcal{H}(t) \\ \Theta^{\bar\alpha}_x(t) \\ C^{\bar\alpha}(t) \end{bmatrix}.
$$

Consider the subspace $\mathcal{V}(t) \subset \mathbb{R}^{n+m+i(t)+j(t)}$ on which $\mathcal{P}^{\bar\alpha}(t)$ is *positive definite*. Let $\pi(S)$ denote the number of positive eigenvalues of a symmetric matrix S. By Theorem 1 in [7], it follows that

$$
\pi(\mathcal{P}^{\bar\alpha}(t)) = \pi(\mathcal{K}^{\bar\alpha}(t)) + \pi(\mathcal{M}(t) - \mathcal{N}^{\bar\alpha}(t)^*\mathcal{K}^{\bar\alpha}(t)^{-1}\mathcal{N}^{\bar\alpha}(t)).
$$

If (II.7) holds, then by stability results for ODEs there exist $\epsilon > 0$ and a Lipschitz continuous matrix function $\widetilde{Q}$ such that the Riccati equation

$$
\dot{\widetilde{Q}}(t) = R^{\bar\alpha}(\widetilde{Q}(t)) + \epsilon I_n
$$

is satisfied along with (4.21), where $R^{\bar\alpha}(\cdot)$ is given by the right-hand side of the Riccati equation (4.20). Putting the matrix $\widetilde{Q}(t)$ into (4.24) we obtain

$$\mathcal{M}(t)-\mathcal{N}^{\bar\alpha}(t)^*\mathcal{K}^{\bar\alpha}(t)^{-1}\mathcal{N}^{\bar\alpha}(t)=\epsilon I_n,\ \text{i.e.,}\ \ \pi(\mathcal{M}(t)-\mathcal{N}^{\bar\alpha}(t)^*\mathcal{K}^{\bar\alpha}(t)^{-1}\mathcal{N}^{\bar\alpha}(t))=n.$$

On the other hand, (II.3) and (II.6) yield

$$\pi(\mathcal{K}^{\bar\alpha}(t))\ge m-i(t)-j(t).$$

Hence we get

$$\pi(\mathcal{P}^{\bar\alpha}(t))\ge n+m-i(t)-j(t); \tag{4.25}$$

i.e., there exists a subspace

$$\mathcal{V}(t)\subset\mathbb{R}^{n+m+i(t)+j(t)},\quad \dim\mathcal{V}(t)\ge n+m-i(t)-j(t),$$

which is generated by appropriate eigenvectors of $\mathcal{P}^{\bar\alpha}(t)$, such that the quadratic form (4.24) is positive definite on $\mathcal{V}(t)$. The right-hand side of (4.24) shows that on $\mathcal{V}(t)$ we must have

$$\begin{array}{lll} & \text{either } w^i=0 \ \text{ or } \ \Theta^i_x(t)y+\Theta^i_u(t)v=0 & \text{for all } i\in I^+_{\bar\alpha}(t)\\ \text{and} & \text{either } z^j=0 \ \text{ or } \ C^j(t)y+\Upsilon^j(t)B(t)v)=0 & \text{for all } j\in J^+_{\bar\alpha}(t).\end{array} \tag{4.26}$$

Let $\bar\lambda\ne0$ be any nonzero eigenvalue of $\mathcal{P}^{\bar\alpha}(t)$ and $(\bar y^*,\bar v^*,\bar w^*,\bar z^*)$ be the corresponding eigenvector, i.e.,

$$\mathcal{P}^{\bar\alpha}(t)\begin{bmatrix}\bar y\\ \bar v\\ \bar w\\ \bar z\end{bmatrix}=\bar\lambda\begin{bmatrix}\bar y\\ \bar v\\ \bar w\\ \bar z\end{bmatrix}.$$

By definition of $\mathcal{P}^{\bar\alpha}(t)$, the last $i(t)+j(t)$ rows of this equation take on the form

$$\Theta^{\bar\alpha}_x(t)\bar y+\Theta^{\bar\alpha}_u(t)\bar v=\bar\lambda\bar w,\qquad C^{\bar\alpha}(t)\bar y+\Upsilon^{\bar\alpha}(t)B(t)\bar v=\bar\lambda\bar z.$$

In view of (4.26), it implies

$$\begin{array}{ll}\Theta^{\bar\alpha}_x(t)\bar y+\Theta^{\bar\alpha}_u(t)\bar v=0 & \text{and } \bar w=0,\\ C^{\bar\alpha}(t)\bar y+\Upsilon^{\bar\alpha}(t)B(t)\bar v=0 & \text{and } \bar z=0.\end{array} \tag{4.27}$$

Therefore, all eigenvectors corresponding to nonzero eigenvalues belong to $(n+m-i(t)-j(t))$-dimensional subspace of $(n+m+i(t)+j(t))$-dimensional vectors satisfying (4.27). In view of (4.25), this subspace coincides with $\mathcal{V}(t)$, which means that all $(n+m-i(t)-j(t))$ nonzero eigenvalues $\bar\lambda\ne0$ of $\mathcal{P}^{\bar\alpha}(t)$ are positive. Moreover, since the lower bounds in (II.3) and (II.6) are uniform with respect to $t\in[0,T]$, we can find a uniform on $[0,T]$ lower bound $\bar\gamma>0$ for all positive eigenvalues of $\mathcal{P}^{\bar\alpha}(t)$. Hence, using (4.21), (4.24), and (4.27) we obtain (4.22) from (4.23). □

PROPOSITION 4.5. *If assumptions* (II.1)–(II.7) *hold, then there exist constants* $\rho_1>0$ *and* $\gamma_1>0$ *such that*

$$F(x,u)\ge F(x,u)+\gamma_1(\|x-x_0\|_2^2+\|u-u_0\|_2^2) \tag{4.28}$$

for all (x,u) *satisfying* (3.2)–(3.5) *and such that*

$$\max\{\|x-x_0\|_\infty, \|u-u_0\|_\infty\} \le \rho_1.$$

Proof. We will follow the idea of the proof of Theorem 2.3.

Note that, for feasible (x,u) the components of the term $(\kappa_0, \theta(x,u))$ in Lagrangian (4.1) can be estimated as follows:

$$\begin{aligned} -\int_0^T \kappa_0^i(t)\theta^i(x(t),u(t))dt &\ge -\int_{L_{\bar\alpha}^i} \kappa_0^i(t)\theta^i(x(t),u(t))dt \\ &\ge \bar\alpha \int_{L_{\bar\alpha}^i} |\theta^i(x(t),u(t))|dt, \end{aligned} \tag{4.29}$$

where

$$L_\alpha^i = \{t\in[0,T] \mid i\in I_\alpha^+(t)\} := \{t\in[0,T] \mid \kappa_0^i(t) > \alpha\}.$$

Similarly, by the regularity of μ_0 (cf. Lemma 4.2) and by (4.6), for any feasible (x,u) and for any component of the last two terms in (4.1) we have

$$\begin{aligned} &-\mu_0^j(0)\vartheta^j(x(0)) - \int_0^T \dot\mu_0^j(t) D_x\vartheta^j(x(t))f(x(t)),u(t))dt \\ &= -\mu_0^j(0)\vartheta^j(x(0)) - \int_0^T \dot\mu_0^j(t)\frac{d}{dt}\vartheta^j(x(t))dt \\ &= -(\mu_0^j(0)-\dot\mu_0^j(0))\vartheta^j(x(0)) + \Sigma_k(\dot\mu(t_k+)-\dot\mu(t_{k+1}-))\vartheta(x(t_k) \\ &-\dot\mu_0^j(T)\vartheta^j(x(T)) + \int_0^T \ddot\mu_0^j(t)\vartheta^j(x(t))dt \\ &\ge \int_{P_{\bar\alpha}^j} \ddot\mu_0^j(t)\vartheta^j(x(t))dt \ge \bar\alpha\int_{P_{\bar\alpha}^j} |\vartheta^j(x(t))|dt, \end{aligned} \tag{4.30}$$

where

$$P_\alpha^j = \{t\in[0,T] \mid j\in J_\alpha^+(t)\} := \{t\in[0,T] \mid -\ddot\mu_0^j(t) > \alpha\}.$$

Estimates (4.29) and (4.30) show that condition (A.2) in the proof of Theorem 2.3 is satisfied with the norm of Y^* substituted by that of $L^1(0,T;\mathbb{R}^k)\times L^1(0,T;\mathbb{R}^l)$. Hence, using (4.22) we can repeat the further steps of that proof and arrive at (4.28). □

Note that it follows from (3.2) that for any feasible (x,u) we have

$$\|\dot x-\dot x_0\|_2 \le c(\|x-x_0\|_2 + \|u-u_0\|_2), \qquad \|\dot x-\dot x_0\|_\infty \le c(\|x-x_0\|_\infty + \|u-u_0\|_\infty).$$

Hence, from Proposition 4.5 we obtain the following SSC result for (O).

THEOREM 4.6. *If assumptions* (II.1)–(II.7) *hold, then there exist constants* $\rho>0$ *and* $\gamma>0$ *such that*

$$F(x,u) \ge F(x_0,u_0) + \gamma(\|x-x_0\|_{1,2}^2 + \|u-u_0\|_2^2) \tag{4.31}$$

for all (x,u) *satisfying* (3.2)–(3.5) *and such that*

$$\max\{\|x-x_0\|_{1,\infty}, \|u-u_0\|_\infty\} \le \rho.$$

Remark 4.7 Theorem 4.6 can be viewed as a generalization of the results due to Dontchev et al. [2], Maurer and Pickenhain [17], and Zeidan [30] to the optimal control problems involving pure state-space constraints of order 1. This generalization is not trivial, since it requires the analysis of regularity of the solutions and Lagrange multipliers, which is not needed if these constraints are not present.

Without state-space constraints, Theorem 4.6 reduces to the SSC derived in [30], where a Riccati equation in the form different from but equivalent to (4.20) was used. In case with pure control constraints only, the result stronger than Theorem 4.6 was obtained in [2], but without characterization of the coercivity condition (4.22) by a Riccati equation.

5. Application to stability analysis. In this section the role of second-order sufficient optimality conditions in stability analysis of solutions to parametric optimal control problems is briefly discussed. Only the main results are presented. For the relevant proofs, the reader is refered to the technical report [14].

Let us introduce a Banach space H of parameters and an open set $G \subset H$ of feasible parameters.

Problem (O) is embedded in a family of parametric problems (O_h) depending upon $h \in G$:

$$
\begin{aligned}
(\mathrm{O}_h) \qquad & \text{Find} (x_h, u_h) \in W^{1,\infty}(0,T;\mathbb{R}^n) \times L^\infty(0,T;\mathbb{R}^m) \text{ such that} \\
& F(x_h, u_h, h) = \min\{F(x,u,h) \\
& \qquad\qquad := \int_0^T f^0(x(t),u(t),h)dt + g(x(T),h)\} \\
& \text{subject to} \\
& \dot{x}(t) - f(x(t),u(t),h) = 0 \qquad \text{for a.a. } t \in [0,T], \\
& \xi(x(0),x(T),h) = 0, \\
& \theta(x(t),u(t),h) \le 0 \qquad \text{for a.a. } t \in [0,T], \\
& \vartheta(x(t),h) \le 0 \qquad \text{for all } t \in [0,T].
\end{aligned}
$$

We assume that

(III.1) functions $f^0(\cdot,\cdot,\cdot), g(\cdot,\cdot), f(\cdot,\cdot,\cdot), \xi(\cdot,\cdot,\cdot), \theta(\cdot,\cdot,\cdot), \vartheta(\cdot,\cdot)$, and $D_x\vartheta(\cdot,\cdot)$ are twice Fréchet differentiable in all arguments, and the respective derivatives are locally Lipschitz continuous in u, x.

(III.2) The space H is independent of t.

(III.3) There exists a possibly local solution $(x_{h_0}, u_{h_0}) := (x_0, u_0)$ of the reference problem (P_{h_0}), where u_0 is a piecewise continuous function with the finite set of discontinuity points denoted by $\{t_k\}$.

In stability analysis we are interested in sufficient conditions under which a neighborhood $G_0 \subset G$ of h_0 exists such that for each $h \in G_0$ there is a locally unique solution (x_h, u_h) of (P_h), which is a Lipschitz continuous function of h.

Such an analysis for a class of optimal control problems very similar to (O_h) was performed in [13]. The main tool of this analysis was a modification of Robinson's implicit function theorem for generalized equations [25] developed in [12]. In this modification additional information on regularity of the solutions and the Lagrange multipliers, as functions of time, is used to overcome the difficulty connected with *two-norm discrepancy.*

In Theorem 7.1 of [13] conditions are formulated under which the solutions to (O_h) are locally Lipschitz continuous functions of h. These conditions consist of con-

straint qualifications, virtually identical to (II.3) and (II.4), and second-order sufficient optimality conditions.

The second-order sufficient optimality condition used therein is much stronger than the SSC derived in section 4. Namely, it is required that the coercivity condition (4.22) is satisfied on the subspace of all pairs $(y,v) \in W^{1,2}(0,T;\mathbb{R}^n) \times L^2(0,T;\mathbb{R}^m)$ satisfying equations (4.14) and (4.15). This is the subspace orthogonal to linearization of equality-type constraints only, and the active inequality type constraints are not taken into account.

It is natural to ask the question, *can the same stability results be obtained under the weaker coercivity condition given in* (4.22)?

This problem is analyzed in part II of [14], where it is shown that (4.22) is *too weak* to repeat the proof of the local stability of the solutions to (P_h) given in [13]. More precisely, equation (4.16) corresponding to the active mixed control-state constraints can be included in the needed coercivity condition; however, it is not possible for equality (4.17) corresponding to the active pure state-space constraints.

This phenomenon is connected with different properties of the Lagrange multipliers associated with those two types of constraints.

Thus, the second-order sufficient optimality condition needed in stability analysis based on Robinson's implicit function theorem takes on the following form:

($\widetilde{\mathrm{SSC}}$) There exist constants $\bar\alpha > 0$ and $\bar\gamma > 0$ such that

$$\left((y,v), \begin{pmatrix} D^2_{xx}\mathcal{L}_0 & D^2_{xu}\mathcal{L}_0 \\ D^2_{ux}\mathcal{L}_0 & D^2_{uu}\mathcal{L}_0 \end{pmatrix}(y,v)\right) \ge \bar\gamma(\|y\|_2^2 + \|v\|_2^2)$$

for all $(y,v) \in W^{1,2}(0,T;\mathbb{R}^n) \times L^2(0,T;\mathbb{R}^m)$ satisfying

$$\begin{aligned} &\dot y(t) - A(t)y(t) - B(t)v(t) = 0,\\ &\Xi_0 y(0) + \Xi_T y(T) = 0,\\ &\Theta^{\bar\alpha}_x(t)y(t) + \Theta^{\bar\alpha}_u(t)v(t) = 0, \end{aligned} \tag{5.1}$$

where the notation is the same as in section 4.

Remark 5.1. Condition ($\widetilde{\mathrm{SSC}}$) is *stronger* than (4.22), but it is *weaker* than the condition used in [13], where the last equation in (5.1) was void.

Note that in the same way as in section 4, condition ($\widetilde{\mathrm{SSC}}$) can be expressed in terms of the modified Legendre–Clebsch condition analogous to (II.6) as well as the Riccati equation analogous to (II.7). The main stability result of [14] (Theorem 3.7 in part II) can be formulated as follows.

THEOREM 5.2. *If* (III.1)–(III.3), (II.4), (II.5)*, and* ($\widetilde{\mathrm{SSC}}$) *are satisfied, then there exist a neighborhood $G_0 \subset G$ of h_0 and a neighborhood $\mathcal{Z}_0 \subset Z$ of z_0 such that for all $h \in G_0$ there exist a unique in $\mathcal{Z}_0$ solution (x_h, u_h) of (O_h) and unique associated Lagrange multipliers $(q_h, \rho_h, \kappa_h, \mu_h)$.*

Moreover, there exists a constant $c > 0$ such that

$$\|z_{h_1} - z_{h_2}\|_{1,2}, \|u_{h_1} - u_{h_2}\|_2, \|q_{h_1} - q_{h_2}\|_{1,2}, |\rho_{h_1} - \rho_{h_2}|,$$

$$\|\kappa_{h_1} - \kappa_{h_2}\|_2, \|\mu_{h_1} - \mu_{h_2}\|_{1,2} \le c\, [\![h_1 - h_2]\!]_H \qquad \textit{for all } h_1, h_2 \in G_0.$$

Appendix A. Proof of Theorem 2.3. Let z be feasible for (P), i.e., $\varphi(z) \in K$. Expanding $\mathcal{L}(\cdot,\lambda_0)$ into Taylor's series at z_0 and using (I.1) and (2.9) we obtain

$$F(z) - F(z_0) = (\lambda_0, \varphi(z))_{\widehat Y} + \frac{1}{2}(D^2_{zz}\mathcal{L}(z_0,\lambda_0)(z-z_0),(z-z_0))_{\widehat Z} + \xi(z,z_0), \tag{A.1}$$

where $\frac{|\xi(z,z_0)|}{\|z-z_0\|^2_{\widehat Z}} \to 0$ as $[\![z - z_0]\!]_Z \to 0$.

To estimate the term $(\lambda_0, \varphi(z))_{\widehat{Y}}$ let us represent any $y \in K$ in the form $y = y^{\perp} + y^{\parallel}$, where $y^{\perp}$ and $y^{\parallel}$ are orthogonal projections onto $\widehat{N}^{\perp}$ and onto its orthogonal complement $\widehat{N} \in \widehat{Y}$, respectively.

By (2.13) we have

$$(\lambda_0 - k, y)_{\widehat{Y}} \geq 0 \qquad \text{for all } k \in (\Lambda_0 - \lambda_0) \cap N.$$

Hence

$$(\lambda_0, y)_{\widehat{Y}} \geq \sup_{k \in (\Lambda_0 - \lambda_0) \cap N} (k, y^{\perp} + y^{\parallel})_{\widehat{Y}} = \sup_{k \in (\Lambda_0 - \lambda_0) \cap N} (k, y^{\parallel})_{\widehat{Y}} = \sigma [\![y^{\parallel}]\!]_{Y^*}.$$

Thus we obtain

$$(\lambda_0, \varphi(z))_{\widehat{Y}} \geq \sigma [\![\varphi(z)^{\parallel}]\!]_{Y^*}. \tag{A.2}$$

Note that by (2.9) and (A.2), we get

$$\varphi(z_0)^{\parallel} = 0. \tag{A.3}$$

On the other hand, by (I.1) we have

$$\varphi(z) = \varphi(z_0) + D_z \varphi(z_0)(z - z_0) + \zeta(z, z_0), \tag{A.4}$$

where $\frac{\|\zeta(z,z_0)\|_Y}{\|z - z_0\|_{\widehat{Z}}} \to 0$ as $[\![z - z_0]\!]_Z \to 0$.

By (2.5) and (2.13) there exists a solution $(z - z_0)^{\parallel}$ of the equation

$$D_z \varphi(z_0)(z - z_0)^{\parallel} = (\varphi(z) - \varphi(z_0) - \zeta(z, z_0))^{\parallel} = \varphi(z)^{\parallel} - \zeta(z, z_0)^{\parallel} \tag{A.5}$$

such that

$$\begin{aligned} \|(z - z_0)^{\parallel}\|_{\widehat{Y}} &\leq c \left(\|\varphi(z)^{\parallel}\|_{\widehat{Y}} + \|\zeta(z, z_0)^{\parallel}\|_{\widehat{Y}} \right) \\ &\leq c \left(\|\varphi(z)^{\parallel}\|_{\widehat{Y}} + \|\zeta(z, z_0)\|_{\widehat{Y}} \right). \end{aligned} \tag{A.6}$$

Let us define

$$(z - z_0)^{\perp} = (z - z_0) - (z - z_0)^{\parallel}.$$

In view of (A.3), (A.4), and (A.5) we have

$$D_z \varphi(z_0)(z - z_0)^{\perp} \in \widehat{N}^{\perp}. \tag{A.7}$$

By (I.5), (2.10), and (A.2) we obtain from (A.1)

$$\begin{aligned} F(z) - F(z_0) \geq{}& \sigma [\![\varphi(z)^{\parallel}]\!]_{Y^*} + \tfrac{\gamma}{2} \|(z - z_0)^{\perp}\|^2_{\widehat{Z}} \\ &- c \|(z - z_0)^{\perp}\|_{\widehat{Z}} \|(z - z_0)^{\parallel}\|_{\widehat{Z}} - \tfrac{c}{2} \|(z - z_0)^{\parallel}\|^2_{\widehat{Z}} - |\xi(z, z_0)|. \end{aligned} \tag{A.8}$$

It follows from (A.1), (A.4), and (A.6) that for each $\epsilon > 0$ there exists $\rho(\epsilon) > 0$ such that for all $[\![z - z_0]\!]_Z \leq \rho(\epsilon)$ we have

$$\begin{aligned} &|\xi(z, z_0)| \leq \epsilon \|z - z_0\|^2_{\widehat{Z}}, \\ &\|(z - z_0)^{\parallel}\|_{\widehat{Z}} \leq c \left(\|\varphi(z)^{\parallel}\|_{\widehat{Y}} + \epsilon \|(z - z_0)^{\perp}\|_{\widehat{Z}} \right). \end{aligned} \tag{A.9}$$

Substituting these estimates into (A.8), using Young's inequality, and performing some elementary calculations, for sufficiently small $\rho(\epsilon)$ we obtain

$$F(z) - F(z_0) \geq \sigma [\![\varphi(z)^{\|}]\!]_{Y^*} + \gamma' \|(z - z_0)^{\perp}\|_{\widehat{Z}}^2 - c' \|\varphi(z)^{\|}\|_{\widehat{Y}}^2. \tag{A.10}$$

On the other hand, for any $y \in Y$ we have

$$\|y\|_{\widehat{Y}}^2 = (y, y)_{\widehat{Y}} \leq [\![y]\!]_{Y^*} [\![y]\!]_Y, \quad \text{i.e., } \frac{\|y\|_{\widehat{Y}}^2}{[\![y]\!]_{Y^*}} \leq [\![y]\!]_Y.$$

Note that by (2.14), (A.3), and continuity of $\varphi(z)$ we have $[\![\varphi(z)^{\|}]\!]_Y \to 0$ as $[\![z - z_0]\!]_Z \to 0$. Hence, shrinking $\rho(\epsilon)$ if necessary, we obtain from (A.10)

$$F(z) - F(z_0) \geq \sigma' [\![\varphi(z)^{\|}]\!]_{Y^*} + \gamma'' \|(z - z_0)^{\perp}\|_{\widehat{Z}}^2.$$

Using (A.9) again we arrive at (2.16).

Acknowledgments. The author would like to express his gratitude to Professor Helmut Maurer for helpful discussions and suggestions.

REFERENCES

[1] A. L. Dontchev and W. W. Hager, *Lipschitz stability in nonlinear control and optimization*, SIAM J. Control Optim., 31 (1993), pp. 569–603.
[2] A. L. Dontchev, W. W. Hager, A. B. Poore, and B. Yang, *Optimality, stability and covergence in nonlinear control*, Appl. Math. Optim., 31 (1995), pp. 297–326.
[3] A. V. Fiacco, *Introduction to Sensitivity and Stability Analysis in Nonlinear Programming*, Academic Press, New York, 1983.
[4] I. M. Gelfand and S. W. Fomin, *Calculus of Variations*, Prentice-Hall, Englewood Cliffs, NJ, 1963.
[5] W. W. Hager, *Lipschitz continuity for constrained processes*, SIAM J. Control Optim., 17 (1979), pp. 321–338.
[6] R. F. Hartl, S. P. Sethi, and R. G. Vickson, *A survey of the maximum principle for optimal control problems with state constraints*, SIAM Rev., 37 (1995), pp. 181–218.
[7] E. V. Haynsworth, *Determination of the inertia of a partitioned Hermitian matrix*, Linear Algebra Appl., 1 (1968), pp. 73–81.
[8] K. Jittorntrum, *Solution point differentiability without strict complementarity in mathematical programming*, Math. Programming Study, 21 (1984), pp. 127–139.
[9] M. Kojima, *Strongly stable solutions in nonlinear programs*, In Analysis and Computation of Fixed Points, S. M. Robinson, ed., Academic Press, New York, 1980, pp. 93–138.
[10] F. Lempio and H. Maurer, *Differential stability in infinite-dimensional nonlinear programming*, Appl. Math. Optim., 6 (1980), pp. 139–152.
[11] K. Malanowski, *Second order conditions and constraint qualifications in stability and sensitivity analysis of solutions to optimization problems in Hilbert spaces*, Appl. Math. Optim., 25 (1992), pp. 51–79.
[12] K. Malanowski, *Two-norm approach in stability and sensitivity analysis of optimization and optimal control problems*, Adv. Math. Sci. Appl., 2 (1993), pp. 397–443.
[13] K. Malanowski, *Stability and sensitivity of solutions to nonlinear optimal control problems*, Appl. Math. Optim., 32 (1995), pp. 111–141.
[14] K. Malanowski, *Sufficient Optimality Conditions in Optimal Control*, Working Paper WP-1-1994, Systems Research Institute, Warszawa, 1994.
[15] K. Malanowski and H. Maurer, *Sensitivity analysis for parametric optimal control problems with control-state constraints*, Comput. Optim. Appl., 5 (1996), pp. 253–283.
[16] H. Maurer, *First- and second-order sufficient optimality conditions in mathematical programming and optimal control*, Math. Programming Study, 14 (1981), pp. 43–62.
[17] H. Maurer and S. Pickenhain, *Second order sufficient conditions for optimal control problems with control-state constraints*, J. Optim. Theory Appl., 86 (1995), pp. 649–667.
[18] H. Maurer and J. Zowe, *First- and second-order conditions in infinite-dimensional programming problems*, Math. Programming, 16 (1979), pp. 98–110.

[19] L. W. Neustadt, *Optimization: A Theory of Necessary Conditions*, Princeton University Press, Princeton, NJ, 1976.
[20] D. Orrell and V. Zeidan, *Another Jacobi sufficiency criterion for optimal control with smooth constraints*, J. Optim. Theory Appl., 58 (1988), pp. 283–300.
[21] J. V. Outrata and Z. Schindler, *An augmented Lagrangian method for a class of convex optimal control problems*, Problems of Control and Information Theory, 10 (1981), pp. 67–81.
[22] S. Pickenhain, *Sufficiency conditions for weak local minimum in multidimensional optimal control problems with mixed control-state restrictions*, Z. Analysis Anwendungen, 11 (1992), pp. 559–568.
[23] S. Pickenhain and K. Tammer, *Sufficient conditions for local optimality with state restrictions*, Z. Analysis Anwendungen, 10 (1991), pp. 397–405.
[24] W. T. Reid, *Riccati Differential Equations*, Academic Press, New York, 1972.
[25] S. M. Robinson, *Strongly regular generalized equations*, Math. Oper. Res., 5 (1980), pp. 43–62.
[26] S. M. Robinson, *Local structure of feasible sets in nonlinear programming, Part* III*: Stability and sensitivity*, Math. Programming Study, 30 (1987), pp. 45–66.
[27] A. Shapiro and J. F. Bonnans, *Sensitivity analysis of parametrized programs under cone constraints*, SIAM J. Control Optim., 30 (1992), pp. 97–116.
[28] V. Zeidan, *First- and second-order sufficient conditions for optimal control and calculus of variations*, Appl. Math. Optim., 11 (1984), pp. 209–226.
[29] V. Zeidan, *Sufficient conditions with minimal regularity assumptions*, Appl. Math. Optim., 20 (1989), pp. 19–31.
[30] V. Zeidan, *The Riccati equation for optimal control problems with mixed state-control constraints; necessity and sufficiency*, SIAM J. Control Optim., 32 (1994), pp. 1297–1321.

SIAM J. CONTROL OPTIM.
Vol. 35, No. 1, pp. 228–242, January 1997

012

OVERLAPPING BLOCK-BALANCED CANONICAL FORMS AND PARAMETRIZATIONS: THE STABLE SISO CASE*

BERNARD HANZON† AND RAIMUND J. OBER‡

Abstract. The balanced canonical form and parametrization of Ober for the case of SISO stable systems are extended to block-balanced canonical forms and related input-normal forms and parametrizations. They form an overlapping atlas of parametrizations of the manifold of stable SISO systems of given order. This extends the usefulness of these parametrizations, e.g., in gradient algorithms for system identification. As an implication of our construction it follows that each of the subsets of the parametrization of [R. Ober, *Internat. J. Control*, 46 (1987), pp. 643–670] corresponding to a choice for the structural indices is in fact an imbedded submanifold of the manifold of stable SISO systems of fixed order.

Key words. linear dynamical systems, differentiable manifolds, stable systems, canonical forms, atlas, system identification

AMS subject classifications. 93XX, 53XX, 15XX

PII. S0363012993260549

1. Introduction.

In [18], [19] a canonical state-space form was presented for the set of asymptotically stable linear systems, with the property that it is balanced; i.e., for each system represented in canonical form, the corresponding observability and controllability Gramians are equal and diagonal (and positive definite). One motivation for studying balanced realizations and balanced canonical forms is their close relation to model reduction (see [19] and the references given there), which is in turn closely related to robust control theory (see, e.g., [20], [3]). Another motivation mentioned in [19] is the potential usefulness of balanced realizations for system identification, as indicated by [15]. In many cases, in system identification as well as in related areas, one can reduce the problem at hand to an optimization problem in which some criterion function is optimized over a set of systems. Very often one cannot solve the optimization problem analytically and has to use search algorithms (e.g., gradient algorithms), in which an initial point in the set of systems is adapted iteratively to give, ideally, a good approximation of the optimal system. In such search algorithms one often uses a parametrization of the set of relevant systems. The balanced parametrization of [19] has the advantage that by construction, problems of identifiability are to a large extent avoided in such a search algorithm. The parametrization has the property that it contains structural indices (i.e., discrete-valued parameters), and to each possible choice of values for these indices corresponds a particular subset of systems, for which a parametrization in terms of real-valued parameters is given. (In fact it will be shown in section 6 that these subsets are in fact submanifolds.) To each system corresponds a unique set of structural indices. Since the structural indices can take a large number of values, even for rather low order systems (the number of possibilities increases fast with increasing order of the system), this means that in a search algorithm one has either to identify the structural indices by other means

*Received by the editors December 27, 1993; accepted for publication (in revised form) November 20, 1995. A first version of this paper was written while the first author visited the Department of Engineering, Cambridge University.

http://www.siam.org/journals/sicon/35-1/26054.html

†De Boelelaan 1105, 1081 HV Amsterdam, the Netherlands (bhanzon@econ.vu.nl).

‡Center for Engineering Mathematics, University of Texas at Dallas, Richardson, TX 75083-06688 (ober@utdallas.edu).

or to apply the search algorithm to a large number of parametrized submanifolds of systems. This is due to the fact that the parametrizations are disjoint.

Several authors (e.g. [4, 2, 10, 11, 20, 5, 6, 21, 22]) have investigated the possibility of using so-called overlapping parametrizations (in differential geometric terms: an atlas of coordinate charts). If one uses overlapping parametrizations, one does not have to search through each and every of the submanifolds but instead can search through the manifold as a whole, using the parametrizations to describe the manifold locally and changing from one parametrization to another when required. In case the search algorithm is of the gradient type, one can make sure that the decision rule for changing from one parametrization to another has little effect on the search algorithm by using a Riemannian gradient with respect to some suitable Riemannian metric on the manifold (cf. [7, 6, 8, 22, 9, 21]).

In view of this it would be very desirable if the balanced parametrization of [19] could be extended to give a set of overlapping parametrizations. In this paper such an extension, will be presented for the case of SISO stable systems. In the extension, balancedness of the realization no longer holds for all realizations. Instead block-balanced realizations and the corresponding input-normal realizations are used. A block-balanced canonical form is a canonical form for which the observability and controllability Gramians are equal and block-diagonal (and of course positive definite).

In section 2 some basic definitions are presented, including the concept of block-balanced realizations. In section 3 we present a Schwarz-like canonical form which will be a building block in the block-balanced canonical forms and the corresponding input-normal canonical forms that are treated in section 4. In section 5 it is shown how this leads to a set of overlapping block-balanced canonical forms and a corresponding atlas for the manifold of stable SISO input-output systems of a fixed order, and remarks are made as to how this atlas can be used if one wants to work with balanced and "almost balanced" realizations in search algorithms in system identification, for example. In section 6 the imbedded submanifolds structure of the original balanced parametrization is analyzed, using the atlas of the previous section.

2. Canonical forms, balanced realizations, and block-balanced realizations. In this section to a large extent the setup of [19] is followed. Let us consider continuous-time SISO systems of the form

$$\dot{x}_t = Ax_t + bu_t, \tag{1}$$

$$y_t = cx_t, \tag{2}$$

with $t \in \mathbf{R}, u_t \in \mathbf{R}, x_t \in \mathbf{R}^n, y_t \in \mathbf{R}, A \in \mathbf{R}^{n\times n}, b \in \mathbf{R}^{1\times n}, c \in \mathbf{R}^{n\times 1}$, and (A, b, c) a minimal triple.

For each $n \in \{1, 2, 3, \ldots\}$ let the set C_n be given by $C_n = \{(A, b, c) \in \mathbf{R}^{n\times n} \times \mathbf{R}^{n\times 1} \times \mathbf{R}^{1\times n} | (A, b, c)$ minimal and the spectrum of A is contained in the open left half plane$\}$.

As is well known, two minimal system representations (A_1, b_1, c_1) and (A_2, b_2, c_2) have the same transfer function, $g(s) = c_1(sI - A_1)^{-1}b_1 = c_2(sI - A_2)^{-1}b_2$, and therefore describe the same input-output behavior iff there exists an $n \times n$ matrix $T \in Gl_n(\mathbf{R})$ such that $A_1 = TA_2T^{-1}, b_1 = Tb_2, c_1 = c_2T^{-1}$. In that case we say that (A_1, b_1, c_1) and (A_2, b_2, c_2) are i/o-equivalent. This is clearly an equivalence relation; write $(A_1, b_1, c_1) \sim (A_2, b_2, c_2)$. A unique representation of a linear system can be obtained by deriving a canonical form.

DEFINITION 2.1. *A canonical form for an equivalence relation $\sim$ on a set X is a map*

$$\Gamma : X \to X$$

which satisfies, for all $x, y \in X$,

(i) $\Gamma(x) \sim x$;

(ii) $x \sim y \Longrightarrow \Gamma(x) = \Gamma(y)$.

Equivalently a canonical form can be given by the image set $\Gamma(X)$; a subset $B \subseteq X$ describes a canonical form if for each $x \in X$ there is precisely one element $b \in B$ such that $b \sim x$. The mapping $X \to B, x \mapsto b$ then describes a canonical form.

Let $(A, b, c) \in C_n$. The controllability Gramian W_c is the positive definite matrix that is given by the integral

$$W_c = \int_0^\infty \exp(At) b b^T \exp(A^T t) \mathrm{d}t.$$

As is well known, W_c can be obtained as the unique solution of the following Lyapunov equation:

$$AW_c + W_c A^T = -bb^T. \tag{3}$$

In a dual fashion, the observability Gramian W_o is the positive definite matrix that is given by the integral

$$W_o = \int_0^\infty \exp(A^T t) c^T c \exp(At) \mathrm{d}t.$$

This matrix is the unique solution of the following Lyapunov equation:

$$A^T W_o + W_o A = -c^T c. \tag{4}$$

DEFINITION 2.2. *Let $(A, b, c) \in C_n$. Then (A, b, c) is called* balanced *if the corresponding observability and controllability Gramians are equal and diagonal; i.e., there exist positive numbers $\sigma_1, \sigma_2, \ldots, \sigma_n$ such that*

$$W_o = W_c = \mathrm{diag}(\sigma_1, \ldots, \sigma_n) =: \Sigma. \tag{5}$$

The numbers $\sigma_1, \ldots, \sigma_n$ are called the (Hankel) singular values of the system.

The singular values are known to be uniquely determined by the input-output behavior of the system.

THEOREM 2.3 (see [17]). *Let $(A, b, c) \in C_n$ with*

$$\Sigma = \mathrm{diag}(\sigma_1 I_{n(1)}, \ldots, \sigma_k I_{n(k)}), \quad \sigma_1 > \sigma_2 > \ldots \sigma_k > 0, \quad \text{and} \quad \sum_{i=1}^{k} n(i) = n.$$

Then (A, b, c) is unique within its i/o-equivalence class up to an orthogonal state-space transformation of the form

$$Q = \mathrm{diag}(Q_1, Q_2, \ldots, Q_k)$$

with orthogonal $Q_i \in \mathbf{R}^{n(i) \times n(i)}, i = 1, \ldots, k$.

DEFINITION 2.4. *Let* $(A,b,c) \in C_n$. *Then* (A,b,c) *is called* input-normal *if* $W_c = I_n$ *and will be called* σ-input-normal *if* $W_c = \sigma I_n$.

Similarly (A,b,c) *is called* output-normal *if* $W_o = I_n$ *and* σ-output-normal *if* $W_o = \sigma I_n$.

It is not difficult to show that an input-normal realization is unique up to an arbitrary orthogonal state-space transformation.

The following definition is basic to our considerations in this paper.

DEFINITION 2.5. *Let* $(A,b,c) \in C_n$. *Then* (A,b,c) *will be called* block-balanced, *with indices* $n(i) \in \mathbf{N}, i = 1,\dots,k$, *adding up to* n, *if the observability Gramian and the controllability Gramian are equal and block-diagonal; i.e., there exist* $n(i) \times n(i)$ *positive definite matrices* $\Sigma_i, i = 1,\dots,k$, *such that*

$$W_o = W_c = \operatorname{diag}(\Sigma_1,\dots,\Sigma_k).$$

It will be convenient to call an arbitrary system representation $(A,b,c) \in \mathbf{R}^{n\times n} \times \mathbf{R}^{n\times 1} \times \mathbf{R}^{1\times n}$ *block-balanced if the pair of Lyapunov equations* $A\Sigma + \Sigma A^T = -bb^T$, $A^T\Sigma + \Sigma A = -c^Tc$ *has a positive definite solution of the form* $\Sigma = \operatorname{diag}(\Sigma_1,\dots,\Sigma_k)$ *(assuming neither asymptotic stability nor minimality).*

Remark. The matrices $\Sigma_i, i = 1,\dots,k$, are in general *not* uniquely determined by the input-output behavior of the system. However, the eigenvalues $\lambda_1(\Sigma_i) \geq \lambda_2(\Sigma_i) \geq \cdots \geq \lambda_{n(i)}(\Sigma_i)$ of the matrices $\Sigma_i, i = 1,\dots,k$, together form the set of Hankel singular values of the system, which are uniquely determined by the input-output behavior of the system, as remarked before.

THEOREM 2.6. *Suppose that* $(A,b,c) \in C_n$ *is block-balanced with indices* $n(j) \in \mathbf{N}, j = 1,\dots,k, \sum_{j=1}^{k} n(j) = n$, *and the additional property* $\lambda_1(\Sigma_1) \geq \lambda_{n(1)}(\Sigma_1) > \lambda_1(\Sigma_2) \geq \lambda_{n(2)}(\Sigma_2) > \cdots > \lambda_1(\Sigma_k) \geq \lambda_{n(k)}(\Sigma_k) > 0$.

This uniquely determines (A,b,c) *within its i/o-equivalence class up to an orthogonal state-space transformation of the form*

$$Q = \operatorname{diag}(Q_1,\dots,Q_k)$$

with orthogonal $Q_i \in \mathbf{R}^{n(i)\times n(i)}, i = 1,\dots,k$.

Proof. First note that if an *orthogonal* state-space transformation Q is applied to the system representation, then both Gramians transform in the same way, and therefore if they were equal before the orthogonal state-space transformation, then they will also be equal after the transformation.

Now consider two i/o-equivalent systems $(A_1,b_1,c_1), (A_2,b_2,c_2)$, which are both block-balanced with the same indices $n(j), j = 1,\dots,k$, and with Gramians $W_o^{(i)} = W_c^{(i)} = \operatorname{diag}(\Sigma_1^{(i)},\dots,\Sigma_k^{(i)}), i = 1,2$, with the property that $\lambda_1(\Sigma_1^{(i)}) \geq \lambda_{n(1)}(\Sigma_1^{(i)}) > \lambda_1(\Sigma_2^{(i)}) \geq \lambda_{n(2)}(\Sigma_2^{(i)}) > \cdots > \lambda_1(\Sigma_k^{(i)}) \geq \lambda_{n(k)}(\Sigma_k^{(i)}) > 0, i = 1,2$.

Because $\Sigma_j^{(i)}$ is symmetric positive definite for any $i = 1,2, j = 1,\dots,k$, there exists an orthogonal matrix $Q_j^{(i)}$ such that $Q_j^{(i)}\Sigma_j^{(i)}(Q_j^{(i)})^T = \operatorname{diag}(\lambda_1(\Sigma_j^{(i)}), \lambda_2(\Sigma_j^{(i)}), \dots, \lambda_{n(j)}(\Sigma_j^{(i)}))$. Therefore, the state-space transformation $Q^{(i)} := \operatorname{diag}(Q_1^{(i)},\dots,Q_k^{(i)})$ applied to the system representation (A_i,b_i,c_i) brings it into *balanced form with non-increasing singular values*, $i = 1,2$. We can therefore apply Theorem 2.3 to the transformed system representations, and it follows that there exists an orthogonal state-space transformation of the form $Q = \operatorname{diag}(Q_1,\dots,Q_k)$ with $Q_i \in \mathbb{R}^{n(i)\times n(i)}, i = 1,2,\dots,k$, that transforms (A_1,b_1,c_1) into (A_2,b_2,c_2) (and vice versa). □

The following theorem will be fundamental for our results.

THEOREM 2.7 (Pernebo and Silverman [24], Kabamba [12]). *Let* $(A,b,c) \in \mathbf{R}^{n\times n} \times \mathbf{R}^{n\times 1} \times \mathbf{R}^{1\times n}$ *be conformally partitioned as*

$$A = \begin{pmatrix} A_{11} & A_{12} \\ A_{21} & A_{22} \end{pmatrix}, \quad b = \begin{pmatrix} b_1 \\ b_2 \end{pmatrix}, \quad c = \begin{pmatrix} c_1 & c_2 \end{pmatrix},$$

with $A_{ii} \in \mathbf{R}^{n(i)\times n(i)}, i = 1,2$, *and let* (A,b,c) *be block-balanced with indices* $n(1), n(2)$ *such that* $\Sigma_1, \Sigma_2 > 0$ *have no eigenvalues in common.*

Then $(A,b,c) \in C_n \Leftrightarrow (A_{ii}, b_i, c_i) \in C_{n(i)}, i = 1,2.$

3. The case *k*=1: A Schwarz-like canonical form for stable SISO systems in continuous time.

THEOREM 3.1. *Consider the set* B_n *of all* $(A,b,c) \in C_n$ *of the following form:*

$$A = \begin{pmatrix} a_{11} & -\alpha_1 & & 0 \\ \alpha_1 & 0 & \ddots & \\ & \ddots & \ddots & -\alpha_{n-1} \\ 0 & & \alpha_{n-1} & 0 \end{pmatrix}, \qquad a_{11} = -\frac{b_1^2}{2} < 0,$$

$$\alpha_i > 0, \qquad i = 1, \ldots, n-1,$$

$$b = \begin{pmatrix} b_1 \\ 0 \\ \vdots \\ 0 \end{pmatrix}, \qquad b_1 > 0,$$

$$c = \begin{pmatrix} c_1 & \gamma_1 & \ldots & \gamma_{n-1} \end{pmatrix}, \qquad c_1 \in \mathbf{R},\ \gamma_j \in \mathbf{R},\ j = 1, \ldots, n-1.$$

Each triple $(A,b,c) \in B_n$ *is input-normal.*

Let S_n *be the set of values of the vector of parameters* $(b_1, \alpha_1, \ldots, \alpha_{n-1}, c_1, \gamma_1, \ldots, \gamma_{n-1})$ *such that the corresponding triple* $(A,b,c) \in B_n$, *i.e., such that* $b_1 > 0, \alpha_i > 0, i = 1, \ldots, n$, *and* $c_1, \gamma_1, \ldots, \gamma_{n-1}$ *such that the pair* (c,A) *is observable.*

The set B_n *describes a real analytic (hence continuous) canonical form, and the parametrization mapping* $S_n \longrightarrow B_n$, *which maps each parameter vector to the corresponding triple* (A,b,c), *is a real analytic diffeomorphism (hence a homeomorphism).*

If $(\gamma_1, \ldots, \gamma_{n-1}) \neq 0 \in \mathbf{R}^{n-1}, n \geq 2$, *then the system has several different singular values.*

Proof. The requirement that a realization is input-normal reduces the freedom of choosing a basis of the state space to the freedom of choosing an *orthonormal* basis, i.e., to the freedom of choosing an element from the orthogonal group.

Now consider the controllability matrix of a triple $(A,b,c) \in B_n$. It is easily seen to be positive upper triangular. According to [19] there is a unique element in the orthogonal group that transforms a controllability matrix to a positive upper triangular matrix. Therefore the form presented here is canonical indeed.

Next let us show the smoothness properties. The mapping $S_n \longrightarrow B_n$, which maps a parameter vector from S_n to its corresponding triple (A,b,c), is polynomial, hence real analytic.

Now consider the mapping $C_n \longrightarrow S_n$, which maps *any* triple $(\tilde{A}, \tilde{b}, \tilde{c}) \in C_n$ to the corresponding parameter vector describing the *canonical form* of the system. Clearly the coefficients of the characteristic polynomial of $\tilde{A}$ depend polynomially on $\tilde{A}$, and therefore the parameters $a_{11}, \alpha_1, \ldots, \alpha_{n-1}$ depend real analytically on $\tilde{A}$, as they are rational functions of these characteristic polynomial coefficients (cf. [18]).

It remains to show that the parameter vector $c = (c_1, \gamma_1, \ldots, \gamma_{n-1})$ depends real analytically on the entries of $(\tilde{A}, \tilde{b}, \tilde{c})$. Let (A, b, c) denote the canonical form of the system and $g(z) := \frac{p(z)}{q(z)} := c(zI - A)^{-1}b = \tilde{c}(zI - \tilde{A})^{-1}\tilde{b}$ denote the (rational) transfer function of the system, with monic polynomial denominator $q(z) := \det(zI - A) = \det(zI - \tilde{A})$ and polynomial numerator $p(z)$. It is easy to see that the coefficients of $p(z)$ depend real analytically on the entries of $(\tilde{A}, \tilde{b}, \tilde{c})$. Let $M(z)$ denote the polynomial matrix of cofactors of $(zI - A)$. Then one has

$$p(z) = cM(z)^T b. \tag{6}$$

Consider $m_{1i}(z)$, which is $(-1)^{1+i}$ times the determinant of the matrix that is obtained from $zI - A$ by leaving out the first row and ith column, $i \in \{1, \ldots, n\}$:

$$m_{1i}(z) = (-1)^{1+i} \times \left|\begin{array}{cccc|cccccc} \alpha_1 & * & \ldots & * & * & \ldots & \ldots & \ldots & \ldots & * \\ 0 & \ddots & & \vdots & \vdots & & & & & \vdots \\ \vdots & & \ddots & * & \vdots & & & & & \vdots \\ 0 & \ldots & 0 & \alpha_{i-1} & * & \ldots & \ldots & \ldots & \ldots & * \\ \hline 0 & \ldots & \ldots & 0 & z & -\alpha_{i+1} & 0 & \ldots & \ldots & 0 \\ \vdots & & & \vdots & \alpha_{i+1} & z & -\alpha_{i+2} & 0 & \ldots & 0 \\ \vdots & & & \vdots & 0 & \ddots & \ddots & \ddots & \ddots & \vdots \\ \vdots & & & \vdots & \vdots & \ddots & \ddots & \ddots & \ddots & 0 \\ \vdots & & & \vdots & \vdots & & \ddots & \ddots & \ddots & -\alpha_{n-1} \\ 0 & \ldots & \ldots & 0 & 0 & 0 & \ldots & 0 & \alpha_{n-1} & z \end{array}\right|$$

$$= (-1)^{1+i} \times \left(\prod_{j=1}^{i-1} \alpha_j\right) z^{n-i} + \text{terms of lower degree in } z,$$

where $i \in \{1, \ldots, n\}$; if $i = 1$, the product $\prod_{j=1}^{i-1} \alpha_j$ is taken to be equal to one by convention. Because $\prod_{j=1}^{i-1} \alpha_j$ is unequal to zero (and in fact positive) for each $i \in \{1, \ldots, n\}$ the polynomials $m_{11}(z), \ldots, m_{1n}(z)$ form a basis of the linear vector space of polynomials of degree $< n$ over $\mathbf{R}$. Therefore (6), which can be rewritten as

$$c_1 m_{11}(z) + \gamma_1 m_{12}(z) + \cdots + \gamma_{n-1} m_{1n}(z) = \frac{p(z)}{b_1}, \tag{7}$$

has a unique solution $c = (c_1, \gamma_1, \gamma_2, \ldots, \gamma_{n-1})$, which depends real analytically on the entries of $(\tilde{A}, \tilde{b}, \tilde{c})$ and the parameters $b_1, \alpha_1, \ldots, \alpha_{n-1}$. Since these parameters themselves depend real analytically on the entries of $(\tilde{A}, \tilde{b}, \tilde{c})$, the real analyticity of all parameters on the entries of $(\tilde{A}, \tilde{b}, \tilde{c})$ follows. This completes the proof of the smoothness properties.

The remaining statements follow from the fact that for $\gamma = 0$, the form is a canonical form for systems with only *one* positive Hankel singular value (i.e., all nonzero Hankel singular values coincide); cf. [19], [18]. □

Remarks. (i) The fact that *if* the asymptotically stable matrix A can be brought into the presented form by a basis change of the state space, then the resulting matrix

is unique, also follows from the fact mentioned in the proof that for $\gamma = 0, c_1 \neq 0$ the form is a canonical form for systems with only one positive Hankel singular value; cf. [19], [18]. Note that here we use a different sign convention for the off-diagonal elements of the matrix A than in those papers. This corresponds to consideration of the dual state-space representation.

(ii) If $c_1 \neq 0$, we define $\sigma := |\frac{c_1}{b_1}| > 0$, which we will call a pseudosingular value. If the vector $\gamma = (\gamma_1, \ldots, \gamma_{n-1})$ is close enough to zero, the pseudosingular value will be close to the true singular values of the system, because of continuity of the singular values as a function of γ and the fact that if $\gamma = 0$, the system has only one singular value and its value is σ. If $c_1 \neq 0$, the system can be brought simply into σ-input-normal form by multiplying c by $\sigma^{-\frac{1}{2}}$ and b by $\sigma^{\frac{1}{2}}$. The resulting σ-input-normal form is a *canonical* form locally around $\gamma = 0$, but not globally because the systems which have $c_1 = 0$ in the previous canonical form cannot be represented in this way. (It would lead to $\sigma = 0$, and therefore one cannot transform back to the input-normal case, etc.) Locally around $\gamma = 0$ it takes the following form:

$$A = \begin{pmatrix} a_{11} & -\alpha_1 & & 0 \\ \alpha_1 & 0 & \ddots & \\ & \ddots & \ddots & -\alpha_{n-1} \\ 0 & & \alpha_{n-1} & 0 \end{pmatrix},$$

$$a_{11} = -\frac{b_1^2}{2\sigma} < 0,$$

$$\alpha_i > 0, \qquad i = 1, \ldots, n-1,$$

$$b = \begin{pmatrix} b_1 \\ 0 \\ \vdots \\ 0 \end{pmatrix}, \qquad b_1 > 0$$

$$c = \begin{pmatrix} sb_1 & \gamma_1 & \ldots & \gamma_{n-1} \end{pmatrix}, \qquad s \in \{-1, 1\},\ \gamma_j \in \mathbf{R},\ j = 1, \ldots, n-1.$$

(iii) Because the canonical form is input-normal, if one starts with an arbitrary input-normal realization $(\tilde{A}, \tilde{b}, \tilde{c})$ of the system, it takes an orthogonal state-space transformation Q in order to obtain the canonical form of the system involved. The same holds for the (local) σ-input-normal canonical form.

(iv) Clearly the canonical forms presented are controllable (because they are input-normal; resp., σ-input-normal), but observability will fail for certain choices of c; the observability Gramian will be singular for such a choice of c. If $\gamma = 0, c_1 \neq 0$, the system is observable, because the observability Gramian will be $\sigma^2 I$ (resp., σI). (In that case the system representation is σ^2-output-normal; resp., σ-output-normal.) Therefore, also in some open neighborhood around such a system, observability will still hold. (This follows from the continuity of the determinant of the observability Gramian as a function of the parameters.)

(v) This canonical form is closely related to the so-called Schwarz canonical form; cf. [13], [14], [25].

(vi) A canonical form can be interpreted as a choice of basis of the state space for each system. In this case the basis can be obtained as follows. Define an inner product on the state space by the inverse of the reachability Gramian. Take the first n columns of the reachability matrix, and apply the Gram–Schmidt orthogonalization procedure to it, with respect to the inner product. With respect to the resulting

set of n vectors as the basis of the state space the system has the canonical form. This observation can in fact be used to obtain an alternative proof of the smoothness properties stated in the theorem.

4. An input-normal and a block-balanced canonical form. Let $n(1), \dots, n(k) \in \{1, 2, \dots, n\}, \sum_{j=1}^{k} n(j) = n$, denote a partition of n as before. Let $C_{n(1),n(2),\dots,n(k)}$ denote the subset of all systems in C_n, with the property that their n Hankel singular values (multiplicities *included*) $\sigma(1) \geq \sigma(2) \geq \cdots \geq \sigma(n) > 0$ can be partitioned into k disjoint sets of singular values (again with multiplicities included) in the following way:

$$\begin{aligned}\sigma(1) &\geq \cdots \geq \sigma(n(1)) > \sigma(n(1)+1)\\ &\geq \cdots \geq \sigma(n(1)+n(2)) > \sigma(n(1)+n(2)+1)\\ &\geq \cdots \geq \sigma\left(\sum_{j=1}^{l} n(j)\right) > \sigma\left(\left(\sum_{j=1}^{l} n(j)\right)+1\right)\\ &\geq \cdots > 0. \qquad (8)\end{aligned}$$

So we require that $\sigma(\sum_{j=1}^{l} n(j)) > \sigma((\sum_{j=1}^{l} n(j)) + 1)$ for $l = 1, 2, \dots, k-1$ and $\sigma(n) > 0$, of course. Note that the notation is consistent with the fact that C_n denotes the set of stable systems which have as their only "restriction" that there are n positive singular values (multiplicities included), i.e., that the order of the system is n.

The other extreme is $C_{1,1,\dots,1}$, which denotes the set of nth-order stable systems with n *distinct* singular values. For this set of systems a balanced canonical form was derived in [12].

Remark. The set $C_{n(1),\dots,n(k)}$ should not be confused with the subset of C_n consisting of the systems which have k distinct singular values $\sigma_1 > \cdots > \sigma_k > 0$ with multiplicities $n(1), \dots, n(k)$. Of course these systems are included in $C_{n(1),\dots,n(k)}$, but they generally form only a (thin) subset.

Next we will present a canonical form on $C_{n(1),\dots,n(k)}$.

THEOREM 4.1. *Consider the set $B_{n(1),\dots,n(k)}$ of triples (A, b, c) of the following form:*

$$\begin{aligned}A &= (A(i,j))_{1\leq i,j\leq k},\\ A(i,j) &\in \mathbf{R}^{n(i)\times n(j)}, \qquad i,j \in \{1,\dots,k\},\\ b &= \begin{pmatrix} b(1)\\ b(2)\\ \vdots\\ b(k)\end{pmatrix}, \qquad b(i) \in \mathbf{R}^{n(i)},\ i = 1,\dots,k,\\ c &= (c(1),\dots,c(k)), \qquad c(j)^T \in \mathbf{R}^{n(j)},\ j = 1,\dots,k,\\ A(i,i) &= \begin{pmatrix} a(i,i)_{11} & -\alpha(i)_1 & 0 & \dots & 0\\ \alpha(i)_1 & 0 & -\alpha(i)_2 & \ddots & \vdots\\ 0 & \alpha(i)_2 & & \ddots & 0\\ \vdots & \ddots & \ddots & & -\alpha(i)_{n(i)-1}\\ 0 & \dots & 0 & \alpha(i)_{n(i)-1} & 0\end{pmatrix},\end{aligned}$$

$$
\begin{aligned}
a(i,i)_{11} &= -\frac{b_i^2}{2}, \\
\alpha(i)_j &> 0, \qquad j = 1, \ldots, n(i)-1, \\
b(i) &= \begin{pmatrix} b_i \\ 0 \\ \vdots \\ 0 \end{pmatrix}, \qquad b_i > 0, \\
c(i) &= \left(c_i, \gamma(i)_1, \ldots, \gamma(i)_{n(i)-1}\right), \qquad i = 1, \ldots, k,
\end{aligned}
$$

where the parameters are to be taken such that the corresponding observability Gramians $\Sigma_i^2, i = 1, \ldots, k$, *which satisfy the observability Lyapunov equations*

$$
\Sigma_i^2 A(i,i) + A(i,i)^T \Sigma_i^2 = -c(i)^T c(i) \tag{9}
$$

are fulfilling the following inequalities:

$$
\lambda_1(\Sigma_1^2) \geq \lambda_{n(1)}(\Sigma_1^2) > \lambda_1(\Sigma_2^2) \geq \lambda_{n(2)}(\Sigma_2^2) > \cdots > \lambda_1(\Sigma_k^2) \geq \lambda_{n(k)}(\Sigma_k^2) > 0. \tag{10}
$$

For each pair $(i,j), i \neq j$, *the matrices* $A(i,j), A(j,i)$ *are determined (uniquely!) from the following pair of linear matrix equations:*

$$
\begin{aligned}
A(i,j) + A(j,i)^T &= -b(i)b(j)^T, \\
\Sigma_i^2 A(i,j) + A(j,i)^T \Sigma_j^2 &= -c(i)^T c(j).
\end{aligned} \tag{11}
$$

The set $B_{n(1),\ldots,n(k)}$ *describes a real analytic (hence continuous) canonical form on* $C_{n(1),\ldots,n(k)}$. *The* $2n$ *"free" parameters of the canonical form are*

$$
b_i, \alpha(i)_1, \ldots, \alpha(i)_{n(i)-1}, c_i, \gamma(i)_1, \ldots, \gamma(i)_{n(i)-1}, \qquad i = 1, \ldots, k.
$$

Let $S_{n(1),\ldots,n(k)} \subset \mathbf{R}^{2n}$ *be the set of all values of the parameter vector for which the corresponding triple* $(A,b,c) \in B_{n(1),\ldots,n(k)}$, *i.e., for all* $i \in \{1, \ldots, k\}$ $b_i > 0$, $\alpha(i)_j > 0$, $j = 1, \ldots, n(i)-1$, *and* $c_i, \gamma(i)_1, \ldots, \gamma(i)_{n(i)-1}$ *such that the matrices* $\Sigma_i, i = 1, \ldots, k$, *found in* (9) *satisfy the inequalities* (10). *The mapping* $S_{n(1),\ldots,n(k)} \longrightarrow B_{n(1),\ldots,n(k)}$ *which maps a parameter vector to the corresponding triple* (A,b,c) *is a real analytic diffeomorphism.*

The form is input-normal, i.e.,

$$
A + A^T = -bb^T, \tag{12}
$$

and has block-diagonal observability Gramian $\Sigma^2 := \operatorname{diag}(\Sigma_1^2, \ldots, \Sigma_k^2) > 0$.

Let $\sigma(1) \geq \sigma(2) \geq \cdots \geq \sigma(n) > 0$ *denote the* n *positive Hankel singular values of the system (with their multiplicities). If for some* $i \in \{1, \ldots, k\}$ *the vector* $\gamma(i) = 0$, *then* Σ_i^2 *is a scalar matrix*

$$
\Sigma_i^2 = \sigma^2 \left(1 + \sum_{j=1}^{i-1} n(j)\right) \times I_{n(i)}, \tag{13}
$$

and

$$\sigma\left(\sum_{j=1}^{i-1} n(j)\right) > \sigma\left(1+\sum_{j=1}^{i-1} n(j)\right) = \sigma\left(2+\sum_{j=1}^{i-1} n(j)\right) = \cdots = \sigma\left(\sum_{j=1}^{i} n(j)\right) > \sigma\left(1+\sum_{j=1}^{i} n(j)\right).$$

The observability Gramian is diagonal if and only if for all $i \in \{1,\ldots,k\}, \gamma(i) = 0$.

Remark. A block-balanced realization can be obtained from the presented canonical form by applying a state-space transformation

$$T := \Sigma^{\frac{1}{2}} = \operatorname{diag}\left(\Sigma_1^{\frac{1}{2}},\ldots,\Sigma_k^{\frac{1}{2}}\right) > 0. \tag{14}$$

The corresponding controllability and observability Gramians will both be equal to

$$\Sigma = \operatorname{diag}(\Sigma_1,\ldots,\Sigma_k) > 0.$$

Proof. (i) To start we will show that the form presented is canonical on $C_{n(1),\ldots,n(k)}$. Consider a system which can be represented by a triple in $C_{n(1),\ldots,n(k)}$. A balanced realization of the system is also in block-balanced form with partitioning indices $n(1),\ldots,n(k)$. So one can find a block-balanced realization (A,b,c) of the system with these partitioning indices. It follows from Theorem 2.6 that the requirement that (A,b,c) is block-balanced with these partitioning indices uniquely determines (A,b,c) up to an orthogonal state-space transformation of the form $Q = \operatorname{diag}(Q_1,Q_2,\ldots,Q_k)$, with orthogonal matrices $Q_i \in \mathbf{R}^{n(i)\times n(i)}$. If (A,b,c) is in block-balanced form, it can be brought into input-normal form with block-diagonal observability Gramian by the state-space transformation T^{-1}, where T is as defined in (14). It follows easily that if (A,b,c) is in input-normal form with block-diagonal controllability Gramian $\Sigma^2 = \operatorname{diag}\left(\Sigma_1^2,\ldots,\Sigma_k^2\right)$, with $\lambda_1(\Sigma_1^2) \geq \lambda_{n(1)}(\Sigma_1^2) > \lambda_1(\Sigma_2^2) \geq \lambda_{n(2)}(\Sigma_2^2) > \cdots > \lambda_1(\Sigma_k^2) \geq \lambda_{n(k)}(\Sigma_k^2) > 0$, $\Sigma_i^2 \in \mathbf{R}^{n(i)\times n(i)}$, then (A,b,c) is uniquely determined up to an orthogonal state-space transformation of the form $Q = \operatorname{diag}(Q_1,Q_2,\ldots,Q_k)$. If such a transformation is applied, then $(A(i,i),b(i),c(i))$ is transformed to $\left(Q_iA(i,i)Q_i^T, Q_ib(i), c(i)Q_i^T\right)$. Note that $(A(i,i),b(i),c(i)) \in C_{n(i)}$ because of Theorem 2.7, and therefore it follows from Theorem 3.1 that there is a unique choice for Q_i which brings $\left(Q_iA(i,i)Q_i^T, Q_ib(i), c(i)Q_i^T\right)$ into the required canonical form.

We need only to check that by using the solutions $A(i,j), A(j,i)$ of (11) the Gramians indeed have the required block structure, which is straightforward and left to the reader.

(ii) Second, we will show the smoothness properties. Clearly the mapping $S_{n(1),\ldots,n(k)} \longrightarrow B_{n(1),\ldots,n(k)}$, which maps any parameter vector in $S_{n(1),\ldots,n(k)}$ to the corresponding triple $(A,b,c) \in B_{n(1),\ldots,n(k)}$, is real analytic.

Now consider the mapping $C_{n(1),\ldots,n(k)} \longrightarrow S_{n(1),\ldots,n(k)}$, which maps a triple $(\tilde{A},\tilde{b},\tilde{c})$ to the parameter vector of the corresponding canonical form.

The map which assigns to $(\tilde{A},\tilde{b},\tilde{c})$ the coefficients of the characteristic polynomial of the product of the Gramians is real analytic. The zeroes of this polynomial are the

squared singular values. Now consider the polynomial

$$a(z) = \prod_{j=n(1)+1}^{n} (z - \sigma(j)^2).$$

Because on $C_{n(1),\ldots,n(k)}$ the inequality $\sigma(n(1)) > \sigma(n(1)+1)$ holds, the coefficients of $a(z)$ depend real analytically on those of the characteristic polynomial of the product of the Gramians (see, e.g., [16]).

Let $\Sigma^2 = W_c^{\frac{1}{2}} W_o W_c^{\frac{1}{2}}$, where W_c and W_o are the controllability and observability Gramians, respectively, of $(\tilde{A}, \tilde{b}, \tilde{c})$; W_c and W_o depend real analytically on $(\tilde{A}, \tilde{b}, \tilde{c})$. The matrix $a(\Sigma^2)$ has as its range space an $n(1)$-dimensional linear subspace of $\mathbf{R}^n$ which clearly depends real analytically on $(\tilde{A}, \tilde{b}, \tilde{c})$. The corresponding orthogonal projection matrix Π_1, which maps an arbitrary vector $x \in \mathbf{R}^n$ to its orthogonal projection in the linear subspace spanned by the columns of $a(\Sigma^2)$ (i.e., the linear subspace which is obtained by taking the direct sum of the eigenspaces of the largest $n(1)$ eigenvalues $\sigma(1)^2, \ldots, \sigma(n(1))^2$ of Σ^2), depends real analytically on $a(\Sigma^2)$.

Now consider $(\Pi_1 W_c^{-\frac{1}{2}} \tilde{A} W_c^{\frac{1}{2}} \Pi_1, \Pi_1 W_c^{-\frac{1}{2}} \tilde{b}, \tilde{c} W_c^{\frac{1}{2}} \Pi_1)$ with corresponding controllability Gramian Π_1 and observability Gramian $\Pi_1 \Sigma^2 \Pi_1 = \Pi_1 \Sigma^2 = \Sigma^2 \Pi_1$. (Because of the way Π_1 is constructed, it commutes with Σ^2.) We can now apply the canonical form of Theorem 3.1 to find a basis for the range space of Π_1 (which corresponds to the state space there) depending real analytically on $(\tilde{A}, \tilde{b}, \tilde{c})$. The first basis vector is

$$\frac{\Pi_1 W_c^{-\frac{1}{2}} \tilde{b}}{\| \Pi_1 W_c^{-\frac{1}{2}} \tilde{b} \|};$$

the second one (Gram–Schmidt orthonormalization) is obtained by normalization of the vector

$$\begin{aligned} &\Pi_1 W_c^{-\frac{1}{2}} \tilde{A} W_c^{\frac{1}{2}} \Pi_1 W_c^{-\frac{1}{2}} \tilde{b} \\ &\quad - \frac{\left(\tilde{b}^T W_c^{-\frac{1}{2}} \Pi_1 W_c^{\frac{1}{2}} \tilde{A}^T W_c^{-\frac{1}{2}} \Pi_1 W_c^{-\frac{1}{2}} \tilde{b} \right)}{\left(\tilde{b}^T W_c^{-\frac{1}{2}} \Pi_1 W_c^{-\frac{1}{2}} \tilde{b} \right)} \times \Pi_1 W_c^{-\frac{1}{2}} \tilde{b}; \end{aligned}$$

and so on. Clearly this choice of basis of the range space of Π_1 is real analytic. With respect to the resulting basis of the $n(1)$-dimensional state space the triple

$$\left(\Pi_1 W_c^{-\frac{1}{2}} \tilde{A} W_c^{\frac{1}{2}} \Pi_1, \Pi_1 W_c^{-\frac{1}{2}} \tilde{b}, \tilde{c} W_c^{\frac{1}{2}} \Pi_1 \right)$$

takes the form $(\tilde{A}(1,1), \tilde{b}(1), \tilde{c}(1))$, as described in Theorem 3.1:

$$\tilde{A}(1,1) = \begin{pmatrix} a(1,1)_{11} & -\alpha(1)_1 & 0 & \ldots & 0 \\ \alpha(1)_1 & 0 & -\alpha(1)_2 & \ddots & \vdots \\ 0 & \alpha(1)_2 & & \ddots & 0 \\ \vdots & \ddots & \ddots & & -\alpha(1)_{n(1)-1} \\ 0 & \ldots & 0 & \alpha(1)_{n(1)-1} & 0 \end{pmatrix},$$

$$a(1,1)_{11} = -\frac{b_1^2}{2},$$
$$\alpha(1)_j > 0, \qquad j = 1, \ldots, n(1)-1,$$
$$\tilde{b}(1) = \begin{pmatrix} b_1 \\ 0 \\ \vdots \\ 0 \end{pmatrix}, \qquad b_1 > 0,$$
$$\tilde{c}(1) = \left(c_1, \gamma(1)_1, \ldots, \gamma(1)_{n(1)-1}\right),$$

and therefore this triple and the parameters describing it depend real analytically on $(\tilde{A}, \tilde{b}, \tilde{c})$. Similarly for any $i \in \{1, \ldots, k\}$ the matrix triple and the parameters describing it depend real analytically on $(\tilde{A}, \tilde{b}, \tilde{c})$. This proves the real analyticity of the mapping which maps $(\tilde{A}, \tilde{b}, \tilde{c})$ to the parameters of the canonical form.

(iii) The remaining statements follow from the results in [19]. □

5. An atlas of overlapping block-balanced canonical forms.

THEOREM 5.1. *Let the state-space dimension n be fixed. The canonical forms $C_{n(1),\ldots,n(k)} \longrightarrow B_{n(1),\ldots,n(k)}$, $n(j) \in \{1, \ldots, n\}, j = 1, \ldots, k$, $\sum_{j=1}^{k} n(j) = n$, $k \in \{1, \ldots, n\}$, form an overlapping set of real analytic (hence continuous) canonical forms covering C_n. Each of the sets $C_{n(1),\ldots,n(k)}$, $\sum_{j=1}^{k} n(j) = n$, is an open subset of C_n, and together they cover C_n.*

Proof. Let $P(n;k) := \{(n(1), \ldots, n(k)) | n(j) \in \{1, \ldots, n\}; j = 1, \ldots, k; \sum_{j=1}^{k} n(j) = n\}$, the set of partitions of n into k parts. It is trivial to show that

$$\bigcup_{k=1}^{n} \bigcup_{(n(1),\ldots,n(k)) \in P(n;k)} C_{n(1),\ldots,n(k)} = C_n, \tag{15}$$

because $C_{n(1),\ldots,n(k)} \subset C_n$ for each partition $(n(1), \ldots, n(k))$ of n and for $k = 1$ one has $n(1) = n$ and $C_{n(1)} = C_n$. Clearly for each partition $(n(1), \ldots, n(k))$ of n the set $C_{n(1),\ldots,n(k)}$ is an open subset of C_n. The remaining properties follow from Theorem 4.1. □

COROLLARY 5.2. *The set of mappings*

$$\phi : C_{n(1),\ldots,n(k)}/\sim \longrightarrow S_{n(1),\ldots,n(k)} \subset \mathbf{R}^{2n},$$
$$(n(1), \ldots, n(k)) \in P(n;k), \qquad k = 1, \ldots, n,$$

which map each equivalence class of triples to the corresponding parameter vector in the canonical form, forms an atlas for the real analytic manifold of stable SISO input-output systems of order n.

Proof. Any input-output system has a minimal state-space realization which is unique up to choice of basis of the state space. Therefore, the equivalence classes of (minimal!) triples in C_n can be identified with stable SISO input-output systems, and the result follows from the theorem. □

Remark. A motivation for using this atlas rather than, for example, just the Schwarz-like canonical form B_n is the following. Suppose one wants to use *balanced realizations.* Then one can use the balanced parametrization of [19]. However, this parametrization is discontinuous at all points of $C_n \setminus C_{1,\ldots,1}$, i.e., in all triples $(\tilde{A}, \tilde{b}, \tilde{c})$ which have two or more coinciding singular values. Also, the complement $C_{1,\ldots,1}$, of the set of discontinuity points consists of 2^n topological components, one component

for each sign pattern of the vector c (which cannot have zero components in this case; cf. (9), (10) with $n(i) = 1, i = 1, 2, \ldots, k$); this should be compared to C_n, which has only $n + 1$ topological components (the Brockett components). It appears that this is a serious disadvantage if one wants to use balanced realizations and balanced parametrizations in, for example, search algorithms for system identification, because one has to find out first which is the right "cell" of the parametrization. Another difficulty is that the balanced parametrization will tend to become numerically ill behaved if two or more of the Hankel singular values of the system are close to each other. For example, for the class of second-order systems, the determinant of the L_2-induced Riemannian metric tensor of the balanced parametrization can be calculated (e.g., using a computer algebra package) to be

$$b_1^2 b_2^2 \left(\frac{s_1\sigma_1 - s_2\sigma_2}{s_1\sigma_1 + s_2\sigma_2} \right)^2$$

in the notation of [19]. Here the s_1 and s_2 are the sign parameters, which are either $+1$ or -1. It follows that if two Hankel singular values come close, for given values of b_1 and b_2, then the parametrization becomes *ill conditioned* in the sense that a small parameter change may lead to a large change in the system (in the L_2-sense) and/or a large parameter change may lead to only a small change in the system (again in the L_2-sense).

In order to overcome these difficulties one could use the overlapping block-balanced canonical forms as follows. If $(\tilde{A}, \tilde{b}, \tilde{c})$ has k distinct Hankel singular values $\sigma_1 > \sigma_2 > \cdots > \sigma_k > 0$ with respective multiplicities $n(1), \ldots, n(k)$, then one can use the block-balanced continuous canonical form on $C_{n(1),\ldots,n(k)}$ *locally around* $(\tilde{A}, \tilde{b}, \tilde{c})$. If one is moving away from $(\tilde{A}, \tilde{b}, \tilde{c})$ in a search algorithm, for example, one has to decide whether the canonical form corresponding to a different partition should be used: if the largest $n(1)$ singular values differ sufficiently from each other, one could use, e.g., $C_{1,\ldots,1,n(2),\ldots,n(k)}$ (where there are $n(1)$ ones in the subindex before $n(2)$), etc. In this way one would use balanced realizations and "almost-balanced" realizations while moving around in the set of nth-order systems, without encountering discontinuity points.

6. On the imbedded submanifolds structure of the balanced canonical form. Consider the balanced canonical form for C_n of [19]. For each $k \in \{1, \ldots, n\}$ and each partition $(n_1, n_2, \ldots, n_k) \in P(n; k)$ let $K_{n_1,\ldots,n_k}$ denote the subset of C_n of systems with k distinct singular values $\sigma_1 > \sigma_2 > \cdots > \sigma_k$, which have multiplicities $n_1, n_2, \ldots, n_k$, respectively. Clearly $K_{n_1,\ldots,n_k} \subset C_{n_1,\ldots,n_k}$ and equality holds only if $k = n,\ n_i = 1,\ i = 1, \ldots, n$. The mapping $K_{n_1,\ldots,n_k} \longrightarrow B_{n_1,\ldots,n_k} \cap K_{n_1,\ldots,n_k}$ is a canonical form on $K_{n_1,\ldots,n_k}$, the restriction of the canonical form $C_{n_1,\ldots,n_k} \longrightarrow B_{n_1,\ldots,n_k}$ to $K_{n_1,\ldots,n_k}$. This canonical form on $K_{n_1,\ldots,n_k}$ is input-normal with diagonal observability Gramian W_o. If one applies the state-space transformation (14) (which is diagonal here), then one obtains the balanced canonical form of [19] restricted to $K_{n_1,\ldots,n_k}$. Clearly on $K_{n_1,\ldots,n_k}$ the balanced canonical form is smooth (real analytic), while it is of course not even continuous on C_n. Both the balanced canonical form and the corresponding input-normal form parametrize $K_{n_1,\ldots,n_k}/\sim$ by the parameters $b_i > 0, \alpha(i)_j > 0, j = 1, \ldots, n_i - 1, c_i \neq 0, i = 1, \ldots, k$. Because $(c_1, \ldots, c_k)$ has 2^k possible sign patterns, it follows that $K_{n_1,\ldots,n_k}/\sim$ has 2^k topological components, each real analytically diffeomorphic to $\mathbf{R}^{n+k}$. It follows clearly that $K_{n_1,\ldots,n_k}/\sim$ is a real analytic manifold. The question arises whether it is a *regular* submanifold of $C_n/\sim$ in the sense of [1] and therefore an imbedded submanifold (cf. [1], esp. Lemma 5.2).

The answer is affirmative and is a direct consequence of the construction developed in the previous sections.

THEOREM 6.1. *For each $k \in \{1, \dots, n\}$ and each partition $(n_1, \dots, n_k) \in P(n; k)$ the subset $K_{n_1,\dots,n_k}/\sim$ is a regular submanifold of $C_n/\sim$ and therefore an imbedded submanifold with the inclusion as the imbedding map.*

Proof. It follows from [1, Chapter III, section 5] that it suffices to show the socalled *$n+k$-submanifold property* for $K_{n_1,\dots,n_k}/\sim$. This property is said to hold if for each point $p \in K_{n_1,\dots,n_k}/\sim$ there exists a coordinate neighborhood U, φ on $C_n/\sim$ with local coordinates $\xi_1, \dots, \xi_{2n}$ such that (i) $\varphi(p) = (0, \dots, 0)$, (ii) $\varphi(U) = \{(\xi_1, \dots, \xi_{2n}) | -\epsilon < \xi_i < \epsilon, i = 1, \dots, 2n\}$, and (iii) $\varphi(U \cap K_{n_1,\dots,n_k}/\sim) = \{\xi \in \varphi(U) | \xi_{n+k+1} = \cdots = \xi_{2n} = 0\}$. The $n+k$-submanifold property can be shown to hold as follows. Suppose that the parameter values of point $p \in K_{n_1,\dots,n_k}/\sim$ are $b_i^0, \alpha(i)_j^0 > 0, j = 1, \dots, n_i - 1, c_i^0 \neq 0$; of course at p, $\gamma(i)_1 = \cdots = \gamma(i)_{n_i-1} = 0$. Now choose the local coordinates $\xi_1, \dots, \xi_{2n}$, as follows: $(\xi_1, \dots, \xi_{n+k}) = (b_1 - b_1^0, \alpha(1)_1 - \alpha(1)_1^0, \dots, \alpha(1)_{n_1-1} - \alpha(1)_{n_1-1}^0, c_1 - c_1^0; b_2 - b_2^0, \alpha(2)_1 - \alpha(2)_1^0, \dots, \alpha(2)_{n_2-1} - \alpha(2)_{n_2-1}^0, c_2 - c_2^0; \dots; b_k - b_k^0, \alpha(k)_1 - \alpha(k)_1^0, \dots, \alpha(k)_{n_k-1} - \alpha(k)_{n_k-1}^0, c_k - c_k^0)$, $(\xi_{n+k+1}, \dots, \xi_{2n}) = (\gamma(1)_1, \dots, \gamma(1)_{n_1-1}, \dots, \gamma(k)_1, \dots, \gamma(k)_{n_k-1})$. Clearly (i) holds. It follows from Theorem 4.1 that there exists a neighborhood U of p such that (ii) holds, and from Theorem 4.1, (iii) follows. ∎

Acknowledgment. Discussions with Dr. J. M. Maciejowski are gratefully acknowledged.

Note added in proof. In a forthcoming article by the present authors in *Linear Algebra and its Applications*, the results presented here are extended to various classes of SISO and multivariable systems.

REFERENCES

[1] W.M. BOOTHBY, *An Introduction to Differentiable Manifolds and Riemannian Geometry*, 2nd ed., Academic Press, New York, 1986.

[2] J.M.C. CLARK, *The consistent selection of local coordinates in linear system identification*, in Proceedings of the Joint Automatic Control Conference, Purdue, 1976, pp. 576–580.

[3] P.A. FUHRMANN AND R.J. OBER, *A functional approach to LQG balancing*, Internat. J. Control, 57 (1993), pp. 627–741.

[4] K. GLOVER AND J.C. WILLEMS, *Parametrizations of linear dynamical systems: Canonical forms and identifiability*, IEEE Trans. Automat. Control, AC-19 (1974), pp. 640–645. (Special issue on system identification and time series analysis.)

[5] R.P. GUIDORZI, *Invariants and canonical forms for systems, structural and parametric identification*, Automatica, 17 (1981), pp. 117–133.

[6] B. HANZON, *On a Gauss-Newton identification method that uses overlapping parametrizations*, in IFAC Identification and System Parameter Estimation, York, U.K., 1985, H.A. Barker and P.C. Young, eds., Pergamon Press, Oxford, pp. 1671–1676.

[7] B. HANZON, *On a coordinate free prediction error algorithm for system identification*, in Modelling, Identification and Control, C.I. Byrnes and A. Lindquist, eds., North-Holland, Amsterdam, 1986, pp. 595–604.

[8] B. HANZON, *Identifiability, Recursive Identification and Spaces of Linear Dynamical Systems*, CWI Tracts 63, 64, CWI, Amsterdam, 1989.

[9] B. HANZON AND R.L.M. PEETERS, *On the Riemannian interpretation of the Gauss-Newton algorithm*, in Mutual Impact of Computing Power and Control Theory, M. Kárny and K. Warwick, eds., Plenum Press, New York, 1993, pp. 111–121.

[10] M. HAZEWINKEL AND R.E. KALMAN, *On invariants, canonical forms and moduli for linear, constant finite dimensional dynamical systems*, in Proc. Udine, Lecture Notes in Econom. and Math. Systems 131, G. Marchesini and S.K. Mitter, eds., Springer-Verlag, Berlin, 1976, pp. 48–60.

[11] M. HAZEWINKEL, *Moduli and canonical forms for linear dynamical systems* II*: The topological case*, Math. Systems Theory, 10 (1977), pp. 363–385.

[12] P.T. Kabamba, *Balanced forms: Canonicity and parametrization*, IEEE Trans. Automat. Control, AC-30 (1985), pp. 1106–1109.
[13] R.E. Kalman and J.E. Bertram, *Control system analysis and design via the "second method" of Lyapunov*, Journal of Basic Engineering, (June 1960), pp. 371–393.
[14] R.E. Kalman, *On partial realizations, transfer functions, and canonical forms*, Acta Polytech. Scand. Math. Comput. Sci. Ser., 31 (1979), pp. 9–32.
[15] J.M. Maciejowski, *Balanced realisations in system identification*, in H.A. Barker and P.C. Young, eds., IFAC Identification and System Parameter Estimation 1985, York, U.K., Pergamon Press, Oxford, pp. 1823–1827.
[16] M. Marden, *The Geometry of the Zeros of a Polynomial in a Complex Variable*, American Mathematical Society, New York, 1949.
[17] B. C. Moore, *Principal component analysis in linear systems: Controllability, observability and model reduction,* IEEE Trans. Automat. Control, AC-26 (1981), pp. 17–32.
[18] R. Ober, *Asymptotically stable allpass transfer functions: Canonical form, parametrization and realization*, in Proc. IFAC World Congress, Munich, 1987.
[19] R. Ober, *Balanced realizations: Canonical form, parametrization, model reduction,* Internat. J. Control, 46 (1987), pp. 643–670.
[20] R. Ober and D. McFarlane, *Balanced canonical forms for minimal systems: A normalized coprime factor approach,* Linear Algebra Appl., 122, 123, 124 (special issue on linear systems and control) (1989), pp. 23–64.
[21] A.J.M. van Overbeek and L. Ljung, *On-line structure selection for multivariable state space models*, Automatica 18, pp. 529–543.
[22] R.L.M. Peeters and B. Hanzon, *The Riemannian Interpretation of Gauss-Newton and Scoring, with Application to System Identification*, Research memorandum 1992-22, Dept. Econometrics, Free University Amsterdam, Amsterdam, 1992.
[23] R.L.M. Peeters, *System Identification Based on Riemannian Geometry: Theory and Algorithms*, Tinbergen Institute Research Series 64, Thesis Publishers, Amsterdam, 1994.
[24] L. Pernebo and L.M. Silverman, *Model reduction via balanced state space representations*, IEEE Trans. Automat. Control, AC-37 (1982), pp. 382–387.
[25] H.R. Schwarz, *A method for determining stability of matrix differential equations*, Z. Angew. Math. Phys., 7 (1956), pp. 473–500 (in German).

SIAM J. CONTROL OPTIM.
Vol. 35, No. 1, pp. 243–253, January 1997

013

STOCHASTIC VERIFICATION THEOREMS WITHIN THE FRAMEWORK OF VISCOSITY SOLUTIONS*

XUN YU ZHOU†, JIONGMIN YONG‡, AND XUNJING LI‡

Abstract. This paper studies controlled systems governed by Ito's stochastic differential equations in which control variables are allowed to enter both drift and diffusion terms. A new verification theorem is derived within the framework of viscosity solutions without involving any derivatives of the value functions. This theorem is shown to have wider applicability than the restrictive classical verification theorems, which require the associated dynamic programming equations to have smooth solutions. Based on the new verification result, optimal stochastic feedback controls are obtained by maximizing the generalized Hamiltonians over *both* the control regions and the superdifferentials of the value functions.

Key words. stochastic optimal control, verification theorem, Hamilton–Jacobi–Bellman equation, viscosity solution, superdifferential, feedback control

AMS subject classifications. 93E20, 49L20, 49L25

PII. S0363012995279973

1. Introduction.

We consider in this paper stochastic optimal control problems of the following kind. For a given $s \in [0,1]$, by the set of admissible controls $U_{ad}[s,1]$ we mean the collection of (i) standard probability spaces $(\Omega, \mathcal{F}, P)$ along with l-dimensional Brownian motions $B = \{B(t) : s \le t \le 1\}$ with $B(s) = 0$ and (ii) Γ-valued $\mathcal{F}_t^s$-adapted measurable processes $u(\cdot) = \{u(t) : s \le t \le 1\}$, where $\mathcal{F}_t^s = \sigma\{B(r) : s \le r \le t\}$ and Γ is a given closed set in some Euclidean space R^m. We denote $(\Omega, \mathcal{F}, P, B; u(\cdot)) \in U_{ad}[s,1]$, but occasionally we will write only $u(\cdot) \in U_{ad}[s,1]$ if no ambiguity arises. Let $(s,y) \in [0,1) \times R^d$ be given. For each $(\Omega, \mathcal{F}, P, B; u(\cdot)) \in U_{ad}[s,1]$, the corresponding cost is

$$J(s,y;u(\cdot)) = E\left[\int_s^1 L(t,x(t),u(t))dt + h(x(1))\right], \tag{1.1}$$

where $x(\cdot) = \{x(t) : s \le t \le 1\}$ is the solution of the following Ito stochastic differential equation (SDE) on the filtered space $(\Omega, \mathcal{F}, P; \mathcal{F}_t^s)$:

$$\begin{cases} dx(t) = f(t,x(t),u(t))dt + \sigma(t,x(t),u(t))dB(t), \\ x(s) = y. \end{cases} \tag{1.2}$$

The solution $x(\cdot)$ of the above SDE is called the response of the control $u(\cdot) \in U_{ad}[s,1]$, and $(x(\cdot), u(\cdot))$ is called an admissible pair. The objective of the optimal control problem is to minimize the cost function $J(s,y;u(\cdot))$, for a given $(s,y) \in [0,1) \times R^d$, over all $u(\cdot) \in U_{ad}[s,1]$. We denote the above problem by $C_{s,y}$ to recall the dependence

*Received by the editors January 18, 1995; accepted for publication (in revised form) November 21, 1995. This research was supported by RGC earmarked grant CUHK 249/94E, the National Nature Science Foundation of China, the Chinese State Education Commission Science Foundation, the Trans-Century Training Program Foundation for the Talents by the State Education Commission of China, the Laboratory of Systems and Control, and Chinese Academy of Science.
http://www.siam.org/journals/sicon/35-1/27997.html
†Department of Systems Engineering and Engineering Management, Chinese University of Hong Kong, Shatin, Hong Kong (xyzhou@se.cuhk.edu.hk).
‡Department of Mathematics, Fudan University, Shanghai, China (jyong@fudan.ihep.ac.cn, xjli@ms.fudan.sh.cn).

on the initial time s and the initial state y. The value function is defined as

$$V(s,y) = \inf_{u(\cdot)\in U_{ad}[s,1]} J(s,y;u(\cdot)). \tag{1.3}$$

An admissible pair $(x^*(\cdot), u^*(\cdot))$ is called *optimal* for $C_{s,y}$ if $u^*(\cdot)$ achieves the minimum of $J(s,y;u(\cdot))$ over $U_{ad}[s,1]$.

As a part of the dynamic programming approach, the so-called verification technique plays an important role in testing for optimality of a given admissible pair and (more importantly) in constructing optimal feedback controls. The classical verification theorem is as follows (see Fleming and Rishel [5, Theorem VI.4.1]):

THEOREM 1.1. *Let $W \in C^{1,2}([0,1]\times R^d)$ be a solution of the following Hamilton–Jacobi–Bellman (HJB) equation:*

$$\begin{cases} -v_t(t,x) + \sup_{u\in\Gamma} G(t,x,u,v_x(t,x),v_{xx}(t,x)) = 0,\ (t,x)\in[0,1)\times R^d, \\ v(1,x) = h(x) \end{cases} \tag{1.4}$$

where the function G is defined as

$$G(t,x,u,q,Q) = -\frac{1}{2}\mathrm{tr}(\sigma^T(t,x,u)Q\sigma(t,x,u)) - q\cdot f(t,x,u) - L(t,x,u) \tag{1.5}$$

for $(t,x,u,q,Q)\in[0,1]\times R^d\times\Gamma\times R^d\times R^{d\times d}$. Then

(a) *$W(s,y)\le J(s,y;u(\cdot))$ for any $(s,y)\in[0,1)\times R^d$ and any $u(\cdot)\in U_{ad}[s,1]$.*

(b) *Suppose that a given admissible pair $(x^*(\cdot),u^*(\cdot))$ for the problem $C_{s,y}$ satisfies*

$$-W_t(t,x^*(t)) + G(t,x^*(t),u^*(t),W_x(t,x^*(t)),W_{xx}(t,x^*(t))) = 0, P\text{-a.s., a.e.}t\in[s,1]; \tag{1.6}$$

then $(x^*(\cdot),u^*(\cdot))$ is an optimal pair for the problem $C_{s,y}$.

Remark 1.1. The function G is called the *generalized Hamiltonian* [17]. By the HJB equation, (1.6) is equivalent to a more familiar form:

$$\begin{aligned} & G(t,x^*(t),u^*(t),W_x(t,x^*(t)),W_{xx}(t,x^*(t))) \\ = \ & \max_{u\in\Gamma} G(t,x^*(t),u,W_x(t,x^*(t)),W_{xx}(t,x^*(t))). \end{aligned}$$

Then, an optimal feedback control $u^*(t,x)$ can be constructed by minimizing $G(t,x,u,W_x(t,x),W_{xx}(t,x))$ over $u\in\Gamma$. For details, see [5].

When practically applying Theorem 1.1, one usually takes the verification function W to be the value function V, as V satisfies the HJB equation if $V\in C^{1,2}([0,1]\times R^d)$. Unfortunately, it is well known that the HJB equation (1.4) does not necessarily admit smooth solutions in general. This makes the applicability of the classical verification theorems very restrictive and is a major deficiency in dynamic programming theory. In recent years, the viscosity solution theory of general nonlinear PDEs, which was launched by Crandall and Lions [4], has been significantly developed. In this theory, all the derivatives involved are replaced by the so-called superdifferentials and subdifferentials, and the solutions in the viscosity sense can be merely continuous functions. The existence and uniqueness of viscosity solutions of the HJB can be guaranteed under very mild and reasonable assumptions, which are satisfied in the great majority of cases arising in optimal control problems. For example, the value function turns out to be the unique viscosity solution of the HJB equation [14]. Since the verification theorems have been playing primary roles in constructing optimal feedback controls, and in many practical problems HJB equations do not have smooth solutions

at all, a natural question arises: do verification theorems still hold, with the solutions of the HJB equation in the classical sense replaced by the ones in the viscosity sense and the derivatives involved replaced by the superdifferentials and/or subdifferentials? For the deterministic case ($\sigma = 0$), the answer to the above question is "yes" [18]. Moreover, based on the new, nonsmooth versions of the verification theorems obtained, a scheme of obtaining feedback controls is proposed in [18], which does not involve any derivative of the value function. For some related works, see [2], [6], and [15].

The present paper proceeds to answer the above question for *stochastic* systems. It should be noted that verification technique is particularly important for stochastic systems because only feedback controls perform well in the uncertain environment. However, the approach for the deterministic case [18] relies heavily on the value function being Lipschitz continuous in both time and spatial variables, which is no longer true for stochastic systems of Ito's type. Indeed, since $\int_0^t \sigma dB$ is only of order $t^{\frac{1}{2}}$, the value function in the stochastic case is Hölder continuous of order $\frac{1}{2}$. This causes a great difficulty in the analysis. In this paper, we shall overcome the difficulty by delicate stochastic analysis.

The paper is organized as follows. In section 2, some preliminary results about viscosity solutions and the associated superdifferentials and subdifferentials will be introduced. In section 3, a new verification theorem in terms of viscosity solutions and the superdifferentials is established. In addition, an example is presented showing that the obtained theorem can test for the optimality of a given control while the classical verification theorems cannot. Section 4 discusses the construction of optimal stochastic feedback controls based on the new verification theorem. Finally, section 5 gives some concluding remarks.

2. Superdifferentials, subdifferentials, and viscosity solutions. We shall use the following basic notation throughout the paper:

A^T : the transpose of any vector or matrix A,

$|A|$: the maximum of the elements of any vector or matrix A,

$R^{n\times k}$: the set of all $n \times k$ matrices,

$S^{n\times n}$: the set of all $n \times n$ symmetric matrices.

Given a probability space $(\Omega, \mathcal{F}, P)$ with a filtration $\{\mathcal{F}_t : a \le t \le b\}$ $(-\infty \le a < b \le +\infty)$, a Hilbert space X with the norm $\|\cdot\|_X$, and $p, 1 \le p \le +\infty$, define the set $L^p_{\mathcal{F}}(a, b; X) = \{\phi(\cdot) = \{\phi(t,\omega) : a \le t \le b\}|\ \phi(\cdot)$ is an $\mathcal{F}_t$-adapted, X-valued measurable process on $[a, b]$, and $E\int_a^b \|\phi(t,\omega)\|_X^p\, dt < +\infty\}$.

DEFINITION 2.1. *Let $v \in C([0,1]\times R^n)$. The right superdifferential (resp., subdifferential) of v at $(t_0, x_0) \in [0,1)\times R^n$, denoted by $D^+_{t+,x}v(t_0,x_0)$ (resp., $D^-_{t+,x}v(t_0,x_0)$), is a set defined by*

$$D^+_{t+,x}v(t_0,x_0) = \{(p,q,Q) \in R^1 \times R^n \times S^{n\times n}| \ \overline{\lim}_{t\to t_0+, x\to x_0} \frac{v(t,x)-v(t_0,x_0)-p(t-t_0)-q\cdot(x-x_0)-\frac{1}{2}(x-x_0)^T Q(x-x_0)}{|t-t_0|+|x-x_0|^2} \le 0\}$$

(*resp.*,

$$D^-_{t+,x}v(t_0,x_0) = \{(p,q,Q) \in R^1 \times R^n \times S^{n\times n}| \ \underline{\lim}\{\cdots\} \ge 0\}).$$

Remark 2.1. To study stochastic control problems, many authors make use of the superdifferential $D^+_{t,x}v(t_0,x_0)$ and subdifferential $D^-_{t,x}v(t_0,x_0)$ obtained by replacing the right-sided limit $t \to t_0+$ in the above definition by the two-sided limit $t \to t_0$ (e.g., [13, 14, 3]). The right-sided differentials have been studied extensively in [17] and proved to be more useful than the two-sided differential in treating some stochastic

control problems (see, e.g., [17, Remark 4.1] and [7]). On the other hand, the following inclusions are clear:

$$D^+_{t,x}v(t_0,x_0) \subseteq D^+_{t+,x}v(t_0,x_0),\ D^-_{t,x}v(t_0,x_0) \subseteq D^-_{t+,x}v(t_0,x_0).$$

DEFINITION 2.2. *A function $v \in C([0,1]\times R^n)$ is called a viscosity solution of the HJB equation* (1.4) *if*

$$\begin{aligned} &-p+\sup_{u\in\Gamma} G(t,x,u,q,Q) \le 0 \quad \forall (p,q,Q)\in D^+_{t+,x}v(t,x)\ \ \forall (t,x)\in[0,1)\times R^d,\\ &-p+\sup_{u\in\Gamma} G(t,x,u,q,Q) \ge 0 \quad \forall (p,q,Q)\in D^-_{t+,x}v(t,x)\ \ \forall (t,x)\in[0,1)\times R^d, \end{aligned}$$

and $v(1,x)=h(x)$.

Remark 2.2. The notion of a viscosity solution in the sense specified in Definition 2.2 is more general than those which involve two-sided differentials in t (cf. [3, 13, 14]) in view of the set inclusions in Remark 2.1. Moreover, the uniqueness of viscosity solutions in our sense holds if the uniqueness holds in the "two-sided" sense.

Now let us turn to the control problem formulated in section 1. We impose the following assumptions throughout this paper.

(A1) f, σ, and L are continuous mappings from $[0,1]\times R^d\times\Gamma$ to R^d, $R^{d\times l}$, and R^1, respectively; moreover, they are continuous with respect to (t,x), uniformly in $u\in\Gamma$.

(A2) There exists a constant $K>0$ which is independent of (t,u) such that

$$\begin{aligned} &|f(t,x,u)-f(t,y,u)|+|\sigma(t,x,u)-\sigma(t,y,u)|\\ +\ &|L(t,x,u)-L(t,y,u)|+|h(x)-h(y)|\le K|x-y| \quad \forall x,y\in R^d,\\ &|f(t,x,u)|+|\sigma(t,x,u)|+|L(t,x,u)|+|h(x)|\le K(1+|x|)\ \ \forall x\in R^d. \end{aligned}$$

The following result can be found in [16, 17].

LEMMA 2.1. *The value function V satisfies*

$$|V(t,x)-V(t',x')|\le C(|t-t'|^{\frac12}+|x-x'|).$$

Moreover, V is a unique viscosity solution of the HJB equation (1.4).

An immediate consequence of Lemma 2.1 is the following.

COROLLARY 2.1. *We have*

$$\inf_{(p,q,Q,u)\in D^+_{t+,x}V(t,x)\times\Gamma}[p-G(t,x,u,q,Q)]\ge 0\ \ \forall (t,x)\in[0,1)\times R^d.$$

We need some technical lemmas.

LEMMA 2.2. *Let $v\in C([0,1]\times R^n)$ be a given function that satisfies*

$$|v(t,x)-v(t',x')|\le C_1(|t-t'|^{\frac12}+|x-x'|).$$

For any $(t_0,x_0)\in[0,1)\times R^n$, if $(p,q,Q)\in D^+_{t+,x}v(t_0,x_0)$ (resp., $(p,q,Q)\in D^-_{t+,x}v(t_0,x_0)$), then there exists a function $\phi:[t_0,1]\times R^n\to R^1$ satisfying

(i) $\phi\in C([t_0,1]\times R^n)\cap C^{1,2}((t_0,1]\times R^n)$,

(ii) $\phi(t_0,x_0)=v(t_0,x_0)$ *and* $\phi(t,x)>v(t,x)$ (*resp.,* $\phi(t,x)<v(t,x)$) *for any* $(t,x)\ne(t_0,x_0)$,

(iii) $\lim_{t\to t_0+,x\to x_0,|x-x_0|\le N|t-t_0|^{\frac12}}\phi_t(t,x)=p$ *for any fixed* $N>0$, $\phi_x(t_0,x_0)=q$ *and* $\phi_{xx}(t_0,x_0)=Q$,

(iv)

$$|\phi_t(t,x)| \le C_2(1+\tfrac{|x-x_0|}{|t-t_0|^{\frac{1}{2}}}) \ \ \forall (t,x)\in (t_0,1]\times R^n,$$
$$|\phi_x(t,x)|+|\phi_{xx}(t,x)| \le C_2(1+|x|+|x|^2+|x|^3) \ \ \forall (t,x)\in [t_0,1]\times R^n.$$

Proof. This lemma was presented and proved in Zhou [17] except that for (iii) the statement there was

$$\lim_{t\to t_0+,x\to x_0}\phi_x(t,x)=q, \qquad \lim_{t\to t_0+,x\to x_0}\phi_{xx}(t,x)=Q.$$

However, it is easily seen by the proof in [17] that

$$\phi_x(t_0,x_0)=q \text{ and } \phi_{xx}(t_0,x_0)=Q$$

hold as well. □

LEMMA 2.3. *Let $g\in C[0,1]$. Suppose that there is $\rho\in L^1[0,1]$ such that for sufficiently small $h>0$,*

$$\frac{g(t+h)-g(t)}{h}\le \rho(t), \ \text{a.e. } t\in[0,1]. \tag{2.1}$$

Then

$$g(t)-g(0)\le \int_0^t \overline{\lim_{h\to 0+}}\frac{g(r+h)-g(r)}{h}dr \ \ \forall t\in[0,1]. \tag{2.2}$$

Proof. First fix $t\in[0,1)$. By (2.1), we can apply Fatou's lemma to get

$$\begin{aligned}\textstyle\int_0^t \overline{\lim}_{h\to 0+}\frac{g(r+h)-g(r)}{h}dr &\ge \textstyle\overline{\lim}_{h\to 0+}\int_0^t \frac{g(r+h)-g(r)}{h}dr\\ &= \textstyle\overline{\lim}_{h\to 0+}\frac{\int_h^{t+h} g(r)dr-\int_0^t g(r)dr}{h}\\ &= \textstyle\overline{\lim}_{h\to 0+}\frac{\int_t^{t+h} g(r)dr-\int_0^h g(r)dr}{h}\\ &= g(t)-g(0).\end{aligned}$$

This proves (2.2) $\forall t\in[0,1)$. Finally, the $t=1$ case is obtained by continuity. □

3. Verification theorems.

THEOREM 3.1. *Let $W\in C([0,1]\times R^d)$ be a viscosity solution of the HJB equation (1.4). Then*

(a) *$W(s,y)\le J(s,y;u(\cdot))$ for any $(s,y)\in[0,1)\times R^d$ and any $u(\cdot)\in U_{ad}[s,1]$.*

(b) *Let $(x^*(\cdot),u^*(\cdot))$ be a given admissible pair for the problem $C_{s,y}$. Suppose that there exists $(p^*,q^*,Q^*)\in L^2_{\mathcal{F}}(s,1;R^1)\times L^2_{\mathcal{F}}(s,1;R^d)\times L^2_{\mathcal{F}}(s,1;S^{d\times d})$ (where the filtration $\mathcal{F}_t=\mathcal{F}_t^s$) such that for a.e. $t\in[s,1]$,*

$$(p^*(t),q^*(t),Q^*(t))\in D^+_{t+,x}W(t,x^*(t)), \ P\text{-a.s.} \tag{3.1}$$

and

$$-p^*(t)+G(t,x^*(t),u^*(t),q^*(t),Q^*(t))=0, \ P\text{-a.s.}; \tag{3.2}$$

then $(x^*(\cdot),u^*(\cdot))$ is an optimal pair for the problem $C_{s,y}$.

Proof. Part (a) is trivial since $W=V$ in view of the uniqueness of the viscosity solutions. (Note that we state our results in the present form purposely in order to

compare with the classical verification theorem.) We prove only part (b) of the theorem. Set $f^*(t) = f(t, x^*(t), u^*(t))$, etc., to simplify the notation. Fix $t \in [s, 1)$ such that (3.1) and (3.2) hold. Choose a test function $\phi \in C([t,1] \times R^d) \cap C^{1,2}((t,1] \times R^d)$ as determined by $(p^*(t), q^*(t), Q^*(t)) \in D^+_{t+,x} W(t, x^*(t))$ and Lemma 2.2. Applying Ito's formula to ϕ, we have for any $h > 0$,

$$\begin{aligned} & W(t+h, x^*(t+h)) - W(t, x^*(t)) \\ \leq\ & \phi(t+h, x^*(t+h)) - \phi(t, x^*(t)) \\ =\ & \textstyle\int_t^{t+h} [\phi_t(r, x^*(r)) + \phi_x(r, x^*(r)) \cdot f^*(r) + \frac{1}{2}\mathrm{tr}(\sigma^{*T}(r)\phi_{xx}(r, x^*(r))\sigma^*(r))]dr. \end{aligned} \tag{3.3}$$

It is well known by the martingale property of stochastic integrals that there are constants $C_3, C_4(\alpha) > 0$, independent of t, such that

$$\begin{aligned} & E|x^*(r) - x^*(t)|^2 \leq C_3|r-t| \ \ \forall r \geq t, \\ & E \sup_{s \leq r \leq 1} |x^*(r)|^\alpha \leq C_4(\alpha) \ \ \forall \alpha \geq 1. \end{aligned} \tag{3.4}$$

Hence, in view of Lemma 2.2 (iv), we have

$$\sup_{t<r\leq 1} E|\phi_t(r, x^*(r))|^2 \leq C_2^2 \sup_{t<r\leq 1} E\left[1 + \frac{|x^*(r) - x^*(t)|^2}{r-t}\right] \leq C_5, \tag{3.5}$$

or

$$\sup_{t<r\leq 1} E|\phi_t(r, x^*(r))| \leq \sqrt{C_5}.$$

Moreover, by Lemma 2.2 (iv) and assumption (A2), one can show that

$$\sup_{t\leq r\leq 1} E|\phi_x(r, x^*(r)) \cdot f^*(r) + \frac{1}{2}\mathrm{tr}(\sigma^{*T}(r)\phi_{xx}(r, x^*(r))\sigma^*(r))| \leq C_6.$$

It then follows from (3.3) that for sufficiently small $h > 0$,

$$\frac{EW(t+h, x^*(t+h)) - EW(t, x^*(t))}{h} \leq \sqrt{C_5} + C_6. \tag{3.6}$$

Now we calculate, for any fixed $N > 0$,

$$\begin{aligned} & \tfrac{1}{h}\textstyle\int_t^{t+h} E[\phi_t(r, x^*(r)) - p^*(t)]dr \\ =\ & \tfrac{1}{h}\textstyle\int_t^{t+h} E[(\phi_t(r, x^*(r)) - p^*(t))\chi_{|x^*(r)-x^*(t)|>N|r-t|^{\frac{1}{2}}}]dr \\ & + \tfrac{1}{h}\textstyle\int_t^{t+h} E[(\phi_t(r, x^*(r)) - p^*(t))\chi_{|x^*(r)-x^*(t)|\leq N|r-t|^{\frac{1}{2}}}]dr \\ =\ & I_1(N,h) + I_2(N,h). \end{aligned}$$

By virtue of (3.4) and (3.5), we have

$$\begin{aligned} I_1(N,h) & \leq \tfrac{1}{h}\textstyle\int_t^{t+h} [E|\phi_t(r, x^*(r)) - p^*(t)|^2]^{\frac{1}{2}} [P(|x^*(r) - x^*(t)| > N|r-t|^{\frac{1}{2}})]^{\frac{1}{2}} dr \\ & \leq \tfrac{C}{N} \to 0 \text{ uniformly in } h > 0 \text{ as } N \to \infty. \end{aligned}$$

On the other hand, for fixed $N > 0$, we apply Lemma 2.2 (iii) to get

$$\sup_{t<r\leq t+h} \left[(\phi_t(r, x^*(r)) - p^*(t))\chi_{|x^*(r)-x^*(t)|\leq N|r-t|^{\frac{1}{2}}}\right] \to 0 \text{ as } h \to 0+, P\text{-a.s.}$$

Thus we conclude by the dominated convergence theorem that

$$I_2(N,h) \to 0 \text{ as } h \to 0+ \text{ for each fixed } N.$$

Therefore, we have proved that $\frac{1}{h}\int_t^{t+h} E\phi_t(r, x^*(r))dr \to Ep^*(t)$ as $h \to 0+$. Similarly (in fact, more easily), we can show that

$$\frac{1}{h}\int_t^{t+h} E[\phi_x(r, x^*(r)) \cdot f^*(r)]dr \to E[\phi_x(t, x^*(t)) \cdot f^*(t)] = E[q^*(t) \cdot f^*(t)]$$

and

$$\frac{1}{h}\int_t^{t+h} E[\frac{1}{2}\mathrm{tr}(\sigma^{*T}(r)\phi_{xx}(r, x^*(r))\sigma^*(r))]dr \to E[\frac{1}{2}\mathrm{tr}(\sigma^{*T}(t)\phi_{xx}(t, x^*(t))\sigma^*(t))]$$

$$= E[\frac{1}{2}\mathrm{tr}(\sigma^{*T}(t)Q^*(t)\sigma^*(t))]$$

as $h \to 0+$. Consequently, (3.3) gives

$$\begin{aligned} & \overline{\lim}_{h\to 0+} \frac{EW(t+h, x^*(t+h)) - EW(t, x^*(t))}{h} \\ \leq\ & E[p^*(t) + q^*(t) \cdot f^*(t) + \frac{1}{2}\mathrm{tr}(\sigma^{*T}(t)Q^*(t)\sigma^*(t))] \\ =\ & -EL^*(t), \end{aligned}$$

where the last equality is due to (3.2). Noting (3.6) and applying Lemma 2.3 to $g(t) = EW(t, x^*(t))$, we arrive at

$$EW(1, x^*(1)) - EW(s, y) \leq \int_s^1 -EL^*(t)dt,$$

which leads to $W(s, y) \geq J(s, y; u^*(\cdot))$. It follows from (a) that $(x^*(\cdot), u^*(\cdot))$ is an optimal pair for $C_{s,y}$. □

Remark 3.1. In view of Corollary 2.1, the condition (3.2) implies that $(p^*(t), q^*(t), Q^*(t), u^*(t))$ achieves the infimum of $p - G(t, x^*(t), u, q, Q)$ over $D^+_{t+,x}V(t, x^*(t)) \times \Gamma$. Meanwhile, it also shows that (3.2) is equivalent to

$$p^*(t) \leq G(t, x^*(t), u^*(t), q^*(t), Q^*(t)).$$

Remark 3.2. The condition (3.2) implies that

$$\max_{u\in\Gamma} G(t, x^*(t), u, q^*(t), Q^*(t)) = G(t, x^*(t), u^*(t), q^*(t), Q^*(t)). \tag{3.7}$$

This is easily seen by recalling the fact that V is the viscosity solution of (1.4):

$$-p^*(t) + \sup_{u\in\Gamma} G(t, x^*(t), u, q^*(t), Q^*(t)) \leq 0,$$

which yields (3.7) under (3.2).

Remark 3.3. By Remark 2.1, the new verification theorem holds if the right superdifferential $D^+_{t+,x}W(t, x^*(t))$ is replaced by the (smaller) two-sided superdifferential $D^+_{t,x}W(t, x^*(t))$.

Theorem 3.1 is a generalization of the classical verification theorem (Theorem 1.1). On the other hand, we do have examples showing that the classical verification theorem may not be able to verify the optimality of a given control, whereas Theorem 3.1 can.

Example 3.1. Consider the following optimal control problem:

$$\begin{aligned} \text{minimize} \quad & E[-x(1)], \\ \text{subject to} \quad & \begin{cases} dx(t) = [x(t)(u(t)+1) - e^t u(t)]dt + |x(t) - e^t|dB(t), \\ x(s) = y \in R^1, \end{cases} \\ \text{control} \quad & u(\cdot) : [0,1] \to \{r \in R^1 | 0 \leq r \leq 1\}. \end{aligned}$$

The HJB equation is

$$\begin{cases} -v_t(t,x) + \sup_{0\le u\le 1}[v_x(t,x)(e^t - x)u] - \frac{1}{2}(x-e^t)^2 v_{xx}(t,x) - xv_x(t,x) = 0, \\ v(1,x) = -x. \end{cases}$$

It is not difficult to verify that the following function is a viscosity solution of the HJB equation

$$V(t,x) = \begin{cases} -e^{1-t}x & \text{if } x \le e^t, \\ (e^{1-t} - e^{1-2t}x - 1)e & \text{if } x > e^t, \end{cases}$$

which, by the uniqueness of the viscosity solutions, turns out to be the value function of the control problem. Let us consider an admissible control $u^*(\cdot) \equiv 0$ for initial time $s = 0$ and initial state $y = 1$. The trajectory under $u^*(\cdot)$ is easily seen to be $x^*(t) = e^t$. Now we want to see if the pair $(x^*(\cdot), u^*(\cdot))$ is optimal. Theorem 1.1 cannot tell us anything, because $V_x(t, x^*(t))$ does not exist on the *whole* trajectory $x^*(\cdot)$. However, we have $D^+_{t+,x}V(t,x^*(t)) = [e,+\infty) \times [-e^{2-2t}, -e^{1-t}] \times \{Q \in S^{d\times d} : Q \ge 0\}$. Now if we take $(p^*(t), q^*(t), Q^*(t)) = (e, -e^{1-t}, 0) \in D^+_{t+,x}V(t,x^*(t))$ for each t, then (3.2) is satisfied. This implies that the pair $(x^*(\cdot), u^*(\cdot))$ is indeed optimal by virtue of Theorem 3.1.

4. Optimal feedback controls. This section describes how to construct optimal feedback controls by the verification theorem obtained. First we recall the definition of admissible feedback controls.

DEFINITION 4.1. *A measurable function* $\mathbf{u}$ *from* $[0,1] \times R^d$ *to* Γ *is called an admissible feedback control if for any* $(s,y) \in [0,1) \times R^d$ *there is a weak solution* $x(\cdot; s, y)$ *of the following equation:*

$$\begin{cases} dx(t) = f(t, x(t), \mathbf{u}(t, x(t)))dt + \sigma(t, x(t), \mathbf{u}(t, x(t)))dB(t), \\ x(s) = y. \end{cases} \tag{4.1}$$

An admissible feedback control $\mathbf{u}^*$ *is called optimal if* $(x^*(\cdot; s, y), \mathbf{u}^*(\cdot, x^*(\cdot; s, y)))$ *is optimal for the problem* $C_{s,y}$ *for* each (s,y)*, where* $x^*(\cdot; s, y)$ *is a solution of* (4.1) *corresponding to* $\mathbf{u}^*$*.*

THEOREM 4.1. *Let* $\mathbf{u}^*$ *be an admissible feedback control and* $\mathbf{p}^*$*,* $\mathbf{q}^*$*, and* $\mathbf{Q}^*$ *be measurable functions satisfying* $(\mathbf{p}^*(t,x), \mathbf{q}^*(t,x), \mathbf{Q}^*(t,x)) \in D^+_{t+,x}V(t,x)$ *for all* (t,x)*. If*

$$\begin{aligned} & \mathbf{p}^*(t,x) - G(t,x,\mathbf{u}^*(t,x), \mathbf{q}^*(t,x), \mathbf{Q}^*(t,x)) \\ = \ & \inf_{(p,q,Q,u)\in D^+_{t+,x}V(t,x)\times\Gamma}[p - G(t,x,u,q,Q)] \\ = \ & 0 \end{aligned} \tag{4.2}$$

for all $(t,x) \in [0,1] \times R^d$*, then* $\mathbf{u}^*$ *is optimal.*

Proof. The result follows readily from (b) of Theorem 3.1. □

By Theorem 4.1, we see that under proper conditions, one can obtain an optimal feedback control by minimizing $p - G(t,x,u,q,Q)$ over $(p,q,Q,u) \in D^+_{t+,x}V(t,x) \times \Gamma$ for each (t,x). Let us investigate the conditions imposed in Theorem 4.1. First of all, (4.2) requires that

$$\inf_{(p,q,Q,u)\in D^+_{t+,x}V(t,x)\times\Gamma}[p - G(t,x,u,q,Q)] = 0, \tag{4.3}$$

and in addition the infimum can be achieved. This condition in fact partially characterizes the existence of an optimal feedback control, although rather implicitly in the sense that the value function is involved. In particular, this condition is satisfied *automatically* if V is smooth. Next, in order to apply Filippov's lemma to obtain a *measurable selector* of $(\mathbf{p}^*(t,x), \mathbf{q}^*(t,x), \mathbf{Q}^*(t,x), \mathbf{u}^*(t,x))$ which achieves the infimum in (4.2), we must study the measurability of the multifunction $(t,x) \mapsto D^+_{t+,x}V(t,x)$. To do this, let us first recall the measurability of the multifunctions (see, e.g., [12], [1], and [8] for details).

DEFINITION 4.2. *Let $X \subset R^n$ be a Lebesgue measurable set, Y be a metric space, and $\Lambda : X \to 2^Y$ be a multifunction. We say that Λ is measurable if for any closed set $F \subset Y$ the set*

$$\Lambda^{-1}(F) \triangleq \Big\{ x \in X \mid \Lambda(x) \bigcap F \neq \emptyset \Big\}$$

is Lebesgue measurable.

Note that in the above we do not need Λ to be closed set valued. It is clear that when Y is a Polish space (i.e., a separable complete metric space), the closed set F in the above definition can be replaced by any open set. Consequently, we have the following simple result.

LEMMA 4.1. *Let $X \subset R^n$ be a Lebesgue measurable set, Y be a Polish space, and $\Lambda : X \to 2^Y$ be a multifunction. Then, Λ is measurable if and only if the multifunction $x \mapsto \overline{\Lambda}(x) \triangleq \overline{\Lambda(x)}$ is measurable.*

Proof. We note that for any open set $U \subset Y$ and $x \in X$,

$$\Lambda(x) \bigcap U \neq \phi \quad \Longleftrightarrow \quad \overline{\Lambda(x)} \bigcap U \neq \phi.$$

Hence,

$$\Lambda^{-1}(U) = \overline{\Lambda}^{-1}(U) \qquad \forall \text{ open set } U \subset Y.$$

Then, by the above observation, we obtain our conclusion. □

PROPOSITION 4.1. *Both the multifunctions $(t,x) \mapsto D^+_{t+,x}V(t,x)$ and $(t,x) \mapsto \overline{D^+_{t+,x}V(t,x)}$ are convex set valued and are measurable.*

Proof. For any $(s,y) \in (0,1] \times R^d$, we define

$$W(t,x,p,q,Q;s,y) = \begin{cases} \frac{V(s,y)-V(t,x)-p(s-t)-q\cdot(y-x)-\frac{1}{2}(y-x)^T Q(y-x)}{|s-t|+|y-x|^2} & \text{if } t \in [0,s), \\ 0 & \text{if } t \in [s,1]. \end{cases}$$

Then, (s,y) being regarded as parameters, the function $(t,x,p,q,Q) \mapsto W(t,x,p,q,Q;s,y)$ is Borel measurable. Hence, the function

$$\overline{\lim_{s \to t+, y \to x}} W(t,x,p,q,Q;s,y) \triangleq \widetilde{W}(t,x,p,q,Q)$$

is also Borel measurable. Then, by [12, Theorem III.2.20], we know that the multifunction

$$D^+_{t+,x}V(t,x) \equiv \{(p,q,Q) \mid \widetilde{W}(t,x,p,q,Q) \leq 0\}$$

is measurable. By Lemma 4.1, we obtain the measurability of the multifunction $(t,x) \mapsto \overline{D^+_{t+,x}V(t,x)}$. The convexity of these two multifunctions is obvious. □

Filippov's lemma (see, e.g., [12], [1], and [8]) says that if Λ is a measurable multifunction defined on some Lebesgue measurable set taking closed set values in a Polish space, then it admits a measurable selection. Therefore, if we assume that $D^+_{t+,x}V(t,x)$ is closed and that the infimum in (4.2) can be achieved, then by Proposition 4.1 and Filippov's lemma, we can find a measurable selection $(\mathbf{p}^*(t,x), \mathbf{q}^*(t,x), \mathbf{Q}^*(t,x), \mathbf{u}^*(t,x)) \in D^+_{t+,x}V(t,x)$ that minimizes $p - G(t,x,u,q,Q)$.

Suppose now we have selected a measurable function $\mathbf{u}^*(t,x)$. It may not be an admissible feedback control. The reason is because the coefficients $\tilde{f}(t,x) \stackrel{\Delta}{=} f(t,x,\mathbf{u}^*(t,x))$ and $\tilde{\sigma}(t,x) \stackrel{\Delta}{=} \sigma(t,x,\mathbf{u}^*(t,x))$ of the SDE (4.1) are only measurable in (t,x), which does not guarantee the existence of a solution. This difficulty occurs in the deterministic case as well. However, for the stochastic case, there are some elegant existence and uniqueness results for SDEs with measurable coefficients. Let us briefly discuss two situations.

Case 1. Assume that $\sigma(t,x,u)$ is a $d \times d$ matrix and is uniformly elliptic, i.e.,

$$\lambda^T \sigma(t,x,u)\lambda \geq \delta|\lambda|^2 \tag{4.4}$$

for some constant $\delta > 0$ for all (t,x,u). Then by Krylov [10, Theorem II.6.1], there exists a solution to SDE (4.1) under $\mathbf{u}^*(t,x)$. By Theorem 4.1, $\mathbf{u}^*(t,x)$ is an optimal feedback control.

Case 2. Assume that σ is a nonsingular $d \times d$ matrix and does not depend on u. Moreover,

$$\sup_{t,x} |\sigma^{-1}(t,x)| < +\infty. \tag{4.5}$$

Then under $\mathbf{u}^*(t,x)$, the existence and uniqueness (in law) of the solutions to (4.1) are obtained by Girsanov's transformation (see, e.g., [9, Theorem IV.4.2]). Once again, $\mathbf{u}^*(t,x)$ is an optimal feedback control by Theorem 4.1.

To summarize the above discussion, we have the following theorem.

THEOREM 4.2. *Assume that*

(i) $\inf_{(p,q,Q,u)\in D^+_{t+,x}V(t,x)\times\Gamma}[p - G(t,x,u,q,Q)] = 0$.

(ii) $D^+_{t+,x}V(t,x)$ *is closed and the infimum above can be achieved.*

(iii) *Either* (4.4) *or* (4.5) *holds.*

Then, there is a measurable selector $(\mathbf{p}^*(t,x), \mathbf{q}^*(t,x), \mathbf{Q}^*(t,x), \mathbf{u}^*(t,x))$ *that minimizes* $p - G(t,x,u,q,Q)$. *Moreover, the fourth component* $\mathbf{u}^*(t,x)$ *is an optimal feedback control.*

Remark 4.1. In the presence of the uniform ellipticity of $\sigma\sigma^T$, Krylov proved the existence of classical solutions to the HJB equation under additional regularity and strong boundness assumptions on the coefficients f, σ, L, h (see [11, Chapter 6, p. 301]). Under the mild assumptions in this paper, one does not know the existence of classical solutions to the HJB equation even with (4.4).

5. Concluding remarks. In this paper we have derived a new verification theorem in the language of viscosity solutions and the associated superdifferentials. The conditions under which the theorem is valid are very mild and reasonable, compared with the restrictive classical verification theorem. We have also discussed the construction of optimal feedback controls based on the verification theorem obtained in this paper. Basically, the verification theorem reduces the original stochastic control problem into a two-phase problem. In the first phase, one has to solve the HJB equations which are fully nonlinear second-order PDEs. In most cases one has to

rely on numerical methods to solve the equations, whereas only in some exceptional cases one may obtain analytical solutions (like the one in Example 3.1). In the second phase, one finds the optimal feedback $\mathbf{u}^*$ by minimizing $p - G(t, x, u, q, Q)$ over *both* superdifferential of V and the control region. The second phase is relatively easy because the superdifferential of V is explicitly known once V is known. However, if V is approximated by numerical solutions V_n, then a natural problem is under what conditions the feedback controls obtained by applying our verification theorem to V_n are good enough. We then need to study the asymptotic behavior of the superdifferentials/subdifferentials of the approximating solutions V_n. These remain very challenging problems.

It should be noted that the results of this paper were derived when there was no state constraint in the optimal control problem. We do not know how to treat the state constraint problems. Indeed, the presence of state constraints causes great difficulty to the analysis; they bring some particular boundary conditions (depending on the particular features of the state constraints imposed) to the associated HJB equations, while the existing viscosity solutions theory on nonlinear PDEs with boundary conditions is far from satisfactory and complete.

REFERENCES

[1] I. P. Aubin and H. Frankowska, *Set-valued Analysis*, Birkhäuser, Boston, 1990.

[2] L. D. Berkovitz, *Optimal feedback controls*, SIAM J. Control Optim., 27 (1989), pp. 991–1007.

[3] M. G. Crandall, H. Ishii, and P. L. Lions, *A user's guide to viscosity solutions*, Bull. Amer. Math. Soc., 27 (1992), pp. 1–67.

[4] M. G. Crandall and P. L. Lions, *Viscosity solutions of Hamilton-Jacobi equations*, Trans. Amer. Math. Soc., 277 (1983), pp. 1–42.

[5] W. H. Fleming and R. W. Rishel, *Deterministic and Stochastic Optimal Control*, Springer-Verlag, New York, 1975.

[6] H. Frankowska, *Lower semicontinuous solutions of Hamilton-Jacobi-Bellman equations*, SIAM J. Control Optim., 31 (1993), pp. 257–272.

[7] U. G. Haussmann, *Generalized solutions of the Hamilton-Jacobi equation of stochastic control*, SIAM J. Control Optim., 32 (1994), pp. 728–743.

[8] C. J. Himmelberg, M. Q. Jacobs, and F. S. van Vleck, *Measurable multifunctions, selectors and Filippov's implicit functions lemma*, J. Math. Anal. Appl., 25 (1969), pp. 276–284.

[9] N. Ikeda and S. Watanabe, *Stochastic Differential Equations and Diffusion Processes*, North-Holland/Kodansha, Tokyo, 1981.

[10] N. V. Krylov, *Controlled Diffusion Processes*, Springer-Verlag, New York, 1980.

[11] N. V. Krylov, *Nonlinear Elliptic and Parabolic Equations of the Second Order*, Reidel, Dordrecht, the Netherlands, 1987.

[12] X. Li and J. Yong, *Optimal Control Theory for Infinite Dimensional Systems*, Birkhäuser, Boston, 1995.

[13] P. L. Lions, *Optimal control of diffusion processes and Hamilton-Jacobi-Bellman equations* I, Comm. Partial Differential Equations, 8 (1983), pp. 1101–1134.

[14] P. L. Lions, *Optimal control of diffusion processes and Hamilton-Jacobi-Bellman equations, part* 2: *Viscosity solutions and uniqueness*, Comm. Partial Differential Equations, 11 (1983), pp. 1229–1276.

[15] J. D. L. Rowland and R. B. Vinter, *Construction of optimal feedback controls*, Systems Control Lett., 16 (1991), pp. 357–367.

[16] X. Y. Zhou, *The connection between the maximum principle and dynamic programming in stochastic control*, Stochastics Stochastics Rep., 31 (1990), pp. 1–13.

[17] X. Y. Zhou, *A unified treatment of maximum principle and dynamic programming in stochastic controls*, Stochastics Stochastics Rep., 36 (1991), pp. 137–161.

[18] X. Y. Zhou, *Verification theorems within the framework of viscosity solutions*, J. Math. Anal. Appl., 177 (1993), pp. 208–225.

SIAM J. CONTROL OPTIM.
Vol. 35, No. 1, pp. 254–264, January 1997

014

BEHAVIORAL CONTROLLABILITY OF DELAY-DIFFERENTIAL SYSTEMS*

PAULA ROCHA† AND JAN C. WILLEMS‡

Abstract. In this paper we will prove that the system described by the delay-differential equation $R(d/dt, \Delta)w = 0$ (with Δ the unit delay operator) is controllable if and only if the rank of $R(\lambda, e^{-\lambda})$ is constant for all $\lambda \in \mathbb{C}$. This condition is compared with the existing results obtained both by the analytic approach and by the algebraic approach to delay-differential systems.

Key words. behavioral systems, controllability, delay-differential systems, polynomial matrices

AMS subject classification. 93B05

PII. S0363012995283054

1. Introduction. The aim of this paper is to analyze controllability for delay-differential (d-d) systems. We will derive a concrete necessary and sufficient condition for controllability of d-d systems in kernel representation. We will use the behavioral approach to dynamical systems [12]. Thus a continuous-time dynamical system is a triple $\Sigma = (\mathbb{R}, \mathbb{R}^q, \mathcal{B})$ with *behavior* $\mathcal{B}$ being a set of trajectories $w : \mathbb{R} \to \mathbb{R}^q$. We will assume that $\mathcal{B}$ is shift invariant, i.e., that $(w(.) \in \mathcal{B}) \Rightarrow (w(t + .) \in \mathcal{B} \ \ \forall t \in \mathbb{R})$. Since the behavior is the most intrinsic feature of a system, it is logical to define the system properties in terms of the set $\mathcal{B}$, i.e., at an external level. This applies in particular for the notion of controllability.

DEFINITION 1.1. *The system Σ is said to be* controllable *if for all $w_1, w_2 \in \mathcal{B}$ there exist a $w \in \mathcal{B}$ and a $T \geq 0$ such that*

$$w(t) = \begin{cases} w_1(t) & for \quad t < 0, \\ w_2(t - T) & for \quad t \geq T. \end{cases}$$

Note that for shift-invariant behaviors the controllability condition of Definition 1.1 is equivalent to the following property. For all $w_1, w_2 \in \mathcal{B}$, w_1 is $\mathcal{B}$-*compatible* with w_2, i.e., for all $t_1 \in \mathbb{R}$ there exist $t_2 \geq t_1$ and $w \in \mathcal{B}$ such that $w^* = w_1 \wedge_{t_1} w \wedge_{t_2} w_2 \in \mathcal{B}$. Here $w^* = w_1 \wedge_{t_1} w \wedge_{t_2} w_2$ stands for the successive concatenation of w_1, w, and w_2, respectively, at times t_1 and t_2 and is defined as follows: $w^*(t) = w_1(t)$ for $t < t_1$, $w^*(t) = w(t)$ for $t_1 \leq t < t_2$, and $w^*(t) = w_2(t)$ for $t \geq t_2$. In other words, controllability requires that every past trajectory can be transferred to any future trajectory. In order to distinguish this property from the classical state controllability and to emphasize the fact that it concerns the system behavior we will refer to it as behavioral controllability.

Behavioral controllability has been widely studied for both continuous- and discrete-time systems, respectively, described by differential and difference equations, see [12, 8]. In this paper we consider continuous-time systems described by differential

*Received by the editors March 15, 1995; accepted for publication (in revised form) November 22, 1995.

http://www.siam.org/journals/sicon/35-1/28305.html

†Department of Mathematics, University of Aveiro, 3800 Aveiro, Portugal (procha@math.ua.pt). The work of this author was supported by the Dutch Network of Systems and Control through a travel grant.

‡Mathematics Institute, University of Groningen, P.O. Box 800, 9700 AV Groningen, the Netherlands (J.C.Willems@math.rug.nl).

equations with delays, i.e., d-d systems. More concretely, we will be concerned with systems whose behavior $\mathcal{B}$ can be described as the kernel of a d-d operator $R(d/dt, \Delta)$ (where $R(z_1, z_2)$ is a two-dimensional (2D) polynomial matrix in z_1 and z_2 and Δ is the delay). This is a very general description which can comprise both the polynomial input-output equations and the pseudostate representations considered in the literature [11], [3]. We will show that $\mathcal{B} = \ker R(d/dt, \Delta)$ is controllable if and only if (iff) $R(\lambda, e^{-\lambda})$ has constant rank for all $\lambda \in \mathbb{C}$. It turns out that this condition reduces to spectral controllability if one considers pseudostate representations as in [3], [6], and [9].

This characterization of behavioral controllability has also been independently obtained in [1], where the author develops an elegant theory for d-d systems in a behavioral framework based on the properties of a suitable ring of entire functions. Here we follow a different approach based on the analysis of the exponential-polynomial trajectories in the system.

2. Delay-differential systems. Let $R(z_1, z_2)$ be a 2D polynomial matrix having g rows and q columns. Now consider the equation

$$R\left(\frac{d}{dt}, \Delta\right) w = 0, \tag{1}$$

where Δ denotes the unit delay operator: $(\Delta f)(t) := f(t-1)$. Equation (1) defines the dynamical system $(\mathbb{R}, \mathbb{R}^q, \mathcal{B})$ with $\mathcal{B} = \ker(R(d/dt, \Delta))$ and $R(d/dt, \Delta)$ viewed as a map from $\mathcal{C}^\infty(\mathbb{R}, \mathbb{R}^q)$ into $\mathcal{C}^\infty(\mathbb{R}, \mathbb{R}^g)$. In other words, the behavior consists of the $\mathcal{C}^\infty$-solutions of (1). We will call (1) a d-d system (even though it would be more appropriate to refer to it as a d-d system in kernel representation).

Note that this kernel representation is more general than the polynomial input-output descriptions considered in [11], as well as than the pseudostate descriptions of [3]. Indeed, any polynomial input-output d-d equation $Py = Qu$ can be regarded as a kernel representation with $R(d/dt, \Delta) = [P(d/dt, \Delta) \mid -Q(d/dt, \Delta)]$ and with $w = col(y, u)$. In turn, the pseudostate description $\{dx/dt = A(\Delta)x + Bu \;\; y = Cx\}$ can also be viewed as a kernel representation with $w = col(x, y, u)$ and $R(d/dt, \Delta) = col([d/dt - A(\Delta) \mid 0 \mid -B], [-C \mid I \mid 0])$. Observe, however, that (1) is a broader class of systems than those mentioned. For example, both the systems defined by

$$w_1 = \frac{d}{dt}\Delta w_2$$

and by

$$\Delta w_1 = (1 + \Delta^2) w_2$$

fit (1) but not the classical input/state/output frameworks.

3. Behavioral controllability of d-d systems. Our problem is to find conditions on the 2D polynomial matrix $R(z_1, z_2)$ such that (1) defines a system which is controllable in the sense of Definition 1.1. The following is the main result of this paper.

THEOREM 3.1. (1) *defines a controllable d-d system iff the rank of the complex matrix* $R(\lambda, e^{-\lambda})$ *is constant for* $\lambda \in \mathbb{C}$.

The above theorem is a natural generalization of the well-known identical result for differential systems $R(d/dt)w = 0$. However, the proof will show that from a mathematical point of view Theorem 3.1 is a much deeper result.

As an alternative to systems (1), consider the following d-d systems. Let $M(z_1, z_2)$ be a 2D polynomial matrix with q rows and l columns. Consider the equation

$$w = M\left(\frac{d}{dt}, \Delta\right) a, \tag{2}$$

where $a \in C^\infty(\mathbb{R}, \mathbb{R}^l)$ corresponds to an auxiliary variable. Equation (2) defines a dynamical system $(\mathbb{R}, \mathbb{R}^q, \mathrm{im}(M(d/dt, \Delta)))$ with $M(d/dt, \Delta)$ viewed as an operator from $\mathcal{C}^\infty(\mathbb{R}, \mathbb{R}^l)$ into $\mathcal{C}^\infty(\mathbb{R}, \mathbb{R}^q)$. We will call (2) a d-d system in image representation. It is easy to prove that (2) defines a dynamical system which is automatically controllable. For differential systems, a system is controllable if and *only if* it admits an image representation. This is, in fact, also the case for d-d systems (1).

THEOREM 3.2. *A d-d system* (1) *is a controllable system iff there exists a* 2*D polynomial matrix* $M(z_1, z_2)$ *such that*

$$\ker R\left(\frac{d}{dt}, \Delta\right) = \mathrm{im} M\left(\frac{d}{dt}, \Delta\right). \tag{3}$$

In order to give a further insight, it is useful to compare our result with the existing results on state controllability for d-d systems. We will first focus on the class of retarded d-d systems Σ considered in [3] which have a pseudostate description of the form

$$\begin{cases} dx/dt &= A(\Delta)x + Bu, \\ y &= Cx, \end{cases}$$

where x is the (n-dimensional) pseudostate, u is the input, y is the output, and $A(z) = A_N z^N + \cdots + A_1 z + A_0$ is a polynomial matrix in z. For the system Σ, the state at time t is defined in [3] as being $z(t) = col(x(t), x_t)$, where $x_t \in L_2[(-N, 0], \mathbb{R}^n]$ is given by $x_t(\tau) = x(t + \tau)$ for all $\tau \in (-N, 0]$. This yields the infinite-dimensional state space $Z = \mathbb{R}^n \times L_2[(-N, 0), \mathbb{R}^n]$. Define, in this state space, the set K_t of all attainable states in time t, and let $K_\infty := \cup_{t>0} K_t$. Then Σ is said to be approximately controllable if K_∞ is dense in Z. The next theorem, providing a characterization of approximate controllability, has been derived in [3].

THEOREM 3.3. Σ *is approximately controllable iff* (1) $\mathrm{rank}[(\lambda I - A(e^{-\lambda}) \mid B] = n \ \ \forall \lambda \in \mathbb{C}$ *and* (2) $\mathrm{rank}[A_N \mid B] = n$.

The first condition of the theorem is known as *spectral controllability*.

Note that the pseudostate description that we have considered here can be regarded as a kernel representation with $R(d/dt, \Delta) = col([d/dt - A(\Delta) \mid 0 \mid -B], [-C \mid I \mid 0])$ if Σ is viewed as a system with external variable vector $w = col(x, y, u)$ and with smooth signals. It turns out from Theorem 3.1 that the behavior of Σ is controllable iff $\mathrm{rank}[(\lambda I - A(e^{-\lambda}) \mid B] = n$ for all $\lambda \in \mathbb{C}$. So behavioral controllability seems to correspond to spectral rather than to approximate controllability. The situation can be illustrated by the following example.

EXAMPLE 3.4. Let $A(z) = A_0 + A_1 z$ with $A_0 = col([0 \mid 1], [0 \mid 0])$, $A_1 = col([0 \mid 0], [-1 \mid 0])$, and $B = col(0, -1)$. Then the corresponding system Σ is not approximately controllable since $\mathrm{rank}[A_1 \mid B] = 1 < 2$. However, it is easy to check that $[\lambda I - A(e^{-\lambda}) \mid B]$ has rank 2 $\forall \lambda \in \mathbb{C}$ and hence the behavior of Σ is controllable. What happens in this case is that the pseudostate components x_1 and x_2 are related by $dx_1/dt = x_2$. This holds in particular in the interval $[-1, 0)$; therefore, not all the

elements in the state space $\mathbb{R}^2 \times L_2([-1,0),\mathbb{R}^2)$ are feasible, which prevents approximate controllability. This obstacle does not arise for behavioral controllability since this property exclusively regards admissible system signals (and hence one does not take into account the signals which do not satisfy $dx_1/dt = x_2$).

The characterization of approximate controllability has been extended to neutral d-d systems in [6] and [9] and later generalized in [13] to the case of (possibly) noncommensurable delays. For systems with a pseudostate description of the form

$$\left\{ \begin{array}{lcl} dx/dt & = & A(\Delta_1,\ldots,\Delta_N,D)x + Bu, \\ y & = & Cx \end{array} \right. \tag{4}$$

(where $A(z_1,\ldots,z_N,z_{N+1}) = A_0 + \sum_{i=1}^{N} A_i(z_{N+1})z_i$, $A_i(z_{N+1})z_i = E_i + F_i z_{N+1}$, and Δ_i represents the delay by h_i units of time, $i = 1,\ldots,N$), the following result has been derived (and formulated in slightly different terms).

THEOREM 3.5 (see [13]). *The system described by* (4) *is approximately controllable iff* (a) $\operatorname{rank}[A(e^{h_1\lambda},\ldots,e^{h_N\lambda},\lambda),B] = n \ \ \forall\lambda \in \mathbb{C}$ *and* (b) $\operatorname{rank}[A_N(\lambda),B] = n$ *for some* $\lambda \in \mathbb{C}$.

As before, the first condition corresponds to spectral controllability and coincides with our characterization of behavioral controllability if the delays are commensurable.

Another interesting issue is the comparison of our notion of controllability with the ones which have been studied in [5] and [2] within an algebraic approach. Here the authors consider systems Σ with pseudostate-space representations of the form

$$\left\{ \begin{array}{lcl} dx/dt & = & A(\Delta)x + B(\Delta)u, \\ y & = & C(\Delta)x + D(\Delta)u, \end{array} \right. \tag{5}$$

where $A(z_2), B(z_2), C(z_2), D(z_2)$ are polynomial matrices in z_2. For such systems the following two notions of controllability are introduced. Let $\mathcal{R}(z_2) := [B(z_2) \mid A(z_2)B(z_2) \mid \ldots \mid (A(z_2))^{n-1}B(z_2)]$, where n is the size of $A(z_2)$. Σ is said to be *weakly controllable* if $\mathcal{R}(z_2)$ has full row rank over the field of fractions $\mathbb{R}(z_2)$. If $\mathcal{R}(\lambda_2)$ has full row rank $\forall\lambda_2 \in \mathbb{C}$, Σ is said to be *strictly controllable*. Theorem 3.6 is shown in [2].

THEOREM 3.6. *With the previous notation,* (1) Σ *is weakly controllable iff* $[z_1 - A(z_2) \mid B(z_2)]$ *is left prime, and* (2) Σ *is strictly controllable iff* $\operatorname{rank}[\lambda_1 - A(\lambda_2) \mid B(\lambda_2)] = n \ \forall(\lambda_1,\lambda_2) \in \mathbb{C}\times\mathbb{C}$.

Regarding the pseudostate representation (5) as a kernel representation, it follows from Theorem 3.1 that the behavior of Σ is controllable iff $\operatorname{rank}[\lambda - A(e^{-\lambda}) \mid B(e^{-\lambda})] = n \ \forall\lambda \in \mathbb{C}$. Thus strict controllability implies behavioral controllability. On the other hand, if $[z_1 - A(z_2) \mid B(z_2)]$ has a left factor $\Phi(z_1,z_2)$ with nontrivial determinant $f(z_1,z_2)$, then $\Phi(\lambda,e^{-\lambda})$ will be a left factor of $[\lambda - A(e^{-\lambda}) \mid B(e^{-\lambda})]$, implying that this matrix drops in rank when λ is a zero of $f(\lambda,e^{-\lambda})$. Therefore we can conclude that behavioral controllability implies weak controllability.

Summarizing the preceding considerations, we have that strict controllability implies behavioral controllability, which in its turn implies weak controllability. The next examples show that the converse implications do not hold true.

EXAMPLE 3.7. Consider the delay-differential system Σ described by

$$\left\{ \begin{array}{lcl} dx/dt & = & (-\Delta + 1)x + (2 - \Delta)u, \\ y & = & x. \end{array} \right.$$

Letting $w := col(u,y,x)$ and $R(z_1,z_2) := col([z_2 - 2 \mid 0 \mid z_1 + (z_2 - 1)],[0 \mid 1 \mid -1])$ this description becomes $R(d/dt,\Delta)w = 0$. Since $R(\lambda,e^{-\lambda}) = col([e^{-\lambda} - 2 \mid 0 \mid$

$\lambda + (e^{-\lambda} - 1)], [0 \mid 1 \mid -1])$ has rank 2 $\forall \lambda \in \mathbb{C}$, the behavior of Σ is controllable. However, $[\lambda_1 - (1-\lambda_2) \mid 2 - \lambda_2]$ clearly drops in rank for $(\lambda_1, \lambda_2) = (-1, 2)$, and hence Σ is not strictly controllable.

EXAMPLE 3.8. Let Σ be described by the following equations:

$$\begin{cases} dx/dt &= (-\Delta + 1)u, \\ y &= x. \end{cases}$$

Proceeding as in the previous example, we have that $R(\lambda, e^{-\lambda}) = col([1 - e^{-\lambda} \mid 0 \mid \lambda], [0 \mid 1 \mid -1])$, which drops in rank for $\lambda = 0$. So the behavior of Σ is not controllable. However $[z_1 \mid z_2 - 1]$ is left prime and hence Σ is weakly controllable.

EXAMPLE 3.9. Consider the system described in image representation by

$$\begin{bmatrix} w_1 \\ w_2 \end{bmatrix} = \begin{bmatrix} 1 - \Delta \\ \frac{d}{dt} \end{bmatrix} a. \tag{6}$$

This is a system with transfer function $w_2 \to w_1$:

$$\frac{\hat{w}_1}{\hat{w}_2} = \frac{1 - e^{-s}}{s}. \tag{7}$$

Obviously, since it is an image representation, it defines a controllable system. The logical candidate for the kernel representation is

$$\frac{d}{dt} w_1 = (1 - \Delta) w_2. \tag{8}$$

However, (8) is not controllable and hence not a faithful representation of (6). This shows that the d-d system (6) *cannot*, in fact, be represented as a kernel representation (1). In particular, this implies that what we call the latent variable elimination theorem [12] does not hold for d-d systems!

4. Proofs. We will show Theorems 3.1 and 3.2 in three main steps, respectively, corresponding to Propositions 4.1, 4.5, and 4.6 below. In the first step we prove that the rank constancy of $R(\lambda, e^{-\lambda})$ implies that (1) has an image representation. In the second step we prove that the existence of an image representation implies controllability. Finally, in the third step we show that if (1) defines a controllable system then $R(\lambda, e^{-\lambda})$ must have constant rank over $\mathbb{C}$. For a question of simplicity in the notation, in this section we will write $D = d/dt$ for the differentiator.

PROPOSITION 4.1. *With the previous notation, if* $\operatorname{rank} R(\lambda, e^{-\lambda}) = r \ \forall \lambda \in \mathbb{C}$ *then there exists a* $2D$ *polynomial matrix* $M(z_1, z_2)$ *such that* $\mathcal{B} := \ker R(D, \Delta) = \operatorname{im} M(D, \Delta)$*, with the operator* $M(D, \Delta)$ *acting on* $\mathcal{C}^\infty(\mathbb{R}, \mathbb{R}^l)$ *for a certain integer* l*.*

Proof. Under the hypothesis, the 2D polynomial matrix $R(z_1, z_2)$ has rank r (over the field of fractions $\mathbb{R}(z_1, z_2)$). Suppose first that $R(z_1, z_2)$ has $q = r$ columns. Then $R(\lambda, e^{-\lambda})$ has full column rank $\forall \lambda \in \mathbb{C}$, and hence $\ker R(D, \Delta)$ does not contain any element with components of the form $t^k e^{\lambda t}$. By [10, Theorem 5] this implies that $\ker R(D, \Delta) = \{0\}$, and the equality $\ker R(D, \Delta) = \operatorname{im} M(D, \Delta)$ is trivially satisfied with $M(z_1, z_2)$ being the $q \times 1$ zero matrix. Suppose now that $R(z_1, z_2)$ has $q > r$ columns. Then Lemma 4.2 follows.

LEMMA 4.2. $R(z_1, z_2)$ *can be factored as* $F(z_1, z_2)\bar{R}(z_1, z_2)$*, where* F *and* $\bar{R}$ *are* $2D$ *polynomial matrices of sizes* $g \times r$ *and* $r \times q$*, respectively, such that* $F(z_1, z_2)$ *has full column rank (over* $\mathbb{R}(z_1, z_2)$*) and* $\mathbb{R}(z_1, z_2)$ *is left prime (i.e.,* $\bar{R}$ *has full row rank and all its left factors are invertible in* $\mathbb{R}^{r \times r}[z_1, z_2]$*).*

Proof. Without loss of generality we may assume that $R(z_1, z_2) = [-Q(z_1, z_2) \mid P(z_1, z_2)]$, where $P(z_1, z_2)$ is a full rank matrix with r columns. Moreover, there exists a rational matrix $G(z_1, z_2)$ such that $PG = Q$. Let $G = \hat{Q}\hat{P}^{-1}$ and $G = \bar{P}^{-1}\bar{Q}$ be, respectively, a right coprime and a left coprime factorization of G [4]. Then the matrix $\bar{R} = [-\bar{Q} \mid \bar{P}]$ is a minimal left annihilator of $H := col(\hat{P}, \hat{Q})$ (cf. [7]); i.e., $\bar{R}H = 0$ and for every 2D polynomial matrix $S(z_1, z_2)$ such that $SH = 0$ there exists a 2D polynomial matrix $L(z_1, z_2)$ satisfying $S = L\bar{R}$. Since, obviously, also $RH = 0$, there exists a polynomial matrix $F(z_1, z_2)$ such that $R = F\bar{R}$. Further, since $\bar{R}$ is a full rank polynomial matrix with r rows, F must have column rank. □

Let then F and $\bar{R}$ be as in the previous lemma. Note that due to the fact that $\text{rank}R(\lambda, e^{-\lambda}) = r \ \forall \lambda \in \mathbb{C}$ neither $F(z_1, z_2)$ nor $\bar{R}(z_1, z_2)$ can have zeros of the form $(z_1, z_2) = (\lambda, e^{-\lambda})$. Now, since $\bar{R}$ is left prime, there exists a polynomial matrix W such that

$$\bar{R}(z_1, z_2)W(z_1, z_2) = N(z_1),$$

with $N(z_1) = \text{diag}(d(z_1), \ldots, d(z_1))$ for a suitable (nonzero) 1D polynomial $d(z_1)$. Let $M(z_1, z_2)$ be a right-prime 2D polynomial matrix such that $\bar{R}M = 0$ (we can take $M = H$ as in Lemma 4.2) and define the matrix $U(z_1, z_2) := [W(z_1, z_2) \mid M(z_1, z_2)]$.

LEMMA 4.3. *The operator $U(D, \Delta) : C^\infty(\mathbb{R}, \mathbb{R}^q) \to C^\infty(\mathbb{R}, \mathbb{R}^q)$ is surjective.*

Proof. We start by showing that $\det U = \det N = d^r(z_1) =: n(z_1)$. Without loss of generality we may assume that $\bar{R}(z_1, z_2)$ can be partitioned as $\bar{R}(z_1, z_2) = [P(z_1, z_2) \mid -Q(z_1, z_2)]$, with $P(z_1, z_2)$ square and nonsingular. Consider the corresponding partitions $\left[\begin{smallmatrix} X \\ Y \end{smallmatrix}\right] =: W$ and $\left[\begin{smallmatrix} \bar{Q} \\ \bar{P} \end{smallmatrix}\right] =: M$ of W and M. It is well known (see [4]) that $\det \bar{P}(z_1, z_2) = \det P(z_1, z_2)$. Now,

$$\det U = \det \begin{bmatrix} X & \bar{Q} \\ Y & \bar{P} \end{bmatrix} = \det \left(\begin{bmatrix} X & \bar{Q} \\ Y & \bar{P} \end{bmatrix} \begin{bmatrix} I & 0 \\ -\bar{P}^{-1}Y & I \end{bmatrix} \right)$$

$$= \det \begin{bmatrix} X - \bar{Q}\bar{P}^{-1}Y & \bar{Q} \\ 0 & \bar{P} \end{bmatrix} = \det \bar{P} \cdot \det(X - \bar{Q}\bar{P}^{-1}Y)$$

$$= \det \bar{P} \cdot \det(X - P^{-1}QY),$$

since $\bar{Q}\bar{P}^{-1} = P^{-1}Q$ (due to the fact that M is a dual basis of $\bar{R}$). Thus,

$$\begin{aligned} \det U &= \det \bar{P} \cdot \det(P^{-1}(PX - QY) \\ &= \det \bar{P}. \det P^{-1} \cdot \det(PX - QY) \\ &= \det \bar{P}.(\det P)^{-1} \cdot \det(PX - QY) \\ &= \det(PX - QY), \end{aligned}$$

and as $N = PX - QY$, we conclude that

$$\det U = \det N = \det(\text{diag}(d(z_1), \ldots, d(z_1))) = d^r(z_1) =: n(z_1).$$

Consider now the equation

$$U(D, \Delta)\alpha = \beta.$$

Given $\beta \in \mathcal{C}^\infty(\mathbb{R}, \mathbb{R}^q)$, define $\bar{\alpha}$ such that

$$\bar{N}(D)\bar{\alpha} = \beta,$$

with $\bar{N}(z_1) := \mathrm{diag}(n(z_1), \ldots, n(z_1))$. Note that $\bar{N}(D)$ is a surjective operator in $\mathcal{C}^\infty(\mathbb{R}, \mathbb{R}^q)$. Define $\alpha := \bar{V}(D, \Delta)\bar{\alpha} \in \mathcal{C}^\infty(\mathbb{R}, \mathbb{R}^q)$, where $\bar{V}(z_1, z_2)$ is such that

$$U(z_1, z_2)\bar{V}(z_1, z_2) = \bar{N}(z_1).$$

Then

$$U(D, \Delta)\alpha = U(D, \Delta)\bar{V}(D, \Delta)\bar{\alpha} = \bar{N}(D)\bar{\alpha} = \beta,$$

showing that $U(D, \Delta)$ is a surjective operator in $\mathcal{C}^\infty(\mathbb{R}, \mathbb{R}^q)$. □

This implies that $\forall w \in \mathcal{B} = \ker R(D, \Delta)$ there exists $\bar{w}$ such that $w = U(D, \Delta)\bar{w}$ and hence $R(D, \Delta)U(D, \Delta)\bar{w} = 0$, i.e., $\bar{w} \in \ker R(D, \Delta)U(D, \Delta)$. So,

$$\mathcal{B} \subseteq U(D, \Delta) \ker(R(D, \Delta)U(D, \Delta)).$$

On the other hand, if $w = U(D, \Delta)\bar{w}$ and $R(D, \Delta)U(D, \Delta)\bar{w} = 0$, then $R(D, \Delta)w = 0$, i.e.,

$$\mathcal{B} \supseteq U(D, \Delta) \ker(R(D, \Delta)U(D, \Delta)).$$

Therefore $\mathcal{B} = U(D, \Delta) \ker(R(D, \Delta)U(D, \Delta))$. Taking into account that $U = [W \mid M]$ and that $RU = [FN \mid 0]$, this yields $\mathcal{B} = [W(D, \Delta) \mid M(D, \Delta)](\ker[F(D, \Delta)N(D, \Delta) \mid 0])$. Thus

$$\mathcal{B} = W(D, \Delta) \ker(F(D, \Delta)N(D, \Delta)) + \mathrm{im}M(D, \Delta).$$

Finally, it turns out that Lemma 4.4 follows.

LEMMA 4.4. $W(\ker FN) \subseteq \mathrm{im}M$.

Proof. Recall that $F(\lambda, e^{-\lambda})$ has full column rank $\forall \lambda \in \mathbb{C}$. This implies that $\ker F(D, \Delta) = \{0\}$, and hence $\ker FN = \ker N$. Therefore, in order to prove the lemma we will show that $W(\ker N) \subseteq \mathrm{im}M$. As is well known,

$$\ker N(D) = \mathrm{span}\{t^j e^{\lambda_i t} e_k : i = 1, \ldots, p, j = 0, \ldots, \mu(\lambda_i) - 1, k = 1, \ldots, r\},$$

where $\lambda_1, \ldots, \lambda_p$ are the distinct roots of $d(z_1)$, $\mu(\lambda_i)(i = 1, \ldots, p)$ are the corresponding multiplicities, and e_k is the kth vector in the canonical basis of $\mathbb{R}^r$. So, $W(\ker N) \subseteq \mathrm{im}M$ iff for every root λ of $d(z_1)$, for every m subject to $0 \leq m \leq \mu(\lambda) - 1$, and for every $k \in \{1, \ldots, r\}$, $W(t^m e^{\lambda t} e_k) \in \mathrm{im}M$; i.e., there is a $\mathcal{C}^\infty$ trajectory x such that

$$Mx(t) = W(t^m e^{\lambda t} e_k). \tag{9}$$

Let then λ be a root of $d(z_1)$ and let m be a positive integer not greater than $\mu(\lambda) - 1$. Without loss of generality we may assume that $\bar{R} = [P \mid -Q]$ with $P(\lambda, e^{-\lambda})$ invertible. Consider the corresponding partitions of M and W as defined in the proof of Lemma 4.3. Note that in this case, $\bar{P}(\lambda, e^{-\lambda})$ is also invertible. Now, (9) can be rewritten as

$$\bar{Q}\, x = X(t^m e^{\lambda t} e_k), \tag{10}$$

$$\bar{P}\, x = Y(t^m e^{\lambda t} e_k). \tag{11}$$

It is not difficult to see that

$$Y(D, \Delta)(t^m e^{\lambda t} e_k) \;=\; Y(\lambda, e^{-\lambda}) e_k t^m e^{\lambda t} + (Y_{m-1} t^{m-1} + \cdots + Y_0) e_k e^{\lambda t}$$

for some suitable matrices $Y_{m-1},\dots,Y_0$. Take x to be of the form $x(t)=(\xi_m t^m+\cdots+\xi_0)e^{\lambda t}$. Then

$$\begin{aligned}(\bar{P}(D,\Delta)x)(t)=\{&\bar{P}(\lambda,e^{-\lambda})\xi_m t^m+[\bar{P}(\lambda,e^{-\lambda})\xi_{m-1}+G_m^{m-1}\xi_m]t^{m-1}\\&+\cdots+[\bar{P}(\lambda,e^{-\lambda})\xi_0+G_1^0\xi_1+\cdots+G_m^0\xi_m]\}e^{\lambda t},\end{aligned}$$

and x satisfies (11) iff

$$\text{(12)}\qquad\begin{cases}\bar{P}(\lambda,e^{-\lambda})\xi_m &= Y(\lambda,e^{-\lambda})e_k,\\ \bar{P}(\lambda,e^{-\lambda})\xi_{m-1}&=Y_{m-1}e_k-G_m^{m-1},\\ \vdots&\\ \bar{P}(\lambda,e^{-\lambda})\xi_0 &= Y_0e_k-(G_1^0\xi_1+\cdots+G_1^0\xi_1).\end{cases}$$

As $\bar{P}(\lambda,e^{-\lambda})$ is invertible, there is a (unique) solution $(\xi_m,\dots,\xi_0)$ to (12), showing that (11) has a $\mathcal{C}^\infty$ solution $x(t)=(\xi_m t^m+\cdots+\xi_0)e^{\lambda t}$. It remains to prove that this solution $x(t)$ also satisfies (10). It follows from (11) that

$$\begin{aligned}Q\bar{P}x&=QY(t^me^{\lambda t}e_k)\ \Leftrightarrow\\ P\bar{Q}x&=QY(t^me^{\lambda t}e_k)\ \Leftrightarrow\\ P\bar{Q}x&=(PX-N)(t^me^{\lambda t}e_k)\ \Leftrightarrow\\ 0&=P(\bar{Q}x-Xt^me^{\lambda t}e_k),\end{aligned}\qquad\text{(13)}$$

since $Q\bar{P}=P\bar{Q}$, $PX+QY=N$, and $t^me^{\lambda t}e_k\in\ker N$. Note that $\bar{Q}(D,\Delta)x-X(D,\Delta)t^me^{\lambda t}e_k=E(t)e^{\lambda t}$, where $E(t)$ is a polynomial column in t containing powers of t of order not greater than m. Assume that $E(t)=E_{\bar{m}}t^{\bar{m}}+\cdots+E_0$, where $E_{\bar{m}}$ is a nonzero column and $\bar{m}\le m$. Then (13) becomes

$$[P(\lambda,e^{-\lambda})E_{\bar{m}}t^{\bar{m}}+(G_{\bar{m}-1}t^{m-1}+\cdots+G_0)]e^{\lambda t}=0,$$

which implies that $P(\lambda,e^{-\lambda})E_{\bar{m}}=0$. This is absurd, since $P(\lambda,e^{-\lambda})$ is an invertible matrix and $E_{\bar{m}}$ is assumed to be nonzero. Thus, $E(t)$ must be zero, i.e.,

$$\bar{Q}x-Xt^me^{\lambda t}e_k=0,$$

which shows that x satisfies equation (10) and hence also (9). □

As a consequence of this lemma we have that $\mathcal{B}=\mathrm{im}M(D,\Delta)$, i.e., (1) has an image representation, proving the proposition. □

Now, it is not difficult to come to the following conclusion.

PROPOSITION 4.5. *If* (1) *has an image representation, then it defines a controllable system.*

Proof. Suppose that (1) has an image representation, i.e., $\mathcal{B}=\mathrm{im}M(D,\Delta)$, and let w_1 and w_2 be two arbitrary signals in $\mathcal{B}$. Then, there exist a_1 and a_2 in $\mathcal{C}^\infty(\mathbb{R},\mathbb{R}^l)$ such that $w_i=Ma_i,(i=1,2)$. Now, it is possible to construct a smooth signal a^* which coincides with a_1 in the past and with a_2 in the (sufficiently far) future. Such signal yields an element $w^*=Ma^*$ in $\mathcal{B}$ which coincides with w_1 in the past and with w_2 in the future. Thus w_1 is $\mathcal{B}$-compatible, with w_2 showing that $\mathcal{B}$ is controllable. □

Finally, if $R(z_1,z_2)$ is a 2D polynomial matrix of rank r and $\mathrm{rank}R(\lambda,e^{-\lambda})<r$ for some $\lambda_0\in\mathbb{C}$ we can show that there exists a signal associated with the frequency λ_0 which is not $\mathcal{B}$-compatible with the identical zero signal and hence $\mathcal{B}$ is not controllable.

PROPOSITION 4.6. *Let $\mathcal{B} := \ker R(D, \Delta)$, where $R(z_1, z_2)$ is a 2D polynomial matrix of rank r. If $\mathcal{B}$ is controllable then* $\text{rank}R(\lambda, e^{-\lambda}) = r \ \forall \lambda \in \mathbb{C}$.

Proof. We start by noting that, formally, $e^{-z_1} = \sum_{k=0}^{+\infty} \frac{(-1)^k}{k!} z_1^k$. Thus, if $z_2 = e^{-z_1}$, $R(z_1, z_2) = \tilde{\Pi}(z_1)$, where $\tilde{\Pi}(z_1)$ is a matrix over the ring $\mathbb{R}[[z_1]]$ of formal power series in z_1. Suppose now that $\text{rank}R(\lambda, e^{-\lambda}) < \text{rank}R(z_1, z_2) = \text{rank}R(z_1, e^{-z_1}) = r$ for a certain $\lambda_0 \in \mathbb{C}$. We consider first the case where $\lambda_0 = 0$; so $\text{rank}R(0, 1) < \text{rank}R(z_1, e^{-z_1})$. This means that $\text{rank}\tilde{\Pi}(0) < \text{rank}\tilde{\Pi}(z_1) = r$, and therefore we may assume without loss of generality that

$$\tilde{\Pi}(z_1) = \text{diag}(z_1^{k_1}, \ldots, z_1^{k_r}, z_1^{k_{r+1}}, \ldots, z_1^{k_s})\tilde{\Gamma}(z_1),$$

where $k_1, \ldots, k_s$ are integers, $k_1 \geq 1$, and

$$\tilde{\Gamma}(0) = \begin{bmatrix} I_{r\times r} & 0 \\ 0 & 0 \end{bmatrix}$$

(with the zero rows possibly void).

Let

$$\Gamma(z_1) = \begin{bmatrix} I_{r\times r} & 0 \\ 0 & 0 \end{bmatrix} + z_1\Gamma_1 + z_1^2\Gamma_2 + \cdots + z_1^{k_1-1}\Gamma_{k_1-1}$$

be such that

$$\tilde{\Gamma}(z_1) = \Gamma(z_1) + \text{higher-order terms.}$$

Then, using the same kind of arguments as in the proof of Lemma 4.4, it is possible to show that there exists a trajectory $w^*(t) = \alpha_{k_1-1}t^{k_1-1} + \cdots + \alpha_0$ such that

$$\Gamma(D)w^*(t) = \begin{bmatrix} 1 \\ 0 \\ \vdots \\ 0 \end{bmatrix} \frac{t^{k_1-1}}{(k_1-1)!}.$$

Now, this trajectory w^* is clearly such that

$$R(D, \Delta)w^* = \text{diag}(D^{k_1}, \ldots, D^{k_s})\Gamma(D)w = D^{k_1}t^{k_1-1} = 0,$$

and hence it belongs to $\mathcal{B}$.

LEMMA 4.7. *With the previous notation, w^* is not $\mathcal{B}$-compatible with the zero trajectory.*

Proof. Suppose that w^* is $\mathcal{B}$-compatible with the zero trajectory, yielding a trajectory $v^* \in \mathcal{B}$ such that $v^*|_{(-\infty,\tau_1)} = w^*|_{(-\infty,\tau_1)}$ and $v^*|_{[\tau_2,+\infty)} = 0|_{[\tau_2,+\infty)}$ for some $\tau_1 < \tau_2$. Since $v^* \in \mathcal{B}$, $R(D, \Delta)v^* = 0$ and therefore also

$$\int_{T_1}^{T_2+1} [R(D, \Delta)v^*]dt = 0.$$

In particular, if $r(z_1, z_2)$ denotes the first row of $R(z_1, z_2)$, we have that

$$\int_{T_1}^{T_2+1} [r(D, \Delta)v^*]dt = 0. \tag{14}$$

Note further that $r(\lambda, e^{-\lambda})$ becomes zero for $\lambda = 0$, and hence $r(z_1, z_2) = r(z_1, e^{-z_1})$ must be of the form

$$r(z_1, z_2) = r(z_1, e^{-z_1}) = z_1 r_0(z_1) + (e^{-z_1} - 1) r_1(z_1, e^{-z_1}),$$

where $r_0(z_1)$ and $r_1(z_1, z_2)$, respectively, are 1D and 2D polynomial rows. So equation (14) is of the form

$$\int_{T_1}^{T_2+1} [(\Delta - 1) r_1(D, \Delta) + D r_0(D)] v^* dt = 0.$$

This is equivalent to

$$\int_{T_1-1}^{T_1} r_1(D, \Delta) v^* dt + \int_{T_2}^{T_2+1} r_1(D, \Delta) v^* dt + [r_0(D) v^*] \Big|_{T_1}^{T_2+1} = 0,$$

which is still equivalent to

$$\int_{T_1-1}^{T_1} r_1(D, \Delta) w^* dt + \int_{T_2}^{T_2+1} r_1(D, \Delta) 0 dt + (r_0(D)0)(T_2 + 1) - (r_0(D) w^*)(T_1) = 0$$

if $T_2 + 1 \gg \tau_2$ and $T_1 \ll \tau_1$ so that in a sufficiently big interval around $T_2 + 1$, v^* coincides with the zero trajectory, and in a sufficiently big interval around T_1, it coincides with w^*. This yields

$$\int_{T_1-1}^{T_1} r_1(D, \Delta) w^* dt - (r_0(D) w^*)(T_1) = 0. \tag{15}$$

Let $\eta = col(\eta_1, \ldots, \eta_q)$ be a trajectory such that $\eta(t) = \frac{\alpha_{k_1-1}}{k_1} t^{k_1} + \cdots + \alpha_0 t$; then $D\eta = w^*$ and we may write (15) as

$$\int_{T_1-1}^{T_1} r_1(D, \Delta) D\eta \, dt - (r_0(D) D\eta)(T_1) = 0,$$

which is equivalent to having

$$[((\Delta - 1) r_1(D, \Delta) + r_0(D) D)\eta](T_1) = 0$$

or still

$$(r(D, \Delta)\eta)(T_1) = 0.$$

Now, it follows from our previous considerations that

$$r(D, \Delta)\eta)(T_1) = (D^{k_1}[1 \;\; 0 \;\; \ldots \;\; 0]\eta)(T_1) = [1 \;\; 0 \;\; \ldots \;\; 0]\alpha_{k_1-1}(k_1 - 1)! = 1,$$

since $[1 \quad 0 \quad \ldots \quad 0]\Gamma(D) w^* = \frac{t^{k_1-1}}{(k_1-1)!}$. In this way we obtain that $0 = 1$, which is absurd. Consequently the hypothesis that w^* is $\mathcal{B}$-concatenable with the zero trajectory cannot hold true. □

It follows from this result that if $R(\lambda, e^{-\lambda})$ drops in rank for $\lambda = 0$, then $\mathcal{B}$ is not controllable. It remains to show that if $\text{rank} R(\lambda, e^{-\lambda}) < \text{rank} R(z_1, z_2)$ for $\lambda = \lambda^* \neq 0$ then $\mathcal{B}$ is not controllable.

Assume now that $R(\lambda, e^{-\lambda})$ drops in rank for $\lambda = \lambda^* \neq 0$, and consider the system Σ^* with behavior $\mathcal{B}^* := \exp_{\lambda^*} \mathcal{B}$, where $\exp_{\lambda^*}$ is defined by $\exp_{\lambda^*}(t) = e^{-\lambda^* t}\ \forall t \in \mathbb{R}$. Then $\mathcal{B}^*$ is described by a polynomial matrix $R^*(z_1, z_2)$ such that $R^*(\lambda, e^{-\lambda}) = R(\lambda + \lambda^*, e^{-(\lambda+\lambda^*)})$. As $\text{rank} R(\lambda, e^{-\lambda})$ drops for $\lambda = \lambda^*$, $R^*(\lambda, e^{-\lambda})$ drops for $\lambda = 0$. Thus, by the foregoing arguments, $\mathcal{B}^*$ is not controllable. This implies that $\mathcal{B}$ is also not controllable, completing the proof of the proposition. □

5. Conclusion. We have presented a necessary and sufficient condition for the controllability of the behavior of d-d systems with kernel representations. Moreover, we have compared the notion of behavioral controllability with the notions of approximate and spectral controllability considered in [3] as well as with other controllability properties (namely, weak and strict controllability) that have been introduced within an algebraic approach to d-d systems [5]. Contrary to what happens with the results of [3], [6], [9], and [5], our results hold for all types of systems with commensurable delays and not only for retarded or neutral systems in pseudostate form.

REFERENCES

[1] H. Glüsing-Lüerssen, *A Behavioral Approach to Delay-Differential Systems*, University of Oldenburg, Oldenburg, Germany, 1995, preprint.

[2] B. C. Lévy, 2-*D Polynomial and Rational Matrices, and Their Applications for the Modelling of* 2-*D Dynamical Systems*, Ph.D. thesis, Stanford University, Stanford, CA, 1981.

[3] A. Manitius, *Necessary and sufficient conditions of approximate controllability for general linear retarded systems*, SIAM J. Control Optim., 19 (1981), pp. 516–532.

[4] M. Morf, B. C. Lévy, and S. Y. Kung, *New results in* 2-*D systems theory, part* I*:* 2-*D polynomial matrices, factorizations and coprimeness*, Proc. IEEE, 65 (1977), pp. 861–872.

[5] A. S. Morse, *Ring models for delay-differential systems*, Automatica, 12 (1976), pp. 529–531.

[6] D. A. O'Connor and T. J. Tarn, *On the function space controllability of linear neutral systems*, SIAM J. Control Optim., 21 (1983), pp. 306–329.

[7] P. Rocha, *Structure and Representation of* 2*D Systems*, Ph.D. thesis, University of Groningen, Groningen, the Netherlands, 1990.

[8] P. Rocha and J. C. Willems, *Controllability of* 2-*D Systems*, IEEE Trans. Automat. Control, 36 (1991), pp. 413–423.

[9] D. Salamon, *Control and Observation of Neutral Systems*, Pitman, Boston, MA, 1984.

[10] L. Schwartz, *Théorie générale des fonctions moyenne-périodiques*, Ann. of Math., 48 (1947), pp. 857–929.

[11] E. D. Sontag and Y. Yamamoto, *On the existence of coprime factorizations for retarded systems*, Systems Control Lett., 13 (1989), pp. 38–53.

[12] J. C. Willems, *Models for dynamics*, Dynam. Report., 2 (1988), pp. 171–269.

[13] Y. Yamamoto, *Reachability of a class of infinite-dimensional linear systems: An external approach with applications to general neutral systems*, SIAM J. Control Optim., 27 (1989), pp. 217–234.

SIAM J. CONTROL OPTIM.
Vol. 35, No. 1, pp. 265–272, January 1997

015

A FARKAS LEMMA WITHOUT A STANDARD CLOSURE CONDITION*

JEAN B. LASSERRE†

Abstract. We present a setting in which we derive a new Farkas lemma for the system $\{Ax = b,\ x \in S\}$ without the standard closure condition on $A(S)$. Further characterization in Hilbert spaces is also presented.

Key words. Farkas lemma, duality, convex cones, Banach spaces, locally convex topological vector spaces, Hilbert spaces

AMS subject classifications. 90C05, 90C08, 90C48, 49N15, 49R20

PII. S0363012994268321

1. Introduction.

The celebrated Farkas lemma is a fundamental tool in optimization to characterize existence of solutions for a system $\{Ax = b,\ x \in S\}$, where A is a linear mapping and S a positive convex cone. It is also the basis of many duality results such as first-order optimality conditions. There have been a great number of papers devoted to various *asymptotic* and *nonasymptotic*, linear and nonlinear versions of Farkas' lemma. The asymptotic version is a weaker notion that replaces feasibility with versions of asymptotic feasibility. For instance, the interested reader is referred to [10] for various Farkas lemmas in arbitrary dual pairs of vector spaces, using weak topologies [4], [3], [12], [13], [14], [16], [18], [19] for linear and nonlinear versions with various applications; the discussion in [11]; and all the references in the above papers as well as the references in [1].

In this paper we are concerned with the linear nonasymptotic version. In this case, Farkas' lemma holds only under a crucial *closure* assumption in some appropriate topology. The closure assumption is concerned with $A(S)$, the image of the positive cone S, and holds for example when the cone S is a *polyhedral* cone in some finite-dimensional space. In general, this closure assumption is very restrictive, and some specific properties of A and S must be invoked. (For example, sufficient conditions are given in [15].) The reader is referred to [3] for simple examples in finite-dimensional spaces where this closure assumption does not hold and only an *asymptotic* version of Farkas' lemma holds.

The purpose of this paper is to provide a setting in which, by introducing another appropriate convex cone (indexed by a scalar parameter and some vector), one may derive a (nonasymptotic) Farkas lemma without this (strong) closure assumption. The underlying idea is that if a solution exists, it must be in this cone for some sufficiently large value of the parameter and also must satisfy some linear constraint related to the vector used in the definition of the new cone. Doing so permits us to use a weak* sequential compactness argument and yields the desired result. The Farkas lemma is stated in terms of the dual of the cone introduced. A more precise characterization is given in Hilbert spaces. This new Farkas lemma is also illustrated for linear systems of matrix equalities involving the cone of positive semidefinite matrices.

*Received by the editors February 7, 1995; accepted for publication (in revised form) November 24, 1995.

http://www.siam.org/journals/sicon/35-1/26832.html

†LAAS-CNRS, 7 Avenue du Colonel Roche, 31 077 Toulouse cédex, France (lasserre@laas.fr).

Finally, thanks to an anonymous referee, we also give a different Farkas lemma that reveals what lies behind the previous one, namely, the existence of an appropriate, compactly generated, convex cone.

2. Notation and definitions. We use notation similar to [10]. Let X and Y be two real vector spaces with a bilinear form $\langle , \rangle$ defined on $X \times Y$. We also assume that

(i) for each $x \neq 0$ in X, $\exists y \in Y$ with $\langle x, y\rangle \neq 0$,

(ii) for each $y \neq 0$ in Y, $\exists x \in X$ with $\langle x, y\rangle \neq 0$,

so that (X, Y) is a *dual* pair. We also equip X with the *weak topology* $\sigma(X, Y)$ so that all the elements of Y are continuous when regarded as linear forms on X. We are also given a convex cone $S \subset X$. Similarly, Y is equipped with the $\sigma(Y, X)$ weak topology and $S^+ \subset Y$ denotes the anticone of S, i.e.,

$$S^+ := \{y \in Y |\ \langle y, x\rangle \geq 0 \ \ \forall x \in S\}.$$

When $Y \equiv X^*$, where X^* is the topological dual of X, S^+ is called the dual cone of S and is denoted by S^*.

Let (Z, W) be some other real dual pair where Z (resp., W) is also equipped with the $\sigma(Z, W)$ (resp., $\sigma(W, Z)$) weak topology. Let A be a linear mapping $A : X \to Z$. A necessary and sufficient condition for A to be weakly continuous is $A^*(W) \subset Y$. The restriction A^+ of A^* to W is then weakly continuous and is called the *adjoint* of A with respect to the dual pairs (X, Y) and (Z, W) (see [10]).

Weak convergence of x_n to x will be denoted by $x_n \overset{w}{\to} x$. Note that with the weak topology, all the vector spaces are *locally convex* [17].

We are interested in existence of solutions to the system

$$Ax = b, \quad x \in S,$$

where b is some vector in Z. We first briefly recall a standard Farkas lemma [10] that addresses this issue.

THEOREM 2.1 (see [10, p. 987]). *Let (X, Y), (Z, W) be real dual pairs, S be a convex cone in X, and $A : X \to Z$ be a weakly continuous linear mapping. If $A(S)$ is weakly closed, then the following are equivalent:*

(a) *The system $Ax = b$ has a solution $x \in S$.*

(b) $A^+w \in S^+ \Rightarrow \langle b, w\rangle \geq 0$.

Conversely, the equivalence of (a) *and* (b) *implies that $A(S)$ is weakly closed.*

As may be noted from the last assertion, the condition "$A(S)$ weakly closed" is crucial. Indeed, without such a condition, only *asymptotic existence* can sometimes be asserted when (b) holds (see [3]). However, in many practical examples, the closure assumption on $A(S)$ is not satisfied.

The purpose of this paper is to provide a setting in which a Farkas lemma holds with no closure assumption on $A(S)$.

3. The main result. In the following discussion, the spaces X, Y, Z, and W are all assumed to be Banach spaces. In addition, we make the following assumption.

Assumption H.

(i) $(X, \| \cdot \|)$ is the topological dual of a separable Banach space Y.

(ii) $\exists y_0 \in S^+$ such that $x \in S$, $\langle y_0, x\rangle = 0, \ \Rightarrow x = 0$.

Remark. H (ii) obviously implies that the cone S is pointed, i.e., $S \cap -S = \{0\}$. Actually, H (ii) merely asserts that the cone S admits a *positive* linear functional $y_0 \in S^+$ (which implies that S has a base). When $Y = X^*$, in which case $S^+ = S^*$,

the dual cone of S, then it is equivalent to assuming that S has a *base* (see, e.g., [5]). In some cases (e.g., when X is a normed separable space and $Y = X^*$), pointedness and existence of a positive functional are equivalent for closed convex cones.

Examples where H holds:

• X is the vector space of (n, n) real symmetric matrices with y_0 the identity matrix, S the cone of positive semidefinite matrices, and $\langle x, y\rangle := \text{trace}(x.y)$.

• Let $(K, \mathcal{B}, \mu)$ be a σ-finite measure space. $X := L^p(K, \mathcal{B}, \mu)$ with $1 < p \leq \infty$. S is the usual positive cone in X. y_0 is any strictly positive function in $Y := L^q(K, \mathcal{B}, \mu)$ with $1/p + 1/q = 1$ and $q = 1$ if $p = \infty$.

• $X := \mathcal{M}(K)$, the Banach space of bounded Borel signed measures on K, with K a locally compact separable metric space. S is again the usual positive cone in X, and $Y := C_0(K)$ the space of continuous functions that vanish at infinity. y_0 is any strictly positive function in $C_0(K)$.

• $X := l^p,\ 1 < p \leq \infty$, with S the usual positive cone in X and y_0 any strictly positive sequence in $Y := l^q$, with $1/p + 1/q = 1$.

• $X := l^1$ with S the usual positive cone and $Y := c_0$ the space of sequences that vanish at infinity. y_0 is any strictly positive sequence in c_0.

For $\lambda > 0$, let $S_{\lambda y_0}$ be the set $\{x \in S|\ \parallel x \parallel \leq \lambda\langle y_0, x\rangle\}$. Obviously, $S_{\lambda y_0}$ is a convex cone. We then have the following result.

THEOREM 3.1. *Let (X, Y), (Z, W) be real dual pairs, S be a weakly closed convex cone in X, and $A: X \to Z$ be a weakly continuous linear mapping. Assume also that* H *holds for some given $y_0 \in Y$. Then the following are equivalent:*

(i) *The system $Ax = b$ has a solution $x \in S$*

(ii) *There exists some positive (λ, δ) such that*

$$A^+w + \gamma y_0 \in S^+_{\lambda y_0}, \qquad \gamma \geq 0 \Rightarrow \langle b, w\rangle + \gamma\delta \geq 0.$$

(iii) *There exists some positive (λ, δ) such that*

$$A^+w \in S^+_{\lambda y_0} \Rightarrow \langle b, w\rangle \geq 0 \quad \textit{and } A^+w + y_0 \in S^+_{\lambda y_0} \Rightarrow \langle b, w\rangle \geq -\delta.$$

Proof. Consider the linear mapping

$$T:\ X \times R \to Z \times R,\ T(x, z) := \begin{bmatrix} Ax \\ \langle y_0, x\rangle + z \end{bmatrix}.$$

Its adjoint $T^+:\ W \times R \to Y \times R$ is given by $T^+(w, \gamma) := (A^+w + \gamma y_0, \gamma)$. Let R^+ denote the nonnegative real numbers. We first prove that $T(S_{\lambda y_0} \times R^+)$ is weakly closed.

For some directed set $\{D,\ \succeq\}$, let $\{(x_\alpha, z_\alpha)\}$ (where $\alpha \in D$) be a net in $S_{\lambda y_0} \times R^+$ such that w-lim $T(x_\alpha, z_\alpha) = (a_1, a_2)$, where w-lim denotes the $(\sigma(Z \times R, W \times R))$ weak limit. In particular, lim $\langle y_0, x_\alpha\rangle + z_\alpha = a_2$ and $a_2 \geq 0$.

If $a_2 = 0$, then lim $\langle y_0, x_\alpha\rangle = 0$ and lim $z_\alpha = 0$ so that since $x_\alpha \in S_{\lambda y_0}$, lim $||x_\alpha|| = 0$, which in turn implies w-lim $x_\alpha = 0$, where the latter w-lim denotes the $\sigma(X, Y)$ weak limit. Thus by continuity of A, we must have $a_1 = 0$. Hence, $T(0, 0) = (0, 0) = (a_1, a_2)$.

If $a_2 \neq 0$, then for some $\alpha_0 \in D$ and all $\alpha \succeq \alpha_0$, $z_\alpha \leq 2a_2$, and $\langle y_0, x_\alpha\rangle \leq 2a_2$, which in turn implies $\parallel x_\alpha \parallel \leq 2\lambda a_2$. Let $\xi_\alpha := \frac{x_\alpha}{2\lambda a_2}$. Then, $\xi_\alpha \in S \cap B(0, 1)$, where $B(0, 1)$ is the unit ball in $(X, \parallel . \parallel)$, compact in the *weak* topology*, which in our setting is precisely the $\sigma(X, Y)$ (weak) topology. (See Alaoglu's theorem in, e.g., [17].) From H (i), $B(0, 1)$ is also weakly sequentially compact (see [17]) so that from

the net $\{(x_\alpha, z_\alpha)\}$ we can extract a countable family and therefore a sequence $\{z_n\}$ that converges to some $z \geq 0$ and a sequence $\{\xi_n\}$ that converges weakly to some $\xi \in S$ ($\xi_n \stackrel{w}{\to} \xi$) since S is weakly closed.

The weak convergence yields

$$\langle y_0, \xi_n \rangle + \frac{z_n}{2\lambda a_2} \to \langle y_0, \xi \rangle + \frac{z}{2\lambda a_2} = \frac{1}{2\lambda} \quad \text{as} \quad n \uparrow \infty$$

so that $\langle y_0, 2\lambda a_2 \xi \rangle + z = a_2$. By weak continuity of A, $A\xi_n \stackrel{w}{\to} a_1/2\lambda a_2 = A\xi$ or equivalently $A(2\lambda a_2 \xi) = a_1$. Hence, $T(2\lambda a_2 \xi, z) = (a_1, a_2)$.

It remains to prove that $x := 2\lambda a_2 \xi \in S_{\lambda y_0}$. Observe that $x \in S$ since $\xi \in S$. Moreover, as $x_n \stackrel{w}{\to} x$ we have $\liminf_{n\uparrow\infty} \| x_n \| \geq \| x \|$ (see, e.g., [9]). Therefore, from $\| x_n \| \leq \lambda \langle y_0, x_n \rangle$ and $x_n \stackrel{w}{\to} x$ we conclude that $\| x \| \leq \lambda \langle y_0, x \rangle$, i.e., $x \in S_{\lambda y_0}$, which in turn implies $(x, z) \in S_{\lambda y_0} \times R^+$ so that, finally, $T(S_{\lambda y_0} \times R^+)$ is weakly closed.

Now the system $Ax = b$ has a solution in S if and only if the system $T(x, z) = (b, \delta)$ has a solution in $S_{\lambda y_0}$ for some λ sufficiently large and some $\delta > 0$. Indeed, by H (ii), given any $x \in S$, $x \in S_{\lambda y_0}$ for all λ sufficiently large.

Therefore, we now can apply the generalized Farkas lemma (cf. Theorem 2.1) to the latter system since $T(S_{\lambda y_0} \times R^+)$ is weakly closed. (Indeed, the weak continuity of T follows easily from the weak continuity of A.) Thus, $T(x, z)$ has a solution in $S_{\lambda y_0} \times R^+$ if and only if

$$T^+(w, \gamma) \in S^+_{\lambda y_0} \times R^+ \Rightarrow \langle b, w \rangle + \gamma\delta \geq 0$$

or equivalently

$$A^+ w + \gamma y_0 \in S^+_{\lambda y_0}, \qquad \gamma \geq 0 \Rightarrow \langle b, w \rangle + \gamma\delta \geq 0,$$

which is the desired result.

(ii) $\Leftrightarrow$ (iii) is trivial. □

We have stated (iii) to express the two cases $\gamma > 0$ and $\gamma = 0$. Indeed, the case $\gamma = 0$ covers the standard Farkas lemma so that the case $\gamma > 0$ (eliminated by dividing w by γ) describes the additional condition required in Farkas' lemma when $A(S)$ is not weakly closed.

Also note that if X is a reflexive Banach space, the weak topology $\sigma(X, Y)$ is also the standard weak topology so that the assumption S *weakly closed* can be replaced by S *closed* since strongly closed and weakly closed convex sets coincide in locally convex topological vector spaces.

Particular case. It may happen that when $x \in S$, $\langle y_0, x \rangle = \| x \|_1$ for some norm $\| \cdot \|_1$ and some appropriate y_0. If this norm is equivalent to $\| \cdot \|$, then $S \equiv S_{\lambda y_0}$ for all λ sufficiently large so that one may replace $S^+_{\lambda y_0}$ by S^+ in Theorem 3.1. An example is the space of (n, n) symmetric matrices with norm $||x||_1 := \sum_i |\sigma_i(x)|$, where the $\{\sigma_i(x)\}$ are the eigenvalues of x. If S is the cone of positive semidefinite matrices, then $\sigma_i(x) \geq 0 \ \forall i$ when $x \in S$ so that $||x||_1 = \langle I, x \rangle$, where I is the identity matrix.

Actually, the same conclusion holds if $\langle y_0, x \rangle \geq ||x||$ when $x \in S$, for some well-chosen y_0.

Similarly to [10] we also can derive a *dual* version of Theorem 3.1 for a linear system $\{Cy = b, \ y \in S^*\}$, where now C is a linear mapping X^* $(:= Y^{**}) :\to Z$ and $S^* \subset X^*$ is the dual cone of $S \subset X$. The $\sigma(X^*, X)$ topology on X^* is the weak* topology. Moreover, if S is a convex cone in X, then S^* is $(\sigma(X^*, X))$ weakly closed,

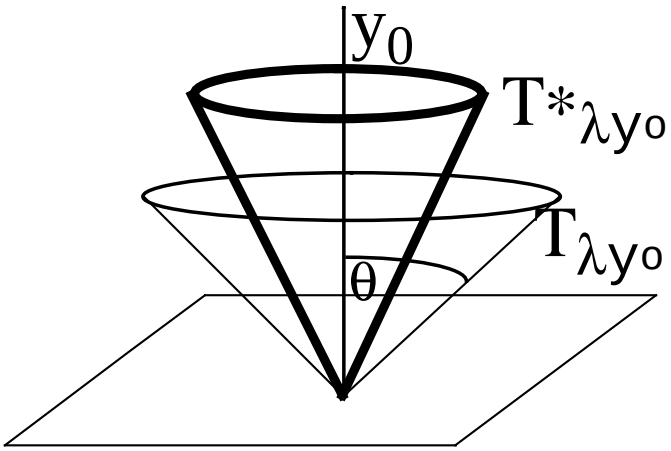


FIG. 1. *Cones* $T_{\lambda y_0}$ *and* $T^*_{\lambda y_0}$.

i.e., weak* closed. Hence it is also closed, and with the natural embedding of X in the bidual X^{**}, $(S^*)^+ = S^{**} \cap X = S$ (see [10]). We also define $S_{\lambda x_0} := \{y \in S^* |\ ||y|| \leq \lambda \langle y, x_0\rangle\}$ for some given $x_0 \in S$ such that $y \in S^*$ and $\langle y, x_0\rangle = 0 \Rightarrow y = 0$. $S^+_{\lambda x_0} = S^*_{\lambda x_0} \cap X$. The dual version of Theorem 3.1 now reads as follows.

THEOREM 3.2. *Let* (X^*, X), (Z, W) *be real dual pairs,* S *be a convex cone in* X*, and* $C : X^* \to Z$ *be a weakly continuous linear mapping. Assume also that* X *is separable and there is some* $x_0 \in S$ *such that* $y \in S^*$ *and* $\langle y, x_0\rangle = 0 \Rightarrow y = 0$*. Then the following are equivalent:*

(i) *The system* $Cy = b$ *has a solution* $y \in S^*$.

(ii) *There exists some positive* (λ, δ) *such that*

$$C^+ w + \gamma x_0 \in S^+_{\lambda x_0}, \qquad \gamma \geq 0 \Rightarrow \langle b, w\rangle + \gamma\delta \geq 0.$$

(iii) *There exists some positive* (λ, δ) *such that*

$$C^+ w \in S^+_{\lambda x_0} \Rightarrow \langle b, w\rangle \geq 0 \quad \textit{and} \quad C^+ x + x_0 \in S^+_{\lambda x_0} \Rightarrow \langle b, w\rangle \geq -\delta.$$

The proof is along the same lines as for Theorem 3.1.

Hilbert spaces. In the case where $(X, \|\ .\ \|)$ is a Hilbert space, the assumption that S is weakly closed is equivalent to S (strongly) closed and one may further characterize the cone $S^+_{\lambda y_0}$.

S^+ and $S^+_{\lambda y_0}$ coincide with the *dual* cones S^* and $S^*_{\lambda y_0}$. We can write $S_{\lambda y_0} = S \cap T_{\lambda y_0}$ with $T_{\lambda y_0} := \{x \in X |\ \|\ x\ \| \leq \lambda \langle y_0, x\rangle\}$ and $T_{\lambda y_0}$ a convex cone. As soon as $\|\ \lambda y_0\ \| > 1$, $T_{\lambda y_0}$ may be interpreted as the cone of vectors x such that their angle θ with y_0 satisfies $\tan(\theta) \leq (\|\ \lambda y_0\ \|^2 - 1)^{1/2}$, i.e., $\theta \leq \theta_0$ (see Figure 1).

LEMMA 3.3. *Assume that* S *is closed and* $0 \in \text{int}\,(S - T_{\lambda y_0})$ *(for the strong topology). Then,* $S^*_{\lambda y_0} = S^* + T^*_{\lambda y_0}$. *In addition, if* $y_0 \in \text{int}\,(S^*)$ *(for the strong topology), then* $S^*_{\lambda y_0} = S^*$ *for* λ *sufficiently large.*

Proof. First note that $T_{\lambda y_0}$ is (strongly) closed and so is S by hypothesis. Therefore, as $0 \in \text{int}\,(S - T_{\lambda y_0})$, then from, e.g., [2],

$$S^*_{\lambda y_0} = (S \cap T_{\lambda y_0})^* = S^* + T^*_{\lambda y_0}.$$

Thus, under the assumptions of the above lemma, one may replace $S^*_{\lambda y_0}$ with $S^* + T^*_{\lambda y_0}$ in Theorem 3.1. In addition,

$$T^*_{\lambda y_0} = \left\{ p \in X \mid \| p \| \leq \frac{\lambda}{\sqrt{\| \lambda y_0 \|^2 - 1}} \langle p, y_0 \rangle \right\}.$$

$T^*_{\lambda y_0}$ is the cone of vectors whose angle θ with y_0 satisfies $\tan(\theta) \leq (\| \lambda y_0 \|^2 - 1)^{-1/2}$, i.e., $\theta \leq \pi/2 - \theta_0$. Therefore, if $y_0 \in \text{int}(S^*)$, then for λ sufficiently large, $T^*_{\lambda y_0} \subset S^*$ so that $S^*_{\lambda y_0} = S^*$. □

4. Illustrative example. As a simple illustrative example, consider the case of a linear system involving positive semidefinite matrices (see [4], [3]).

In this case, $X \equiv Y$ is the vector space R^n of (N, N) real symmetric matrices identified as a vector in R^n with $n = N(N+1)/2$. X is a finite-dimensional Hilbert space with norm induced from the bracket $\langle x, y \rangle := \text{trace}(x.y)$, where $x.y$ denotes the standard (N, N)-matrix product. S is the closed convex cone of positive semidefinite matrices. We want to check existence of solutions to the system $\{Ax = b;\ x \in S\}$, where A is a linear mapping $A : X \to Z$ and $Z \equiv R^m$ with the usual scalar product.

Choose $y_0 := I$ the identity matrix in X. Assumption H trivially holds. Then, if $\{\sigma_i(x)\}$ denotes the eigenvalues of x, all nonnegative when $x \in S$,

$$0 = \langle I, x \rangle = \text{trace}(x) = \sum_{i=1}^{N} \sigma_i(x) \Rightarrow \sigma_i(x) = 0 \quad \forall i \ \Rightarrow x = 0.$$

Moreover (see last paragraph in the previous section), $S \equiv S_{\lambda I}$ for all $\lambda \geq 1$. Indeed, when $x \in S$,

$$\langle I, x \rangle := \sum_{i=1}^{N} \sigma_i(x) \geq \sqrt{\sum_{i=1}^{N} \sigma_i^2(x)} = ||x||.$$

Thus, one may derive a Farkas lemma for existence of solutions to systems of matrix equalities with no closure assumption on $A(S)$. In general, without very specific properties of the mapping A, $A(S)$ is not closed and the standard Farkas lemma states only a property of *asymptotic consistency* (see [3], [4]). In this context, our theorem specializes to the following.

THEOREM 4.1. *Consider X, the Hilbert space of (n, n) symmetric real matrices with norm $||x||^2 := \langle x, x \rangle := \text{trace}(x^2)$. Let Z be R^m and $A : X \to Z$ be a linear mapping.*

The system $Ax = b$ has a solution $x \in S$ if and only if there is some $\delta > 0$ such that

$$A^*u \in S^* \Rightarrow \langle b, u \rangle \geq 0 \text{ and } A^*u + I \in S^* \Rightarrow \langle b, u \rangle \geq -\delta.$$

In the example given in [3, p. 378] the system $\{Ax = b,\ x \in S\}$ is inconsistent whereas $A^*u \in S^* \Rightarrow \langle b, u \rangle \geq 0$. However, as indicated in the theorem, one may check that $A^*u + y_0 \in S^* \Rightarrow \langle b, u \rangle \geq -\delta$ does not hold for any positive δ.

5. Another approach. In this section we present another Farkas theorem with a short proof kindly provided by a referee. This theorem (Theorem 5.1 below) reveals what lies behind Theorem 3.1, which appears as a simple corollary.

THEOREM 5.1. *Let X, Y, Z, W, S, and A be as in Theorem* 3.1. *Then the following are equivalent:*

(1) *The system* $Ax = b$ *has a solution* $x \in S$.

(2) *There exists a compact, convex set* $0 \notin B \subset S$ *and a* $\delta > 0$ *such that for* $K := \text{cone } B$, $y_0 \in \text{int } K^+$, *we have*

$$A^+ w + y_0 \in K^+ \Rightarrow \langle b, w \rangle \geq -\delta.$$

Proof. (1) $\Rightarrow$ (2). Suppose that $x \in S$ and $Ax = b$, $x \neq 0$. (The case $b = 0$ is trivial.) Choose an appropriate B such that it is convex and compact and $0 \notin B \subset S$, with $x \in B$ (e.g., $B := \{x\}$). Therefore, the generated dual cone K^+ has nonempty interior. Choose any $y_0 \in \text{int } K^+$. Then, the program

$$(P) \quad \mu = \{\inf \langle b, w \rangle \mid A^+ w + y_0 \in K^+\}$$

satisfies Slater's condition with $w := 0$. The dual program

$$(D) \quad \nu = \{\sup \langle y_0, x \rangle \mid Ax = b,\ x \in K\}$$

is consistent. Therefore, by weak duality, $\mu \geq \nu > -\infty$; i.e., we can choose $\delta = -\mu$. Therefore, (2) holds.

(2) $\Rightarrow$ (1). Conversely, if (2) holds, then Slater's condition holds for (P) and $\mu > -\infty$. Therefore, by strong duality, (D) must be consistent. □

This theorem can also be derived from closely related results in [6], [7], [8]. Theorem 3.1 is a corollary of Theorem 5.1. This follows from the Banach–Alaoglu theorem; i.e., we can choose λ in the definition of $S_{\lambda y_0}$ so that this cone contains the x from the consistent system. A compact and convex base for the cone is obtained by taking the intersection with the set $\{x|\ \langle y_0, x \rangle = 1\}$. The definition of the cone implies that this set is norm bounded and thus $w*$-compact.

Finally, note that Theorem 4.1 can be proved directly and trivially, since Slater's condition holds for (P) with $y_0 = I$ and $K = S$.

6. Conclusion. We have presented a new nonasymptotic Farkas lemma in Banach spaces for a linear system $Ax = b,\ x \in S$, where S is some positive cone. Under some assumptions on the spaces involved, this Farkas lemma holds without the usual restrictive closure assumption on $A(S)$. The assumptions are rather weak and concern many examples. Theorem 5.1 reveals what lies behind Theorem 3.1, i.e., the existence of an appropriate, compactly generated, convex cone. Using this new Farkas lemma to develop first-order optimality conditions in mathematical programming is a topic for further research.

Acknowledgment. The author wishes to thank the anonymous referees for very helpful comments and suggestions to improve the original version. In particular, Theorem 5.1 and its short and elegant proof were kindly provided by one of them.

REFERENCES

[1] E. J. Anderson and P. Nash, *Linear programming in infinite dimensional spaces*, John Wiley & Sons, New York, 1987.

[2] J. P. Aubin, *L'analyse non linéaire et ses motivations économiques*, Masson, Paris, 1984.

[3] A. Ben-Israel, *Linear inequalities and inequalities on finite dimensional, real or complex vector spaces: A unified theory*, J. Math Anal. Appl., 27 (1969), pp. 376–389.

[4] A. Berman and A. Ben-Israel, *More on linear inequalities with applications to matrix theory*, J. Math. Anal. Appl., 33 (1971), pp. 482–496.

[5] J. M. Borwein, *Continuity and differentiability properties of convex operators*, Proc. London Math. Soc. (3), 44 (1982), pp. 420–444.

[6] J. M. Borwein and H. Wolkowicz, *Characterizations of optimality for the abstract convex program with finite dimensional range*, J. Austral. Math. Soc. Ser. A, 30 (1981), pp. 390–411.
[7] J. M. Borwein and H. Wolkowicz, *Characterizations of optimality without constraint qualification for the abstract convex program*, Math. Programming Studies, 19 (1982), pp. 77–100.
[8] J. M. Borwein and H. Wolkowicz, *A simple constraint qualification in infinite dimensional programming*, Math. Programming 35 (1986), pp. 83–96.
[9] H. Brezis, *Analyse Fonctionnelle*: *Théorie et Applications*, Masson, Paris, 1983.
[10] B. D. Craven and J. J. Koliha, *Generalizations of Farkas' theorem*, SIAM J. Math. Anal., 8 (1977), pp. 983–997.
[11] B. D. Craven, *Mathematical Programming and Control Theory*, Chapman and Hall, London, 1978.
[12] B. M. Glover, *A generalized Farkas lemma with applications to quasidifferentiable programming*, Z. Oper. Res., 26 (1982), pp. 125–141.
[13] B. M. Glover, *Differentiable programming in Banach spaces*, Optimization, 14 (1983), pp. 499–508.
[14] B. M. Glover, V. Jeyakumar, and W. Oettli, *A Farkas lemma for difference sublinear systems and quasi-differentiable mappings*, Math. Programming, 63 (1994), pp. 109–125.
[15] J. Gwinner, *Closed images of convex multivalued mappings in linear topological spaces with applications*, J. Math. Anal. Appl., 60 (1977), pp. 75–86.
[16] V. Jeyakumar, *Nonlinear alternative theorems and nondifferentiable programming*, Z. Oper. Res., 28 (1984), pp. 175–187.
[17] H. L. Royden, *Real Analysis*, 3rd ed., Macmillan, New York, 1988.
[18] C. Zalinescu, *A generalization of Farkas lemma and applications to convex programming*, J. Math. Anal. Appl., 66 (1978), pp. 651–678.
[19] C. Zalinescu, *Solvability results for sublinear functions and operators*, Z. Oper. Res., 31 (1987), pp. 79–101.

SIAM J. CONTROL OPTIM.
Vol. 35, No. 1, pp. 273–284, January 1997

016

EQUIVALENT UNCONSTRAINED MINIMIZATION AND GLOBAL ERROR BOUNDS FOR VARIATIONAL INEQUALITY PROBLEMS*

NOBUO YAMASHITA† AND MASAO FUKUSHIMA†

Abstract. New merit functions for variational inequality problems are constructed through the Moreau–Yosida regularization of some gap functions. The proposed merit functions constitute *unconstrained* optimization problems equivalent to the original variational inequality problem under suitable assumptions. Conditions are studied for those merit functions to be differentiable and for any stationary point of those functions to be a solution of the original variational inequality problem. Moreover, those functions are shown to provide global error bounds for general variational inequality problems under the strong monotonicity assumption only.

Key words. variational inequality problem, merit function, unconstrained optimization problem, global error bounds

AMS subject classifications. 47H05, 49J40, 65K10, 90C30, 90C33

PII. S0363012994277645

1. Introduction.

Merit functions for variational inequality and complementarity problems have recently drawn much attention (see [8] for a survey). Such a function naturally constitutes an equivalent optimization formulation and has turned out to be very useful in designing new globally convergent algorithms and analyzing the rate of convergence of some algorithms [2, 6, 7, 9, 10, 12, 13, 14, 15, 16, 17, 20, 21, 22, 19, 28, 26, 27, 31, 29, 30, 32, 33]. For nonlinear complementarity problems, both constrained and unconstrained differentiable optimization formulations are already known and various interesting results on iterative methods and error bounds have been established [9, 12, 17, 19, 31, 30, 33]. For variational inequality problems, constrained differentiable optimization formulations have been proposed [6, 7, 13, 20, 28, 32]. To the authors' knowledge, however, an unconstrained differentiable optimization formulation has yet to be discussed in the literature. (During the review of this paper, Peng proposed an unconstrained optimization formulation in a technical report [23], which is based on the regularized gap function [7] and the implicit Lagrangian [19]. His approach is, however, quite different from the one adopted in this paper.) The purpose of this paper is to present unconstrained differentiable optimization formulations for general variational inequality problems and to give some error bound results based on them.

The variational inequality problem (VIP) is to find a vector x in a closed convex subset S of R^n such that

$$\langle F(x), y - x\rangle \geq 0 \quad \text{for all } y \in S,$$

where F is a mapping from R^n into itself and $\langle \cdot, \cdot \rangle$ denotes the inner product in R^n. Throughout the paper, we shall assume that the mapping F is continuous.

*Received by the editors November 28, 1994; accepted for publication (in revised form) November 26, 1995. The research of the second author was supported in part by a Scientific Research Grant-in-Aid from the Japanese Ministry of Education, Science and Culture.

http://www.siam.org/journals/sicon/35-1/27764.html

†Department of Applied Mathematics and Physics, Graduate School of Engineering, Kyoto University, Kyoto 606-01, Japan (nobuo@kuamp.kyoto-u.ac.jp, fuku@kuamp.kyoto-u.ac.jp).

First, we review some well-known merit functions for VIP. Let $\Psi : R^n \times R^n \to R$ be defined by

$$\Psi(x, y) = \langle F(x), x - y \rangle. \tag{1}$$

Using Ψ, define the class of functions $f(\,\cdot\,;\alpha) : R^n \to R \cup \{+\infty\}$ by

$$f(x;\alpha) = \sup_{y \in S} \left\{ \Psi(x, y) - \alpha \|x - y\|^2 \right\}, \tag{2}$$

where α is a nonnegative parameter. This function was first introduced by Auslender [2] for $\alpha = 0$ (see also [11]) and by Fukushima [7] for $\alpha > 0$. The latter case was also considered by Auchmuty [1] independently in a general form (see also [13]). The function $f(\,\cdot\,;0)$ is commonly called the gap function, while the function $f(\,\cdot\,;\alpha)$ with $\alpha > 0$ is called the regularized gap function. For each $\alpha \geq 0$, the function $f(\,\cdot\,;\alpha)$ is nonnegative on S and becomes zero at any solution of VIP [11, 7]. Hence, VIP is equivalent to finding a global minimizer of $f(\,\cdot\,;\alpha)$ on S. If F is a continuous mapping, then $f(\,\cdot\,;\alpha)$ is lower semicontinuous for $\alpha = 0$ [5, Lemma 4.1] and continuous for $\alpha > 0$ [7, Theorem 3.1]. The gap function $f(\,\cdot\,;0)$ has the serious drawback, however, that it is in general nondifferentiable even if F is differentiable, and, even worse, it may not be finite valued. On the other hand, the regularized gap function $f(\,\cdot\,;\alpha)$ with $\alpha > 0$ has such desirable properties that it is finite valued everywhere and is differentiable whenever F is differentiable [7, Theorem 3.1]. (See [7, 13] for further properties of the regularized gap function.) A modification of the regularized gap function has recently been proposed in [26, 27].

Closely related to the functions $f(\,\cdot\,;\alpha)$ is the class of functions $h(\,\cdot\,;\beta) : R^n \to R \cup \{+\infty\}$ given by

$$h(x;\beta) = \sup_{y \in S} \left\{ -\Psi(y, x) + \beta \|x - y\|^2 \right\}, \tag{3}$$

where β is a nonnegative parameter. This function has been studied in [20] for the case $\beta = 0$. In particular, if F is pseudomonotone on S [4, 25], i.e.,

$$\langle F(y), x - y \rangle \geq 0 \Longrightarrow \langle F(x), x - y \rangle \geq 0 \quad \text{for all } x, y \in S,$$

then the function $h(\,\cdot\,;0)$ is nonnegative on S and vanishes at any solution of VIP, implying that VIP is equivalent to minimizing $h(\,\cdot\,;0)$ over S [20, Proposition 1]. We will see that this property is extended to the case $\beta > 0$. It is easily seen that, for any $\beta \geq 0$, the function $h(\,\cdot\,;\beta)$ is convex. Moreover, if F is continuous, then $h(\,\cdot\,;\beta)$ is lower semicontinuous, as shown in the next section.

It should be noted that the above-mentioned merit functions all constitute constrained optimization problems equivalent to VIP. In this paper, we explore the possibility of constructing unconstrained differentiable optimization problems equivalent to VIP. Specifically, we consider the following functions derived from the Moreau–Yosida regularization:

$$\phi_f(x;\alpha,\lambda) = \inf_{z \in S} \left\{ f(z;\alpha) + \lambda \|x - z\|^2 \right\} \tag{4}$$

and

$$\phi_h(x;\beta,\lambda) = \inf_{z \in S} \left\{ h(z;\beta) + \lambda \|x - z\|^2 \right\}, \tag{5}$$

where λ is a positive parameter and $f(\,\cdot\,;\alpha)$ and $h(\,\cdot\,;\beta)$ are defined by (2) and (3), respectively. In general, the functions $\phi_f(\,\cdot\,;\alpha,\lambda)$ and $\phi_h(\,\cdot\,;\beta,\lambda)$ may not be easy to

evaluate in practice unless VIP has a certain special structure. (In section 3, we show that the function $\phi_f(\,\cdot\,;\alpha,\lambda)$ is actually computable for monotone affine VIP with polyhedral convex set S.) Nevertheless these functions enjoy some nice theoretical properties that existing merit functions for VIP do not have. Under some assumptions, the functions $\phi_f(\,\cdot\,;\alpha,\lambda)$ and $\phi_h(\,\cdot\,;\beta,\lambda)$ are shown to be differentiable even if F is not differentiable. We shall investigate conditions under which any stationary point of these functions solves the original VIP. We also discuss how the foregoing results are specialized to affine VIP and illustrate how the function $\phi_f(\,\cdot\,;\alpha,\lambda)$ is calculated for the monotone affine VIP with linear constraints. Finally, we will show that $\phi_f(\,\cdot\,;\alpha,\lambda)$ and $\phi_h(\,\cdot\,;\beta,\lambda)$ provide global error bounds for the strongly monotone VIP. We remark that those results are valid for problems with general convex constraints S and do not rely on the Lipschitz continuity of F, while the well-known error bound for VIP obtained by Pang [21] assumes that the set S is polyhedral and F is both strongly monotone and Lipschitz continuous.

2. VIP as unconstrained minimization. In this section, we show that the unconstrained minimization of the function $\phi_f(\,\cdot\,;\alpha,\lambda)$ defined by (4) is equivalent to VIP, and that, under some additional assumptions, the unconstrained minimization of the function $\phi_h(\,\cdot\,;\beta,\lambda)$ defined by (5) is also equivalent to VIP. Further, we shall investigate the differentiability and the convexity of these functions. Recall that the mapping $F : R^n \to R^n$ is said to be monotone on S if

$$\langle F(x) - F(y), x - y\rangle \geq 0 \quad \text{for all } x, y \in S,$$

and strongly monotone with modulus $\mu > 0$ on S if

$$\langle F(x) - F(y), x - y\rangle \geq \mu\|x - y\|^2 \quad \text{for all } x, y \in S.$$

First, we state some lemmas which will be useful in proving the subsequent theorems.

LEMMA 2.1. *For any $\alpha \geq 0$, $f(\,\cdot\,;\alpha)$ is nonnegative on S. Moreover, $f(x;\alpha) = 0$ if and only if x is a solution of VIP.*

Proof. See [11] for the case $\alpha = 0$ and [7, Theorem 3.1] for the case $\alpha > 0$. □

LEMMA 2.2. *For any $\beta \geq 0$, the function $h(\,\cdot\,;\beta)$ is a closed convex function.*

Proof. The convexity of $h(\,\cdot\,;\beta)$ follows from the definition (3) directly, since $-\Psi(y,\,\cdot\,)+\beta\|\,\cdot\,-y\|^2$ is convex for every y. The closedness of $h(\,\cdot\,;\beta)$ follows from the fact that a pointwise supremum of continuous functions yields a lower semicontinuous function. □

LEMMA 2.3.

(a) *For any $\beta \geq 0$, $h(\,\cdot\,;\beta)$ is nonnegative on S.*

(b) *Suppose that F is pseudomonotone on S. Then x is a solution of VIP if and only if $h(x;0) = 0$ and $x \in S$.*

(c) *Suppose that F is strongly monotone on S with modulus μ and that β is chosen to satisfy $0 \leq \beta \leq \mu$. Then x is a solution of VIP if and only if $h(x;\beta) = 0$ and $x \in S$.*

Proof. (a) This is obvious from the definition (3) of $h(\,\cdot\,;\beta)$. (b) See [20, Proposition 1].

(c) By the definition (1) of Ψ, we have

$$\begin{aligned}\Psi(x,y) &\geq -\Psi(y,x) + \mu\|x-y\|^2\\ &\geq -\Psi(y,x) + \beta\|x-y\|^2\\ &\geq -\Psi(y,x) \qquad \text{for all } x, y \in S,\end{aligned}$$

implying that

$$f(x;0) \geq h(x;\beta) \geq h(x;0) \quad \text{for all } x \in S. \tag{6}$$

Let x^* be a solution of VIP. Then, since $f(x^*;0) = 0$ by Lemma 2.1 and $h(x^*;0) = 0$ by part (b) of this lemma, we have $h(x^*;\beta) = 0$ by (6). Conversely, let $h(x;\beta) = 0$. Then, since $h(\,\cdot\,;0)$ is nonnegative on S by part (a), it follows from (6) that $h(x;0) = 0$. Thus, x is a solution of VIP by part (b). □

Using these lemmas, we establish the equivalence between VIP and the unconstrained minimization of $\phi_f(\,\cdot\,;\alpha,\lambda)$ and $\phi_h(\,\cdot\,;\beta,\lambda)$. The next theorem shows that these functions are nonnegative on R^n and that, under suitable conditions, the sets of unconstrained minima of these functions coincide with the set of solutions of VIP.

THEOREM 2.4.

(a) *For any $\alpha \geq 0$, $\beta \geq 0$, and $\lambda > 0$, the functions $\phi_f(\,\cdot\,;\alpha,\lambda)$ and $\phi_h(\,\cdot\,;\beta,\lambda)$ are nonnegative on R^n.*

(b) *For any $\alpha \geq 0$ and $\lambda > 0$, x^* is a solution of VIP if and only if $\phi_f(\,\cdot\,;\alpha,\lambda)$ attains its global minimum at x^* and the minimum value is zero.*

(c) *Suppose that F is pseudomonotone on S. Then, for any $\lambda > 0$, x^* is a solution of VIP if and only if $\phi_h(\,\cdot\,;0,\lambda)$ attains its global minimum at x^* and the minimum value is zero.*

(d) *Suppose that F is strongly monotone on S with modulus μ and that β is chosen to satisfy $0 \leq \beta \leq \mu$. Then, for any $\lambda > 0$, x^* is a solution of VIP if and only if $\phi_h(\,\cdot\,;\beta,\lambda)$ attains its global minimum at x^* and the minimum value is zero.*

Proof. We consider the function $\phi_f(\,\cdot\,;\alpha,\lambda)$. (a) Since $f(\,\cdot\,;\alpha)$ is nonnegative on S by Lemma 2.1, we can easily deduce from the definition (4) of $\phi_f(\,\cdot\,;\alpha,\lambda)$ that $\phi_f(x;\alpha,\lambda)$ is nonnegative for all $x \in R^n$. (b) Suppose that x^* is a solution of VIP. Then, we have

$$\begin{aligned}\phi_f(x^*;\alpha,\lambda) &= \inf_{z\in S}\left\{f(z;\alpha) + \lambda\|x^* - z\|^2\right\} \\ &\leq f(x^*;\alpha) + \lambda\|x^* - x^*\|^2 \\ &= 0,\end{aligned}$$

where the last equality follows from $f(x^*;\alpha) = 0$. Since $\phi_f(x;\alpha,\lambda) \geq 0$ for all x as shown above, we obtain $\phi_f(x^*;\alpha,\lambda) = 0$. Conversely, suppose $\phi_f(x^*;\alpha,\lambda) = 0$. Then since $f(z;\alpha) \geq 0$ for all $z \in S$, it follows from the definition (4) of $\phi_f(\,\cdot\,;\alpha,\lambda)$ that there exists a sequence $\{z^k\}$ in S such that $f(z^k;\alpha) \to 0$ and $\|z^k - x^*\| \to 0$. Since S is closed, $z^k \to x^*$ and $z^k \in S$ imply $x^* \in S$. Moreover, since $f(\,\cdot\,;\alpha)$ is lower semicontinuous for $\alpha = 0$ [5, 13] and continuous for $\alpha > 0$ [7], we have

$$0 \leq f(x^*;\alpha) \leq \lim_{k\to\infty} f(z^k;\alpha) = 0.$$

Thus, x^* must be a solution of VIP.

By using Lemmas 2.2 and 2.3, the proof for the function $\phi_h(\,\cdot\,;\beta,\lambda)$ can be done analogously. □

By this theorem, the unconstrained minimization problem

$$\min_{x\in R^n} \quad \phi_f(x;\alpha,\lambda) \tag{7}$$

is equivalent to VIP, and the unconstrained minimization problem

$$\min_{x\in R^n} \quad \phi_h(x;\beta,\lambda) \tag{8}$$

is equivalent to VIP under suitable assumptions on F and the parameters involved.

In order for these minimization problems to be practically useful, it is desirable that the objective functions $\phi_f(\,\cdot\,;\alpha,\lambda)$ and $\phi_h(\,\cdot\,;\beta,\lambda)$ are everywhere differentiable. For the discussions to follow, it will be convenient to define the functions $\Phi_f(\,\cdot\,,\,\cdot\,;\alpha,\lambda) : R^{2n} \to R \cup \{+\infty\}$ and $\Phi_h(\,\cdot\,,\,\cdot\,;\beta,\lambda) : R^{2n} \to R \cup \{+\infty\}$ by

$$\Phi_f(x,z;\alpha,\lambda) = f(z;\alpha) + \lambda\|x-z\|^2$$

and

$$\Phi_h(x,z;\beta,\lambda) = h(z;\beta) + \lambda\|x-z\|^2,$$

respectively. By (4) and (5), we have

$$\phi_f(x;\alpha,\lambda) = \inf_{z\in S} \Phi_f(x,z;\alpha,\lambda)$$

and

$$\phi_h(x;\beta,\lambda) = \inf_{z\in S} \Phi_h(x,z;\beta,\lambda).$$

Whenever the functions $\Phi_f(x,\,\cdot\,;\alpha,\lambda)$ and $\Phi_h(x,\,\cdot\,;\beta,\lambda)$ are assumed to attain their minima uniquely in the set S, we shall denote those minima by $z_f(x;\alpha,\lambda)$ and $z_h(x;\beta,\lambda)$, respectively. The following propositions are consequences of [2, Chapter 4, Theorem 1.7]. Note that these propositions do not rely upon the differentiability of F.

PROPOSITION 2.5. *Let $\alpha \geq 0$ and $\lambda > 0$. If the function $\Phi_f(x,\,\cdot\,;\alpha,\lambda)$ attains its unique minimum $z_f(x;\alpha,\lambda)$ on S for each $x \in R^n$, then $\phi_f(\,\cdot\,;\alpha,\lambda)$ is differentiable on R^n and*

$$\nabla\phi_f(x;\alpha,\lambda) = 2\lambda\left(x - z_f(x;\alpha,\lambda)\right).$$

Proof. This follows immediately from Theorem 1.7 in [2, Chapter 4]. □

PROPOSITION 2.6. *For any $\beta \geq 0$ and $\lambda > 0$, the function $\Phi_h(x,\,\cdot\,;\beta,\lambda)$ attains its minimum $z_h(x;\beta,\lambda)$ uniquely. Moreover, $\phi_h(\,\cdot\,;\beta,\lambda)$ is a differentiable convex function on R^n and*

$$\nabla\phi_h(x;\beta,\lambda) = 2\lambda\left(x - z_h(x;\beta,\lambda)\right).$$

Proof. By Lemma 2.2, $h(\,\cdot\,;\beta)$ is a closed convex function. Thus, by Theorem 31.5 in [24], $\Phi_h(x,\,\cdot\,;\beta,\lambda)$ uniquely attains its minimum on S for any x and hence, by Theorem 1.7 in [2, Chapter 4], $\phi_h(\,\cdot\,;\beta,\lambda)$ is differentiable and its gradient is represented as indicated in the proposition. Moreover, it follows from the convexity of $h(\,\cdot\,;\beta)$ that $\phi_h(\,\cdot\,;\beta,\lambda)$ is also convex (see the proof of Proposition 4.1 in [3]). □

Theorem 2.4 is concerned with the equivalence between VIP and global minimization of $\phi_f(\,\cdot\,;\alpha,\lambda)$ and $\phi_h(\,\cdot\,;\beta,\lambda)$. In general, however, there may exist local minima or stationary points which do not solve VIP. In the remainder of the section, we study conditions under which any stationary point of problems (7) and (8) is a solution of VIP.

The following proposition gives a condition for $\phi_f(\,\cdot\,;\alpha,\lambda)$ to be convex.

PROPOSITION 2.7. *Suppose that $\Psi(\,\cdot\,,y)$ is convex for each fixed $y \in S$. Then, for any $\lambda > 0$, $\phi_f(\,\cdot\,;0,\lambda)$ is a differentiable convex function on R^n. Moreover, suppose*

that $\Psi(\,\cdot\,, y)$ is strongly convex with modulus μ for each fixed $y \in S$. Then, for any $\lambda > 0$ and any α such that $0 \le \alpha \le \mu$, $\phi_f(\,\cdot\,; \alpha, \lambda)$ is a differentiable convex function on R^n.

Proof. We prove the first half only; the latter half can be proved by a similar argument. Since $\Psi(\,\cdot\,, y)$ is finite valued and convex for each fixed $y \in S$, $f(\,\cdot\,; 0)$ is a closed proper convex function. So we can prove the convexity of $\phi_f(\,\cdot\,; 0, \lambda)$ in a way similar to the proof of Proposition 4.1 in [3]. Moreover, since $\Phi_f(x, \,\cdot\,; 0, \lambda)$ uniquely attains its minimum on S for any x by Theorem 31.5 in [24], $\phi_f(\,\cdot\,; 0, \lambda)$ is differentiable by Proposition 2.5. □

Note that, as shown in [11], if S is polyhedral convex, $\langle F(x), x\rangle$ is convex on S and each component of F is concave on S, then $f(\,\cdot\,; 0)$ is convex. Thus it follows that, under these assumptions, $\phi_f(\,\cdot\,; 0, \lambda)$ is a differentiable convex function.

The following result is a consequence of Proposition 2.7.

THEOREM 2.8. *Assume that VIP has a solution. If $\Psi(\,\cdot\,, y)$ is convex for each fixed $y \in S$, then for each $\lambda > 0$ any stationary point of $\phi_f(\,\cdot\,; 0, \lambda)$ is a solution of VIP. Moreover, if $\Psi(\,\cdot\,, y)$ is strongly convex with modulus μ for each fixed $y \in S$, then for each $\lambda > 0$ and α such that $0 \le \alpha \le \mu$ any stationary point of $\phi_f(\,\cdot\,; \alpha, \lambda)$ is a solution of VIP.*

Proof. Under given assumptions, $\phi_f(\,\cdot\,; \alpha, \lambda)$ is a differentiable convex function for each $\alpha \ge 0$ by Proposition 2.5. Therefore $\nabla\phi_f(x; \alpha, \lambda) = 0$ if and only if $\phi_f(\,\cdot\,; \alpha, \lambda)$ attains its global minimum at x. The desired results then follow from Theorem 2.4. □

Next we restrict our attention to the function $\phi_f(\,\cdot\,; \alpha, \lambda)$ with $\alpha > 0$. The following theorem gives conditions under which any stationary point of $\phi_f(\,\cdot\,; \alpha, \lambda)$ is a solution of VIP.

THEOREM 2.9. *Assume that VIP has a solution. Let $\alpha > 0$ and $\lambda > 0$. Suppose that the function $\Phi_f(x, \,\cdot\,; \alpha, \lambda)$ attains its unique minimum $z_f(x; \alpha, \lambda)$ on S for any fixed $x \in R^n$. If F is differentiable and $\nabla F(x)$ is positive definite on S, then any stationary point of $\phi_f(\,\cdot\,; \alpha, \lambda)$ is a solution of VIP.*

Proof. Let $\hat{x}$ be an arbitrary stationary point of $\phi_f(\,\cdot\,; \alpha, \lambda)$. Then by Proposition 2.5 it satisfies

$$\nabla\phi_f(\hat{x}; \alpha, \lambda) = 2\lambda\,(\hat{x} - z_f(\hat{x}; \alpha, \lambda)) = 0.$$

Hence, we have $\hat{x} = z_f(\hat{x}; \alpha, \lambda) \in S$. On the other hand, since $z_f(x; \alpha, \lambda)$ is the minimizer of $\Phi_f(x, \,\cdot\,; \alpha, \lambda)$ on S and since $f(\,\cdot\,; \alpha)$ is differentiable [7], the first-order optimality condition yields

$$\langle \nabla_z \Phi_f(x, z_f(x; \alpha, \lambda); \alpha, \lambda), y - z_f(x; \alpha, \lambda)\rangle \ge 0 \quad \text{for all } y \in S;$$

that is,

$$\langle \nabla f(z_f(x; \alpha, \lambda); \alpha) + 2\lambda(z_f(x; \alpha, \lambda) - x), y - z_f(x; \alpha, \lambda)\rangle \ge 0 \quad \text{for all } y \in S. \tag{9}$$

Substituting $\hat{x}$ for x in (9) and using $\hat{x} = z_f(\hat{x}; \alpha, \lambda)$, we have

$$\langle \nabla f(\hat{x}; \alpha), y - \hat{x}\rangle \ge 0 \quad \text{for all } y \in S.$$

Namely, $\hat{x}$ is a stationary point of $f(\,\cdot\,; \alpha)$ on S. Then, by Theorem 3.3 in [7], the positive definiteness of $\nabla F(x)$ ensures that $\hat{x}$ is a solution of VIP. □

Finally, we consider the function $\phi_h(\,\cdot\,; \beta, \lambda)$. Proposition 2.6 and Theorem 2.4 immediately yield the following theorem.

THEOREM 2.10. *Assume that VIP has a solution. Let $\lambda > 0$. If F is pseudomonotone on S, then any stationary point of $\phi_h(\,\cdot\,;0,\lambda)$ is a solution of VIP. Moreover, if F is strongly monotone on S with modulus μ and if β is chosen to satisfy $0 \le \beta \le \mu$, then any stationary point of $\phi_h(\,\cdot\,;\beta,\lambda)$ is a solution of VIP.*

Proof. First note that for each $\beta \ge 0$ and $\lambda > 0$, $\phi_h(\,\cdot\,;\beta,\lambda)$ is a differentiable convex function by Proposition 2.6. Thus, $\nabla\phi_h(x;\beta,\lambda) = 0$ if and only if $\phi_h(\,\cdot\,;\beta,\lambda)$ attains its global minimum at x. The theorem then follows from Theorem 2.4 (c) and (d). □

3. Monotone affine VIP. In this section, we consider the special case of VIP where F is the affine mapping

$$F(x) = Mx + a.$$

Here the $n \times n$ matrix M is assumed to be positive semidefinite, that is, the mapping F is monotone. The next two theorems show some properties of $\phi_f(\,\cdot\,;\alpha,\lambda)$ and $\phi_h(\,\cdot\,;\beta,\lambda)$ for the monotone affine VIP.

THEOREM 3.1. *Let μ be a nonnegative constant such that*

$$\langle x, Mx\rangle \ge \mu\|x\|^2 \quad \text{for all } x \in R^n.$$

Then for any $\lambda > 0$ and $\alpha \ge 0$ such that $\alpha \le \mu$, $\phi_f(\,\cdot\,;\alpha,\lambda)$ is a differentiable convex function. Moreover, any stationary point of $\phi_f(\,\cdot\,;\alpha,\lambda)$ is a solution of the affine VIP.

Proof. If $0 \le \alpha \le \mu$, then the function

$$\Psi(z,y) - \alpha\|z-y\|^2 = \langle Mz + a, z - y\rangle - \alpha\|z-y\|^2$$

is convex in z for each fixed y. Hence, by the definition (2), $f(\,\cdot\,;\alpha)$ is a closed convex function. So it follows from the proof of Proposition 4.1 in [3] that $\phi_f(\,\cdot\,;\alpha,\lambda)$ is a closed convex function. Moreover, by Theorem 31.5 in [24], $\Phi_f(x,\,\cdot\,;\alpha,\lambda)$ uniquely attains its minimum on S for any x. Hence, by Proposition 2.5, $\phi_f(\,\cdot\,;\alpha,\lambda)$ is differentiable. This proves the first half of the theorem. Since $\nabla\phi_f(x;\alpha,\lambda) = 0$ if and only if x is a global minimizer of $\phi_f(\,\cdot\,;\alpha,\lambda)$, the last half of the theorem follows from Theorem 2.4 (b). □

THEOREM 3.2. *Let μ be a nonnegative constant such that*

$$\langle x, Mx\rangle \ge \mu\|x\|^2 \quad \text{for all } x \in R^n.$$

Then, for any $\lambda > 0$ and $\beta \ge 0$ such that $\beta \le \mu$, $\phi_h(\,\cdot\,;\beta,\lambda)$ is a differentiable convex function and any stationary point of $\phi_h(\,\cdot\,;\beta,\lambda)$ is a solution of the affine VIP.

Proof. By Proposition 2.6 and Theorem 2.10, we get the desired results using a similar argument to the proof of Theorem 3.1. □

Note that if M is not positive definite but simply positive semidefinite, then the constant μ in Theorems 3.1 and 3.2 is zero, and hence α and β must be zero, too. Namely, the theorems then apply to the functions $\phi_f(\,\cdot\,;0,\lambda)$ and $\phi_h(\,\cdot\,;0,\lambda)$.

In the rest of this section, we show that the function $\phi_f(\,\cdot\,;\alpha,\lambda)$ can be practically computed for the monotone affine VIP with linear constraints: Find $x \in S$ such that

$$\langle Mx + a, y - x\rangle \ge 0 \quad \text{for all } y \in S, \tag{10}$$

where $S = \{x \in R^n \mid x \ge 0, Ax \ge b\}$ with $m \times n$ matrix A and m-vector b.

To evaluate the function $\phi_f(\,\cdot\,;\alpha,\lambda)$, we need to solve the minimization problem

$$\min_{z\in S} f(z;\alpha)+\lambda\|x-z\|^2, \tag{11}$$

where $f(z;\alpha)$ is given by

$$f(z;\alpha)=\sup_{y\in S}\left\{\langle Mz+a,z-y\rangle-\alpha\|z-y\|^2\right\}.$$

Since

$$\nabla f(z;\alpha)=Mz+a+(M^T-2\alpha I)(z-[z-\tfrac{1}{2\alpha}(Mz+a)]_S^+),$$

where $[\,\cdot\,]_S^+$ denotes the orthogonal projection onto S, the Karush–Kuhn–Tucker conditions for the minimization problem (11) are written as

$$\begin{gathered} Mz+a+(M^T-2\alpha I)(z-[z-\tfrac{1}{2\alpha}(Mz+a)]_S^+)+2\lambda(z-x)-u-A^Tv=0,\\ s=Az-b,\\ z\ge 0,\ \ s\ge 0,\ \ u\ge 0,\ \ v\ge 0,\\ \langle z,u\rangle=0,\ \ \langle s,v\rangle=0, \end{gathered} \tag{12}$$

where $u\in R^n$ and $v\in R^m$ are the vectors of Lagrange multipliers and $s\in R^m$ is the vector of slack variables. Moreover, the vector $[z-\frac{1}{2\alpha}(Mz+a)]_S^+$ is a solution of the linear complementarity problem (LCP) of finding $y\in R^n$ such that

$$\begin{gathered} y-z+\tfrac{1}{2\alpha}(Mz+a)-p-A^Tq=0,\\ t=Ay-b,\\ y\ge 0,\ \ t\ge 0,\ \ p\ge 0,\ \ q\ge 0,\\ \langle y,p\rangle=0,\ \ \langle t,q\rangle=0. \end{gathered}$$

Hence, the Karush–Kuhn–Tucker conditions (12) can be rewritten as the following LCP: find $(z,v,y,q)\in R^{2m+2n}$ such that

$$\begin{gathered} \begin{pmatrix} u\\ s\\ p\\ t \end{pmatrix}=\begin{pmatrix} Mz+a+(M^T-2\alpha I)(z-y)+2\lambda(z-x)-A^Tv\\ Az-b\\ y-z+\frac{1}{2\alpha}(Mz+a)-A^Tq\\ Ay-b \end{pmatrix},\\ z\ge 0,\ \ s\ge 0,\ \ y\ge 0,\ \ t\ge 0,\ \ u\ge 0,\ \ v\ge 0,\ \ p\ge 0,\ \ q\ge 0,\\ \langle z,u\rangle=0,\ \ \langle s,v\rangle=0,\ \ \langle y,p\rangle=0,\ \ \langle t,q\rangle=0. \end{gathered} \tag{13}$$

Now let us put $\alpha=\frac{1}{2}$. Then the LCP (13) becomes

$$\begin{gathered} \begin{pmatrix} u\\ s\\ p\\ t \end{pmatrix}=\begin{pmatrix} N & -A^T & -M^T+I & 0\\ A & 0 & 0 & 0\\ M-I & 0 & I & -A^T\\ 0 & 0 & A & 0 \end{pmatrix}\begin{pmatrix} z\\ v\\ y\\ q \end{pmatrix}+\begin{pmatrix} a-2\lambda x\\ -b\\ a\\ -b \end{pmatrix},\\ z\ge 0,\ \ s\ge 0,\ \ y\ge 0,\ \ t\ge 0,\ \ u\ge 0,\ \ v\ge 0,\ \ p\ge 0,\ \ q\ge 0,\\ \langle z,u\rangle=0,\ \ \langle s,v\rangle=0,\ \ \langle y,p\rangle=0,\ \ \langle t,q\rangle=0, \end{gathered} \tag{14}$$

where $N=M+M^T+(2\lambda-1)I$. If $2\lambda\ge 1$, then the positive semidefiniteness of M ensures that (14) is a monotone LCP. Thus, the function $\phi_f(x;\alpha,\lambda)$ for the monotone affine VIP (10) can be computed by solving the monotone LCP (14), provided that α and λ are chosen such that $\alpha=\frac{1}{2}$ and $\lambda\ge\frac{1}{2}$.

4. Global error bounds. Error bounds play an important role in analyzing the convergence rate of iterative algorithms [16, 29]. In particular, the natural residual $r : R^n \to R$, which is defined by

$$r(x) = \|x - [x - F(x)]_S^+\|, \tag{15}$$

is known to provide a global error bound for linearly constrained VIP under the Lipschitz continuity and strong monotonicity of F [21]. In this section, we show that $\phi_f(\,\cdot\,;\alpha,\lambda)$ and $\phi_h(\,\cdot\,;\beta,\lambda)$ provide global error bounds for VIP under the strong monotonicity of F.

To begin with, we point out that, when F is strongly monotone, the functions $f(\,\cdot\,;\alpha)$ and $h(\,\cdot\,;\beta)$ provide error bounds on S, provided that the parameters α and β are small enough. Note that the strong monotonicity of F guarantees the existence of a unique solution to VIP.

LEMMA 4.1. *If F is strongly monotone on S with modulus μ and if α is chosen to satisfy $0 \le \alpha < \mu$, then we have $f(x^*;\alpha) = 0$ and*

$$f(x;\alpha) \ge (\mu - \alpha)\,\|x - x^*\|^2 \qquad \text{for all } x \in S,$$

where x^* is the unique solution of VIP.

Proof. For this proof see Proposition 3.4 in [28] and Theorem 4.5 in [13]. □

LEMMA 4.2. *If F is strongly monotone on S with modulus μ and if β is chosen to satisfy $0 < \beta \le \mu$, then we have $h(x^*;\beta) = 0$ and*

$$h(x;\beta) \ge \beta\|x - x^*\|^2 \qquad \text{for all } x \in S,$$

where x^* is the unique solution of VIP.

Proof. By Lemma 2.3, we have $h(x^*,\beta) = 0$. Let $x \in S$ be arbitrary. Then, we have

$$\begin{aligned} h(x;\beta) &= \sup_{y\in S}\left\{\langle F(y), x - y\rangle + \beta\|x - y\|^2\right\} \\ &\ge \langle F(x^*), x - x^*\rangle + \beta\|x - x^*\|^2 \\ &\ge \beta\|x - x^*\|^2, \end{aligned}$$

where the last inequality follows from the fact that x^* is a solution of VIP. □

Note that the error bounds given in the above lemmas are valid on the constraint set S. Using these lemmas, we prove below that $\phi_f(\,\cdot\,;\alpha,\lambda)$ and $\phi_h(\,\cdot\,;\beta,\lambda)$ provide global error bounds for VIP on the whole space R^n, whenever F is strongly monotone.

The next theorem shows that the growth rate of $\phi_f(\,\cdot\,;\alpha,\lambda)$ is in the order of the squared distance to the unique solution of VIP, provided that the parameter α is chosen sufficiently small.

THEOREM 4.3. *If F is strongly monotone on S with modulus μ and if α is chosen to satisfy $0 \le \alpha < \mu$, then for each $\lambda > 0$,*

$$\frac{1}{2}\min\{\mu - \alpha, \lambda\}\,\|x - x^*\|^2 \le \phi_f(x;\alpha,\lambda) \le \lambda\|x - x^*\|^2 \qquad \text{for all } x \in R^n,$$

where x^* is the unique solution of VIP.

Proof. First we consider the right-hand inequality. Since $x^* \in S$ and $f(x^*;\alpha) = 0$ by Lemma 2.1, we have

$$\begin{aligned} \phi_f(x;\alpha,\lambda) &= \inf_{z\in S}\left\{f(z;\alpha) + \lambda\|z - x\|^2\right\} \\ &\le f(x^*;\alpha) + \lambda\|x - x^*\|^2 \\ &= \lambda\|x - x^*\|^2. \end{aligned}$$

Next, we prove the left-hand inequality. It follows from Lemma 4.1 that

$$\begin{aligned}
\phi_f(x;\alpha,\lambda) &= \inf_{z\in S}\left\{f(z;\alpha)+\lambda\|z-x\|^2\right\}\\
&\geq \inf_{z\in S}\left\{(\mu-\alpha)\|z-x^*\|^2+\lambda\|z-x\|^2\right\}\\
&\geq \min\{\mu-\alpha,\lambda\}\inf_{z\in S}\left\{\|z-x^*\|^2+\|z-x\|^2\right\}\\
&= \min\{\mu-\alpha,\lambda\}\left(\left\|\left[\frac{x+x^*}{2}\right]_S^+-x^*\right\|^2+\left\|\left[\frac{x+x^*}{2}\right]_S^+-x\right\|^2\right)\\
&\geq \frac{1}{2}\min\{\mu-\alpha,\lambda\}\|x-x^*\|^2,
\end{aligned}$$

where the last inequality follows from the inequality

$$\|a\|^2+\|b\|^2\geq\frac{\|a-b\|^2}{2}\qquad\text{for all } a,b\in R^n. \qquad \square$$

Now we turn our attention to the function $\phi_h(\,\cdot\,;\beta,\lambda)$. The next theorem shows that the quadratic growth rate of $\phi_h(\,\cdot\,;\beta,\lambda)$ is ensured, provided that the parameter β is chosen sufficiently small.

THEOREM 4.4. *If F is strongly monotone on S with modulus μ and if β is chosen to satisfy $0<\beta\leq\mu$, then for each $\lambda>0$,*

$$\frac{1}{2}\min\{\beta,\lambda\}\|x-x^*\|^2\leq\phi_h(x;\beta,\lambda)\leq\lambda\|x-x^*\|^2\quad\text{for all } x\in R^n,$$

where x^* is the unique solution of VIP.

Proof. Noting that $x^*\in S$ and $h(x^*;\beta)=0$, the right-hand inequality can be proved in a way similar to the first part of the proof of Theorem 4.3. Moreover, by using Lemma 4.2, we can prove the left-hand inequality analogously to the last part of the proof of Theorem 4.3. $\square$

Theorems 4.3 and 4.4 demonstrate that $\phi_f(\,\cdot\,;\alpha,\lambda)$ and $\phi_h(\,\cdot\,;\beta,\lambda)$ provide global error bounds for VIP without the Lipschitz continuity of F. Recall that for the natural residual r defined by (15) to provide a global error bound not only the strong monotonicity but also the Lipschitz continuity of F is required [21]. In this respect, $\phi_f(\,\cdot\,;\alpha,\lambda)$ and $\phi_h(\,\cdot\,;\beta,\lambda)$ are more favorable than the natural residual r.

5. Concluding remarks. When S is given by

$$S=\{x\in R^n_+\mid g_i(x)\leq 0,\ \ i=1,\ldots,m\},\tag{16}$$

where R^n_+ denotes the nonnegative orthant in R^n and $g_i: R^n\to R$ are twice differentiable convex functions, the Karush–Kuhn–Tucker conditions for VIP yield the nonlinear complementarity problem: find $(x,\nu)\in R^n_+\times R^m_+$ such that

$$\langle(x,\nu),H(x,\nu)\rangle=0,\quad H(x,\nu)\geq 0,\tag{17}$$

where

$$H(x,\nu)=\begin{bmatrix}F(x)+\nabla g(x)\nu\\ -g(x)\end{bmatrix}$$

with $g(x)=(g_1(x),g_2(x),\ldots,g_m(x))^T$. For nonlinear complementarity problems various equivalent formulations as a system of equations or an unconstrained optimization

problem have been proposed and studied extensively [9, 12, 18, 19, 31, 29, 33]. So we might first reformulate VIP with the inequality constraints (16) as the nonlinear complementarity problem (17) and then convert it into one of the known equivalent unconstrained optimization problems. However, such a transformation into the nonlinear complementarity problem (17) is not necessarily a panacea for dealing with the inequality-constrained VIP. For instance, global error bounds for nonlinear complementarity problems are obtained under the strong monotonicity of the mapping involved [21, 33]. Unfortunately, it is unlikely that the nonlinear complementarity problem (17) satisfies the strong monotonicity condition, because the Jacobian

$$\nabla H(x,\nu) = \begin{bmatrix} \nabla F(x) + \sum_{i=1}^{m} \nu_i \nabla^2 g_i(x) & \nabla g(x) \\ -\nabla g(x)^T & 0 \end{bmatrix}$$

is never positive definite for any (x,ν), even if F is strongly monotone. On the other hand, as shown in the preceding sections, the functions $\phi_f(\,\cdot\,;\alpha,\lambda)$ and $\phi_h(\,\cdot\,;\beta,\lambda)$ enjoy various favorable properties, particularly when F is strongly monotone.

Acknowledgments. The authors would like to thank the associate editor and the referees for their constructive comments. They also thank Professor Paul Tseng for his helpful comment.

REFERENCES

[1] G. Auchmuty, *Variational principles for variational inequalities,* Numer. Funct. Anal. Optim., 10 (1989), pp. 863–874.
[2] A. Auslender, *Optimisation: Méthodes Numériques*, Masson, Paris, 1976.
[3] D. P. Bertsekas and J. N. Tsitsiklis, *Parallel and Distributed Computation: Numerical Methods*, Prentice-Hall, Englewood Cliffs, NJ, 1989.
[4] R. W. Cottle and J. C. Yao, *Pseudo-monotone complementarity problems in Hilbert space*, J. Optim. Theory Appl., 75 (1992), pp. 281–295.
[5] L. C. Dunn, *Convergence rates for conditional gradient sequences generated by implicit step length rules*, SIAM J. Control Optim., 18 (1980), pp. 473–487.
[6] A. Friedlander, J. M. Martinez, and S. A. Santos, *A New Strategy for Solving Variational Inequalities in Bounded Polytopes,* Technical report RP 02/94, IMECC, Universidade Estadual de Campinas, Campinas-São Paulo, Brasil, 1994.
[7] M. Fukushima, *Equivalent differentiable optimization problems and descent methods for asymmetric variational inequality problems,* Math. Programming, 53 (1992), pp. 99–110.
[8] M. Fukushima, *Merit functions for variational inequality and complementarity problems,* in Nonlinear Optimization and Applications, G. DiPillo and F. Giannessi, eds., Plenum Press, New York, 1996, pp. 155–170.
[9] C. Geiger and C. Kanzow, *On the resolution of monotone complementarity problems,* Comput. Optim. Appl., 5 (1996), pp. 155–173.
[10] P. T. Harker and J.-S. Pang, *Finite-dimensional variational inequality and nonlinear complementarity problem: A survey of theory, algorithms and applications*, Math. Programming, 48 (1990), pp. 161–220.
[11] D. W. Hearn, *The gap function of a convex program*, Oper. Res. Lett., 1 (1982), pp. 67–71.
[12] C. Kanzow, *Nonlinear complementarity as unconstrained optimization,* J. Optim. Theory Appl., 88 (1996), pp. 139–155.
[13] T. Larsson and M. Patriksson, *A class of gap functions for variational inequalities*, Math. Programming, 64 (1994), pp. 53–79.
[14] Z.-Q. Luo and J.-S. Pang, *Error bounds for analytic systems and their applications,* Math. Programming, 67 (1994), pp. 1–28.
[15] Z.-Q. Luo and P. Tseng, *On a global error bound for a class of monotone affine variational inequality problems*, Oper. Res. Lett., 11 (1992), pp. 159–165.
[16] Z.-Q. Luo and P. Tseng, *Error bounds and convergence analysis of feasible descent methods: A general approach*, Ann. Oper. Res., 46 (1993), pp. 157–178.
[17] X.-D. Luo and P. Tseng, *On global projection-type error bound for the linear complementarity problem to be global*, Linear Algebra Appl., to appear.

[18] O. L. Mangasarian, *Equivalence of the complementarity problem to a system of nonlinear equations,* SIAM J. Appl. Math., 31 (1976), pp. 89–92.
[19] O. L. Mangasarian and M. V. Solodov, *Nonlinear complementarity as unconstrained and constrained minimization*, Math. Programming, 62 (1993), pp. 277–297.
[20] S. Nguyen and C. Dupuis, *An efficient method for computing traffic equilibria in networks with asymmetric transportation costs,* Transportation Sci., 18 (1984), pp. 185–202.
[21] J.-S. Pang, *A posteriori error bounds for the linearly-constrained variational inequality problem,* Math. Oper. Res., 12 (1987), pp. 474–484.
[22] J.-S. Pang, *Complementarity problems,* in Handbook of Global Optimization, R. Horst and P. Pardalos, eds., Kluwer Academic Publishers, Norwell, MA, 1994, pp. 271–338.
[23] J. M. Peng, *Equivalence of variational inequality problems to unconstrained optimization,* Technical report, State Key Laboratory of Scientific and Engineering Computing, Academia Sinica, Beijing, China, 1995.
[24] R. T. Rockafellar, *Convex Analysis*, Princeton University Press, Princeton, NJ, 1970.
[25] S. Schaible, *Generalized monotonicity—concepts and uses*, in Variational Inequality and Network Equilibrium Problems, F. Giannessi and A. Maugeri, eds., Plenum Press, New York, 1995, pp. 289–299.
[26] K. Taji and M. Fukushima, *A new merit function and a successive quadratic programming algorithm for variational inequality problems*, SIAM J. Optim., 6 (1996), pp. 704–713.
[27] K. Taji and M. Fukushima, *A globally convergent Newton method for solving variational inequality problems with inequality constraints*, in Recent Advances in Nonsmooth Optimization, D.-Z. Du, L. Qi, and R.S. Womersley, eds., World Scientific Publishers, River Edge, NJ, 1995, pp. 405–417.
[28] K. Taji, M. Fukushima, and T. Ibaraki, *A globally convergent Newton method for solving strongly monotone variational inequalities,* Math. Programming, 58 (1993), pp. 369–383.
[29] P. Tseng, *On linear convergence of iterative methods for the variational inequality problem*, J. Comput. Appl. Math., 60 (1995), pp. 237–252.
[30] P. Tseng, *Growth behavior of a class of merit functions for the nonlinear complementarity problem*, J. Optim. Theory Appl., 89 (1996), pp. 17–37.
[31] P. Tseng, N. Yamashita, and M. Fukushima, *Equivalence of complementarity problems to differentiable minimization: A unified approach*, SIAM J. Optim., 6 (1996), pp. 446–460.
[32] J. H. Wu, M. Florian, and P. Marcotte, *A general descent framework for the monotone variational inequality problem*, Math. Programming, 61 (1993), pp. 281–300.
[33] N. Yamashita and M. Fukushima, *On stationary points of the implicit Lagrangian for nonlinear complementarity problems*, J. Optim. Theory Appl., 84 (1995), pp. 653–663.

SIAM J. CONTROL OPTIM.
Vol. 35, No. 1, pp. 285–314, January 1997

017

STABLITY OF SET-VALUED MAPPINGS IN INFINITE DIMENSIONS: POINT CRITERIA AND APPLICATIONS*

BORIS S. MORDUKHOVICH† AND YONGHENG SHAO†

Abstract. This paper deals with effective characterizations of stability and regularity properties of set-valued mappings in infinite dimensions, which are of great importance for applications to many aspects in optimization and control. The main purpose is to obtain verifiable necessary and sufficient conditions for these properties that are expressed in terms of constructive generalized differential structures at reference points and are convenient for applications. Based on advanced techniques in nonsmooth analysis, new dual criteria are proven in this direction under minimal assumptions. Applications of such point conditions are given to sensitivity analysis for parametric constraint and variational systems which describe sets of feasible and optimal solutions to various optimization and related problems.

Key words. infinite-dimensional systems, stability, metric regularity, optimization, nonsmooth analysis, set-valued solution maps, generalized differentiation, sensitivity

AMS subject classifications. 49J52, 58C06, 49K40, 90C31, 58C20

PII. S0363012994278171

1. Introduction.

It has long been recognized that many principal aspects in optimization and control (e.g., optimality conditions, controllability, sensitivity, numerical methods, and so on) are related to studying stability/regularity properties of corresponding set-valued mappings (multifunctions). Such properties are known under different names (metric regularity, openness, covering, surjection, Lipschitzian stability, etc.), which are often equivalent to each other. We refer the reader to [1, 2, 3, 6, 7, 8, 9, 10, 11, 13, 14, 15, 17, 18, 19, 20, 25, 26, 27, 28, 29, 30, 31, 32, 37, 38, 40, 41, 42] and bibliographies therein for various results in this direction as well as for applications of the mentioned properties to optimization and control.

The main goal of the paper is to obtain effective characteristics and applications of such stability properties by means of appropriate generalized derivatives in nonsmooth analysis. In finite dimensions, complete *dual* characterizations of openness, metric regularity, and Lipschitzian behavior for closed-graph multifunctions were established by Mordukhovich [25, 26] on the basis of nonconvex generalized derivativelike objects developed in [24, 25]. Various applications of those results to optimization, sensitivity, and optimal control can be found in [27, 28, 29, 30, 31]. However, using only finite-dimensional characteristics, one cannot cover a broad range of significant problems arising, e.g., in optimization and sensitivity analysis of constrained control systems governed by ordinary and partial differential equations/inclusions as well as variational inequalities. This is one of our primary motivations: to develop proper infinite-dimensional extensions of the previous theory.

We have been already concerned with this topic in [32], where dual differential characterizations of stability and associated properties of multifunctions have been obtained in Banach spaces. In contrast to [26], corresponding criteria and constants in [32] are expressed not in terms of generalized differential constructions at reference

*Received by the editors December 5, 1994; accepted for publication (in revised form) November 26, 1995. This research was supported in part by National Science Foundation grants DMS-9206989 and DMS-9404128 and NATO contract CRG-950360.

http://www.siam.org/journals/sicon/35-1/27817.html

†Department of Mathematics, Wayne State University, Detroit, MI 48202 (boris@math.wayne.edu).

points but in terms of their perturbations in neigborhoods. Some results in this direction are related to those in Ioffe [15] and Kruger [20]; see [32] for more details. Note that such *neighborhood conditions* are rather complicated and generally are not very convenient for applications.

In this paper we establish effective differential *point conditions* for the fundamental properties considered under appropriate assumptions on multifunctions and spaces in question. For this purpose we use generalized differential constructions introduced by Kruger and Mordukhovich on the base of *sequential limits* of the so-called *Fréchet ε-normals* [21]. These simply defined *nonconvex* objects are infinite-dimensional extensions of the corresponding constructions [24] in finite dimensions and possess some useful properties in general Banach space settings; see [33]. However, full calculus (at the same level of perfection as in finite dimensions) is available for them in the class of *Asplund spaces*; see [34, 35, 36] and section 2 below for more details. The latter subclass of Banach spaces is sufficiently broad and includes, in particular, all spaces with Fréchet differentiable renorms (therefore, all reflexive spaces) as well as those with separable duals; see, e.g., [39].

We employ the sequential differential constructions to obtain effective point criteria for openness, stability, and related properties of infinite dimensional multifunctions under *partial normal compactness* (p.n.c.) assumptions on their graphs introduced in this paper. These assumptions support the limiting procedure to prove point criteria in terms of the sequential constructions in general settings when the latter may not even be closed. In section 3 we discuss the p.n.c. assumptions in detail and compare them with previous results in this direction.

Employing the p.n.c. assumptions, we establish a number of *sufficient* point conditions for openness, metric regularity, and stability properties of nonsmooth mappings and multifunctions between Asplund spaces with *point estimates* of the exact bounds for corresponding moduli. Moreover, those conditions are proven to be also *necessary* for the fundamental properties under consideration even without p.n.c. and Asplundity assumptions. The conditions obtained completely cover the finite-dimensional case in [26] and turn out to be useful in many infinite-dimensional settings important for applications to optimization, sensitivity, control, etc. In particular, in the last section we provide some effective applications of the main point criteria to *sensitivity analysis* of constraint and variational systems.

One of the principal results in the paper is a sufficient point condition for both openness and metric regularity properties of p.n.c. multifunctions $\Phi : X \rightrightarrows Y$ between Asplund spaces, which is obtained in the form

$$[(0, y^\star) \in N((\bar{x}, \bar{y}); \operatorname{gph} \Phi)] \Longrightarrow y^\star = 0, \tag{1.1}$$

where $N((\bar{x}, \bar{y}); \operatorname{gph} \Phi)$ is our basic sequential normal cone in Definition 2.1(ii) to the graph

$$\operatorname{gph} \Phi := \{(x, y) \in X \times Y |\ y \in \Phi(x)\}$$

at $(\bar{x}, \bar{y})$; see Theorem 4.2. We also establish the necessity of (1.1) for the mentioned properties and provide two-sided point estimates for corresponding moduli.

These results were first obtained and presented at the workshop "Convexity, Monotonicity, and Differentiability" (Waterloo, Ontario, Canada, March 1993) under the assumption that the graph of Φ is compactly epi-Lipschitzian [5]. Then we received the paper [18] by Jourani and Thibault, which contains another proof of the fact that an analogue of (1.1) is a sufficient condition for metric regularity of Φ in

general Banach spaces under the less restrictive "partial compact epi-Lipschitzian" assumption. The latter one implies the p.n.c. assumption imposed here and actually has a different nature: primal vs. dual; see section 3.

Note that, in contrast to (1.1), the point condition in [18] is expressed in terms of an approximate normal cone defined by Ioffe, who proved a similar sufficient condition for the case of $\dim Y < \infty$ [15]. The latter construction is also an infinite-dimensional generalization of that in [24], which is well defined in any Banach space [16]. On the other hand, it is more complicated and may be bigger (but never smaller) than the Kruger–Mordukhovich sequential extension in infinite dimensions (see [35, section 9] for more details). Moreover, in the case of Asplund spaces, approximate normals used in [16, 18] can be obtained by adding points of the weak-star *topological closure* to the sequential normal cone in (1.1). One can observe that such a procedure may substantially enlarge the set in the left-hand side of (1.1).

The rest of the paper is organized as follows. Section 2 contains basic definitions and preliminaries in nonsmooth analysis needed in later discussion. Section 3 is concerned with normal compactness conditions important for limiting procedures. In section 4 we obtain the main point characterizations of openness properties for set-valued mappings with corresponding modulus estimates. Section 5 is devoted to point criteria for metric regularity and Lipschitzian stability. In section 6 we present some applications of the main criteria to sensitivity analysis of constraint and variational systems.

Throughout the paper we use standard notation, with some special symbols introduced where they are defined. Unless otherwise stated, all spaces considered are *Banach* whose norms are always denoted by $\|\cdot\|$. For any space X we consider its dual space $X^\star$ equipped with the weak-star topology $w^\star$, where $\langle\cdot,\cdot\rangle$ means the canonical pairing. Recall that cl Ω stands for the closure of $\Omega \subset X$, while cl* is used for the weak-star topological closure in $X^\star$. The *distance function* to the set Ω is denoted by

$$\operatorname{dist}(x,\Omega) := \inf\{\|x-\omega\| \ \text{ s.t. } \ \omega \in \Omega\}.$$

In contrast to the case of *single-valued mappings* $f : X \to Y$, the symbol $\Phi : X \rightrightarrows Y$ stands for a *multifunction* from X into Y with the *domain* and *kernel* denoted, respectively, by

$$\operatorname{Dom}\Phi := \{x \in X |\ \Phi(x) \neq \emptyset\} \ \text{ and } \ \operatorname{Ker}\Phi := \{x \in X |\ 0 \in \Phi(x)\}.$$

The *inverse* multifunction $\Phi^{-1} : Y \rightrightarrows X$ to Φ satisfies the relationships

$$x \in \Phi^{-1}(y) \Longleftrightarrow y \in \Phi(x) \Longleftrightarrow (x,y) \in \operatorname{gph}\Phi,$$

and the *norm* of any positive homogeneous multifunction is defined by

$$\|\Phi\| := \sup\{\|y\| \ \text{ s.t. } \ y \in \Phi(x) \ \text{ and } \ \|x\| \le 1\}. \tag{1.2}$$

For multifunctions $\Phi : X \rightrightarrows X^\star$, the expression

$$\limsup_{x\to\bar{x}} \Phi(x) := \{x^\star \in X^\star |\ \exists \text{ sequences } \ x_k \to \bar{x} \ \text{ and } \ x_k^\star \overset{w^\star}{\to} x^\star \ \text{ with } \ x_k^\star \in \Phi(x_k) \ \text{ for all } \ k = 1,2,\ldots\}$$

always means the *sequential* Kuratowski–Painlevé upper limit with respect to the norm topology in X and the weak-star topology in $X^\star$.

As usual, we denote by B and $B^\star$ the unit closed balls in the space and dual space in question; $A^\star$ stands for the adjoint operator to a linear bounded operator $A : X \to Y$. The symbol $B_r(x)$ means the closed ball with center x and radius r, while

$$B_r(\Phi(x)) := \bigcup_{y\in\Phi(x)} B_r(y) \ \text{ and } \ \Phi(B_r(x)) := \bigcup_{z\in B_r(x)} \Phi(z)$$

for any multifunction $\Phi : X \rightrightarrows Y$.

If $\varphi : X \to \bar{\mathbf{R}} := [-\infty, \infty]$ is an *extended-real-valued function*, then

$$\operatorname{dom}\varphi := \{x \in X \ \text{ with } \ |\varphi(x)| < \infty\} \ \text{ and } \ \operatorname{epi}\varphi := \{(x,\mu) \in X \times \mathbf{R} |\ \mu \ge \varphi(x)\}.$$

In this case, limsup $\varphi(x)$ and liminf $\varphi(x)$ denote the upper and lower limits in the classical (scalar) sense. Depending on context, the symbols $x \overset{\varphi}{\to} \bar{x}$ and $x \overset{\Omega}{\to} \bar{x}$ mean, respectively, that $x \to \bar{x}$ with $\varphi(x) \to \varphi(\bar{x})$ and $x \to \bar{x}$ with $x \in \Omega$. Thoughout the paper we use the convention that $\inf\emptyset = \infty$, $\sup\emptyset = -\infty$, $\|\emptyset\| = \infty$, and $a + \emptyset = \emptyset + b = \emptyset$ for any elements a and b.

2. Basic definitions and preliminaries. This section is mostly concerned with preliminary material on the basic generalized differential constructions used in the paper. Let us begin with the definitions of normal elements to arbitrary sets in Banach spaces, as appeared in [21].

DEFINITION 2.1. (i) *Let $\Omega \subset X$ and $\varepsilon \ge 0$. Given $x \in \operatorname{cl}\Omega$, the nonempty set*

$$\hat{N}_\varepsilon(x;\Omega) := \left\{ x^\star \in X^\star |\ \limsup_{u \overset{\Omega}{\to} x} \frac{\langle x^\star, u - x\rangle}{\|u - x\|} \le \varepsilon \right\} \tag{2.1}$$

is called the set of (Fréchet) ε-normals *to Ω at x. When $\varepsilon = 0$, the set* (2.1) *is a cone which is called the* prenormal cone *or* Fréchet normal cone *to Ω at x and is denoted by $\hat{N}(x;\Omega)$. If $x \notin \operatorname{cl}\Omega$, we set $\hat{N}_\varepsilon(x;\Omega) = \emptyset$ for all $\varepsilon \ge 0$.*

(ii) *Let $\bar{x} \in \operatorname{cl}\Omega$. The nonempty cone*

$$N(\bar{x};\Omega) := \limsup_{x\to\bar{x},\ \varepsilon\downarrow 0} \hat{N}_\varepsilon(x;\Omega) \tag{2.2}$$

is called the normal cone *to Ω at $\bar{x}$. We set $N(\bar{x};\Omega) = \emptyset$ for $\bar{x} \notin cl\,\Omega$.*

(iii) *The set Ω is called* regular *at $\bar{x}$ if $N(\bar{x};\Omega) = \hat{N}(\bar{x};\Omega)$.*

The reader may consult with the recent papers [33, 35] and their references for basic properties of the normal cone (2.2) and related subdifferential constructions. It is shown in [35] that for Asplund spaces X one can always let $\varepsilon = 0$ in (2.2); i.e., the normal cone is represented by

$$N(\bar{x};\Omega) = \limsup_{x\to\bar{x}} \hat{N}(x;\Omega). \tag{2.3}$$

Let us mention [35] that, for the case of Asplund spaces, the weak-star *topological closure* in $X^\star$ of the sequential normal cone (2.2) coincides with the G-normal cone in Ioffe [16], while the weak-star closure of its *convexification* gives Clarke's normal cone [9]. Note also that the normal cone (2.2) coincides with the cone of "limiting proximal normals" when X is Hilbert; cf. [23].

Now we consider a derivativelike object for *multifunctions* which is used for formulations and proofs of the main results in the paper.

DEFINITION 2.2. *Let $\Phi : X \rightrightarrows Y$ be a multifunction between Banach spaces and $(\bar{x}, \bar{y}) \in \text{cl gph } \Phi$. The multifunction $D^\star\Phi(\bar{x}, \bar{y})$ from $Y^\star$ into $X^\star$ defined by*

$$D^\star\Phi(\bar{x}, \bar{y})(y^\star) := \{x^\star \in X^\star | \ (x^\star, -y^\star) \in N((\bar{x}, \bar{y}); \text{gph } \Phi)\} \tag{2.4}$$

is called the coderivative *of Φ at $(\bar{x}, \bar{y})$. We let $D^\star\Phi(\bar{x}, \bar{y})(y^\star) := \emptyset$ if $(\bar{x}, \bar{y}) \notin \text{cl gph } \Phi$. The symbol $D^\star\Phi(\bar{x})(y^\star)$ is used in* (2.4) *when Φ is single valued at $\bar{x}$ and $\bar{y} = \Phi(\bar{x})$.*

The coderivative introduced can be treated as a generalized concept of *adjoint mapping* that coincides with the adjoint operator to the derivative in the classical framework. Indeed, let $\Phi = f : X \to Y$ be *strictly differentiable* at $\bar{x}$ with the derivative $f'(\bar{x})$, i.e.,

$$\lim_{x \to \bar{x}, u \to \bar{x}} \frac{f(x) - f(u) - f'(\bar{x})(x - u)}{\|x - u\|} = 0.$$

Then it is easy to show (see [36]) that for any Banach spaces X and Y one has

$$D^\star f(\bar{x})(y^\star) = (f'(\bar{x}))^\star y^\star. \tag{2.5}$$

In general, the coderivative (2.4) is a positive homogeneous multifunction with respect to $y^\star$ whose values may be nonconvex and even not closed in $X^\star$. Nevertheless, this construction possesses a rich calculus, especially in the Asplund space setting. We refer the reader to our paper [36], which contains complete infinite-dimensional extensions of the previous finite-dimensional results in [28].

Now let us consider subdifferential constructions for *extended-real-valued functions* related to the normal and prenormal cones in Definition 2.1.

DEFINITION 2.3. *Let $\varphi : X \to \bar{\mathbf{R}}$ and $\bar{x} \in \text{dom } \varphi$. The set*

$$\partial\varphi(\bar{x}) := \{x^\star \in X^\star | \ (x^\star, -1) \in N((\bar{x}, \varphi(\bar{x})); \text{epi } \varphi)\} \tag{2.6}$$

is called the subdifferential *of φ at $\bar{x}$, while the set*

$$\hat{\partial}\varphi(\bar{x}) := \{x^\star \in X^\star | \ (x^\star, -1) \in \hat{N}(\bar{x}, \varphi(\bar{x})); \text{epi } \varphi)\} \tag{2.7}$$

is called the presubdifferential *or* Fréchet subdifferential *of φ at this point. We let $\partial\varphi(\bar{x}) = \hat{\partial}\varphi(\bar{x}) := \emptyset$ when $\bar{x} \notin \text{dom } \varphi$.*

When φ is convex, both the subdifferential and the presubdifferential coincide with the subdifferential of convex analysis. In general, the set (2.7) is always convex but is frequently empty (as, e.g., for $\varphi(x) = -|x|$ at $0 \in \mathbf{R}$), while the subdifferential (2.6) is nonempty at least for locally Lipschitzian functions but may be nonconvex in common situations (as in the example above). Note that when φ is *lower semicontinuous* (l.s.c.) around $\bar{x} \in \text{dom } \varphi$, one has

$$\hat{\partial}\varphi(\bar{x}) = \left\{x^\star \in X^\star | \ \liminf_{x \to \bar{x}} \frac{\varphi(x) - \varphi(\bar{x}) - \langle x^\star, x - \bar{x}\rangle}{\|x - \bar{x}\|} \geq 0\right\}$$

for any Banach space X and

$$\partial\varphi(\bar{x}) = \limsup_{x \xrightarrow{\varphi} \bar{x}} \hat{\partial}\varphi(x) \tag{2.8}$$

if X is Asplund; see [35]. One can easily check that

$$\hat{\partial}\delta(\bar{x}, \Omega) = \hat{N}(\bar{x}; \Omega) \quad \text{and} \quad \partial\delta(\bar{x}, \Omega) = N(\bar{x}; \Omega) \tag{2.9}$$

for $\bar{x} \in \Omega$, where $\delta(\cdot, \Omega)$ is the *indicator function* of Ω.

The next "zero fuzzy calculus" result for Fréchet subdifferentials in Asplund spaces follows from Fabian [12]. We prove in [34] that such a calculus rule is equivalent to the *extremal principle* (generalized Euler equation for extremal points) [21, 25].

PROPOSITION 2.4. *Let X be an Asplund space and $\varphi_i : X \to \bar{\mathbf{R}}$, $i = 1, \ldots, n$ ($n \geq 2$), be a collection of l.s.c. functions, all but one of which are Lipschitz continuous around $\bar{x}$. Assume that $\bar{x}$ is a local minimum point for the sum $\varphi_1 + \cdots + \varphi_n$. Then for any positive numbers ε and γ one has*

$$0 \in \bigcup\{\hat{\partial}\varphi_1(x_1) + \cdots + \hat{\partial}\varphi_n(x_n) |\ x_i \in B_\gamma(\bar{x}),\ |\varphi_i(x_i) - \varphi_i(\bar{x})| \leq \gamma,\ i = 1, \ldots, n\} + \varepsilon B^\star.$$

In conclusion of this section we discuss relations between the coderivative (2.4) and the subdifferential (2.6). It follows from the definitions that

$$\partial\varphi(\bar{x}) = D^\star E_\varphi(\bar{x}, \varphi(\bar{x}))(1), \quad \text{where} \quad E_\varphi(x) := \{\mu \in \mathbf{R} |\ \mu \geq \varphi(x)\} \tag{2.10}$$

for any $\varphi : X \to \bar{\mathbf{R}}$ and $\bar{x} \in \operatorname{dom}\varphi$. Moreover, $\partial\varphi(\bar{x}) = D^\star\varphi(\bar{x})(1)$ if φ is continuous around $\bar{x}$. On the other hand, for some classes of single-valued mappings f between Banach spaces, the coderivative of f can be expressed in terms of the subdifferential of its *Lagrange scalarization* $\langle y^\star, f\rangle(x) := \langle y^\star, f(x)\rangle$.

Recall that a mapping $f : X \to Y$ Lipschitz continuous around $\bar{x}$ is said to be *strictly Lipschitzian* at $\bar{x}$ [35] if there exists a neighborhood V of the origin in X such that the sequence

$$[f(x_k + t_k v) - f(x_k)]/t_k, \quad k = 1, 2, \ldots,$$

has a convergent subsequence in the norm topology of Y for each $v \in V$, $x_k \to \bar{x}$, and $t_k \downarrow 0$ as $k \to \infty$. (As we were recently informed by Lionel Thibault, this definition is equivalent to a variant of his concept of compactly Lipschitzian mappings [43].)

Obviously, every mapping strictly differentiable at $\bar{x}$ is strictly Lipschitzian at this point. When $\dim Y < \infty$, there is no difference between locally Lipschitzian and strictly Lipschitzian mappings. Furthermore, any locally Lipschitzian mapping between Banach spaces is strictly Lipschitzian at $\bar{x}$ if it has a norm-compact-valued "strict prederivative" in the sense of Ioffe [16] that includes many important applications, particularly in optimal control [13]. The following scalarization result is proven in [35].

PROPOSITION 2.5. *Let X and Y be Asplund and Banach spaces, respectively, and let $f : X \to Y$ be strictly Lipschitzian at $\bar{x}$. Then one has*

$$D^\star f(\bar{x})(y^\star) = \partial\langle y^\star, f\rangle(\bar{x}) \neq \emptyset \ \ \forall y^\star \in Y^\star.$$

3. Normal compactness conditions. In this section we study effective conditions on sets and multifunctions that are widely used to justify limiting procedures in proving the main results of the paper.

Let us start with arbitrary closed sets Ω in a Banach space Z, which will be mainly considered in the form $X \times Y$ later on. When Z is finite dimensional, the normal cone (2.2) at $\bar{z} \in \Omega$ has closed values, and moreover, its graph is closed. (It is the so-called *robustness* or upper semicontinuity property.) In infinite dimensions, these facts are no longer true both in topological and in sequential senses. This appears because the normal cone is defined *sequentially*, while the weak-star topology of $Z^\star$ is not always sequential; see [4, 22, 23] for more discussions. To overcome these difficulties, Loewen introduced in [22] a local compactness condition of the following kind, which we call here *normal compactness* of a set around a given point.

DEFINITION 3.1. *A closed set $\Omega \subset Z$ is said to be* normally compact *around $\bar{z} \in \Omega$ if there exist positive numbers γ, σ, and a compact subset C of Z such that*

$$\hat{N}(z;\Omega) \subset K_\sigma(C) := \left\{ z^\star \in Z^\star \ \ s.t. \ \ \sigma\|z^\star\| \le \max_{c\in C} |\langle z^\star, c\rangle| \right\} \ \ \forall z \in B_\gamma(\bar{z}) \cap \Omega. \tag{3.1}$$

Note that condition (3.1) is always valid if Z is finite dimensional. When Z is Asplund, the prenormal cone $\hat{N}(z;\Omega)$ in Definition 3.1 can be equivalently replaced by the normal cone (2.2). This follows from (2.3) and the fact that the cone $K_\sigma(C)$ in (3.1) is weak-star closed in $Z^\star$; see the proof of Proposition 3.4 below.

Let us observe that condition (3.1) is *dual* in the sense that it is expressed in the dual space setting. Loewen shows [22] that (3.1) always holds when Ω is *compactly epi-Lipschitzian* around $\bar{x}$ in the sense of Borwein and Strojwas [5]. The latter is a primal condition on the set Ω that generalizes Rockafellar's original concept of epi-Lipschitzian sets; see [5, 22].

In [22], Loewen demonstrates that the graph of the normal-cone multifunction $N(\cdot;\Omega)$ is closed near $\bar{z}$ in the norm$\times$weak-star topology of $Z \times Z^\star$ if Ω satisfies the normal compactness condition (3.1) around $\bar{z}$. He establishes this fact for the case of reflexive spaces Z that is essential in his proof. Let us obtain this robustness property of (2.2) in a more general setting based on recent developments in Borwein and Fitzpatrick [4].

Recall that a Banach space Z is *weakly compactly generated* (WCG), provided that there is a weakly compact set K such that $Z = \mathrm{cl}(\mathrm{span}\, K)$. This class of spaces includes, in particular, all reflexive spaces as well as all separable Banach spaces; see [39]. We will use the following result proven in [4].

PROPOSITION 3.2. *Let Z be a WCG space and $\{A_k\}$ be a sequence of subsets of $Z^\star$ such that $A_{k+1} \subset A_k$ for all $k = 1, 2, \ldots$. Then one has*

$$\bigcup_{m=1}^{\infty} \bigcap_{k=1}^{\infty} \mathrm{cl}^\star(A_k \cap mB^\star) = \left\{ \lim_{k\to\infty} z_k^\star |\ z_k^\star \in A_k \ \text{ for all } \ k \right\}, \tag{3.2}$$

where $\lim z_k^\star$ is taken in the weak-star topology of $Z^\star$.

Now we can establish extensions of Loewen's main results in [22]. Say that a set $K \subset Z^\star$ is *weak-star locally bounded*, provided that each point of K has a weak-star neighborhood U such that $U \cap K$ is norm bounded in $Z^\star$.

PROPOSITION 3.3. *Let $(M,\ \rho)$ be a metric space, Z be a WCG space, and $\Phi : M \rightrightarrows Z^\star$ be an arbitrary multifunction. Equip $M \times Z^\star$ with the $\rho\times$weak-star topology, and assume that there exists a weak-star locally bounded, weak-star closed subset K of $Z^\star$ such that*

$$\Phi(z) \subset K \ \ \text{for any} \ \ z \in M. \tag{3.3}$$

Then $(\bar{z}, z^\star) \in \mathrm{cl\ gph}\ \Phi$ if and only if $z^\star = \lim_{k\to\infty} z_k^\star$ for some sequence $z_k^\star \in \Phi(z_k)$ with $z_k \to \bar{z}$ as $k \to \infty$.

Proof. Let $\{(z_\alpha, z_\alpha^\star)\}_{\alpha\in\Lambda} \subset M \times Z^\star$ be a net such that $z_\alpha \to \bar{z}$ and $z_\alpha^\star \overset{w^\star}{\to} z^\star$ with $z_\alpha^\star \in \Phi(z_\alpha)$ for all $\alpha \in \Lambda$. The weak-star closedness of K and condition (3.3) ensure $z^\star \in K$. Now taking into account the weak-star local boundedness of K, we find a natural number m and a subnet $\{(z_\beta, z_\beta^\star)\}_{\beta\in\tilde{\Lambda}}$ ($\tilde{\Lambda} \subset \Lambda$) of $\{(z_\alpha, z_\alpha^\star)\}_{\alpha\in\Lambda}$ such that $\|z_\beta^\star\| \le m$ for all $\beta \in \tilde{\Lambda}$. Letting

$$A_k = \bigcup \{\Phi(z) |\ \rho(z,\bar{z}) \le 1/k\}, \ \ k = 1, 2, \ldots,$$

one concludes that $z^\star$ belongs to the left-hand side set in (3.2). Finally employing Proposition 3.2, we finish the proof. □

The result obtained ensures the following *robustness property* of the normal cone (2.2).

PROPOSITION 3.4. *Let Z be a WCG Asplund space and Ω be a closed subset of Z satisfying the normal compactness condition* (3.1) *around $\bar{z} \in \Omega$. Then the multifunction $N(\cdot;\Omega)$ has closed graph near $\bar{z}$; i.e., for some $\gamma > 0$ the set*

$$(\text{gph } N(\cdot;\Omega)) \cap ((\bar{z} + \gamma B) \times Z^\star) \tag{3.4}$$

is closed in the norm×weak-star topology of $Z \times Z^\star$. In particular, for any sequences $z_k \to \bar{z}$ and $z_k^\star \overset{w^\star}{\to} z^\star$ with $z_k^\star \in N(z_k;\Omega)$, $k = 1, 2, \ldots$, one has $z^\star \in N(\bar{z};\Omega)$.

Proof. Following the proof in [22, Proposition 3.5], one can conclude that the cone $K_\sigma(C)$ in (3.1) is both weak-star closed and weak-star locally bounded in $Z^\star$. The only change we need to do is to observe that the set

$$\{z^\star \in Z^\star |\ \sigma\|z^\star\| - |\langle z^\star, c\rangle| \le 0\}$$

is weak-star closed in $Z^\star$ for any $c \in Z$. The latter follows directly from the well-known lower semicontinuity of the norm function $\|z^\star\|$ and continuity of the function $|\langle z^\star, c\rangle|$ in the weak-star topology of $Z^\star$.

Now applying Proposition 3.3 with $(M,\rho) = (\Omega \cap B_\gamma(\bar{z}), \|\cdot\|_Z)$ and $\Phi(\cdot) = \hat{N}(\cdot;\Omega)$, we conclude that the topological closure of gph $\hat{N}(\cdot;\Omega)$ in $M \times Z^\star$ coinsides with its sequential closure. Due to (2.3) the latter set is equal to the graph of $N(\cdot;\Omega)$ near $\bar{z}$. This proves the closedness of the set (3.4) in the norm×weak-star topology of $Z \times Z^\star$. □

Remarks 3.5. (i) The robustness property in Proposition 3.4 does not hold, and moreover, values of the normal cone $N(\cdot;\Omega)$ may *not* be closed in the weak-star topology of $Z^\star$ if one has all the assumptions of the theorem except that Z is a WCG space. This happens, in particular, for epigraphical sets $\Omega = \text{epi}\,\varphi \subset X \times \mathbf{R}$ generated by *Lipschitz* continuous functions φ on spaces X with Fréchet differentiable renorms; see examples in Borwein and Fitzpatrick [4].

(ii) Recently Fitzpatrick showed (personal communication) that the normal cone (2.2) may not be closed even in the *norm* topology of the *Hilbert* space l^2.

(iii) It is worth mentioning that the normal compactness assumption (3.1) has been employed in [35] to establish important calculus results for the normal cone (2.2) to general closed sets in arbitrary Asplund (not just WCG) spaces. What turns out to be essential for those purposes is *not* robustness of (2.2) but the following *limiting property*, which is easily implied by (3.1): for any sequences

$$z_k^\star \in \hat{N}(z_k;\Omega) \ \text{ with } \ z_k \to \bar{z} \ \text{ and } \ z_k^\star \overset{w^\star}{\to} 0 \ \text{ as } \ k \to \infty$$

one has $z_k^\star \to 0$ in the norm topology of $Z^\star$.

Now let us consider normal compactness conditions for *multifunctions* $\Phi : X \rightrightarrows Y$ between Banach spaces. It is clear that the case of multifunctions can be reduced to the case of sets in the space $Z = X \times Y$ taking $\Omega = \text{gph}\,\Phi$. According to Definition 3.1, a closed-graph multifunction Φ is said to be *normally compact* around $(\bar{x},\bar{y}) \in \text{gph}\,\Phi$ if there are positive numbers γ and σ as well as compact sets $P \subset X$ and $S \subset Y$ such that

$$\sigma(\|x^\star\| + \|y^\star\|) \le \max_{p\in P} |\langle x^\star, p\rangle| + \max_{s\in S} |\langle y^\star, s\rangle| \tag{3.5}$$

for any $(x^\star, y^\star) \in X^\star \times Y^\star$ satisfying

$$(x^\star, y^\star) \in \hat{N}((x,y); \operatorname{gph} \Phi) \text{ with } (x,y) \in [B_\gamma(\bar{x}) \times B_\gamma(\bar{y})] \cap \operatorname{gph} \Phi. \tag{3.6}$$

It follows from the previous discussions that the normal compactness of Φ around $(\bar{x}, \bar{y})$ implies that for any sequence $(x_k, y_k, x_k^\star, y_k^\star)$ with

$$(x_k^\star, y_k^\star) \in \hat{N}((x_k, y_k); \operatorname{gph} \Phi), \ (x_k, y_k) \to (\bar{x}, \bar{y}), \text{ and } (x_k^\star, y_k^\star) \overset{w^\star}{\to} (0,0) \tag{3.7}$$

one has $\|x_k^\star\| \to 0$ and $\|y_k^\star\| \to 0$ as $k \to \infty$.

One can observe that for many applications including those in this paper, it is sufficient that (3.7) implies the norm convergence of only *one* component (depending on the context). This allows us to weaken the normal compactness condition (3.5) in the following way (considering for definiteness the case of the y-component): there exist positive numbers γ, σ and compact sets $P \subset X$, $S \subset Y$ such that

$$\sigma \|y^\star\| \le \max_{p \in P} |\langle x^\star, p \rangle| + \max_{s \in S} |\langle y^\star, s \rangle| \tag{3.8}$$

for any $(x^\star, y^\star)$ satisfying (3.6). Moreover, requirement (3.7) can be also relaxed to achieve our purposes in this paper. Namely, we need $\|y_k^\star\| \to 0$ not for all $(x_k^\star, y_k^\star) \overset{w^\star}{\to} (0,0)$ satisfying the inclusion in (3.7) but only for those with $\|x_k^\star\| \to 0$. Let us introduce an *intrinsic* dual condition on Φ that generalizes the above normal compactness and ensures the required limiting behavior (see Proposition 3.8).

DEFINITION 3.6. (i) *A multifunction $\Phi : X \rightrightarrows Y$ of closed graph is called* partially normally compact (p.n.c.) with respect to y (image) *around $(\bar{x}, \bar{y}) \in \operatorname{gph} \Phi$ if there exist a weak-star closed subspace $L^\star \subset Y^\star$ of finite codimension, positive numbers γ and σ, and a compact set $S \subset Y$ such that*

$$\|x^\star\| + \max_{s \in S} |\langle y^\star, s \rangle| \ge \sigma \tag{3.9}$$

for any $(x^\star, y^\star)$ satisfying (3.6) *with*

$$\|y^\star\| = 1 \text{ and } \operatorname{dist}(y^\star, L^\star) \le \gamma. \tag{3.10}$$

(ii) *Φ is said to be* p.n.c. with respect to x (domain) *around $(\bar{x}, \bar{y})$ if the inverse multifunction $\Phi^{-1} : Y \rightrightarrows X$ is p.n.c. with respect to its image.*

Remark 3.7. In contrast of the normal compactness conditions discussed previously, the p.n.c. conditions in Definition 3.6 may become stronger (even in the case of Asplund spaces) if the prenormal cone in (3.6) is replaced by the basic normal cone $N((x,y); \operatorname{gph} \Phi)$ related to the coderivative (2.4). However, an important advantage of the latter modification is a *rich calculus* available for this coderivative in infinite dimensions; see [36]. Based on such a calculus, one can employ the p.n.c. conditions for various combinations of nonsmooth mappings and multifunctions.

In the rest of this section we deal with the p.n.c. condition for Φ with respect to its image. The next proposition establishes the mentioned *sequential limiting property* of p.n.c. multifunctions, which is what we need to prove the main results of the paper.

PROPOSITION 3.8. *Let $\Phi : X \rightrightarrows Y$ be p.n.c. with respect to y around $(\bar{x}, \bar{y})$. Then any sequence $(x_k, y_k, x_k^\star, y_k^\star)$ satisfying*

$$(x_k^\star, y_k^\star) \in \hat{N}((x_k, y_k); \operatorname{gph} \Phi), \ (x_k, y_k) \to (\bar{x}, \bar{y}), \ \|x_k^\star\| \to 0, \text{ and } y_k^\star \overset{w^\star}{\to} 0 \tag{3.11}$$

as $k \to \infty$ contains a subsequence with $\|y_{k_m}^\star\| \to 0$ as $m \to \infty$.

Proof. Let $E^\star \subset Y^\star$ be a finite-dimensional subspace complementary to $L^\star$ from Definition 3.6(i). Taking $y_k^\star$ in (3.11), we have the unique representation

$$y_k^\star = l_k^\star + e_k^\star \ \text{ with } \ l_k^\star \in L^\star \ \text{ and } \ e_k^\star \in E^\star \ \ \forall k = 1, 2, \ldots.$$

Since $y_k^\star \overset{w^\star}{\to} 0$ and $\dim E^\star < \infty$, one gets $l_k^\star \overset{w^\star}{\to} 0$ and $\|e_k^\star\| \to 0$ as $k \to \infty$.

Proving by contradiction, we assume that the result of the proposition does not hold. In this case there is $\alpha > 0$ such that $\|y_k^\star\| \ge \alpha$ for all k. Then taking $\gamma > 0$ in Definition 3.6(i), one derives from the inclusion in (3.11) that

$$(\tilde{x}_k^\star, \tilde{y}_k^\star) \in \hat{N}((x_k, y_k); \operatorname{gph} \Phi) \ \text{ with } \ \|\tilde{y}_k^\star\| = 1 \ \text{ and } \ (x_k, y_k) \in [B_\gamma(\bar{x}) \times B_\gamma(\bar{y})] \cap \operatorname{gph} \Phi,$$

where $\tilde{x}_k^\star := x_k^\star / \|y_k^\star\|$ and $\tilde{y}_k^\star := y_k^\star / \|y_k^\star\|$ for all k. It follows from

$$\operatorname{dist}(\tilde{y}_k^\star, L^\star) \le \|y_k^\star\|^{-1} \|e_k^\star\| \le \|e_k^\star\| / \alpha \to 0 \ \text{ as } \ k \to \infty$$

that $\operatorname{dist}(\tilde{y}_k^\star, L^\star) \le \gamma$ for k sufficiently large. Employing (3.9), we arrive at

$$\|\tilde{x}_k^\star\| + \max_{s \in S} |\langle \tilde{y}_k^\star, s \rangle| \ge \sigma,$$

which implies the estimate

$$\|y_k^\star\| \le \sigma^{-1} \left(\|x_k^\star\| + \max_{s \in S} |\langle y_k^\star, s \rangle| \right). \tag{3.12}$$

Using the compactness of the set S in (3.12), we conclude that $\langle y_k^\star, s \rangle \to 0$ *uniformly* in $s \in S$. Therefore, (3.12) implies that $\|y_k^\star\| \to 0$ as $k \to \infty$. The contradiction obtained ends the proof of the proposition. □

Now we present some sufficient conditions ensuring the p.n.c. property of Φ in Definition 3.6(i). First let us observe that this property always holds (with $L^\star = \{0\}$) when Y is *finite dimensional*. One can also check that the previously discussed condition (3.8) implies the p.n.c. property with $L^\star = Y^\star$.

Next let us consider an important class of multifunctions generated by a single-valued mapping $f : X \to Y$ and closed sets $\Lambda,\ \Omega$ in the form

$$\Phi(x) := \begin{cases} f(x) + \Lambda & \text{if } x \in \Omega, \\ \emptyset & \text{otherwise,} \end{cases} \tag{3.13}$$

which frequently appears in applications to optimization and related problems. The following result provides the p.n.c. condition for (3.13) in the general case of Banach spaces X and Y.

PROPOSITION 3.9. *Let $f : X \to Y$ be Lipschitz continuous around $\bar{x} \in \Omega$ and $\Lambda \subset Y$ be normally compact around $\bar{y} := -f(\bar{x}) \in \Lambda$. Then multifunction* (3.13) *is p.n.c. with respect to the image around $(\bar{x}, 0)$.*

Proof. According to Definition 3.1 we can find positive numbers $\sigma,\ \gamma$ and a compact set $S \subset Y$ such that

$$\sigma \|y^\star\| \le \max_{s \in S} |\langle y^\star, s \rangle| \ \ \forall y^\star \in \hat{N}(v; \Lambda) \ \text{ with } \ v \in B_\gamma(\bar{y}) \cap \Lambda. \tag{3.14}$$

Let $l \ge 0$ be a Lipschitz modulus of f in some (fixed) neighborhood of $\bar{x}$. Consider any pair $(x^\star, y^\star)$ satisfying

$$(x^\star, y^\star) \in \hat{N}((x, y); \operatorname{gph} \Phi) \ \text{ with } \ \|x - \bar{x}\| \le \gamma/(l+1), \ \|y\| \le \gamma/(l+1) \tag{3.15}$$

for Φ defined in (3.13). We are going to show that (3.15) implies the existence of $v \in B_\gamma(y) \cap \Lambda$ such that $y^\star \in \hat{N}(v;\Lambda)$. In this way we obtain condition (3.8) for multifunction (3.13) around $(\bar{x},0)$ with the given compact set $S \subset Y$ and $P = \{0\}$. The latter automatically ensures the p.n.c. property of Φ under consideration.

To justify the mentioned fact, let us employ the definition of the prenormal cone in (3.15). For any $\varepsilon > 0$ we find $\eta > 0$ such that $B_{2\eta}(\bar{x})$ lies in the given neighborhood of $\bar{x}$ where f is locally Lipschitzian, and moreover,

$$\langle x^\star, u - x\rangle + \langle y^\star, w - y\rangle \le \varepsilon(\|u - x\| + \|w - y\|)$$

when $\|u - x\| \le \eta$ and $\|w - y\| \le \eta$ with $w \in \Phi(u)$ and $u \in \Omega$. Letting $u = x$ in the latter formula and taking (3.13) into account, one has

$$\langle y^\star, w - y\rangle \le \varepsilon\|w - y\| \ \ \forall w \in B_\eta(y) \cap [f(x) + \Lambda] \tag{3.16}$$

with $x \in B_{\gamma/(l+1)}(\bar{x}) \cap \Omega$ and $\|y\| \le \gamma/(l+1)$. Now for any such (x,y) we consider $v := y - f(x) \in \Lambda$ which satisfies

$$\|v - \bar{y}\| \le \|y\| + l\|x - \bar{x}\| \le \gamma,$$

i.e., $v \in B_\gamma(\bar{y}) \cap \Lambda$. Letting $w = f(x) + \vartheta$ in (3.16), we get from here that

$$\langle y^\star, \vartheta - v\rangle \le \varepsilon\|\vartheta - v\| \ \ \forall \vartheta \in B_\eta(v) \cap \Lambda.$$

The latter implies that $y^\star \in \hat{N}(v;\Lambda)$ and $\sigma\|y^\star\| \le \max_{s\in S}\langle y^\star, s\rangle$ due to (3.14). This completes the proof of the proposition. □

An important special case of multifunctions (3.13) occurs when $\Lambda = \{0\}$; i.e., one considers in fact a single-valued mapping defined on a closed set. It turns out that Proposition 3.9 cannot be employed in this case for $\dim Y = \infty$ since the set $\Lambda = \{0\}$ is *not* normally compact in infinite dimensions. The case of (3.13) with $\Lambda = \{0\}$ was studied in Ioffe [15] and Ginsburg and Ioffe [13] on the base of the so-called *finite codimension property* for f with respect to Ω. Such a property provides sufficient amounts of compactness to obtain point criteria in the case considered. Moreover, it appears to be inherent in a broad class of *Fredholm operators* important for applications in optimal control; see [13, 15] for more details and discussions.

Two versions of the finite codimension property were introduced in [13, 15] in terms of topologically closed approximate subdifferentials and normal cones. Now we consider a *sequential* version of this property in terms of the basic constructions in section 2.

DEFINITION 3.10. *Let Ω be a closed subset of X with $\bar{x} \in \Omega$ and $f : X \to Y$ be a single-valued mapping continuous around $\bar{x}$. We say that f has the* finite codimension property *with respect to Ω around $\bar{x}$ if there exist a weak-star closed subspace $L^\star \subset Y^\star$ of finite codimension and positive numbers γ, σ such that for any $x \in B_\gamma(\bar{x})$, $u^\star \in N(x;\Omega)$, and $x^\star \in D^\star f(x)(y^\star)$ with $y^\star$ satisfying (3.10), one has $\|x^\star + u^\star\| \ge \sigma$.*

Let us show that, in the case of Asplund spaces X and Y, this finite codimension property implies the p.n.c. condition in Definition 3.6(i) for multifunctions defined in (3.13) when $\Lambda = \{0\}$. Moreover, in this case one always has $S = \{0\}$ in (3.9).

PROPOSITION 3.11. *Let both X and Y be Asplund and $f : X \to Y$ be Lipschitz continuous around $\bar{x} \in \Omega$. If f has the finite codimension property with respect to Ω around $\bar{x}$ and $\Lambda = \{0\}$ in (3.13), then the latter multifunction is p.n.c. with respect to y around $(\bar{x}, f(\bar{x}))$ when $S = \{0\}$ in (3.9).*

Proof. According to Thibault [44, Proposition 2.7], for any closed-graph multifunction $\Phi : X \rightrightarrows Y$ between Banach spaces and any point $(\bar{x}, \bar{y}) \in \text{gph } \Phi$ one has

$$N(\bar{x}, \bar{y}); \text{gph } \Phi) = \bigcup_{\lambda>0} \lambda \partial \text{dist}(\cdot, \Phi(\cdot))(\bar{x}, \bar{y}). \tag{3.17}$$

For the multifunction Φ defined in (3.13) with $\Lambda = \{0\}$, we easily get the representation

$$\text{dist}(v, \Phi(u)) = \|v - f(u)\| + \delta((u, v), \Omega \times Y) \ \ \forall (u, v) \in X \times Y, \tag{3.18}$$

where the first function is Lipschitz continuous. Then employing in (3.17), (3.18) the subdifferential sum and chain rules in Asplund spaces [35], we arrive at the inclusion

$$N(x, f(x)); \text{gph } \Phi) \subset \bigcup_{\lambda>0} \bigcup_{v^\star \in B^\star} [D^\star f(x)(\lambda v^\star) \times \{-\lambda v^\star\}] + N((x; \Omega) \times \{0\} \ \ \forall x \in \Omega.$$

The latter yields

$$D^\star \Phi(x, f(x))(y^\star) \subset D^\star f(x)(y^\star) + N(x; \Omega) \ \ \forall x \in \Omega \ \text{ and } \ y^\star \in Y^\star. \tag{3.19}$$

According to (3.19), for any $(x^\star, y^\star) \in N((x, f(x)); \text{gph } \Phi)$ with $x \in \Omega$ there exist $x_1^\star \in D^\star f(x)(-y^\star)$ and $x_2^\star \in N(x; \Omega)$ such that $x^\star = x_1^\star + x_2^\star$. Therefore, the finite codimension property in Definition 3.10 implies the p.n.c. condition in Definition 3.6(i) with $S = \{0\}$ in (3.9). □

Now let us compare the (dual) p.n.c. property of multifunctions with their (primal) epi-Lipschitzian kind of behavior discussed in the beginning of this section. Note that the notion of compactly epi-Lipschitzian sets mentioned above immediately induces the corresponding notion for multifunctions being applied to their graphs. In [18], Jourani and Thibault introduce a useful generalization of the latter property in the following way: A closed-graph multifunction $\Phi : X \rightrightarrows Y$ is said to be *partially compactly epi-Lipschitzian* (with respect to y) around $(\bar{x}, \bar{y}) \in \text{gph } \Phi$ if there exist neighborhoods U of $\bar{x}$, V of $\bar{y}$, and O of the origin in Y as well as a number $r > 0$ and compact sets $P \subset X$ and $S \subset Y$ such that

$$(\text{gph}) \cap (U \times V) + \lambda(\{0\} \times O) \subset \text{gph } \Phi + \lambda(P \times S) \ \ \forall \lambda \in (0, r). \tag{3.20}$$

When $P = \{0\}$ and S is a singleton in (3.20), this property corresponds to that considered by Kruger [20] under the name "uniformly epi-Lipschitzian" multifunctions. In [18, 20] one can find useful examples of multifunctions with such a behavior. Let us show that (3.20) always implies estimate (3.8) and, therefore, the p.n.c. property in Definition 3.6(i). One may observe that this result can be derived from [18, Proposition 3.5], taking into account that the normal cone (2.2) is always included in the so-called nucleus of the G-normal cone [16].

PROPOSITION 3.12. *Let $\Phi : X \rightrightarrows Y$ be a multifunction between Banach spaces that is partially compactly epi-Lipschitzian around $(\bar{x}, \bar{y})$. Then Φ is p.n.c. with respect to y around this point.*

Proof. Let us use (3.20) with $B_\gamma(\bar{x}) \subset U$, $B_\gamma(\bar{y}) \subset V$, and $\sigma B \subset O$ for some positive numbers γ and σ. Then for any $(x, y) \in (\text{gph } \Phi) \cap [B_\gamma(\bar{x}) \times B_\gamma(\bar{y})]$, $e \in B$, and a sequence $\lambda_k \in (0, r)$ with $\lambda_k \downarrow 0$ as $k \to \infty$, there are $p_k \in P$ and $s_k \in S$ such that

$$(x, y) + \lambda_k \sigma(0, e) - \lambda_k(p_k, s_k) \in \text{gph } \Phi \ \ \forall k = 1, 2, \ldots.$$

Due to compactness of P and S one can select a subsequence of $\{(p_k, s_k)\}$ which converges to some $(p, s) \in P \times S$. This implies that the difference $\sigma(0, e) - (p, s)$ belongs to the (Bouligand) *contingent cone* $K((x, y); \text{gph } \Phi)$; see [2]. It is well known that the Fréchet normal cone is always contained in the polar to the contingent cone. Therefore,

$$\min_{(p,s)\in P\times S} \langle (x^\star, y^\star), \sigma(0, e) - (p, s)\rangle \le 0 \ \ \forall (x^\star, y^\star) \in \hat{N}((x, y); \text{gph } \Phi)$$

with $(x, y) \in (\text{gph } \Phi) \cap [B_\gamma(\bar{x}) \times B_\gamma(\bar{y})]$. The latter implies (3.8) and completes the proof of the proposition. □

4. Point criteria and estimates for openness properties of multifunctions. In the rest of the paper $\Phi : X \rightrightarrows Y$ is a *closed-graph* multifunction between Banach spaces. In this section we pay the most attention to the following *openness property* of Φ in a *neighborhood* of a given point from its graph.

DEFINITION 4.1. *A multifunction Φ is said to be* open at a linear rate *around $(\bar{x}, \bar{y}) \in \text{gph } \Phi$ if there exist neighborhoods U of $\bar{x}$ and V of $\bar{y}$ as well as a positive number a such that*

$$B_{ar}(\Phi(x) \cap V) \subset \Phi(B_r(x)) \ \text{ for any } \ (x, r) \ \text{ with } \ B_r(x) \subset U.$$

Each of such numbers a (corresponding to different neighborhoods) is called an openness modulus *for Φ around $(\bar{x}, \bar{y})$. The supremum of all openness moduli is called the* openness bound *for Φ around $(\bar{x}, \bar{y})$ and is denoted by* $(\text{ope } \Phi)(\bar{x}, \bar{y})$.

Note that, for the case of linear bounded operators, this property goes back to the classical (Banach) open mapping principle. Let us emphasize two essential features of the given definition: *linear rate* of openness and *uniformity* of the openness property around the point under consideration; see [26, 32] for more discussion and references.

Using the coderivative (2.4) of Φ at $(\bar{x}, \bar{y})$, we introduce the main *openness constant* in the general setting,

$$a(\Phi, \bar{x}, \bar{y}) := \inf\{\|x^\star\| \ \text{ s.t. } \ x^\star \in D^\star\Phi(\bar{x}, \bar{y})(y^\star) \ \text{ and } \ \|y^\star\| = 1\}, \tag{4.1}$$

and formulate the principal result of the paper.

THEOREM 4.2. *For any multifunction $\Phi : X \rightrightarrows Y$ and point $(\bar{x}, \bar{y}) \in \text{gph } \Phi$, let us consider the properties:*

(a) *Φ is open at a linear rate around $(\bar{x}, \bar{y})$;*

(b) *$a(\Phi, \bar{x}, \bar{y}) > 0$ for the openness constant* (4.1)*;*

(c) *the coderivative* (2.4) *is injective at $(\bar{x}, \bar{y})$, i.e.,*

$$\text{Ker } D^\star\Phi(\bar{x}, \bar{y}) = \{0\}.$$

Then (b)$\Longrightarrow$(c) *and the following results hold.*

(I) *When Φ is p.n.c. with respect to y around $(\bar{x}, \bar{y})$ and both spaces X and Y are Asplund, one has*

$$\text{(c)} \Longrightarrow \text{(a)}; \tag{4.2}$$

i.e., each of the conditions (b) *and* (c) *is sufficient for Φ to be open at a linear rate around $(\bar{x}, \bar{y})$. Moreover,*

$$(\text{ope } \Phi)(\bar{x}, \bar{y}) \ge a(\Phi, \bar{x}, \bar{y}) \tag{4.3}$$

if Y is finite dimensional.

(II) *When Φ is an arbitrary multifunction from a finite-dimensional space X into a Banach space Y, one has*

$$\text{(a)} \Longrightarrow \text{(b)}; \tag{4.4}$$

i.e., both conditions (b) *and* (c) *are necessary for Φ to be open at a linear rate around $(\bar{x},\bar{y})$. Moreover, in this case*

$$(\text{ope}\,\Phi)(\bar{x},\bar{y}) \le a(\Phi,\bar{x},\bar{y}). \tag{4.5}$$

Proof. Implication (b)$\Longrightarrow$(c) is obvious. Let us establish assertion (I) and first justify (4.2) under the assumptions made. Proving by contradiction, we suppose that Φ is *not* open at a linear rate around $(\bar{x},\bar{y})$. Therefore, for any $a_k \downarrow 0$ one can find sequences $\{x_k\}$, $\{y_k\}$, $\{r_k\}$, and $\{z_k\}$ such that

$$x_k \to \bar{x},\ y_k \to \bar{y},\ r_k \downarrow 0 \ \text{ as } \ k \to \infty \ \text{ with } \ y_k \in \Phi(x_k) \cap B_{a_k}(\bar{y}), \tag{4.6}$$

$$\|z_k - y_k\| \le a_k r_k, \ \text{ and } \ z_k \notin \Phi(x) \ \ \forall x \in B_{r_k}(x_k), \ \ k = 1,2,\ldots. \tag{4.7}$$

Now let us employ the Ekeland variational principle that has become a conventional tool for this kind of study; see, e.g., [3, 14, 18, 20, 25]. For each k we consider an l.s.c. function $f_k : E_k \to \mathbf{R}$ defined by

$$f_k(x,y) := \|y - z_k\| \ \text{ on } \ E_k := (\text{gph}\,\Phi) \cap B_{r_k}((x_k,y_k)) \tag{4.8}$$

with the metric on E_k induced by the norm $\|(x,y)\| := \|x\| + \|y\|$ on $X \times Y$. Due to (4.6) and (4.7) one has

$$f_k(x_k,y_k) \le a_k r_k \ \text{ and } \ f_k(x,y) > 0 \ \ \forall (x,y) \in E_k.$$

Applying Ekeland's principle to the function f_k on the metric space E_k, we obtain the following: for the given numbers $\epsilon_k := a_k r_k$, $\lambda_k := (1/2)r_k$ and point (x_k,y_k) there exists $(\tilde{x}_k,\tilde{y}_k) \in \text{gph}\,\Phi \cap B_{(1/2)r_k}((x_k,y_k))$ such that

$$0 < \|\tilde{y}_k - z_k\| \le \|y_k - z_k\| \le \epsilon_k \ \text{ and}$$
$$\|\tilde{y}_k - z_k\| \le \|y - z_k\| + (\epsilon_k/\lambda_k)\|(x,y) - (\tilde{x}_k,\tilde{y}_k)\| \ \ \forall (x,y) \in E_k.$$

The latter means that $(\tilde{x}_k,\tilde{y}_k)$ is a local minimizer of the function

$$\varphi_k(x,y) := \|y - z_k\| + 2a_k\|(x,y) - (\tilde{x}_k,\tilde{y}_k)\| + \delta((x,y),\text{gph}\,\Phi) \tag{4.9}$$

defined on $X \times Y$. It is well known that the Cartesian product of two Asplund spaces is also Asplund, so one can apply Proposition 2.4 to the sum of three functions in (4.9) at $(\tilde{x}_k,\tilde{y}_k)$. Using this result with $\varepsilon = \gamma < \min\{a_k,\ \|z_k - \tilde{y}_k\|\}$ and taking into account the first formula in (2.9), we find $(x_{ik},y_{ik}) \in \text{gph}\,\Phi$ such that $\|x_{ik} - \tilde{x}_k\| \le a_k$, $\|y_{ik} - \tilde{y}_k\| \le a_k$, $y_{ik} \ne z_k$, $i = 1,2,3$, and

$$0 \in \hat{\partial}(\|\cdot - z_k\|)(x_{1k},y_{1k}) + \hat{\partial}(2a_k\|(\cdot,\cdot) - (\tilde{x}_k,\tilde{y}_k)\|)(x_{2k},y_{2k})$$
$$+ \hat{N}((x_{3k},y_{3k});\text{gph}\,\Phi) + a_k(B^\star \times B^\star).$$

Employing the well-known convex subdifferential formula for the norm function and taking into account that $y_{ik} \ne z_k$, we obtain a triple $(\tilde{x}_k^\star,\tilde{y}_k^\star,\tilde{z}_k^\star) \in X^\star \times Y^\star \times Y^\star$ with

$$(\tilde{x}_k^\star,-\tilde{y}_k^\star) \in \hat{N}((x_{3k},y_{3k});\text{gph}\,\Phi), \ \ \|\tilde{z}_k^\star\| = 1, \ \text{ and } \ \|\tilde{z}_k^\star - \tilde{y}_k^\star\| \le 3a_k.$$

The latter yields $\|\tilde{y}_k^\star\| \ge 1 - 3a_k > 1/2$ for all big k. Then letting $x_k := x_{3k}$, $y_k := y_{3k}$, $x_k^\star := \tilde{x}_k^\star/\|\tilde{y}_k^\star\|$, and $y_k^\star := \tilde{y}_k^\star/\|\tilde{y}_k^\star\|$, we conclude that $x_k \to \bar{x}$ and $y_k \to \bar{y}$ as $k \to \infty$ while

$$(x_k^\star, -y_k^\star) \in \hat{N}((x_k, y_k); \operatorname{gph}\Phi) \ \text{ with } \ \|y_k^\star\| = 1 \ \text{ and } \ \|x_k^\star\| \le 6a_k. \tag{4.10}$$

Since Y is Asplund, the unit ball of the dual space $Y^\star$ is sequentially weak-star compact. Taking into account the boundedness of $\{y_k^\star\}$, one may assume that $y_k^\star \xrightarrow{w^\star} \tilde{y}^\star$ as $k \to \infty$ for some $\tilde{y}^\star \in Y^\star$. On the other hand, relationships (4.10) imply that $x_k^\star \to 0$ as $k \to \infty$ in the norm topology of $X^\star$ and $0 \in D^\star\Phi(\bar{x}, \bar{y})(\tilde{y}^\star)$ by virtue of Definition 2.2. Moreover, due to the p.n.c. property of Φ with respect to y around $(\bar{x}, \bar{y})$ one can conclude that $\tilde{y}^\star \ne 0$. Indeed, otherwise we employ Proposition 3.8 and arrive at the contradiction with $\|y_k^\star\| = 1$ for all k. Finally letting $y^\star := \tilde{y}^\star/\|\tilde{y}^\star\|$, one has

$$0 \in D^\star\Phi(\bar{x}, \bar{y})(y^\star) \ \text{ with } \ \|y^\star\| = 1.$$

This contradicts condition (c) in the theorem and completes the proof of implication (4.2).

Now let us establish estimate (4.3) assuming that Y is finite dimensional. To furnish this, we make some changes in the previous procedure similarly to the related proof of [32, Theorem 3.2] in a somewhat different situation.

Suppose that (4.3) is not true; i.e., there exist a positive number $a < a(\Phi, \bar{x}, \bar{y})$ as well as sequences $\{x_k\}$, $\{y_k\}$, $\{r_k\}$, and $\{z_k\}$ such that

$$\begin{gathered} x_k \to \bar{x}, \ \ y_k \to \bar{y}, \ \ r_k \downarrow 0 \ \text{ as } \ k \to \infty, \ \text{ and} \\ y_k \in \Phi(x_k), \ \ \|z_k - y_k\| \le ar_k, \ \ z_k \notin \Phi(x) \ \ \forall x \in B_{r_k}(x_k). \end{gathered} \tag{4.11}$$

For any $\varepsilon > a$ we take some $\alpha \in (a/\varepsilon, 1)$ and pick a sequence $\{\eta_k\}$ with

$$0 < \eta_k < \min\left\{ r_k, \ \frac{1}{2(\varepsilon\alpha + 1)}, \ \frac{\varepsilon(1-\alpha)}{1 + \varepsilon(\varepsilon\alpha + 1)} \right\} \ \ \forall k = 1, 2, \ldots. \tag{4.12}$$

For each fixed k, let us consider function (4.8) on the space E_k whose metric is induced by the norm $\|(x, y)\|_{\eta_k} := \|x\| + \eta_k\|y\|$ on $X \times Y$, in contrast to the proof of (4.2). Now we again apply the Ekeland variational principle to this function f_k at the point (x_k, y_k), but with different parameters, namely, $\epsilon_k := ar_k$ and $\lambda_k := ar_k/\varepsilon\alpha$. Noting that $f_k(x_k, y_k) \le \epsilon_k$ due to (4.11), we find a point $(\tilde{x}_k, \tilde{y}_k) \in E_k$ such that the function

$$\psi_k(x, y) := \|y - z_k\| + \varepsilon\alpha\|(x, y) - (\tilde{x}_k, \tilde{y}_k)\|_{\eta_k} + \delta((x, y), \operatorname{gph}\Phi) \tag{4.13}$$

attains its unconditional local minimum on $X \times Y$ at the point $(\tilde{x}_k, \tilde{y}_k)$.

Letting $\rho_k := \|\tilde{y}_k - z_k\|$, we apply Proposition 2.4 with $\varepsilon = \eta_k$ and $\gamma = \rho_k\eta_k/2$ to the sum of three functions in (4.13) at $(\tilde{x}_k, \tilde{y}_k)$. In this way we find three pairs (x_{ik}, y_{ik}) such that

$$\begin{gathered} \|(x_{ik}, y_{ik}) - (\tilde{x}_k, \tilde{y}_k)\| \le \rho_k\eta_k/2 \ \text{ with } \ y_{ik} \ne z_k \ \text{ for } \ i = 1, 2, 3 \ \text{ and} \\ 0 \in \hat{\partial}(\|\cdot - z_k\|)(x_{1k}, y_{1k}) + \hat{\partial}(\varepsilon\alpha\|(\cdot, \cdot) - (\tilde{x}_k, \tilde{y}_k)\|_{\eta_k})(x_{2k}, y_{2k}) \\ + \hat{N}((x_{3k}, y_{3k}); \operatorname{gph}\Phi) + \eta_k(B^\star \times B^\star). \end{gathered} \tag{4.14}$$

Now computing subdifferentials of the norm functions in (4.14) and taking into account (4.12), we get a pair $(x_k^\star, y_k^\star) \in X^\star \times Y^\star$ satisfying

$$(x_k^\star, -y_k^\star) \in \hat{N}((x_{3k}, y_{3k});\ \text{gph}\ \Phi),\ \ \|y_k^\star\| = 1,\ \ \|x_k^\star\| \le \frac{\varepsilon\alpha + \eta_k}{1 - \eta_k(\varepsilon\alpha + 1)} < \varepsilon. \tag{4.15}$$

By the constructions above we have $x_{3k} \to \bar{x}$ and $y_{3k} \to \bar{y}$ as $k \to \infty$. Moreover, due to Asplundity of X and dim $Y < \infty$ one can select a subsequence of $\{k\}$ along which $x_k^\star \overset{w^\star}{\to} x^\star \in X^\star$ and $y_k^\star \to y^\star$ with $\|y^\star\| = 1$. Finally passing to the limit in (4.15) when $k \to \infty$ and using (2.3), (2.4), (4.1) as well as the l.s.c. of the norm function in the weak-star topology of $X^\star$, we conclude that $a(\Phi, \bar{x}, \bar{y}) \le \varepsilon$. Since $\varepsilon > a$ was chosen arbitrary, the latter yields $a(\Phi, \bar{x}, \bar{y}) \le a$ that contradicts the choice of a. Therefore, one arrives at (4.3), which completes the proof of assertion (I).

Let us prove assertion (II) under the assumptions made therein. We are going to show that if Φ possesses the openness property with a modulus $a > 0$, then one always has $a \le a(\Phi, \bar{x}, \bar{y})$. The latter implies both (4.4) and (4.5).

Proving by contradiction, we assume that $a > a(\Phi, \bar{x}, \bar{y})$. Then there is a number $\alpha > 0$ such that $a(\Phi, \bar{x}, \bar{y}) < a - \alpha$, i.e.,

$$\|x^\star\| < a - \alpha \ \text{ for some } \ x^\star \in D^\star\Phi(\bar{x}, \bar{y})(y^\star) \ \text{ and } \ y^\star \in Y^\star \ \text{ with } \ \|y^\star\| = 1.$$

Now using the basic definitions (2.4) and (2.2) as well as dim $X < \infty$, one gets sequences $\{(x_k, y_k)\} \subset \text{gph}\ \Phi$, $\{x_k^\star\} \subset X^\star$, $\{y_k^\star\} \subset Y^\star$, and $\{\varepsilon_k\} \subset \mathbf{R}_+$ such that

$$\|x_k^\star\| < a - \alpha \ \text{ and } \ (x_k^\star, -y_k^\star) \in \hat{N}_{\varepsilon_k}((x_k, y_k); \text{gph}\ \Phi) \ \ \forall k = 1, 2, \ldots, \tag{4.16}$$

where $x_k \to \bar{x}$, $y_k \to \bar{y}$, $\varepsilon_k \downarrow 0$, $x_k^\star \to x^\star$ and $y_k^\star \overset{w^\star}{\to} y^\star$ as $k \to \infty$. Due to (2.1) and the inclusion in (4.16), we find a sequence $\gamma_k \downarrow 0$ with

$$\langle y_k^\star, y - y_k\rangle + 2\varepsilon_k(\|x - x_k\| + \|y - y_k\|) \ge \langle x_k^\star, x - x_k\rangle \tag{4.17}$$

for all $(x, y) \in \text{gph}\ \Phi$ satisfying $\|x - x_k\| \le \gamma_k$ and $\|y - y_k\| \le \gamma_k$ as $k = 1, 2, \ldots$.

Next let us choose a positive number β, a sequence of positive numbers $\{r_k\}$, and the sequence $\{\varepsilon_k\}$ in (4.17) such that

$$\beta \le \min\{a, \alpha/2\},\ \ r_k \le \min\{\gamma_k, \gamma_k/a\},\ \text{ and } \ \varepsilon_k \le (\alpha - \beta)/2(1 + a) \tag{4.18}$$

for all k. Due to $y_k^\star \overset{w^\star}{\to} y^\star$ with $\|y^\star\| = 1$ and the l.s.c. of $\|\cdot\|$ in the weak-star topology of $Y^\star$, one can assume that

$$\|y_k^\star\| > 1 - \beta/a \ \ \forall k = 1, 2, \ldots.$$

Therefore, there is $v_k \in Y$ such that

$$\|v_k\| = 1 \ \text{ and } \ \langle y_k^\star, v_k\rangle > 1 - \beta/a \ \ \forall k = 1, 2, \ldots. \tag{4.19}$$

Let us define $z_k := y_k - ar_kv_k$ for each $k = 1, 2, \ldots$. Using (4.18) and (4.19), we have $\|z_k - y_k\| = ar_k \le \gamma_k$ and the chain of estimates

$$\begin{aligned}
&\langle y_k^\star, z_k - y_k\rangle + 2\varepsilon_k(\|x - x_k\| + \|z_k - y_k\|)\\
&< -ar_k + \beta r_k + \frac{\alpha - \beta}{1 + a}(\|x - x_k\| + ar_k) \le -(a - \alpha)r_k < \langle x_k^\star, x - x_k\rangle,
\end{aligned}$$

which are valid for any $x \in B_{r_k}(x_k)$. By virtue of (4.17) the latter means that $z_k \notin \Phi(x)$ whatever $x \in B_{r_k}(x_k)$. Therefore, we have sequences $x_k \to \bar{x},\ y_k \to \bar{y},\ z_k \to \bar{y}$, and $r_k \downarrow 0$ as $k \to \infty$ such that (4.11) holds. Due to Definition 4.1 this contradicts the assumption that the number a is an openness modulus for Φ around $(\bar{x}, \bar{y})$. The contradiction obtained shows that $a \le a(\Phi, \bar{x}, \bar{y})$ and completes the proof of the theorem. □

COROLLARY 4.3. *Let $\Phi : X \rightrightarrows Y$ be a multifunction from a finite-dimensional space X into an Asplund space Y. Assume that Φ is p.n.c. with respect to y around $(\bar{x}, \bar{y}) \in \operatorname{gph} \Phi$. Then each of the conditions (b) and (c) in Theorem* 4.2 *is necessary and sufficient for Φ to be open at a linear rate around $(\bar{x}, \bar{y})$.*

Proof. This follows directly from the comparison of assertions (I) and (II) in Theorem 4.2. □

Remark 4.4. Another approach to prove the results in Theorem 4.2 consists of directly using neighborhood criteria established in our paper [32] in terms of the Fréchet coderivative of Φ and then employing the limiting arguments developed above. Following these arguments, one can observe that *when Y is finite dimensional*, a sufficient condition for the openness property of Φ around $(\bar{x}, \bar{y})$ can be expressed in the form

$$\operatorname{Ker} D^\star_s \Phi(\bar{x}, \bar{y}) = \{0\}, \tag{4.20}$$

where $D^\star_s \Phi(\bar{x}, \bar{y})$ is an analogue of the coderivative (2.4) corresponding to the *strong* upper limit (with respect to the norm topology on $X^\star \times Y^\star$) in the normal cone definition (2.2); cf. Penot [38, section 5]. Although the coderivative object in (4.20) may be smaller than the basic construction (2.4), no useful calculus rules are available for the former one in contrast to (2.4). Note also that the necessity of condition (4.20) for the openness of Φ is a weaker result in comparison with (II) in Theorem 4.2, but it does not need the assumption about $\dim X < \infty$; see [38, Remark 4.5].

Now employing Theorem 4.2 and the robustness property in Proposition 3.4, we obtain a useful *neighborhood criterion* for openness of multifunctions that is expressed in terms of the coderivative (2.4). Note that this criterion and its inverse counterpart (d) in Theorem 5.4 are important for applications in optimal control; cf. [25, 31].

COROLLARY 4.5. *Let both X and Y be WCG Asplund spaces and $\Phi : X \rightrightarrows Y$ be a multifunction normally compact around $(\bar{x}, \bar{y}) \in \operatorname{gph} \Phi$. Then each of the conditions* (b) *and* (c) *in Theorem* 4.2 *is equivalent to the following one:*

(d) *There exist positive numbers μ, γ, and η such that*

$$\|y^\star\| \le \mu \|x^\star\| \ \ \forall x^\star \in D^\star \Phi(x, y)(y^\star),\ x \in B_\gamma(\bar{x}),\ y \in \Phi(x) \cap B_\eta(\bar{y}).$$

Therefore, condition (d) *is sufficient for the openness property of Φ around $(\bar{x}, \bar{y})$ and is also necessary for this property when* $\dim X < \infty$.

Proof. Since implications (d)$\Longrightarrow$(b)$\Longrightarrow$(c) are always true, it remains to establish that (c)$\Longrightarrow$(d).

Proving by contradiction, let us assume that (d) fails. Then one has sequences $\{(x_k, y_k)\} \subset \operatorname{gph} \Phi$ and $\{(x_k^\star, y_k^\star)\} \subset X^\star \times Y^\star$ such that

$$x_k^\star \in D^\star \Phi(x_k, y_k)(y_k^\star), \ \ \|y_k^\star\| > k \|x_k^\star\| \ \ \forall k = 1, 2, \ldots$$

and $(x_k, y_k) \to (\bar{x}, \bar{y})$ as $k \to \infty$. Since $\|y_k^\star\| > 0$, we set $\tilde{y}_k^\star := y_k^\star / \|y_k^\star\|,\ \tilde{x}_k^\star := x_k^\star / \|y_k^\star\|$ and get

$$\tilde{x}_k^\star \in D^\star \Phi(x_k, y_k)(\tilde{y}_k^\star), \ \text{ with } \ \|\tilde{y}_k^\star\| = 1 \ \text{ and } \ \|\tilde{x}_k^\star\| < 1/k \ \ \forall k = 1, 2, \ldots. \tag{4.21}$$

Taking into account the sequential weak-star compactness of the unit ball in $Y^\star$ (since Y is Asplund), one may assume that $\tilde{y}_k^\star \xrightarrow{w^\star} \tilde{y}^\star$ as $k \to \infty$ for some $\tilde{y}^\star \in Y^\star$. On the other hand, (4.21) implies that $\tilde{x}_k^\star \to 0$ as $k \to \infty$ in the norm topology of $X^\star$. Since Φ is normally compact around $(\bar{x}, \bar{y})$ and both X and Y are Asplund, we can use the limiting property (3.7) with the normal cone $N((x,y); \operatorname{gph}\Phi)$ replacing the prenormal one. This yields $\tilde{y}^\star \neq 0$. Moreover, Proposition 3.4 ensures that $0 \in D^\star\Phi(\bar{x}, \bar{y})(\tilde{y}^\star)$. Denoting $y^\star := \tilde{y}^\star/\|\tilde{y}^\star\|$, we arrive at

$$0 \in D^\star\Phi(\bar{x}, \bar{y})(y^\star) \ \text{ with } \ \|y^\star\| = 1. \tag{4.22}$$

This contradicts condition (c) in Theorem 4.2 and completes the proof of the corollary. □

The next corollary contains effective characteristics for the openness property of multifunctions belonging to class (3.13), which is important for applications in optimization and optimal control. In particular, results in this vein directly induce necessary optimality conditions in various problems of scalar and/or vector optimization; cf. [17, 18, 25].

COROLLARY 4.6. *Assume that both spaces X and Y are Asplund, the sets $\Omega \subset X$ and $\Lambda \subset Y$ are closed, and the function $f : X \to Y$ is strictly Lipschitzian around $\bar{x} \in X$ with $\bar{y} := -f(\bar{x}) \in \Lambda$. In addition, let Λ be normally compact around $\bar{y}$. Then the multifunction Φ defined by* (3.13) *is open at a linear rate around $(\bar{x}, 0)$ if*

$$[y^\star \in \partial\operatorname{dist}(\cdot, \Lambda)(\bar{y}) \ \text{ and } \ 0 \in \partial\langle y^\star, f\rangle(\bar{x}) + N(\bar{x}; \Omega)] \Longrightarrow y^\star = 0. \tag{4.23}$$

Proof. Following the proof of Proposition 3.11, we get the respresentation

$$N((\bar{x}, 0); \operatorname{gph}\Phi) = \bigcup_{\lambda > 0} \lambda\partial\varphi(\bar{x}, 0), \tag{4.24}$$

where the distance function $\varphi(x, y) := \operatorname{dist}(y, \Phi(x))$ is expressed in the form

$$\varphi(x, y) = \operatorname{dist}(y - f(x), \Lambda) + \delta((x, y), \Omega \times Y) \ \ \forall (x, y) \in X \times Y. \tag{4.25}$$

One can easily observe that the mapping $(x, y) \to y - f(x)$ is strictly Lipschitzian around $(\bar{x}, 0)$. Now employing subdifferential calculus rules [35] and Proposition 2.5 in (4.24) and (4.25), we find a number $\lambda > 0$ such that

$$N((\bar{x}, 0); \operatorname{gph}\Phi) \subset \bigcup_{\lambda > 0} \bigcup_{y^\star \in \partial dist(\cdot, \Lambda)(\bar{y})} [\partial\langle \lambda y^\star, f\rangle(\bar{x}) \times \{-\lambda y^\star\}] + N(\bar{x}; \Omega) \times \{0\}. \tag{4.26}$$

Therefore, condition (4.23) implies criterion (c) in Theorem 4.2 due to (4.26) and the coderivative construction (2.4). Moreover, Proposition 3.9 ensures that the multifunction Φ in (3.13) is p.n.c. with respect to y around $(\bar{x}, 0)$ under the assumptions made. In this way we deduce the corollary from assertion (I) of Theorem 4.2. □

In conclusion of this section let us consider a *global* (relative to the image) counterpart of the openness property for closed-graph multifunctions.

DEFINITION 4.7. *We say that Φ enjoys the* covering property *around $\bar{x} \in \operatorname{Dom}\Phi$ if there exist a number $a > 0$ and a neighborhood U of $\bar{x}$ such that for any (x, r) with $B_r(x) \subset U$ one has*

$$B_{ar}(\Phi(x)) \subset \Phi(B_r(x)).$$

Each of such numbers a is called the covering modulus *for Φ around $\bar{x}$. The supremum of all covering moduli is called the* covering bound *for Φ around $\bar{x}$ and is denoted by* $(\operatorname{cov}\Phi)(\bar{x})$.

Let us consider the *covering constant*

$$a(\Phi,\bar{x}) := \inf\{\|x^\star\| \text{ s.t. } x^\star \in D^\star\Phi(\bar{x},\bar{y})(y^\star),\ \ \|y^\star\|=1, \quad \text{and } \ \bar{y}\in\Phi(\bar{x})\} \tag{4.27}$$

related to (4.1) and formulate the covering characterization result that follows from Theorem 4.2 and Corollary 4.5. Recall that a multifunction $\Phi : X \rightrightarrows Y$ between Banach spaces is said to be *locally compact* around $\bar{x}$ if there are a neighborhood U of $\bar{x}$ and a compact set $V \subset Y$ such that $\Phi(U) \subset V$.

THEOREM 4.8. *Let the multifunction $\Phi : X \rightrightarrows Y$ be locally compact around $\bar{x} \in \operatorname{Dom}\Phi$. Consider the following statements:*

(a) *Φ enjoys the covering property around $\bar{x}$;*

(b) *$a(\Phi,\bar{x}) > 0$ for the covering constant* (4.27)*;*

(c) *the coderivative* (2.4) *at $(\bar{x},\bar{y})$ is injective for all $\bar{y}\in\Phi(\bar{x})$, i.e.,*

$$\operatorname{Ker} D^\star\Phi(\bar{x},\bar{y}) = \{0\} \ \ \forall \bar{y}\in\Phi(\bar{x});$$

(d) *there exist positive numbers μ and γ such that*

$$\|y^\star\| \le \mu\|x^\star\| \ \ \forall x^\star \in D^\star\Phi(x,y)(y^\star),\ \ x\in B_\gamma(\bar{x}),\ \ y\in\Phi(x).$$

Then (d) $\Longrightarrow$ (b) $\Longrightarrow$ (c) *and the following results hold:*

(I) *When Φ is p.n.c. with respect to y around $(\bar{x},\bar{y})$ for any $\bar{y}\in\Phi(\bar{x})$ and both spaces X and Y are Asplund, one has* (c)$\Longrightarrow$(a)*; i.e., each of the conditions* (b), (c), *and* (d) *is sufficient for Φ to enjoy the covering property around $\bar{x}$. Moreover,*

$$(\operatorname{cov}\Phi)(\bar{x}) \ge a(\Phi,\bar{x})$$

when Y is finite dimensional.

(II) *When Φ is an arbitrary multifunction from a finite-dimensional space X into a Banach space Y, one has* (a)$\Longrightarrow$(b)*; i.e., both conditions* (b) *and* (c) *are necessary for Φ to enjoy the covering property around $\bar{x}$. Moreover, in this case*

$$(\operatorname{cov}\Phi)(\bar{x}) \le a(\Phi,\bar{x}).$$

(III) *Each of the conditions* (b) *and* (c) *is equivalent to* (d) *when both spaces X and Y are WCG Asplund and the multifunction Φ is normally compact around $\bar{x}$ for any $\bar{y}\in\Phi(\bar{x})$.*

Proof. Using the local compactness of Φ around $\bar{x}$ and the compactness arguments in [26, Theorem 3.9] (cf. also [42, Theorem 2.2]), one can establish that the covering property of Φ around $\bar{x}$ is *equivalent* in this case to the openness property of Φ around $(\bar{x},\bar{y})$ for *every* $\bar{y}\in\Phi(\bar{x})$. Moreover,

$$a(\Phi,\bar{x}) = \inf\{a(\Phi,\bar{x},\bar{y})|\ \bar{y}\in\Phi(\bar{x})\} \ \text{ and } \ (\operatorname{cov}\Phi)(\bar{x}) = \inf\{(\operatorname{ope}\Phi)(\bar{x},\bar{y})|\ \bar{y}\in\Phi(\bar{x})\}.$$

In this way one can also check that properties (b)–(d) in the theorem are equivalent to the fulfilment of the corresponding properties in Theorem 4.2 and Corollary 4.5 for all $\bar{y}\in\Phi(\bar{x})$ (respectively, all $y\in\Phi(x)$ in (d)). Therefore, the results formulated follow from the corresponding results of Theorem 4.2 and Corollary 4.5. □

5. Point characterizations of metric regularity and Lipschitzian stability. This section is concerned with other significant properties of closed-graph multifunctions $\Phi : X \rightrightarrows Y$ related to the openness and covering considered above. We start with definitions of the (local) metric regularity and pseudo-Lipschitzian properties initiated, respectively, by Robinson [40] and Aubin [1].

DEFINITION 5.1. (i) Φ *is said to be* local-metrically regular *around* $(\bar{x},\bar{y}) \in \operatorname{gph}\Phi$ *with modulus* $c > 0$ *if there exist a neighborhood* U *of* $\bar{x}$*, a neighborhood* V *of* $\bar{y}$*, and a number* $\alpha > 0$ *such that*

$$\operatorname{dist}(x, \Phi^{-1}(y)) \;\le\; c\operatorname{dist}(y, \Phi(x))$$

for any $x \in U$ *and* $y \in V$ *satisfying* $\operatorname{dist}(y,\Phi(x)) \le \alpha$*. The infimum of all regularity moduli* c *is called the* bound of local-metric regularity *for* Φ *around* $(\bar{x},\bar{y})$ *and is denoted by* $(\operatorname{lreg}\Phi)(\bar{x},\bar{y})$.

(ii) Φ *is said to be* pseudo-Lipschitzian *around* $(\bar{x},\bar{y}) \in \operatorname{gph}\Phi$ *with modulus* $l > 0$ *if there exist a neighborhood* U *of* $\bar{x}$ *and a neighborhood* V *of* $\bar{y}$ *such that*

$$\Phi(x_1) \cap V \subset \Phi(x_2) + l\|x_1 - x_2\|B \;\; \forall x_1, x_2 \in U. \tag{5.1}$$

The infimum of all such moduli l *is called the* bound of pseudo-Lipschitzness *for* Φ *around* $(\bar{x},\bar{y})$ *and is denoted by* $(\operatorname{plip}\Phi)(\bar{x},\bar{y})$.

The interrelations between the properties in Definitions 4.1 and 5.1 can be obtained from Borwein and Zhuang [7] and Penot [37]; cf. also [26].

PROPOSITION 5.2. (I) Φ *is local-metrically regular around* $(\bar{x},\bar{y})$ *if and only if* Φ *is open at a linear rate around* $(\bar{x},\bar{y})$*. Moreover,* $(\operatorname{lreg}\Phi)(\bar{x},\bar{y}) = 1/(\operatorname{ope}\Phi)(\bar{x},\bar{y})$.

(II) Φ *is pseudo-Lipschitzian around* $(\bar{x},\bar{y}) \in \operatorname{gph}\Phi$ *with modulus* l *if and only if* Φ^{-1} *is local-metrically regular around* $(\bar{y},\bar{x})$ *with the same modulus* $c = l$.

Following [26], let us introduce the *regularity constant*

$$c(\Phi,\bar{x},\bar{y}) := \inf\{c > 0 \;\text{ s.t. }\; \|y^\star\| \le c\|x^\star\| \;\text{ when }\; x^\star \in D^\star\Phi(\bar{x},\bar{y})(y^\star)\} \tag{5.2}$$

and observe the relationships between constants (5.2), (4.1), and the norm (1.2) of the coderivative (2.4):

$$c(\Phi,\bar{x},\bar{y}) = 1/a(\Phi,\bar{x},\bar{y}) \;\text{ when }\; a(\Phi,\bar{x},\bar{y}) > 0; \tag{5.3}$$

$$c(\Phi^{-1},\bar{y},\bar{x}) = \|D^\star\Phi(\bar{x},\bar{y})\|. \tag{5.4}$$

Now using Proposition 5.2 and relationships (5.3) and (5.4), one can deduce effective characterizations of the metric regularity and Lipschitzian behavior of multifunctions from our main results in Theorem 4.2 and Corollary 4.5.

THEOREM 5.3. (I) *Let both spaces* X *and* Y *be Asplund and the multifunction* $\Phi : X \rightrightarrows Y$ *be p.n.c. with respect to* y *around* $(\bar{x},\bar{y}) \in \operatorname{gph}\Phi$*. Then each of the conditions*

$$c(\Phi,\bar{x},\bar{y}) < \infty \tag{5.5}$$

and (c) *in Theorem* 4.2 *is sufficient for* Φ *to be local-metrically regular around* $(\bar{x},\bar{y})$*. Moreover, one has*

$$(\operatorname{lreg}\Phi)(\bar{x},\bar{y}) \le c(\Phi,\bar{x},\bar{y})$$

when Y *is finite dimensional.*

(II) *Let Φ be an arbitrary closed-graph multifunction from a finite-dimensional space X into a Banach space Y. Then both conditions* (5.5) *and* (c) *in Theorem* 4.2 *are necessary for Φ to be local-metrically regular around $(\bar{x},\bar{y})$. In this case one has*

$$(\operatorname{lreg}\Phi)(\bar{x},\bar{y})\geq c(\Phi,\bar{x},\bar{y}).$$

One can observe that in the classical cases where Φ either is strictly differentiable at $\bar{x}$ or has convex graph, the coderivative criteria of Theorems 4.2 and 5.3 are reduced to the corresponding surjectivity and interiority conditions of the celebrated Ljusternik–Graves and Robinson–Ursescu theorems; see, e.g., [2] and references therein. Note that those conditions turn out to be *necessary and sufficient* for the metric regularity/openness (at a linear rate) properties under the assumptions made. The latter fact holds in more general infinite-dimensional settings; cf. [10, 11, 32]. Let us emphasize that we also get effective modulus estimates.

Next we formulate results on the pseudo-Lipschitzian property that are *inverse* to criteria in Theorem 4.2 and Corollary 4.5. One can easily observe from the definitions that

$$D^\star\Phi^{-1}(\bar{y},\bar{x})(x^\star)=\{y^\star\in Y^\star|\ x^\star\in -D^\star\Phi(\bar{x},\bar{y})(-y^\star)\} \tag{5.6}$$

for any multifunction $\Phi: X\Rightarrow Y$ between Banach spaces.

THEOREM 5.4. *Let $\Phi: X\Rightarrow Y$ be a closed-graph multifunction with $(\bar{x},\bar{y})\in \operatorname{gph}\Phi$. Consider the following properties:*

(a) *Φ is pseudo-Lipschitzian around $(\bar{x},\bar{y})$;*

(b) *the coderivative $D^\star\Phi(\bar{x},\bar{y})(\cdot)$ is bounded, i.e., $\|D^\star\Phi(\bar{x},\bar{y})\|<\infty$;*

(c) *the coderivative satisfies the null-condition at $(\bar{x},\bar{y})$:*

$$D^\star\Phi(\bar{x},\bar{y})(0)=\{0\}; \tag{5.7}$$

(d) *there are numbers $\gamma>0$, $\eta>0$, and $l>0$ such that the coderivative satisfies the uniform linear estimate around $(\bar{x},\bar{y})$:*

$$\sup\{\|x^\star\|\ s.t.\ x^\star\in D^\star\Phi(x,y)(y^\star)\}\leq l\|y^\star\|\ \ \forall x\in B_\gamma(\bar{x}),\ y\in\Phi(x)\cap B_\eta(\bar{y}),\ \text{and}\ y^\star\in Y^\star.$$

Then (d)$\Longrightarrow$(b)$\Longrightarrow$ (c) *and the following assertions hold:*

(I) *One has* (c)$\Longrightarrow$(a) *when Φ is p.n.c. with respect to x around $(\bar{x},\bar{y})$ and both spaces X and Y are Asplund. Thus in this case each of the conditions* (b), (c), *and* (d) *is sufficient for Φ to be pseudo-Lipschitzian around $(\bar{x},\bar{y})$. Moreover,*

$$(\operatorname{plip}\Phi)(\bar{x},\bar{y})\leq\|D^\star\Phi(\bar{x},\bar{y})\| \tag{5.8}$$

when X is finite dimensional.

(II) *One has* (a)$\Longrightarrow$(b) *when Φ is an arbitrary multifunction from a Banach space X into a finite-dimensional space Y. Thus in this case both conditions* (b) *and* (c) *are necessary for Φ to be pseudo-Lipschitzian around $(\bar{x},\bar{y})$ and, in addition,*

$$(\operatorname{plip}\Phi)(\bar{x},\bar{y})\geq\|D^\star\Phi(\bar{x},\bar{y})\|.$$

(III) *Conditions* (b), (c), *and* (d) *are equivalent when both X and Y are WCG Asplund spaces and Φ is normally compact around $(\bar{x},\bar{y})$.*

COROLLARY 5.5. *Let X be Asplund, let* $\dim Y<\infty$, *and let $\Phi: X\Rightarrow Y$ be p.n.c. with respect to x around $(\bar{x},\bar{y})$. Then each of the conditions* (b) *and* (c) *in Theorem* 5.4 *is necessary and sufficient for Φ to be pseudo-Lipschitzian around $(\bar{x},\bar{y})$.*

In the next section we present some applications of Theorem 5.4 to Lipschitzian stability of parametric constraint and variational systems. Now let us show how the results obtained allow one to characterize Lipschitzian behavior of an extended-real-valued function $\varphi : X \to \bar{\mathbf{R}}$ in terms of its *singular subdifferential*

$$\partial^\infty \varphi(\bar{x}) := \{x^\star \in X^\star |\ (x^\star, 0) \in N((\bar{x}, \varphi(\bar{x})); \operatorname{epi} \varphi)\} \tag{5.9}$$

at $\bar{x} \in \operatorname{dom} \varphi$. Along with φ we consider the *epigraphical multifunction* E_φ associated with φ by virtue of (2.10). We say that φ is *normally epi-compact* around $\bar{x}$ if the set $\Omega := \operatorname{epi} \varphi$ (or the multifunction E_φ) is normally compact around $(\bar{x}, \varphi(\bar{x}))$. The latter property always holds if φ is compactly epi-Lipschitzian (in particular, Lipschitz continuous or directionally Lipschitzian) around $\bar{x}$; see [22, 35] for more details.

COROLLARY 5.6. *Let X be a Banach space and $\varphi : X \to \bar{\mathbf{R}}$ be l.s.c. around $\bar{x} \in \operatorname{dom} \varphi$. Consider the following properties:*

(a) *φ is Lipschitz continuous around $\bar{x}$;*

(b) *E_φ is pseudo-Lipschitzian around $(\bar{x}, \varphi(\bar{x}))$;*

(c) $\partial^\infty \varphi(\bar{x}) = \{0\}$.

Then one always has (a)$\Longleftrightarrow$(b)$\Longrightarrow$(c). *Moreover,* (c) *is equivalent to* (a) *and* (b) *when X is Asplund and φ is normally epi-compact around $\bar{x}$.*

Proof. It follows directly from the definitions that (a)$\Longrightarrow$(b) and, conversely, (b)$\Longrightarrow$(a) if φ is continuous around $\bar{x}$. On the other hand, it is easy to show by contradiction that the pseudo-Lipschitzian property of E_φ around $(\bar{x}, \varphi(\bar{x}))$ automatically implies the upper semicontinuity of φ around $\bar{x}$. This yields the equivalence between (a) and (b) in general Banach spaces. Further taking (2.4), (2.10), and (5.9) into account, we observe that

$$D^\star E_\varphi(\bar{x}, \varphi(\bar{x}))(0) = \partial^\infty \varphi(\bar{x}).$$

Therefore, the implication (b)$\Longrightarrow$(c) follows from Theorem 5.4(II) with $\Phi = E_\varphi$.

It remains to prove that (c)$\Longrightarrow$(b) when X is Asplund and φ is normally epi-compact around $\bar{x}$. But this is a direct corollary of Theorem 5.4(I) since the normal epi-compactness of φ around $\bar{x}$ obviously implies the partial normal compactness of the multifunction E_φ with respect to x around $(\bar{x}, \varphi(\bar{x}))$. □

Remark 5.7. Following the line of [26], one may consider the so-called *global-metric regularity* property of a multifunction Φ around $\bar{x} \in \operatorname{Dom} \Phi$ that is a global counterpart of the regularity property in Definition 5.1. (A local condition $y \in V$ for a neighborhood V of $\bar{y}$ is replaced by $y \in Y$.) Similarly to Theorem 4.8 we can derive the corresponding characteristics of global-metric regularity from their local analogues in Theorem 5.3. Moreover, these results follow directly from Theorem 4.8 due to the equivalence between the covering and global-metric regularity properties established in [32, Proposition 5.2] for the case of general Banach spaces.

In the same way we get effective characteristics for the *(Hausdorff) local Lipschitz continuity* of Φ around $\bar{x}$ which corresponds to $V = Y$ in (5.1). Indeed, when Φ is locally compact around $\bar{x}$, its local Lipschitz continuity around this point is equivalent to the pseudo-Lipschitzian property of Φ around $(\bar{x}, \bar{y})$ for every $\bar{y} \in \Phi(\bar{x})$. This allows us to obtain analogues of the results in Theorem 5.4 for such a local Lipschitz continuity where one should take *every* $\bar{y} \in \Phi(\bar{x})$ in all the criteria and replace $B_\eta(\bar{y})$ with Y in (d); cf. the proof of Theorem 4.8 above and [27, Theorem 3.5]. When Φ happens to be locally *single valued* around $\bar{x}$, the results obtained provide dual criteria for the *classical local Lipschitzian property* of continuous mappings with effective modulus estimates.

Remark 5.8. It immediately follows from the proofs given above that the p.n.c. property in Theorems 4.2, 4.8, 5.3, and 5.4 and their corollaries can be replaced with the weaker sequential limiting property established in Proposition 3.8 and called *partial sequential normal compactness* in [36].

6. Applications to sensitivity analysis for constraint and variational systems. In this section we consider a class of multifunctions $\Phi : X \rightrightarrows Y$ given in the form

$$\Phi(x) = \{y \in Y |\ g(x,y) \in \Lambda,\ (x,y) \in \Omega\}, \tag{6.1}$$

where $g : X \times Y \to Z$ is a mapping between Banach spaces and Λ and Ω are subsets of the spaces Z and $X \times Y$, respectively. Following Rockafellar [42], we call (6.1) *constraint systems* depending on a parameter $x \in X$. One can treat (6.1) as, e.g., a natural generalization of the *feasible solution sets* to perturbed problems in nonlinear programming with equality and inequality constraints described by

$$\Phi(x) = \{y |\ \varphi_i(x,y) \le 0 \text{ for } i = 1,\dots,r \text{ and } \varphi_i(x,y) = 0 \text{ for } i = r+1,\dots,q\}, \tag{6.2}$$

which corresponds to (6.1) when $g = (\varphi_1,\dots,\varphi_q)$, $\Omega = X \times Y$, $Z = \mathbf{R}^q$, and

$$\Lambda = \{(\mu_1,\dots,\mu_q) |\ \mu_i \le 0 \text{ for } i = 1,\dots,r \text{ and } \mu_i = 0 \text{ for } i = r+1,\dots,q\}. \tag{6.3}$$

A special case of (6.1) with $\Lambda = \{0\}$ and $\Omega = X \times Y$ is addressed by the classical implicit function theorem when the mapping (6.1) is single valued and smooth. In general we have an *implicit multifunction* in this case and are interested in properties of *Lipschitz continuity*.

Another important class of multifunctions described by (6.1) is related to solution sets for parametrized *generalized equations*

$$\Phi(x) = \{y \in Y |\ 0 \in f(x,y) + Q(y)\}, \tag{6.4}$$

where $f : X \times Y \to W$ and $Q : Y \rightrightarrows W$. One can see that (6.4) corresponds to (6.1) with

$$g(x,y) = (y, -f(x,y)),\ \ \Lambda = \operatorname{gph} Q,\ \ \Omega = X \times Y,\ \ Z = Y \times W. \tag{6.5}$$

Generalized equations were introduced by Robinson [41] and turned out to be a very convenient model for developing sensitivity analysis and numerical methods in problems of optimization, control, complementarity, mathematical economics, equilibrium, etc.; see, e.g., [11, 19, 30, 41] and references therein. When $Q(y) = N(y;\Omega)$ is the normal-cone operator for a convex set Ω, the generalized equation in (6.4) is reduced to the parametric *variational inequality*

$$\text{find } \ y \in \Omega \ \text{ s.t. } \ \langle f(x,y), \omega - y\rangle \ge 0 \ \ \forall \omega \in \Omega,$$

which is of particular interest for applications. As an important special case, generalized equations/variational inequalities include sets of all *optimal solutions* with associated Lagrange multipliers satisfying first-order necessary optimality conditions in nonlinear programming and optimal control.

In what follows we obtain point conditions ensuring the pseudo-Lipschitzian property for parametric constraint systems (6.1) and generalized equations (6.4) in infinite dimensions. To furnish this, we develop the approach and results in Mordukhovich

[27, 30] where such conditions were obtained in the case of finite dimensional spaces X, Y, Z, and W. This approach is based on using the null-condition (5.7) in Theorem 5.4 valid for arbitrary closed-graph multifunctions and then on effective calculus rules available for our basic generalized differential constructions. In this way we are able to express *sufficient* as well as *necessary and sufficient* conditions for Lipschitzian stability of parametric systems under consideration in terms of their initial data. Moreover, we provide lower and upper estimates for the *exact bounds* of associated Lipschitz moduli that appears to be even more practical in some situations.

THEOREM 6.1. *Let Φ be defined by (6.1), where $g : X \times Y \to Z$ is a continuous function and $\Lambda \subset Z$ and $\Omega \subset X \times Y$ are closed sets with $\bar{z} := g(\bar{x}, \bar{y}) \in \Lambda$ and $(\bar{x}, \bar{y}) \in \Omega$. Then one has the following results:*

(I) *Assume that* $\dim X < \infty$, *Ω is normally compact around $(\bar{x}, \bar{y})$, and either*

(h1) *both Y and Z are Asplund while Λ is normally compact around $\bar{z}$, or*

(h2) *both Y and Z are Banach while g is strictly differentiable at $(\bar{x}, \bar{y})$ with $g'(\bar{x}, \bar{y})$ invertible.*

Then Φ is pseudo-Lipschitzian around $\bar{x}$ if the following three conditions are fulfilled simultaneously:

$$[(x^\star, 0) \in D^\star g(\bar{x}, \bar{y})(z^\star) + N((\bar{x}, \bar{y}); \Omega) \ \text{ and } \ z^\star \in N(\bar{z}; \Lambda)] \Longrightarrow x^\star = 0, \tag{6.6}$$

$$D^\star g(\bar{x}, \bar{y})(z^\star) \cap (-N((\bar{x}, \bar{y}); \Omega)) = \{0\} \ \ \forall z^\star \in N(\bar{z}; \Lambda), \tag{6.7}$$

$$\text{Ker } D^\star g(\bar{x}, \bar{y}) \cap N(g(\bar{x}, \bar{y}); \Lambda) = \{0\}. \tag{6.8}$$

Moreover, under these assumptions one has the upper estimate

$$\begin{aligned}(\text{plip } \Phi)(\bar{x}, \bar{y}) \le \bar{l} := \sup \Big\{ \|x^\star\| \ s.t. \ (x^\star, -y^\star) \in \bigcup [D^\star g(\bar{x}, \bar{y})(z^\star) \ \ with \\ z^\star \in N(\bar{z}; \Lambda)] + N((\bar{x}, \bar{y}); \Omega), \ \ \|y^\star\| \le 1 \Big\}.\end{aligned} \tag{6.9}$$

(II) *Let* $\dim Y < \infty$, *and let one of the following groups of hypotheses hold:*

(h3) *the spaces X and Z are Asplund; the sets Ω and Λ are regular at the points $(\bar{x}, \bar{y})$ and $\bar{z}$, respectively; the qualification conditions* (6.7) *and* (6.8) *are fulfilled; and either g is strictly differentiable at $(\bar{x}, \bar{y})$ or* $\dim Z < \infty$ *and g is Lipschitz continuous at this point with* gph g *regular at $(\bar{x}, \bar{y}, \bar{z})$;*

(h4) *both X and Z are Banach, $\Omega = X \times Y$, and g is strictly differentiable at $(\bar{x}, \bar{y})$ with $g'(\bar{x}, \bar{y})$ invertible.*

Then condition (6.6) *is necessary for the pseudo-Lipschitzian property of Φ around $(\bar{x}, \bar{y})$. Moreover, under these assumptions one has the lower estimate (*$\text{plip } \Phi)(\bar{x}, \bar{y}) \ge \bar{l}$, *where the number $\bar{l}$ is defined in* (6.9).

Proof. Let us observe that gph $\Phi = g^{-1}(\Lambda) \cap \Omega$ for the multifunction Φ in (6.1). Therefore, its coderivative (2.4) is represented in the form

$$D^\star \Phi(\bar{x}, \bar{y})(y^\star) = \{x^\star \in X^\star | \ (x^\star, -y^\star) \in N((\bar{x}, \bar{y}); g^{-1}(\Lambda) \cap \Omega)\}. \tag{6.10}$$

Using the calculus of normal cones [35, Corollary 4.5], we obtain the inclusion

$$N((\bar{x}, \bar{y}); g^{-1}(\Lambda) \cap \Omega) \subset N((\bar{x}, \bar{y}); g^{-1}(\Lambda)) + N((\bar{x}, \bar{y}); \Omega), \tag{6.11}$$

provided that Ω is normally compact around $(\bar{x}, \bar{y})$ and the qualification condition

$$N((\bar{x}, \bar{y}); g^{-1}(\Lambda)) \cap (-N((\bar{x}, \bar{y}); \Omega)) = \{0\} \tag{6.12}$$

is fulfilled. Moreover, equality holds in (6.11) when either $\Omega = X \times Y$ or both sets Ω and $g^{-1}(\Lambda)$ are regular at $(\bar{x}, \bar{y})$.

Next let us use [35, Corollary 6.9] for representing the normal cone to $g^{-1}(\Lambda)$ at $(\bar{x}, \bar{y})$. According to this result the qualification condition (6.8), Asplundity of X, and assumptions (h1) in the theorem ensure the inclusion

$$N((\bar{x}, \bar{y}); g^{-1}(\Lambda)) \subset \bigcup [D^\star g(\bar{x}, \bar{y})(z^\star) |\ z^\star \in N(\bar{z}; \Lambda)], \tag{6.13}$$

where, in addition, equality holds and $g^{-1}(\Lambda)$ is regular at $(\bar{x}, \bar{y})$ under assumptions (h3) concerning g, Λ, and Z. Moreover, equality also holds in (6.13) (but no regularity of $g^{-1}(\Lambda)$ is guaranteed) under assumptions (h2) with a Banach space X; see [33, Corollary 4.4]. Thus (6.12) follows from (6.7) under the assumptions made in (I).

Now substituting (6.13) into (6.10)–(6.12), we arrive at the inclusion

$$D^\star \Phi(\bar{x}, \bar{y})(y^\star) \subset \Big\{ x^\star \in X^\star |\ (x^\star, -y^\star) \in \bigcup [D^\star g(\bar{x}, \bar{y})(z^\star) \ \text{ with } \ z^\star \in N(\bar{z}; \Lambda)] + N((\bar{x}, \bar{y}); \Omega) \Big\}, \tag{6.14}$$

which is valid when Ω is normally compact around $(\bar{x}, \bar{y})$, the qualification conditions (6.7) and (6.8) are fulfilled, and either X is Asplund and assumptions (h1) hold or X is Banach and one has assumptions (h2). (Note that (6.8) is automatic in the latter case.) Therefore, (6.6) implies the null-condition (5.7) in Theorem 5.4(I), while the upper estimate (6.9) follows from (5.8) and (1.2). This proves the sufficiency assertion (I) of the theorem.

To establish the necessity part (II), we use the assumptions above ensuring equality in the coderivative formula (6.14). Finally employing Theorem 5.4(II), we come to all the conclusions (II) of the theorem under the assumptions made therein. □

COROLLARY 6.2. *Let Φ be defined in* (6.1)*, where* $\dim X < \infty$*, Y and Z are Asplund, g is strictly Lipschitzian at $(\bar{x}, \bar{y})$, and the sets Ω and Λ are normally compact around $(\bar{x}, \bar{y})$ and $\bar{z}$, respectively. Then the condition*

$$[(x^\star, 0) \in \partial \langle z^\star, g \rangle (\bar{x}, \bar{y}) + N((\bar{x}, \bar{y}); \Omega) \text{ and } z^\star \in N(\bar{z}; \Lambda)] \Longrightarrow z^\star = 0, \quad x^\star = 0, \tag{6.15}$$

is sufficient for Φ to be pseudo-Lipschitzian around $(\bar{x}, \bar{y})$, and one can replace $D^\star g(\bar{x}, \bar{y})(z^\star)$ with $\partial \langle z^\star, g \rangle (\bar{x}, \bar{y})$ in the upper estimate (6.9).

Proof. First we observe that Proposition 2.5 allows us to replace $D^\star g(\bar{x}, \bar{y})(z^\star)$ with $\partial \langle z^\star, g \rangle (\bar{x}, \bar{y})$ in all conditions (6.6)–(6.9) when g is strictly Lipschitzian at $(\bar{x}, \bar{y})$ and both X and Y are Asplund. Then following [27, Corollary 4.2], one can show that in this case the simultaneous fulfillment of conditions (6.6)–(6.8) is *equivalent* to (6.15). Thus we obtain all the conclusions of the corollary from Theorem 6.1(I). □

Following [27, 30], one can derive various consequences of Theorem 6.1 and Corollary 6.2 for special parametric constraint systems. Let us present effective results for the classical constraint system (6.2) in *nonlinear programming* that seem to be new in infinite dimensions.

COROLLARY 6.3. *Let Φ be given by* (6.2)*, where real-valued functions φ_i are strictly differentiable at $(\bar{x}, \bar{y})$ for all $i = 1, \ldots, q$. Then the following hold:*

(I) *The Mangasarian–Fromovitz condition*

$$[\lambda_1 (\varphi_1)'_y(\bar{x}, \bar{y}) + \cdots + \lambda_q (\varphi_q)'_y(\bar{x}, \bar{y}) = 0] \Longrightarrow \lambda_i = 0 \ \text{ for } \ i = 1, \ldots, q$$
$$\text{if } \ \lambda_i \ge 0 \ \text{ and } \ \lambda_i \varphi_i(\bar{x}, \bar{y}) = 0 \ \text{ for } \ i = 1, \ldots, r$$

is sufficient for Φ *to be pseudo-Lipschitzian around* $(\bar{x},\bar{y})$ *when* $\dim X<\infty$ *and* Y *is Asplund. Moreover, in this case one has the upper modulus estimate*

$$(\text{plip }\Phi)(\bar{x},\bar{y})\le\sup\left\{\left|\sum_{i=1}^q\lambda_i(\varphi_i)'_x(\bar{x},\bar{y})\right|\ \ s.t.\ \ \left|\sum_{i=1}^q\lambda_i(\varphi_i)'_y(\bar{x},\bar{y})\right|\le 1,\right.$$
$$\left.\lambda_i\ge 0,\ \ and\ \ \lambda_i\varphi_i(\bar{x},\bar{y})=0\ \ for\ \ i=1,\dots,r\right\}.$$

(II) *Let* X *be Asplund and let* $\dim Y<\infty$. *Then the condition*

$$\left[\sum_{i=1}^q\lambda_i(\varphi_i)'_y(\bar{x},\bar{y})=0\right]\Longrightarrow\left[\sum_{i=1}^q\lambda_i(\varphi_i)'_x(\bar{x},\bar{y})=0\right]$$
$$if\ \ \lambda_i\ge 0\ \ and\ \ \lambda_i\varphi_i(\bar{x},\bar{y})=0\ \ for\ \ i=1,\dots,r$$

is necessary for the pseudo-Lipschitzian property of Φ *around* $(\bar{x},\bar{y})$ *provided that*

$$[\lambda_1\varphi_1'(\bar{x},\bar{y})+\cdots+\lambda_q\varphi_q'(\bar{x},\bar{y})]\Longrightarrow\lambda_i=0\ \ for\ \ i=1,\dots,q$$
$$if\ \ \lambda_i\ge 0\ \ and\ \ \lambda_i\varphi_i(\bar{x},\bar{y})=0\ \ for\ \ i=1,\dots,r.$$

Proof. This follows from Theorem 6.1 with $g=(\varphi_1,\dots,\varphi_q):X\times Y\to\mathbf{R}^q$, $\Omega=X\times Y$, and Λ defined in (6.3) by taking into account representations (2.5) and

$$N(g(\bar{x},\bar{y});\Lambda)=\{(\lambda_1,\dots,\lambda_q)\in\mathbf{R}^q|\ \lambda_i\ge 0\ \text{ with }\ \lambda_i\varphi_i(\bar{x},\bar{y})=0\ \text{ for }\ i=1,\dots,r\}.\quad\square$$

Next we provide a local sensitivity analysis for *generalized equations* and find effective conditions ensuring the pseudo-Lipschitzian property of the parametric solution sets (6.4).

THEOREM 6.4. *Let* Φ *be defined by* (6.4), *where* $f:X\times Y\to W$ *is continuous around* $(\bar{x},\bar{y})\in\text{gph }\Phi$ *and where* $Q:Y\rightrightarrows W$ *has closed graph around* $(\bar{y},\bar{w})$ *with* $\bar{w}:=-f(\bar{x},\bar{y})$. *Assume that both* Y *and* W *are Asplund, that* $\dim X<\infty$, *and that* Q *is normally compact around* $(\bar{y},\bar{w})$. *Then the condition*

$$[(x^\star,-y^\star)\in D^\star f(\bar{x},\bar{y})(w^\star)\ \ and\ \ y^\star\in D^\star Q(\bar{y},\bar{w})(w^\star)]\Longrightarrow x^\star=0,\ y^\star=0,\ w^\star=0\tag{6.16}$$

is sufficient for Φ *to be pseudo-Lipschitzian around* $(\bar{x},\bar{y})$. *Moreover, in this case one has the upper modulus estimate*

$$(\text{plip }\Phi)(\bar{x},\bar{y})\le\sup\{\|x^\star\|\ \ s.t.\ \ (x^\star,-y^\star)\in D^\star f(\bar{x},\bar{y})(y^\star)+(0,D^\star Q(\bar{y},\bar{w})(w^\star))\ \ with\ \ (y^\star,w^\star)\in Y^\star\times W^\star,\ \|y^\star\|\le 1\}.\tag{6.17}$$

Proof. Let us represent (6.4) in form (6.1) with data (6.5) and evaluate the coderivative of this Φ using inclusion (6.14). In order to do it, we first compute the coderivative of g in (6.5). For this g one obviously has

$$g(x,y)=g_1(x,y)+g_2(x,y),\tag{6.18}$$

where $g_1(x,y):=(y,0)$ and $g_2(x,y):=(0,-f(x,y))$. It is easy to see that

$$D^\star g_1(\bar{x},\bar{y})(u^\star,-w^\star)=(0,u^\star)\ \text{ and }\ D^\star g_2(\bar{x},\bar{y})(u^\star,-w^\star)=D^\star f(\bar{x},\bar{y})(w^\star)$$

for all $(u^\star, w^\star) \in Y^\star \times W^\star$. Now applying the coderivative sum rule [36, Theorem 3.6] in (6.18), we get

$$D^\star g(\bar{x}, \bar{y})(u^\star, -w^\star) = (0, u^\star) + D^\star f(\bar{x}, \bar{y})(w^\star).$$

According to (6.14) this yields the inclusion

$$D^\star \Phi(\bar{x}, \bar{y})(y^\star) \subset \{x^\star \in X^\star |\ (x^\star, -y^\star) \in D^\star f(\bar{x}, \bar{y})(w^\star) + (0, D^\star Q(\bar{y}, \bar{w})(w^\star))\} \tag{6.19}$$

for the coderivative of (6.4). Finally substituting (6.19) into (5.7) and (5.8), we derive conclusions (6.16) and (6.17) of the theorem from the corresponding results of Theorem 5.4(I). □

COROLLARY 6.5. *Let f be strictly Lipschitzian at $(\bar{x}, \bar{y})$, in addition to the other assumptions in Theorem* 6.4. *Then the solution map* (6.4) *is pseudo-Lipschitzian around $(\bar{x}, \bar{y})$ if one has*

$$[0 \in \mathrm{proj}_y \partial\langle w^\star, f\rangle(\bar{x}, \bar{y}) + D^\star Q(\bar{y}, \bar{w})(w^\star)] \Longrightarrow w^\star = 0, \tag{6.20}$$

where $\mathrm{proj}_y \partial\langle w^\star, f\rangle(\bar{x}, \bar{y})$ *denotes the projection of the set* $\partial\langle w^\star, f\rangle(\bar{x}, \bar{y}) \subset X^\star \times Y^\star$ *on the space* $Y^\star$.

Proof. When f is strictly Lipschitzian at $(\bar{x}, \bar{y})$, condition (6.20) is equivalent to (6.16) due to Proposition 2.5. □

Finally let us consider the generalized equation in (6.4), where f is *strictly differentiable* at $(\bar{x}, \bar{y})$. In this case we introduce the adjoint relationship of the same (but linearized/homogenized and unperturbed) form

$$0 \in (f'_y(\bar{x}, \bar{y}))^\star w^\star + D^\star Q(\bar{y}, \bar{w})(w^\star), \tag{6.21}$$

which is called the *adjoint generalized equation* to (6.4) at $(\bar{x}, \bar{y})$. Now we are able to obtain *sufficient* as well as *necessary* conditions for the pseudo-Lipschitzian property of the original solution map (6.4) in the form of *Fredholm's alternative*.

THEOREM 6.6. *Let Φ be given by* (6.4), *where f is strictly differentiable at $(\bar{x}, \bar{y})$ in the framework of Theorem* 6.4. *Then the following hold:*

(I) *Assume that* $\dim X < \infty$ *and either*

(h1) *both Y and W are Asplund while Q is normally compact around $(\bar{y}, \bar{w})$, or*

(h2) *Y is Banach and the operator $f'_x(\bar{x}, \bar{y}) : X \to W$ is invertible (hence* $\dim W < \infty$).

Then Φ is pseudo-Lipschitzian around $(\bar{x}, \bar{y})$ if the adjoint generalized equation (6.21) *has only the trivial solution, i.e.,*

$$[0 \in (f'_y(\bar{x}, \bar{y}))^\star w^\star + D^\star Q(\bar{y}, \bar{w})(w^\star)] \Longrightarrow w^\star = 0. \tag{6.22}$$

Moreover, under these assumptions one has the upper modulus estimate

$$\begin{aligned}(\mathrm{plip}\ \Phi)(\bar{x}, \bar{y}) \le \sup\{\|(f'_x(\bar{x}, \bar{y}))^\star w^\star\| \ \ &s.t.\ \exists y^\star \in D^\star Q(\bar{y}, \bar{w})(w^\star)\\ &with\ \ \|(f'_y(\bar{x}, \bar{y}))^\star w^\star + y^\star\| \le 1\}.\end{aligned} \tag{6.23}$$

(II) *Assume that* $\dim Y < \infty$ *and either*

(h3) *both X and W are Asplund,* $\mathrm{Ker}(f'_x(\bar{x}, \bar{y}))^\star = \{0\}$ *while the graph of Q is regular at $(\bar{y}, \bar{w})$, or*

(h4) *both X and W are Banach while the operator $f'_x(\bar{x}, \bar{y})$ is invertible.*

Then condition (6.22) *is necessary for* Φ *to be pseudo-Lipschitzian around* $(\bar{x},\bar{y})$*, and one has the opposite inequality in* (6.23)*.*

Proof. It is easy to see that the sufficiency part (I) of the theorem follows from Corollary 6.5 under assumptions (h1). To establish (I) under assumptions (h2), we observe that the operator $g : X \times Y \to Y \times W$ defined in (6.5) has the invertible strict derivative at $(\bar{x},\bar{y})$ if the linear operator $f'_x(\bar{x},\bar{y}) : X \times Y \to W$ is invertible. In this case conditions (6.7) and (6.8) are automatic while (6.6) is equivalent to (6.22) and yields the pseudo-Lipschitzian property of (6.4) due to Theorem 6.1(I). The upper estimate (6.23) follows from (6.17) due to formula (2.5). This ends the proof of (I).

The necessity part (II) of the theorem follows from Theorem 6.1(II) for the special structure (6.5). □

COROLLARY 6.7. *Condition* (6.22) *is necessary and sufficient for the pseudo-Lipschitzian property of the solution map* (6.4) *around* $(\bar{x},\bar{y})$ *when spaces* X *and* Y *are finite dimensional and one of the following groups of hypotheses is fulfilled:*

(h1) W *is Asplund,* $\mathrm{Ker}(f'_x(\bar{x},\bar{y}))^\star = \{0\}$*,* Q *is normally compact around* $(\bar{y},\bar{w})$ *while the graph of* Q *is regular at this point;*

(h2) $\dim W < \infty$ *and* $f'_x(\bar{x},\bar{y})$ *is invertible.*

In both cases (h1) *and* (h2) *equality holds in* (6.23)*, where the supremum is attained.*

Proof. This follows directly from Theorem 6.6, combining assertions (I) and (II) therein. We can conclude that the supremum is attained in (6.23) because both spaces X and Y are finite dimensional; cf. [30]. □

Remark 6.8. Similarly to [27, 29, 30] in the finite-dimensional case, we can consider various concretizations and refinements of the results obtained when the multifunction Q admits some special representations. In particular, let $Q : Y \rightrightarrows Y^\star$ be a *subdifferential mapping*, i.e.,

$$Q(y) = \begin{cases} \partial\varphi(y) & \text{if } |\varphi(y)| < \infty, \\ \emptyset & \text{otherwise,} \end{cases} \tag{6.24}$$

in terms of the subdifferential (2.6) of an extended-real-valued function. Then the pseudo-Lipschitzian property of the solution map (6.4) generated by (6.24) can be characterized by the *second-order subdifferential* of φ at $(\bar{y},\bar{w}) \in \mathrm{gph}\,\partial\varphi$ defined as

$$\partial^2\varphi(\bar{y},\bar{w})(w^\star) := (D^\star\partial\varphi)(\bar{y},\bar{w})(w^\star).$$

Note that subdifferential mappings (6.24) cover the case of *variational inequalities and complementarity problems* in (6.4) when φ is the indicator function of a convex set. Note also that another approach [29] can be developed to obtain refined sufficient conditions for Lipschitzian stability of infinite-dimensional variational systems like (6.4).

Remark 6.9. Results of this paper related to *sufficient* conditions for openness, metric regularity, and Lipschitzian stability of set-valued mappings as well as their applications to sensitivity analysis can be obtained in broader classes of Banach spaces using different subdifferential structures. Indeed, one can observe that the proof of the main Theorem 4.2(I) holds true for *sequential* limits of any subdifferentials satisfying the "zero fuzzy calculus" rule of Proposition 2.4 in appropriate Banach spaces. Such a rule appears to be important for all reasonable subdifferentials, and now it is known for most subdifferential constructions used in applications; see, e.g., the recent paper [6] and its references. In this connection let us note that, for an arbitrary Banach space, our basic subdifferential (2.6) is included in the sequential closure of *any* subdifferential satisfying the "zero fuzzy calculus" rule mentioned above; see [35, Theorem 9.7].

REFERENCES

[1] J.-P. Aubin, *Lipschitz behavior of solutions to convex minimization problems*, Math. Oper. Res., 9 (1984), pp. 87–111.

[2] J.-P. Aubin and H. Frankowska, *Set-Valued Analysis*, Birkhäuser, Boston, 1990.

[3] J. M. Borwein, *Stability and regular points of inequality systems*, J. Optim. Theor. Appl., 48 (1986), pp. 9–52.

[4] J. M. Borwein and S. P. Fitzpatrick, *Weak-star sequential compactness and bornological limit derivatives*, J. Convex Anal., 2 (1995), pp. 59–68.

[5] J. M. Borwein and H. M. Strojwas, *Tangential approximations*, Nonlinear Anal., 9 (1985), pp. 1347–1366.

[6] J. M. Borwein and Q. J. Zhu, *Viscosity solutions and viscosity subderivatives in smooth Banach spaces with applications to metric regularity*, SIAM J. Control Optim., 34 (1996), pp. 1568–1591.

[7] J. M. Borwein and D. M. Zhuang, *Verifiable necessary and sufficient conditions for regularity of set-valued and single-valued maps*, J. Math. Anal. Appl., 134 (1988), pp. 441–459.

[8] J. V. Burke, *An exact penalization viewpoint of constraint optimization*, SIAM J. Control Optim., 29 (1991), pp. 968–998.

[9] F. H. Clarke, *Optimization and Nonsmooth Analysis*, Wiley, New York, 1983.

[10] R. Cominetti, *Metric regularity, tangent sets, and second order optimality conditions*, Appl. Math. Optim., 21 (1990), pp. 265–287.

[11] A. L. Dontchev and W. W. Hager, *Implicit functions, Lipschitz maps, and stability in optimization*, Math. Oper. Res., 19 (1994), pp. 753–768.

[12] M. Fabian, *Subdifferentiability and trustworthiness in the light of a new variational principle of Borwein and Preiss*, Acta Univ. Carolin. Math. Phys., 30 (1989), pp. 51–56.

[13] B. Ginsburg and A. D. Ioffe, *The maximum principle in optimal control of systems governed by semilinear equations*, in Nonlinear Analysis and Geometric Methods in Deterministic Optimal Control, IMA Vol. Math. Appl. 78, B. S. Mordukhovich and H. J. Sussmann, eds., Springer-Verlag, New York, 1996, pp. 81–110.

[14] A. D. Ioffe, *Regular points of Lipschitz mappings*, Trans. Amer. Math. Soc., 251 (1979), pp. 61–69.

[15] A. D. Ioffe, *On the local surjection property*, Nonlinear Anal., 11 (1987), pp. 565–592.

[16] A. D. Ioffe, *Approximate subdifferentials and applications,* III: *The metric theory*, Mathematika, 36 (1989), pp. 1–38.

[17] A. Jourani and L. Thibault, *Approximations and metric regularity in mathematical programming in Banach spaces*, Math. Oper. Res., 18 (1993), pp. 390–401.

[18] A. Jourani and L. Thibault, *Verifiable conditions for openness and regularity of multivalued mappings in Banach spaces*, Trans. Amer. Math. Soc., 347 (1995), pp. 1255–1268.

[19] A. J. King and R. T. Rockafellar, *Sensitivity analysis for nonsmooth generalized equations*, Math. Programming, 55 (1992), pp. 193–212.

[20] A. Y. Kruger, *A covering theorem for set-valued mappings*, Optimization, 19 (1988), pp. 736–780.

[21] A. Y. Kruger and B. S. Mordukhovich, *Extremal points and the Euler equation in nonsmooth optimization*, Dokl. Akad. Nauk BSSR, 24 (1980), pp. 684–687.

[22] P. D. Loewen, *Limits of Fréchet normals in nonsmooth analysis*, in Optimization and Nonlinear Analysis, Pitman Res. Notes Math. Ser. 244, A. D. Ioffe et al., eds., Longman, Harlow, UK, 1992, pp. 178–188.

[23] P. D. Loewen, *Optimal Control via Nonsmooth Analysis*, CRM-AMS Lecture Notes in Mathematics, American Mathematical Society, Providence, RI, 1993.

[24] B. S. Mordukhovich, *Maximum principle in problems of time optimal control with nonsmooth constraints*, J. Appl. Math. Mech., 40 (1976), pp. 960–969.

[25] B. S. Mordukhovich, *Approximation Methods in Problems of Optimization and Control*, Nauka, Moscow, 1988.

[26] B. S. Mordukhovich, *Complete characterization of openness, metric regularity, and Lipschitzian properties of multifunctions*, Trans. Amer. Math. Soc., 340 (1993), pp. 1–35.

[27] B. S. Mordukhovich, *Lipschitzian stability of constraint systems and generalized equations*, Nonlinear Anal., 22 (1994), pp. 173–206.

[28] B. S. Mordukhovich, *Generalized differential calculus for nonsmooth and set-valued mappings*, J. Math. Anal. Appl., 183 (1994), pp. 250–288.

[29] B. S. Mordukhovich, *Stability theory for parametric generalized equations and variational inequalities via nonsmooth analysis*, Trans. Amer. Math. Soc., 343 (1994), pp. 609–658.

[30] B. S. Mordukhovich, *Sensitivity analysis for constraint and variational systems by means of set-valued differentiation*, Optimization, 31 (1994), pp. 13–46.
[31] B. S. Mordukhovich, *Discrete approximations and refined Euler-Lagrange conditions for nonconvex differential inclusions*, SIAM J. Control Optim., 33 (1995), pp. 882–915.
[32] B. S. Mordukhovich and Y. Shao, *Differential characterizations of covering, metric regularity, and Lipschitzian properties of multifunctions between Banach spaces*, Nonlinear Anal., 24 (1995), pp. 1401–1424.
[33] B. S. Mordukhovich and Y. Shao, *On nonconvex subdifferential calculus in Banach spaces*, J. Convex Anal., 2 (1995), pp. 211–228.
[34] B. S. Mordukhovich and Y. Shao, *Extremal characterizations of Asplund spaces*, Proc. Amer. Math. Soc., 124 (1996), pp. 197–205.
[35] B. S. Mordukhovich and Y. Shao, *Nonsmooth sequential analysis in Asplund spaces*, Trans. Amer. Math. Soc., 348 (1996), pp. 1235–1280.
[36] B. S. Mordukhovich and Y. Shao, *Generalized differential calculus for infinite dimensional multifunctions*, Set-Valued Anal., 4 (1996), pp. 205–236.
[37] J.-P. Penot, *Metric regularity, openness, and Lipschitzian behavior of multifunctions*, Nonlinear Anal., 13 (1989), pp. 629–643.
[38] J.-P. Penot, *Inverse functions theorems for mappings and multimappings*, SEA Bull. Math., 19 (1995), pp. 1–16.
[39] R. R. Phelps, *Convex Functions, Monotone Operators and Differentiability*, 2nd ed., Lecture Notes in Math. 1364, Springer-Verlag, Berlin, 1993.
[40] S. M. Robinson, *Stability theory for systems of inequalities, part* II: *Differentiable nonlinear systems*, SIAM J. Numer. Anal., 13 (1976), pp. 497–513.
[41] S. M. Robinson, *Generalized equations and their solutions, part* I: *Basic theory*, Math. Programming Stud., 10 (1979), pp. 128–141.
[42] R. T. Rockafellar, *Lipschitzian properties of multifunctions*, Nonlinear Anal., 9 (1985), pp. 867–885.
[43] L. Thibault, *Subdifferentials of compactly Lipschitzian vector-valued functions*, Ann. Mat. Pura Appl. (4), 125 (1980), pp. 157–192.
[44] L. Thibault, *On subdifferentials of optimal value functions*, SIAM J. Control Optim., 29 (1991), pp. 1019–1036.

SIAM J. CONTROL OPTIM.
Vol. 35, No. 1, pp. 315–340, January 1997

018

FIRST- AND SECOND-ORDER SUFFICIENT OPTIMALITY CONDITIONS FOR BANG-BANG CONTROLS*

ANDREI V. SARYCHEV†

Abstract. We study L_1-local optimality of a given control $\tilde{u}(\cdot)$ in the time-optimal control problem for an affine control system. We start with the necessary optimality condition—the Pontryagin maximum principle, which selects the candidates for minimizers, the extremal controls. Generally the corresponding Pontryagin extremals consist of bang-bang and singular subarcs, separated by switching points. In the present paper we treat only pure bang-bang extremals. We introduce extended first and second variations along a bang-bang extremal and establish first- and second-order sufficient optimality conditions for the bang-bang extremal controls.

Key words. optimal control problem, Pontryagin maximum principle, bang-bang extremals, sufficient optimality conditions

AMS subject classifications. 49K30, 93C10

PII. S0363012993246191

1. Introduction.

We consider a nonlinear time-optimal control problem:

$$t \to \min, \tag{1.1}$$

$$\dot{q} = f(q) + G(q)u(\tau),\ q(0) = q_0,\ q \in M,\ u \in U, \tag{1.2}$$

$$q(t) = q_1, \tag{1.3}$$

for an affine control system (1.2) with end-point condition (1.3) on a C^∞-smooth n-dimensional manifold M. Here $G(q) = (g^1(q), \dots, g^r(q))$ and $f(q), g^1(q), \dots, g^r(q)$ are C^∞-smooth vector fields on M; admissible controls $u(\tau) = (u_1(\tau), \dots, u_r(\tau))$ are measurable and take their values in a convex compact polyhedron $U \subset R^r$.

We set the problem of L_1-*local optimality* according to the following definition.

DEFINITION 1.1. *A pair* $(\tilde{u}(\cdot), \tilde{q}(\cdot))$ *meeting* (1.2)–(1.3) *for* $t = T$ *is called* L_1 *locally optimal if there exist* $\Delta > 0$ *and a ball* $\mathcal{U} \supset \tilde{u}(\cdot)$ *in* $L_1^r[0, T]$ *such that no admissible control from* $\mathcal{U}$ *can steer the system* (1.2) *from* q_0 *to* q_1 *in time* $T' \in [T - \Delta, T)$.

A first-order optimality condition for the problem (1.1)–(1.3) is provided by the *Pontryagin maximum principle* (see [7]). If a pair $(\tilde{u}(\cdot), \tilde{q}(\cdot))$ meets this principle for some covector function (Hamiltonian multiplier) $\tilde{\zeta}(\cdot)$, then the triple $(\tilde{u}(\cdot), \tilde{q}(\cdot), \tilde{\zeta}(\cdot))$ is called a *Pontryagin extremal* and $\tilde{u}(\cdot)$ is called the *extremal control.* There can exist different Pontryagin extremals with different $\tilde{\zeta}(\cdot)$ corresponding to the same extremal control $\tilde{u}(\cdot)$.

In what follows we assume that the extremal control $\tilde{u}(\cdot)$ is a piecewise C^1-smooth function of τ. Then due to the Pontryagin maximum principle the domain $[0, T]$ of $\tilde{u}(\cdot)$ can be subdivided into subintervals $0 = \tau_0 < \tau_1 < \cdots < \tau_m < \tau_{m+1} = T$ in such way that for each $\tau \in (\tau_i, \tau_{i+1})$ the maximality condition of the Pontryagin maximum principle is fulfilled on a k_i-dimensional ($k_i \geq 0$) face W_i of the polyhedron U. The

*Received by the editors March 26, 1993; accepted for publication (in revised form) November 27, 1995. This research was supported in part by the A. von Humboldt Foundation, Germany. A version of this paper was presented at the 35th IEEE Conference on Decision and Control, Kobe, Japan, December 11–13, 1996.

http://www.siam.org/journals/sicon/35-1/24619.html

†Department of Mathematics, University of Aveiro, 3810, Aveiro, Portugal (ansar@mat.ua.pt).

subinterval (τ_i, τ_{i+1}) is called *bang-bang* if $k_i = 0$; i.e., the maximum is achieved at a vertex of U and is called *singular* if k_i is positive. The points τ_i $(i = 1, \dots, m)$ are called *switching points*.

As is well known, the extremality of control $\tilde{u}(\cdot)$ does not imply its optimality. To ascertain optimality one should at least investigate the *second variation* of the control system (1.2) on singular subintervals. But even *bang-bang Pontryagin extremals*, which have no singular arcs, may happen to be nonoptimal. Corresponding examples, as well as some high-order necessary optimality conditions for bang-bang extremals, can be found in [4, 15, 16, 17].

One should note a characteristic feature of bang-bang Pontryagin extremals: the first variation of the system (1.2) along these extremals cannot be nontrivially nullified. Due to this fact the traditional approach of the calculus of variations and optimal control theory is no longer valid in the bang-bang case, since the first-order conditions do not guarantee optimality, while high-order variations, which according to this approach are to be defined on the kernel of the first variation, simply do not exist because this kernel is trivial.

To overcome the difficulties we shall introduce below an extension of the first variation by adding to the space of admissible variations of the extremal control $\tilde{u}(\cdot)$ some finite-dimensional space of Dirac measures, located at the switching points of the extremal. This extension gives new addends for the first and the second variations, which are called *first* and *second variations of the system at switching points of extremal*.

Studying the first variation at switching points we derive a *first-order sufficient condition of L_1-local optimality* for bang-bang Pontryagin extremals (Theorem 6.1). When this condition is not met for a bang-bang extremal, we bring into consideration a corresponding *second variation at switching points*. It is a finite-dimensional quadratic form and its *negative definiteness* is the crucial point for setting *second-order sufficient conditions of optimality* for bang-bang Pontryagin extremals (Theorems 7.1 and 7.2).

In a forthcoming paper we are going to present a Legendre–Jacobi–Morse-type theory of the second variation for Pontryagin extremals, containing both bang-bang and singular arcs, and establish for these extremals second-order sufficient optimality conditions. Most of these results has been published in the preprint [13]; part of the formulations have been presented in [12, 14].

2. Preliminaries. Below we introduce some notation of *chronological calculus* developed by Agrachev and Gamkrelidze. The details are to be found in [1, 3].

Let $C^\infty(M)$ be an algebra of infinitely differentiable or smooth functions on M. The value of $\varphi \in C^\infty(M)$ at a point $q \in M$ is denoted by $q \circ \varphi$. The correspondence $\varphi \mapsto q \circ \varphi$ defines a multiplicative functional on M. A diffeomorphism $P : M \to M$ is identified with the corresponding automorphism of $C^\infty(M)$: $\varphi(\cdot) \mapsto P \circ \varphi(\cdot) = \varphi(P(\cdot))$. The group of diffeomorphisms $P : M \to M$ is denoted by $\mathrm{Diff}M$. The value of $P \in \mathrm{Diff}M$ at a point $q \in M$ is denoted $q \circ P$. Smooth vector fields on M are arbitrary derivations of the algebra $C^\infty(M)$, or R-linear mappings $Y : C^\infty(M) \to C^\infty(M)$ satisfying the Leibnitz rule: $Y(\varphi\psi) = (Y\varphi)\psi + \varphi(Y\psi)$. The value of a vector field Y at a point $q \in M$ is denoted $q \circ Y$; it belongs to the tangent space T_qM to M at point q. We denote by $[Y, Z]$ the *commutator* or *Lie bracket* of vector fields Y, Z. In local coordinates on M this Lie bracket is calculated as

$$[Y, Z] = \left[\sum_{i=1}^{m} Y_i \partial/\partial x_i, \sum_{i=1}^{m} Z_i \partial/\partial x_i\right] = \sum_{i=1}^{m} (\partial Z_i/\partial x Y - \partial Y_i/\partial x Z)\partial/\partial x_i.$$

The Lie algebra of smooth vector fields on M is denoted by $\mathrm{Vect}M$. Let us note that defining diffeomorphisms and vector fields as operators of $C^\infty(M)$ we embed them in some *linear* space of linear operators $\mathcal{L}(C^\infty(M), C^\infty(M))$.

For $P \in \mathrm{Diff}M$ the symbol $\mathrm{Ad}P$ denotes the following inner automorphism of the Lie algebra. $\mathrm{Vect}M$: $\mathrm{Ad}PY = P \circ Y \circ P^{-1} = (\partial P^{-1}/\partial x Y)(P(\cdot)) = P_*^{-1}Y$. (We denote by $P_*^{-1}Y$ the result of translation of the vector field Y by the diffeomorphism P^{-1}.) For $Y \in \mathrm{Vect}M$ the inner derivation $\mathrm{ad}Y$ of $\mathrm{Vect}M$ is defined as $(\mathrm{ad}Y)Z = [Y, Z]$ $\forall Z \in \mathrm{Vect}M$.

The *Whitney topology* in $C^\infty(M)$ is introduced by means of a family of seminorms $\|\cdot\|_{s,K}$, where $s \geq 0$, K is a compact set ($K \subset M$), and the seminorm $\|\cdot\|_{s,K}$ introduces the topology of uniform convergence of all derivatives of order $\leq s$ on the compact set K. The Whitney topology in the space of vector fields is defined by means of the family of seminorms

$$\|Y\|_{s,K} = \sup\{\|Y\varphi\|_{s,K} : \ \|\varphi\|_{s+1,K} = 1\} \ \forall Y \in \mathrm{Vect}M.$$

The Whitney topology introduces the structure of a Frechet space in $\mathrm{Vect}M$.

A *flow* on M is an absolutely continuous curve $\tau \mapsto P_\tau$ in $\mathrm{Diff}\, M$ ($P_0 = I$—the identical isomorphism) such that $\varphi(P_\tau(\cdot))$ is absolutely continuous with respect to τ for every $\varphi \in C^\infty(M)$. A *time-dependent vector field* on M is a locally integrable curve $\tau \mapsto Y_\tau$ in $\mathrm{Vect}M$ such that $\forall \varphi \in C^\infty(M)$, the function $(Y_\tau \varphi)(q)$ is measurable with respect to τ for every $q \in M$ and

$$\int_{t_1}^{t_2} \|Y_\tau \varphi\|_{s,K} d\tau < +\infty \ \forall t_1 \leq t_2, \ \forall s, K.$$

A time-dependent vector field $\tau \mapsto Y_\tau$ defines the ordinary differential equation $\dot{q}(\tau) = q(\tau) \circ Y_\tau$ on the manifold M; if every solution of the differential equation is defined $\forall \tau \in R$, then the vector field Y_τ is called complete. A complete vector field Y_τ defines a flow P_τ ($\tau \in R$) on M, the unique solution of the operator differential equation

$$dP_\tau/d\tau = P_\tau \circ Y_\tau, \ P_0 = I. \tag{2.1}$$

This solution can be represented (see [1, 3]) as *right chronological exponential* in Y_τ, denoted by $P_t = \overrightarrow{\exp} \int_0^t Y_\tau d\tau$. If the vector field Y_τ is time independent, i.e., $Y_\tau \equiv Y$, then the corresponding flow is denoted by $P_t = e^{tY}$.

Let us also introduce the *Volterra expansion*, or *Volterra series*, for the chronological exponential $\overrightarrow{\exp} \int_0^t Y_\tau d\tau$. It is expressed as follows (see [1, 3]):

$$\overrightarrow{\exp} \int_0^t Y_\tau d\tau \asymp I + \sum_{i=1}^{\infty} \int_0^t d\tau_1 \int_0^{\tau_1} d\tau_2 \ldots \int_0^{\tau_{i-1}} d\tau_i (Y_{\tau_i} \circ \cdots \circ Y_{\tau_1}) = I + \int_0^t Y_{\tau_1} d\tau_1$$
$$+ \int_0^t \int_0^{\tau_1} (Y_{\tau_2} \circ Y_{\tau_1}) d\tau_2 d\tau_1 + \int_0^t \int_0^{\tau_1} \int_0^{\tau_2} (Y_{\tau_3} \circ Y_{\tau_2} \circ Y_{\tau_1}) d\tau_3 d\tau_2 d\tau_1 + \cdots. \tag{2.2}$$

One can prove that the Volterra series (2.2) provides an asymptotic approximation for $\overrightarrow{\exp} \int_0^t Y_\tau d\tau$. Namely, according to [1], $\forall \varphi \in C^\infty(M)$

$$\left\| \overrightarrow{\exp} \int_0^t Y_\tau d\tau - \left(I + \sum_{i=1}^{\ell-1} \int_0^t d\tau_1 \int_0^{\tau_1} d\tau_2 \ldots \int_0^{\tau_{i-1}} d\tau_i (Y_{\tau_i} \circ \cdots \circ Y_{\tau_1}) \right) \varphi \right\|_{s,K}$$
$$\leq C e^{(c_2 \int_0^t \|Y_\tau\|_{s,\tilde{K}} d\tau)} \left(\int_0^t \|Y_\tau\|_{s+\ell-1,\tilde{K}} d\tau \right)^\ell \|\varphi\|_{s+\ell,\tilde{K}}, \tag{2.3}$$

where $\tilde{K}$ is some compact neighborhood of the compact set K. Since further on we deal with some neighborhood of a continuous curve $\tilde{q}(\cdot) : [0,T] \to M$, then without lack of generality we may assume M to be compact or, all the same, ignore dependence of $\|\cdot\|_{s,K}$ on K.

It follows also from the results of [1] that

$$\left\|\left(\int_0^t d\tau_1 \int_0^{\tau_1} d\tau_2 \ldots \int_0^{\tau_{\ell-1}} d\tau_\ell (Y_{\tau_\ell} \circ \cdots \circ Y_{\tau_1})\right)\varphi\right\|_s \le C\left(\int_0^t \|Y_\tau\|_{s+\ell-1} d\tau\right)^\ell \|\varphi\|_{s+\ell}.$$

We put

$$\left\|\left(\int_0^t d\tau_1 \int_0^{\tau_1} d\tau_2 \ldots \int_0^{\tau_{\ell-1}} d\tau_\ell (Y_{\tau_\ell} \circ \cdots \circ Y_{\tau_1})\right)\right\|$$
$$= \sup\left\{\left\|\int_0^t d\tau_1 \int_0^{\tau_1} d\tau_2 \ldots \int_0^{\tau_{\ell-1}} d\tau_\ell (Y_{\tau_\ell} \circ \cdots \circ Y_{\tau_1})\varphi\right\|_0 \ : \ \|\varphi\|_\ell = 1\right\}$$

and

$$\left\|q_0 \circ \left(\int_0^t d\tau_1 \int_0^{\tau_1} d\tau_2 \ldots \int_0^{\tau_{\ell-1}} d\tau_\ell (Y_{\tau_\ell} \circ \cdots \circ Y_{\tau_1})\right)\right\|$$
$$= \inf_{\mathcal{V}} \sup\left\{\left\|\int_0^t d\tau_1 \int_0^{\tau_1} d\tau_2 \ldots \int_0^{\tau_{\ell-1}} d\tau_\ell (Y_{\tau_\ell} \circ \cdots \circ Y_{\tau_1})\varphi\right\|_0 \ | \right.$$
$$\left. \operatorname{supp}\varphi \subset \mathcal{V},\ \|\varphi\|_\ell = 1\right\} \tag{2.4}$$

with the infimum taken over the set of all neighborhoods $\mathcal{V}$ of the point $q_0 \in M$.

Finally, if $\zeta \in \mathcal{T}^*_{q_0}M$ is a covector, then we put

$$\left\|\left\langle \zeta, q_0 \circ \left(\int_0^t d\tau_1 \int_0^{\tau_1} d\tau_2 \ldots \int_0^{\tau_{\ell-1}} d\tau_\ell (Y_{\tau_\ell} \circ \cdots \circ Y_{\tau_1})\right)\right\rangle\right\|$$
$$= \inf_{\mathcal{V}} \sup\left\{\left\|\int_0^t d\tau_1 \int_0^{\tau_1} d\tau_2 \ldots \int_0^{\tau_{\ell-1}} d\tau_\ell (Y_{\tau_\ell} \circ \cdots \circ Y_{\tau_1})\varphi\right\|_0 \ | \right.$$
$$\left. \operatorname{supp}\varphi \subset \mathcal{V},\ d\varphi|_{q_0} = \zeta,\ \|\varphi\|_\ell = 1\right\}. \tag{2.5}$$

If one considers a family of operators $\mathrm{Ad}P_t$ produced by a flow $\overrightarrow{\exp}\int_0^t Y_\tau d\tau$, then differentiating $\mathrm{Ad}P_t Z = P_t \circ Z \circ P_t^{-1}$ with respect to t, one obtains

$$d(\mathrm{Ad}P_t Z)/dt = \mathrm{Ad}P_t \circ \mathrm{ad}Y_t Z \ \forall Z \in \mathrm{Vect}M,$$

or after omission of Z : $d(\mathrm{Ad}P_t)/dt = \mathrm{Ad}P_t \circ \mathrm{ad}Y_t$. This means, that $\mathrm{Ad}P_t$ satisfies an operator differential equation similar to (2.1) and justifies the notation

$$\mathrm{Ad}\left(\overrightarrow{\exp}\int_0^t Y_\tau d\tau\right) = \overrightarrow{\exp}\int_0^t \mathrm{ad}Y_\tau d\tau.$$

This last chronological exponential also admits the Volterra expansion

$$\overrightarrow{\exp}\int_0^t \mathrm{ad}Y_\tau d\tau \asymp I + \sum_{i=1}^\infty \int_0^t d\tau_1 \int_0^{\tau_1} d\tau_2 \ldots \int_0^{\tau_{i-1}} d\tau_i (\mathrm{ad}Y_{\tau_i} \circ \cdots \circ \mathrm{ad}Y_{\tau_1}).$$

Let us assume for the moment that time-dependent vector fields Y_τ and $Y_\tau + Z_\tau$ are complete. The following variant of the variation of constants formula appeared in [1]; it represents the flow $\overrightarrow{\exp}\int_0^t (Y_\tau + Z_\tau)d\tau$ as a perturbation of the flow $\overrightarrow{\exp}\int_0^t Y_\tau d\tau$:

$$\overrightarrow{\exp}\int_0^t (Y_\tau + Z_\tau)d\tau = \overrightarrow{\exp}\int_0^t \mathrm{Ad}\left(\overrightarrow{\exp}\int_0^\tau Y_\theta d\theta\right) Z_\tau d\tau \circ \overrightarrow{\exp}\int_0^t Y_\tau d\tau.$$

By virtue of the aforesaid one may also write

$$\overrightarrow{\exp}\int_0^t (Y_\tau + Z_\tau)d\tau = \overrightarrow{\exp}\int_0^t \overrightarrow{\exp}\int_0^\tau \mathrm{ad}Y_\theta d\theta Z_\tau d\tau \circ \overrightarrow{\exp}\int_0^t Y_\tau d\tau. \tag{2.6}$$

3. Local properties of end-point mapping and optimality. Let $\tilde{u}(\tau)$ be an admissible control which steers the system (1.2) from q_0 to q_1 in time T. In what follows we always assume that the final moment T is a generic (Lebesgue) point of $\tau \mapsto \tilde{u}(\tau)$ (a non-Lebesgue point would require minor modifications to the transversality condition of the Pontryagin maximum principle and of the high-order conditions which are presented below).

Let us put $\tilde{f}_\tau(q) = f(q) + G(q)\tilde{u}(\tau)$ and denote by $\tilde{P}_t = \overrightarrow{\exp}\int_0^t \tilde{f}_\tau d\tau$ $(t \in [0,T])$ the solution of ordinary differential equation

$$\partial P_\tau / d\tau = P_\tau \circ \tilde{f}_\tau,\ \tau \in [0,T],\ P_0 = I.$$

Let $u(\cdot)$ be an admissible variation of $\tilde{u}(\cdot)$, i.e., $\tilde{u}(\tau)+u(\tau) \in U$ for every $\tau \in [0,T]$. We consider a family of mappings $F_t : L^r_\infty \to M$ defined on the space of admissible variations. For a given t the mapping F_t maps $u(\cdot)|_{[0,t]}$ into the point

$$q(t) = q_0 \circ \overrightarrow{\exp}\int_0^t (\tilde{f}_\tau + Gu(\tau))d\tau$$

of the trajectory $q(\cdot)$ of the system (1.2) driven by $\tilde{u}(\cdot) + u(\cdot)$. When $t = T$ we will call F_T an *end-point mapping*.

It is known (see [1, 3]) that F_t is C^∞-smooth with respect to $u(\cdot)$ in some neighborhood of the origin of $L^r_\infty[0,T]$. By virtue of the variation of constants formula (2.6) one can represent F_t as

$$\begin{aligned} F_t(u(\cdot)) &= q_0 \circ \overrightarrow{\exp}\int_0^t (\tilde{f}_\tau + Gu(\tau))d\tau \\ &= q_0 \circ \overrightarrow{\exp}\int_0^t \left(\overrightarrow{\exp}\int_0^\tau \mathrm{ad}\tilde{f}_\theta d\theta\right) Gu(\tau)d\tau \circ \overrightarrow{\exp}\int_0^t \tilde{f}_\tau d\tau. \end{aligned} \tag{3.1}$$

It is often more suitable to use a family of mappings $\Phi(u(\cdot)) = F_t(u(\cdot)) \circ \tilde{P}_t^{-1}$ in place of F_t. To calculate Φ_t one should map $u(\cdot)$ into $q(t)$ by means of F_t and then pull the result back by $\tilde{P}_t^{-1} = (\overrightarrow{\exp}\int_0^t \tilde{f}_\tau d\tau)^{-1}$. It follows from (3.1) that Φ_t can be represented as

$$\Phi_t(u(\cdot)) = q_0 \circ \overrightarrow{\exp}\int_0^t X_\tau u_\tau d\tau, \tag{3.2}$$

where

$$X_\tau = (X^1_\tau, \ldots, X^r_\tau),\ X^i_\tau = \overrightarrow{\exp}\int_0^\tau \mathrm{ad}\tilde{f}_\theta d\theta g^i\ (i = 1, \ldots, r). \tag{3.3}$$

This means that Φ_t is determined by a time-dependent differential equation

$$\dot{q}(\tau) = q(\tau) \circ X_\tau u(\tau),\ q(0) = q_0, \tag{3.4}$$

which is linear (homogeneous) with respect to the control u.

Obviously $\Phi_t(0) = q_0$ for any t. We call Φ_T the *pulled-back end-point mapping* (below the words pulled-back are omitted for the sake of brevity).

Taking the Volterra expansion (see (2.2) for the chronological exponential (3.2) (with $t = T$), we obtain Taylor expansion for Φ_T:

$$\begin{aligned}\Phi_T(u(\cdot)) = q_0 + q_0 \circ \int_0^T X_\tau d\tau + q_0 \circ \int_0^T \int_0^\tau X_\xi u(\xi) d\xi \circ X_\tau u(\tau) d\tau \\ + q_0 \circ \int_0^T \int_0^\tau \int_0^\xi X_\theta u(\theta) d\theta \circ X_\xi u(\xi) d\xi \circ X_\tau u(\tau) d\tau + \cdots .\end{aligned} \tag{3.5}$$

(Let us recall that the vector fields and their compositions, appearing in this formula, belong to the *linear* space of operators over $C^\infty(M)$.)

We are going to establish the relations between the optimality of $\tilde{u}(\cdot)$ and the properties of the end-point mappings F_T and Φ_T. We shall present some (almost trivial) results which provide *sufficient optimality conditions* for $\tilde{u}(\cdot)$ in terms of local properties of F_T and Φ_T. In fact one can hardly apply these results directly, but they are useful as auxiliary tools. To prove optimality we shall verify the conditions of one of these auxiliary propositions.

Let us for a moment consider $F_t(u(\cdot))$ as the mapping defined on the pairs $(t, u(\cdot))$. Define a small (semi)neighborhood W_Δ of $(T, \tilde{u}(\cdot))$ as $W \times (T-\Delta, T]$, where W is small in terms of the L_1-metric neighborhood of $\tilde{u}(\cdot)$ in the space of admissible controls.

For a subset $A \subset M$ let us define a *set of tangent vectors to A at a point $a \in \bar{A}$* as a set of tangent vectors to the C^1-curves $\gamma : [0, \varepsilon) \to M$, starting at $\gamma(0) = a$ and such that $\gamma(\tau) \in A$ for $\tau \in (0, \varepsilon)$.

PROPOSITION 3.1. *Let $F_T(\tilde{u}(\cdot)) = q_1$. If for some small (in terms of the metric of $R \times L_1^r[0, T]$) neighborhood W_Δ of $(T, \tilde{u}(\cdot))$ the vector $q_1 \circ \tilde{f}_T$ does not belong to the set of tangent vectors (at the point q_1) to the image $F_\cdot(W_\Delta)$, then the control $\tilde{u}(\cdot)$ is optimal for the problem* (1.1)–(1.3).

Proof. Indeed, if $\tilde{u}(\cdot)$ is nonoptimal, this means that for some $\varepsilon \in (0, \Delta]$,

$$q_1 \in F_{T-\varepsilon}(W) \subseteq F_\cdot(W_\varepsilon) \subseteq F_\cdot(W_\Delta).$$

Let us consider the trajectory of the control system (1.2), starting at q_1, which is driven by the constant control $u = \tilde{u}(T)$ (or, equivalently, the trajectory of the vector field $\tilde{f}_T$). It follows from the definition of F_t that $q_1 \circ e^{\eta \tilde{f}_T} \in F_{T-\varepsilon+\eta}(W) \subseteq F_\cdot(W_\Delta)$ for $\eta \in [0, \varepsilon]$. Therefore the tangent vector $q_1 \circ \tilde{f}_T$ to the curve $q_1 \circ e^{\eta \tilde{f}_T}$ at q_1 belongs to the set of tangent vectors, and we get a contradiction. □

In what follows we proceed from a stronger hypothesis than that of Proposition 3.1. Namely we will assume the existence of a nonzero covector $\zeta_1 \in \mathcal{T}_{q_1}^* M$, such that the tangent vectors to the image $F_\cdot(W_\Delta)$ belong to the semispace $\{y :\ \langle \zeta_1, y \rangle \le 0\}$ of $\mathcal{T}_{q_1} M$, while $\langle \zeta_1, q_1 \circ \tilde{f}_T \rangle > 0$. When the first-order (with respect to the variation of the final moment T) tangent vector $dF_{T-\varepsilon}(\tilde{u}(\cdot))/d\varepsilon|_{\varepsilon=0} = -q_1 \circ \tilde{f}_T$ belongs to the *open* semispace $\{y :\ \langle \zeta_1, y \rangle < 0\}$, one can easily prove that whenever the tangent vectors to the image $F_T(W)$ (with fixed T!) belong to the semispace $\{y :\ \langle \zeta_1, y \rangle \le 0\}$, then the same holds for the tangent vectors to $F_\cdot(W_\Delta)$. So in this case one can dispense with the variation of the final moment T. This gives the following.

PROPOSITION 3.2. *If there exists a neighborhood (in terms of L_1-metric) W of the control $\tilde{u}(\cdot)$ in the space of admissible controls such that the vector $q_1 \circ \tilde{f}_T$ can be strictly separated by a nonzero covector $\zeta_1 \in \mathcal{T}^*_{q_1}M$ from the set of tangent vectors to the image $F_T(W)$ at the point q_1, then $\tilde{u}(\cdot)$ is optimal for the problem* (1.1)–(1.3).

Finally we will reformulate the Proposition 3.2 by passing from F_T to Φ_T and pulling back the whole consideration from q_1 to q_0.

PROPOSITION 3.3 (auxiliary lemma on optimality). *If there exists a neighborhood (in terms of the L_1-metric) W of the control $\tilde{u}(\cdot)$ in the space of admissible controls such that the vector $Y_T = q_0 \circ Ad(\overrightarrow{\exp} \int_0^T \tilde{f}_\tau d\tau)\tilde{f}_T$ can be strictly separated by a nonzero covector $\zeta_0 \in \mathcal{T}^*_{q_0}M$ from the set of tangent vectors to the image $\Phi_T(W)$ at the point q_0, then $\tilde{u}(\cdot)$ is optimal for the problem* (1.1)–(1.3).

4. First and second variations of control system. Pontryagin maximum principle. In the previous section we have established the relation between optimality of $\tilde{u}(\cdot)$ and local properties of the end-point mappings F_T and Φ_T. These local properties are essentially determined by Taylor expansions of F_T and Φ_T. In this section we define first and second variations of the control system (1.2) which correspond to the first and second differentials of Φ_T. We also formulate the Pontryagin maximum principle, which provides *first-order necessary optimality condition* for $\tilde{u}(\cdot)$.

From the Volterra expansion (3.5) of Φ_T one derives an expression

$$\Phi'_T u(\cdot) = \int_0^T q_0 \circ X_\tau u(\tau) d\tau \tag{4.1}$$

for the (first) differential of Φ_T at the origin of $L^r_\infty[0,T]$; here X_τ is defined according to (3.3). Obviously (4.1) defines linear mapping from $L^r_\infty[0,T]$ into $\mathcal{T}_{q_0}M$.

Let $\tilde{u}(\cdot) \in \mathcal{U}$, where $\mathcal{U}$ is the set of admissible controls of the system (1.2). A *cone of admissible variations* of $\tilde{u}(\cdot)$ is by definition the (convex) conic hull of the set $\mathcal{U} - \tilde{u}(\cdot)$. We denote this cone by $K_{\tilde{u}}\mathcal{U}$.

DEFINITION 4.1. *The restriction of the first differential Φ'_T to the cone $K_{\tilde{u}}\mathcal{U}$ is called the first variation of the system* (1.2) *on $[0,T]$ along the control $\tilde{u}(\cdot)$. The image $\Phi'_T(K_{\tilde{u}}\mathcal{U})$ is called the first variational cone along $\tilde{u}(\cdot)$.*

It is clear, that (for any $\varepsilon > 0$) the first variational cone along $\tilde{u}(\cdot)$ is a subset of the set of tangent vectors to the image $\Phi_T(\mathcal{U}_\varepsilon)$, where $\mathcal{U}_\varepsilon = \{u(\cdot) \in \mathcal{U} : \ \|u - \tilde{u}\|_{L_1} < \varepsilon\}$. The auxiliary lemma on optimality implies that the strict separability of the vector $Y_T = q_0 \circ \mathrm{Ad}(\overrightarrow{\exp} \int_0^T \tilde{f}_\tau d\tau)\tilde{f}_T$ from the set of tangent vectors is sufficient for optimality of $\tilde{u}(\cdot)$. The Pontryagin maximum principle implies (see [7]) that separability of Y_T from the first variational cone along $\tilde{u}(\cdot)$ is necessary for optimality of $\tilde{u}(\cdot)$.

DEFINITION 4.2. *A control $\tilde{u}(\cdot)$ is called an extremal control for the problem* (1.1)–(1.3) *if the vector $Y_T = q_0 \circ Ad(\overrightarrow{\exp} \int_0^T (f + G\tilde{u}(\tau))d\tau)(f + G\tilde{u}(T))$ can be separated from the first variational cone along $\tilde{u}(\cdot)$.*

According to the definition and the Pontryagin maximum principle, any optimal control for the problem (1.1)–(1.3) must be an extremal one.

The separability of the vector Y_T from the first variational cone means the existence of a (possibly nonunique) covector $\zeta_0 \in \mathcal{T}^*_{q_0}M$ such that $\forall u(\cdot) \in K_{\tilde{u}}\mathcal{U}$ the following inequalities hold:

$$\int_0^T \langle \zeta_0, q_0 \circ X_\tau u(\tau)\rangle d\tau \le 0, \tag{4.2}$$

$$\langle \zeta_0, Y_T \rangle \ge 0. \tag{4.3}$$

The last inequality is the so called *transversality condition of Pontryagin maximum principle*. We call its strengthened form,

$$\langle \zeta_0, Y_T \rangle > 0, \tag{4.4}$$

the *strong transversality condition*.

The inequality (4.2) implies the inequalities

$$\langle \zeta_0, q_0 \circ X_\tau(u - \tilde{u}(\tau)) \rangle \le 0 \ \forall u \in U \text{ for almost every } \tau \in [0, T],$$

so that for almost every $\tau \in [0, T]$ we have

$$\tilde{u}(\tau) \in \mathrm{Argmax}_{u \in U} \langle \zeta_0, q_0 \circ X_\tau u \rangle. \tag{4.5}$$

Introducing linear operator $\Xi_\tau : R^r \to \mathcal{T}_{q_0} M$, which maps $u \in R^r$ to $q_0 \circ X_\tau u$, we consider its adjoint operator $\Xi_\tau^* : \mathcal{T}_{q_0}^* M \to R^{r^*}$, and denoting by $\chi_\tau = \Xi_\tau^* \zeta_0$, we transform (4.5) into

$$\tilde{u}(\tau) \in \mathrm{Argmax}_{u \in U} \langle \chi_\tau, u \rangle. \tag{4.6}$$

The covector function χ_τ is called the *switching function*.

The conditions (4.3) and (4.6) can be easily transformed into a standard form of Pontryagin maximum principle. Indeed by the definition of X_τ, the switching function χ_τ can be represented as $\chi_\tau = \tilde{\zeta}(\tau) G(\tilde{q}(\tau))$, where the covector function $\tilde{\zeta}(\cdot)$ is a solution of the *adjoint equation of the Hamiltonian system* with the Hamiltonian

$$H(q, \zeta, u) = \langle \zeta, f(q) + G(q)u \rangle. \tag{4.7}$$

In local coordinates this (linear) adjoint equation has the form

$$\dot{\zeta} = -\partial H / \partial q(\tilde{q}(\tau), \zeta, \tilde{u}(\tau)). \tag{4.8}$$

The initial condition for $\tilde{\zeta}(\cdot)$ is $\zeta(0) = \zeta_0$.

Now we formulate the following.

THEOREM 4.1 (Pontryagin maximum principle [5, 7]). *If $(\tilde{q}(\cdot), \tilde{u}(\cdot), T)$ is a solution of the optimal control problem* (1.1)–(1.3)*, then there exists a nonzero absolutely continuous covector function $\tilde{\zeta} : R \to \mathcal{T}^* M$ meeting the condition $\tilde{\zeta}(t) \in \mathcal{T}_{\tilde{q}(t)} M$ and satisfying in local coordinates the adjoint equation* (4.8) *with the Hamiltonian* (4.7) *such that for $(\tilde{q}(\cdot), \tilde{u}(\cdot), \tilde{\zeta}(\cdot), T)$ the following conditions hold:*

(i) *maximality condition* (*equivalent to* (4.6))*:*

$$H(\tilde{q}(t), \tilde{\zeta}(t), \tilde{u}(t)) = \max\{H(\tilde{q}(t), \tilde{\zeta}(t), u) : \ u \in U\} \ \textit{a.e. on } [0, T]; \tag{4.9}$$

(ii) *transversality condition* (*equivalent to* (4.3))*:*

$$H(\tilde{q}(T), \tilde{\zeta}(T), \tilde{u}(T)) \ge 0. \qquad \square \tag{4.10}$$

We shall call the strong transversality condition the strengthened version of (4.10):

$$H(\tilde{q}(T), \tilde{\zeta}(T), \tilde{u}(T)) > 0. \tag{4.11}$$

Remark 4.1. Since variations of $\tilde{u}(\cdot)$ on subsets $\Omega \subset [0, T]$ of zero measure have no effect, then without loss of generality we may assume that the maximality condition (4.9) is satisfied at *all* points of $[0, T]$.

It is evident that for any $\tau \in [0,T]$ maximum is attained at some (perhaps 0-dimensional) face W_τ of the polyhedron U. We will denote by V_τ a directing subspace of this face; $V_\tau = \operatorname{span}\{W_\tau - w\}$, where $w \in W$ can be chosen arbitrarily. Evidently $\forall v \in V_\tau,\ \tau \in [0,T]:\ \langle \chi_\tau, v\rangle = 0$.

In what follows we assume the mapping $\tau \to V_\tau$ to be piecewise constant with $0 = \tau_0 < \tau_1 < \cdots < \tau_m < \tau_{m+1} = T$ determining the intervals of constancy. We will call $(\tau_i, \tau_{i+1}]$ a *bang-bang interval of the extremal control* $\tilde{u}(\cdot)$ if $\dim V_\tau = 0$, or equivalently W_τ is a vertex of U for $\tau \in (\tau_i, \tau_{i+1})$, and a *singular interval* if $\dim V_\tau > 0$ on (τ_i, τ_{i+1}). The points τ_i are called *switching points*; if τ_i separates two bang-bang intervals, then we shall call it a *bang-bang switching point*. It is obvious that $W_{\tau_i \pm 0} \subseteq W_{\tau_i}$ for any τ_i and $\dim W_{\tau_i} \geq 1$. The face W_{τ_i} is called *face of switching at* τ_i.

In what follows the extremal $(\tilde{u}(\cdot), \tilde{q}(\cdot), \tilde{\zeta}(\cdot))$ (with its switching points) is fixed and we will simplify the notation, putting $X_i, W_i, V_i, X_{i\pm 0}, W_{i\pm 0}, V_{i\pm 0}$ instead of $X_{\tau_i}, W_{\tau_i}, V_{\tau_i},\ X_{\tau_i \pm 0}, W_{\tau_i \pm 0}, V_{\tau_i \pm 0}$.

Let us consider two faces $W \subseteq W'$ of the polyhedron $U \subset R^r$. Obviously, for all points $w \in \operatorname{relint} W$ the conic hulls of the sets $W' - w$ coincide. We shall call any of them the *tangent cone* to W' at W and denote this cone by $\mathcal{K}_W W'$. These tangent cones are closed.

Any of these cones is naturally imbedded into the directing linear space V' of the face W'. The cone $\mathcal{K}_W W'$ is pointed if and only if W is a vertex of U; otherwise the directing space V of the face W is the maximal linear subspace of $\mathcal{K}_W W':\ V = \mathcal{K}_W W' \cap (-\mathcal{K}_W W')$.

Now we introduce certain *genericity assumptions*, which will be involved in the optimality conditions.

Strong genericity assumption for bang-bang switchings. For every bang-bang switching point τ_i the maximum $\max\{H(\tilde{q}(t), \tilde{\zeta}(t), u) :\ u \in U\}$ is achieved on an edge (= 1-dimensional face) of U; the left and right derivatives $\dot{\chi}_{\tau_j \pm 0}$ of the switching function χ_τ are not orthogonal to this edge of switching.

The nonorthogonality of $\dot{\chi}_{\tau_i \pm 0}$ to the edge of switching means that the Pontryagin maximum principle definitely forces the switching. It can be better visualized in the case of scalar u's, when the segment U (for example, $U = [-1,1]$) is the edge itself. Then by virtue of the Pontryagin maximum principle the sign of extremal control $\tilde{u}(\tau)$ must coincide with the sign of the (scalar-valued) function χ_τ. The nonorthogonality turns out in this case to be $\dot{\chi}_{\tau_i \pm 0} \neq 0$, and it implies the transversality of the graph $\tau \mapsto \chi_\tau$ to the τ-axis at the point $(\tau_i, 0)$.

Weak genericity assumption for bang-bang switchings. For every switching point τ_i the maximum $\max\{H(\tilde{q}(t), \tilde{\zeta}(t), u) :\ u \in U\}$ is achieved on a face W_i of U. For the cones $K_i^- = \mathcal{K}_{W_{i-0}} W_i,\ K_i^+ = \mathcal{K}_{W_{i+0}} W_i$ and the switching function χ_τ we have

$$\dot{\chi}_{\tau_i - 0}\xi \neq 0 \quad \forall \xi \in K_i^-; \quad \dot{\chi}_{\tau_i + 0}\xi \neq 0 \quad \forall \xi \in K_i^+. \tag{4.12}$$

Remark 4.2. Actually the inequalities (4.12) together with the maximality condition (4.9) imply more

$$\dot{\chi}_{\tau_i - 0}\xi > 0 \quad \forall \xi \in K_i^-; \quad \dot{\chi}_{\tau_i + 0}\xi < 0 \quad \forall \xi \in K_i^+. \tag{4.13}$$

Remark 4.3. For a bang-bang switching point τ_i the cones K_i^-, K_i^+ are the tangent cones to the face W_j at the vertices $\tilde{u}(\tau_i - 0)$ and $\tilde{u}(\tau_i + 0)$, correspondingly.

Remark 4.4. It is clear that for bang-bang switchings the strong genericity assumption is a particular case of the weak one.

In what follows we assume both the weak genericity assumption and the strong transversality condition to hold for any extremal under consideration.

To introduce the second variation of the system (1.2) along an extremal $(\tilde{u}(\cdot), \tilde{q}(\cdot), \tilde{\zeta}(\cdot))$ on $[0,T]$ we denote by $\ker \Phi'_T$ the kernel of the first variation along $\tilde{u}(\cdot)$, i.e., the set of $u(\cdot) \in K_{\tilde{u}}\mathcal{U}$ such that

$$\int_0^T q_0 \circ X_\tau u(\tau) d\tau = 0. \tag{4.14}$$

Note that the conditions (4.2) and (4.14) imply an inclusion

$$u(\tau) \in V_\tau \text{ a.e. on } [0,T]$$

for any control $u(\cdot) \in \ker \Phi'_T$; in particular, $u(\tau)$ must vanish on the bang-bang intervals of $\tilde{u}(\cdot)$. Obviously $\ker \Phi'_T$ is linear subspace of L^r_∞, which is trivial if $\tilde{u}(\cdot)$ is a bang-bang extremal control.

Returning to the Taylor expansion (3.5) of the mapping Φ_T let us consider its quadratic term. We denote by Φ''_T the restriction of this vector-valued quadratic form to $\ker \Phi'_T$. It is known (see [1, 3, 8]) that Φ''_T can be represented as

$$\Phi''_T(u(\cdot)) = \int_0^T q_0 \circ \left[\int_0^\tau X_\theta u(\theta) d\theta, X_\tau u(\tau) \right] d\tau, \ u(\cdot) \in \ker \Phi'_T.$$

Definition 4.3. *The second variation of the system* (1.2) *along the extremal* $(\tilde{u}(\cdot), \tilde{q}(\cdot), \tilde{\zeta}(\cdot))$ *on* $[0,T]$ *is the projection of* Φ''_T *on the covector* $\zeta_0 = \tilde{\zeta}(0)$*, i.e., the scalar valued quadratic form*

$$\zeta_0 \Phi''_T(u(\cdot)) = \int_0^T \left\langle \zeta_0, q_0 \circ \left[\int_0^\tau X_\theta u(\theta) d\theta, X_\tau u(\tau) \right] \right\rangle d\tau, \ u(\cdot) \in \ker \Phi'_T. \tag{4.15}$$

5. Extended first and second variations of the control system. From now on we deal with a fixed bang-bang Pontryagin extremal $(\tilde{u}(\cdot), \tilde{q}(\cdot), \tilde{\zeta}(\cdot))$. As was already shown, the first variation of the system (1.2) along the bang-bang control $\tilde{u}(\cdot)$ has trivial kernel. This fact on one hand obstructs the construction of high-order variations while on the other hand does not guarantee optimality of $\tilde{u}(\cdot)$. Our idea is to define first variation with extended domain in such a way that the triviality of the kernel of the extended first variation along the bang-bang extremal control $\tilde{u}(\cdot)$ implies L_1-local optimality of $\tilde{u}(\cdot)$. If the extended first variation can be nullified, then we will define on its kernel an extended second variation and formulate second-order sufficient conditions of L_1-local optimality for the bang-bang extremal control $\tilde{u}(\cdot)$.

The extended first and second variations are to be defined via the end-point mapping Φ_T. Let us recall that Φ_T can be represented via chronological exponential $\Phi_T(u(\cdot) = q_0 \circ \overrightarrow{\exp} \int_0^t X_\tau u(\tau) d\tau$, which is in turn defined by the control system

$$\dot{q} = q(\tau) \circ X_\tau u(\tau), \ q(0) = q_0, \tag{5.1}$$

$$X_\tau = (X^1_\tau, \dots, X^r_\tau), \ X^i_\tau = \overrightarrow{\exp} \int_0^\tau \operatorname{ad} \tilde{f}_\theta d\theta g^i \ (i = 1, \dots, r),$$

with admissible controls being admissible variations of $\tilde{u}(\cdot)$.

Let us extend the domain of Φ_T by adding some Dirac measures located at the switching points of the bang-bang extremal. To introduce them let us consider the

directing subspaces $V_j = V_{\tau_j}$ of the switching faces $W_j = W_{\tau_j}$ $(j = 1, \dots, m)$ and consider the generalized functions of the form

$$\omega = \sum_{j=1}^{m} \omega_j \delta(\tau - \tau_j),\ \omega_j \in V_j\ (j = 1, \dots, m), \tag{5.2}$$

where τ_j $(j = 1, \dots, m)$ are the switching points of the extremal. We will denote by $\Delta\{\tau_j, V_j\}$ the set of the generalized controls (5.2).

A natural way to construct extended first and second variations would be by extending the end-point mapping Φ_T onto the extended domain, which includes the Dirac measures (5.2), and then calculating first and second differentials of the extension. To realize this scheme one has to define the generalized trajectories of the control system (5.1), driven by the generalized controls (5.2). These trajectories cannot be constructed in the classical way. Indeed, it is evident that the trajectories should have discontinuities at the points τ_j $(j = 1, \dots, m)$. Therefore when substituting (5.2) into (5.1) and transforming the differential equation into an integral one, we obtain an integral of a discontinous function with respect to the measure (5.2), which has atoms just at the points of discontinuity.

An approach to constructing generalized trajectories for a class of generalized controls, which is much wider than (5.2), has been developed in [9, 10, 11]. (See especially [9, Section 4] and [10, Section 5], where the trajectory of the control system (5.1), driven by a generalized control (5.2), has been calculated explicitly.) Here we do not go into details of the topic, only introducing the definitions of extended first and second variations along the extremal.

DEFINITION 5.1. *Let $\Delta\{\tau_j, V_j\}$ be the set of Dirac measures* (5.2). *The linear operator $\Phi_T^{e'} : \mathcal{K}_{\tilde{u}}\mathcal{U} \oplus \Delta\{\tau_j, V_j\} \to \mathcal{T}_{q_0}M$, defined by*

$$\Phi_T^{e'}(u(\cdot) \oplus \omega) = q_0 \circ \left(\int_0^T X_\tau u(\tau) d\tau + \sum_{j=1}^{m} X_j \omega_j \right), \tag{5.3}$$

is called an extended first variation along the extremal. The summand $q_0 \circ \sum_{j=1}^{m} X_j \omega_j$ is called the first variation at the switching points of the extremal.

This definition is justified by the following proposition.

LEMMA 5.1. *For every $j = 1, \dots, m$ let a δ-sequence of controls $w_k^j(\cdot)$ $(k = 1, \dots)$ tend weakly to the Dirac measure $\omega_j \delta(\tau - \tau_j)$. Put $w_k(\cdot) = \sum_{j=1}^{m} w_k^j(\cdot)$. Then as $k \to \infty$ the values of the first variation*

$$\Phi'_T(u(\cdot) + w_k(\cdot)) = q_0 \circ \left(\int_0^T X_\tau (u(\tau) + w_k(\tau)) \right) d\tau$$

tend to the value of the extended first variation (5.3). □

(Note that we are not able to approximate Dirac measures by *admissible* controls, since the latter are bounded.)

In the next section we will prove that if the extended first variation along the bang-bang extremal has trivial kernel, then the bang-bang extremal control $\tilde{u}(\cdot)$ is optimal.

If the extended first variation along $\tilde{u}(\cdot)$ can be nullified, then we define on its kernel an extended second variation along the extremal.

DEFINITION 5.2. *The extended second variation of the system* (1.2) *along the Pontryagin extremal* $(\tilde{u}(\cdot), \tilde{q}(\cdot), \tilde{\zeta}(\cdot))$ *is the quadratic form*

$$\zeta_0 \Phi_T^{e''} = \frac{1}{2}\left\langle \zeta_0, \int_0^T q_0 \circ \left[\int_0^\tau X_\xi u(\xi)d\xi + \sum_{\tau_j \le \tau} X_j\omega_j, X_\tau u(\tau)\right] d\tau \right.$$

$$\left. + \sum_{i=1}^m q_0 \circ \left[\int_0^{\tau_i} X_\xi u(\xi)d\xi + \sum_{j=1}^i X_j\omega_j, X_i\omega_i\right]\right\rangle, \tag{5.4}$$

where $u(\cdot) \oplus \omega \in \ker \Phi_T^{e'}$, *i.e.,* $u(\cdot) \oplus \omega \in \mathcal{K}_{\tilde{u}}\mathcal{U} \oplus \Delta\{\tau_j, V_j\}$ *and*

$$q_0 \circ \left(\int_0^T X_\tau u(\tau)d\tau + \sum_{j=1}^m X_j\omega_j\right) = 0. \tag{5.5}$$

Putting $u(\cdot) \equiv 0$ in (5.4) and (5.5), one obtains the quadratic form

$$\frac{1}{2}\left\langle \zeta_0, q_0 \circ \left[\sum_{j=1}^i X_j\omega_j, X_i\omega_i\right]\right\rangle \tag{5.6}$$

defined on the kernel of the first variation at the switching points, which is a finite-dimensional subspace:

$$\left\{\omega = (\omega_1, \dots, \omega_m) \in \oplus_{j=1}^m V_j : \ q_0 \circ \sum_{j=1}^m X_j\omega_j = 0\right\}. \tag{5.7}$$

DEFINITION 5.3. *The quadratic form* (5.6) *defined on the subspace* (5.7) *is called the second variation of the system* (1.2) *at the switching points of the extremal* $(\tilde{u}(\cdot), \tilde{q}(\cdot), \tilde{\zeta}(\cdot))$.

Agrachev and Gamkrelidze introduced in [2] a different type of second variation which corresponds to the variation of the moments of switchings. Using this form they established necessary second-order optimality condition for a bang-bang extremal.

6. First-order sufficient optimality condition for bang-bang Pontryagin extremals. As has been already established, the first variation along a bang-bang Pontryagin extremal has trivial kernel; at the same time a bang-bang Pontryagin extremal is not necessarily optimal. We have introduced above the extended first and second variations along a Pontryagin extremal. In this section we shall establish that if the kernel of the extended first variation is also trivial or, equivalently, if the kernel of the first variation at the switching points is trivial, then the Pontryagin extremal is optimal (under few additional assumptions).

DEFINITION 6.1. *We say that second-order horizontality conditions for switchings of the Pontryagin extremal* $(\tilde{u}(\cdot), \tilde{q}(\cdot), \tilde{\zeta}(\cdot))$ *hold if*

$$\langle \tilde{\zeta}(\tau_j), \tilde{q}(\tau_j) \circ [Gv, Gv']\rangle = 0 \ \forall v, v' \in V_j, \ \forall j = 1, \dots, m, \tag{6.1}$$

or, equivalently,

$$\langle \zeta_0, q_0 \circ [X_j v, X_j v']\rangle = 0 \ \forall v, v' \in V_j, \ \forall j = 1, \dots, m \tag{6.2}$$

for all its switching points τ_j $(j = 1, \dots, m)$.

Remark 6.1. If the equality $\langle\tilde{\zeta}(\tau_j), y\rangle = 0$ defines a "horizontal hyperplane" in $T_{\tilde{q}(\tau_j)}M$ (correspondingly, $\langle\tilde{\zeta}_0, y\rangle = 0$ defines a horizontal hyperplane in $T_{q_0}M$), then the equalities (6.1), (6.2) mean that for those controlled vector fields whose values are horizontal, their second-order Lie brackets also have horizontal values. Everywhere below we assume this condition to hold.

Remark 6.2. Goh has established [6] that the fulfillment of the second-order horizontality condition along a singular subarc of extremal is necessary for optimality of this subarc. We require this condition to hold along the degenerate singular subarcs, which are the switching points.

Remark 6.3. If the strong genericity assumption holds, then $\dim V_j = 1$ $(j = 1, \ldots, m)$, and the second-order horizontality condition (6.1) for switchings holds automatically.

THEOREM 6.1 (first-order sufficient optimality condition for bang-bang extremals). *Let the bang-bang Pontryagin extremal* $(\tilde{u}(\cdot), \tilde{q}(\cdot), \tilde{\zeta}(\cdot))$ *meet the strong transversality condition* (4.11)*, the weak genericity assumption* (4.12)*, and the second-order horizontality condition for switchings* (6.1)*. If the first variation at the switching points of the bang-bang extremal has a trivial kernel, i.e.,*

$$q_0 \circ \sum_{i=1}^{m} X_i v_i = 0,\ v_i \in V_i\ (i = 1, \ldots, m) \Rightarrow v_i = 0\ (i = 1, \ldots, m), \tag{6.3}$$

then the bang-bang extremal control $\tilde{u}(\cdot)$ *is optimal for the problem* (1.1)–(1.3). □

Remark 6.4. The triviality of the kernel of the first variation at switching points can be better visualized when the controls u are scalar. Then it amounts to the linear independence of the values of the vector fields $q_0 \circ X_{\tau_j}$ computed for the switching points τ_j $(j = 1, \ldots, m)$. Recall that $q_0 \circ X_{\tau_j}$ (see (5.1) and (3.3)) is the result of pulling back the value of the (unique) controlled vector field g from $\tilde{q}(\tau_j)$ to q_0 by means of $\tilde{P}^{-1}_{\tau_j}$. Obviously the triviality of the kernel limits the number of switchings by $n = \dim M$.

Proof of Theorem 6.1. To establish optimality of $\tilde{u}(\cdot)$ we shall verify the conditions of the auxiliary lemma on optimality (section 4).

Let us choose some positive $\Delta < 1$ in such a way that the intervals $[\tau_j - \Delta, \tau_j + \Delta]$ $(j = 1, \ldots, m)$ are mutually nonintersecting. Consider an arbitrary admissible variation $u(\cdot)$ of the bang-bang control $\tilde{u}(\cdot)$, and take the restriction of $u(\cdot)$ to a subinterval $[\tau_j - \Delta, \tau_j)$. Since $\tilde{u}(\cdot)$ is constant on $[\tau_j - \Delta, \tau_j)$ and takes its value at some vertex η_j^- of the polyhedron U, the admissible variations of $\tilde{u}(\cdot)$ on this interval must take their values in the polyhedral tangent cone $K_{\eta_j^-}$ to U at η_j^- (the convex conic hull of the set $U - \eta_j^-$). Let us consider the tangent cone K_j^- to the face of switching W_j at the point $\tilde{u}(\tau_j - 0) = \eta_j^-$ (see (4.12)). Obviously K_j^- is one of the faces of the cone $K_{\eta_j^-}$. Let us denote by $K_j^{-\natural}$ the convex conic hull of those edges of $K_{\eta_j^-}$ which do not belong to K_j^-. The admissible variation $u(\cdot)|_{[\tau_j - \Delta, \tau_j)}$ can be represented (nonuniquely!) as a sum $u(\tau) = \bar{u}(\tau) + v(\tau)$, with $v(\tau) \in K_j^-$ and $\bar{u}(\tau) \in K_j^{-\natural}$. We can repeat a similar procedure for $u(\cdot)|_{(\tau_j, \tau_j + \Delta]}$ and for all $j = 1, \ldots, m$. Taking the concatenation of all $\bar{u}(\cdot)$'s and $v(\cdot)$'s constructed on the intervals $[\tau_j - \Delta, \tau_j + \Delta]$, we obtain the functions $\bar{u}(\cdot), v(\cdot)$, whose support is located in the set $I^\Delta = \cup_{j=1}^m [\tau_j - \Delta, \tau_j + \Delta]$; $u(\tau) = \bar{u}(\tau) + v(\tau)\ \forall \tau \in I^\Delta$. Finally we will put $\bar{u}(\tau) = u(\tau)$, $v(\tau) = 0$ on $[0, T] \setminus I^\Delta$. Then $u(\tau) = \bar{u}(\tau) + v(\tau)\ \forall \tau \in [0, T]$.

Setting $v_j = \int_{\tau_j-\Delta}^{\tau_j+\Delta} v(\tau)d\tau$ for $j = 1, \dots, m$, we put $v = (v_1, \dots, v_m)$ and denote by $v^s(t)$ the piecewise constant function which equals $v_j/2\Delta$ on $[\tau_j - \Delta, \tau_j + \Delta]$ and vanishes outside I^Δ. We put $|v| = \sum_{j=1}^m |v_j|$. Obviously $|v| = \|v^s(\cdot)\|_{L_1}$. Finally define the function $w(\cdot) = v(\cdot) - v^s(\cdot)$. Evidently $w(\cdot)$ vanishes outside I^Δ and $\int_{\tau_j-\Delta}^{\tau_j+\Delta} w(\tau)d\tau = 0$. We put the triple $(\bar{u}(\cdot), v, w(\cdot))$ into correspondence with an admissible variation $u(\cdot)$ and consider it as a (nonuniquely defined) "coordinatization" of $u(\cdot)$.

Let us introduce the vertical direction in $T_{q_0}M$, a vector $z \in T_{q_0}M$ such that $\langle \zeta_0, z\rangle = 1$. Let us put $\mathcal{V}^1(u(\cdot)) = q_0 \circ \int_0^T X_\tau u(\tau)d\tau$ and denote by $\mathcal{V}^1_v(u(\cdot))$ its vertical projection $\langle \zeta_0, \mathcal{V}^1(u(\cdot))\rangle z$ and by $\mathcal{V}^1_h(u(\cdot)) = \mathcal{V}^1(u(\cdot)) - \mathcal{V}^1_v(u(\cdot))$ its projection onto the horizontal hyperplane $\{y \mid \langle \zeta_0, y\rangle = 0\}$. Let us denote by $\mathcal{V}^2(u(\cdot)) = \Phi_T(u(\cdot)) - \mathcal{V}^1(u(\cdot))$ the first-order remainder term of the Taylor expansion for Φ_T and introduce by analogy its vertical and horizontal components $\mathcal{V}^2_v(u(\cdot))$ and $\mathcal{V}^2_h(u(\cdot))$.

Now we shall derive a lower estimate for the length of $\mathcal{V}^1(u(\cdot)) + \mathcal{V}^2_h(u(\cdot)) \in T_{q_0}M$.

LEMMA 6.2. *If the weak genericity assumption holds for the bang-bang Pontryagin extremal $(\tilde{u}(\cdot), \tilde{q}(\cdot), \tilde{\zeta}(\cdot))$, then there exists a positive constant α and for any sufficiently small $\Delta > 0$ there exists a neighborhood $\mathcal{U}^\Delta$ of the origin by the L_1-norm, such that for all admissible variations $u(\cdot) \in \mathcal{U}^\Delta \cap L^r_\infty[0, T]$ and all possible coordinatizations the inequality*

$$|\mathcal{V}^1(u(\cdot)) + \mathcal{V}^2_h(u(\cdot))| \geq \alpha(\|\bar{u}(\cdot)\|_{L_1} + \Delta^{-2}|v|^2 + \|w(\cdot)\|^2_{L_1}) \tag{6.4}$$

holds. □

The proof of this lemma will be presented below together with the proof of another technical lemma, which provides an upper estimate for the vertical component $|\mathcal{V}^2_v(u(\cdot))|$ of the first-order remainder term.

LEMMA 6.3. *Provided that the second-order horizontality condition for switchings holds, there exists a positive constant A and for any sufficiently small $\Delta > 0$ there exists a neighborhood $\mathcal{U}^\Delta$ of the origin by the L_1-norm such that for any admissible variation $u(\cdot) \in \mathcal{U}^\Delta \cap L^r_\infty[0, T]$ and all possible coordinatizations the inequality*

$$|\mathcal{V}^2_v(u(\cdot))| \leq A\Delta(\|\bar{u}(\cdot)\|_{L_1} + \Delta^{-2}|v|^2 + \|w(\cdot)\|^2_{L_1}) \tag{6.5}$$

holds. □

On the basis of these lemmas one can prove Theorem 6.1 easily. First let us fix some $\Delta_0 > 0$ meeting the condition $1 - \Delta_0 A/\alpha > 0$ and denote by $\mathcal{U}^{\Delta_0}$ the neighborhood of the origin by L_1-metric for which the conclusions of the Lemmas 6.2 and 6.3 hold. Let us denote A/α by κ; $0 < \kappa\Delta_0 < 1$. Then (6.4)–(6.5) imply $\forall u(\cdot) \in \mathcal{U}^{\Delta_0}$

$$\|\mathcal{V}^2_v(u(\cdot))\| \leq \kappa\Delta_0\|\mathcal{V}^1(u(\cdot)) + \mathcal{V}^2_h(u(\cdot))\|$$

and

$$\begin{aligned}\|\Phi_T(u(\cdot))\| &= \|\mathcal{V}^1(u(\cdot)) + \mathcal{V}^2_h(u(\cdot)) + \mathcal{V}^2_v(u(\cdot))\| \\ &\geq (1 - \kappa\Delta_0)\alpha(\|\bar{u}(\cdot)\|_{L_1} + \Delta_0^{-2}|v|^2 + \|w(\cdot)\|^2_{L_1}).\end{aligned}$$

This means that

$$\{u^m(\cdot)\} \in \mathcal{U}^{\Delta_0},\ \Phi_T(u^m(\cdot)) \overset{m\to+\infty}{\longrightarrow} 0 \Longrightarrow \|u^m(\cdot)\|_{L_1} \overset{m\to+\infty}{\longrightarrow} 0 \tag{6.6}$$

and also

$$\Phi_T(u(\cdot)) = 0, u(\cdot) \in \mathcal{U}^{\Delta_0} \Longrightarrow u(\cdot) = 0.$$

The bang-bang extremality of $\tilde{u}(\cdot)$ (see the Pontryagin maximum principle) implies that the vector $\mathcal{V}^1(u(\cdot))$, and hence also the vector $\mathcal{V}^1(u(\cdot)) + \mathcal{V}^2_h(u(\cdot))$, belongs to the open "lower" semispace of $T_{q_0}M$, defined by the inequality $\langle \zeta_0, y\rangle < 0$.

Therefore for any $\Delta < \Delta_0$ $\forall u(\cdot) \in \mathcal{U}^{\Delta}$:

$$|\langle \zeta_0, (\mathcal{V}^1+\mathcal{V}^2)(u(\cdot))\rangle| = |\langle \zeta_0, \Phi_T(u)\rangle| \le |\mathcal{V}^2_v(u(\cdot))| \le A(\Delta\|\bar{u}(\cdot)\|_{L_1}+\Delta^{-1}|v|^2+\Delta\|w(\cdot)\|^2_{L_1})$$

and

$$\|\Phi_T(u(\cdot))\| \ge (1-\kappa\Delta)\alpha(\|\bar{u}(\cdot)\|_{L_1} + \Delta^{-2}|v|^2 + \|w(\cdot)\|^2_{L_1}).$$

Hence

$$\langle \zeta_0, \Phi_T(u)\rangle / \|\Phi_T(u(\cdot))\| \le A\Delta/(1-\kappa\Delta)\alpha = \kappa\Delta/(1-\kappa\Delta) \tag{6.7}$$

for $u(\cdot)$ which are sufficiently close to the origin in the L_1-metric.

Let us consider an arbitrary C^1-smooth curve $\tau \to \gamma(\tau)$ ($\tau \in [0,t]$) starting at $q_0 = \gamma(0)$ and belonging to $\Phi_T(\mathcal{U}^{\Delta_0})$. By virtue of (6.6)–(6.7) its tangent vector at $q_0 = \gamma(0)$ belongs to the nonconvex subcone of $T_{q_0}M$ defined by the inequality

$$\langle \zeta_0, y\rangle \le \frac{\kappa\Delta}{1-\kappa\Delta}\|y\|. \tag{6.8}$$

Recall that $\Delta > 0$ in (6.8) can be chosen arbitrarily small. When $\Delta \to +0$, the right-hand side of this inequality tends to $+0$ and we may conclude that the tangent vectors to the set $\Phi_T(\mathcal{U}^{\Delta_0})$ belong to the closed lower semispace of $T_{q_0}M$ determined by the inequality $\langle \zeta_0, y\rangle \le 0$. Applying the auxiliary lemma on optimality we establish L_1-local optimality of $\tilde{u}(\cdot)$. □

Proof of Lemma 6.2. Using the coordinatization $(\bar{u}(\cdot), v, w(\cdot))$ of the arbitrary admissible variation $u(\cdot)$ of the extremal control $\tilde{u}(\cdot)$ we represent the first variation $\int_0^T q_0 \circ X_\tau u(\tau)d\tau$ as

$$\begin{aligned} q_0 \circ \int_0^T X_\tau u(\tau)d\tau &= q_0 \circ \left(\int_0^T X_\tau \bar{u}(\tau)d\tau \right. \\ &\left. + \sum_{i=1}^m \left(\int_{\tau_i-\Delta}^{\tau_i+\Delta} X_\tau d\tau \right) v_i/2\Delta + \sum_{i=1}^m \int_{\tau_i-\Delta}^{\tau_i+\Delta} X_\tau w(\tau)d\tau \right) \\ = q_0 \circ \left(\int_0^T X_\tau \bar{u}(\tau)d\tau + \sum_{i=1}^m X_i v_i + \sum_{i=1}^m \int_{\tau_i-\Delta}^{\tau_i+\Delta} (X_\tau - X_i)(v_i/2\Delta + w(\tau))d\tau \right). \end{aligned} \tag{6.9}$$

Recall that the notation X_i is for X_{τ_i} $(i = 1, \dots, m)$.

The first and the third summands in the right-hand side of (6.9) belong to the open lower semispace, while the second summand belongs to the horizontal hyperplane.

In accordance with the definition of the coordinatization $u(\cdot) \mapsto (\bar{u}(\cdot), v, w(\cdot))$ of the admissible variations of $\tilde{u}(\cdot)$, the values $\bar{u}(\tau)$ belong to $K^{\natural}_{\tilde{u}(\tau)}U$. By virtue of the maximality condition (4.2) of the Pontryagin maximum principle there exists $a_1 > 0$ such that

$$\max_{u \in K^{\natural}_{\tilde{u}(\tau)}U} \langle \zeta_0, q_0 \circ X_\tau \bar{u}(\tau)\rangle \le -a_1|\bar{u}(\tau)| \ \forall \tau \in [0,T].$$

This gives the estimate for the projection on ζ_0 of the first summand in the right-hand side of (6.9):

$$\left\langle \zeta_0, q_0 \circ \int_0^T X_\tau \bar{u}(\tau) d\tau \right\rangle \leq -a_1 \|\bar{u}(\cdot)\|_{L_1}. \tag{6.10}$$

Since the first variation at the switching points of the extremal control $\tilde{u}(\cdot)$ has a trivial kernel, then the second addend of (6.9) admits a lower estimate:

$$\left| \sum_{i=1}^m q_0 \circ X_i v_i \right| \geq b_3 |v|. \tag{6.11}$$

Since $\dot{X}_\tau$ is continuous with regard to τ on $[\tau_i - \Delta, \tau_i)$ and on $(\tau_i, \tau_i + \Delta]$, then we may assume by virtue of the weak genericity assumption (and decreasing Δ, if necessary), that for some $c_1 > 0$,

$$\begin{aligned} &\langle \zeta_0, q_0 \circ \dot{X}_\tau v \rangle \geq c_1 |v| \ \forall \tau \in [\tau_i - \Delta, \tau_i), \ \forall v \in K_i^-, \ \forall i = 1, \ldots, m, \\ &\langle \zeta_0, q_0 \circ \dot{X}_\tau v \rangle \leq -c_1 |v| \ \forall \tau \in (\tau_i, \tau_i + \Delta], \ \forall v \in K_i^+, \ \forall i = 1, \ldots, m. \end{aligned} \tag{6.12}$$

Therefore, since $v(\tau) \in K_i^-$ for $\tau \in [\tau_i - \Delta, \tau_i)$, $v(\tau) \in K_i^+$ for $\tau \in (\tau_i, \tau_i + \Delta]$, then

$$\left\langle \zeta_0, q_0 \circ \sum_{i=1}^m \int_{\tau_i - \Delta}^{\tau_i + \Delta} (X_\tau - X_i) v(\tau) d\tau \right\rangle \leq -c_1 \sum_{i=1}^m \int_{\tau_i - \Delta}^{\tau_i + \Delta} |\tau - \tau_i| |v(\tau)| d\tau. \tag{6.13}$$

To estimate the integral $\int_{\tau_i}^{\tau_i + \Delta} (\tau - \tau_i) |v(\tau)| d\tau$ we integrate it by parts obtaining

$$\int_{\tau_i}^{\tau_i + \Delta} (V(\tau_i + \Delta) - V(\tau)) d\tau, \text{ where } V(\tau) = \int_{\tau_i}^{\tau} |v(\xi)| d\xi.$$

As long as $|v(\tau)| \leq M$, $V(\tau_i) = 0$, and $dV(\tau) = |v(\tau)| d\tau$, we have

$$\begin{aligned} \int_{\tau_i}^{\tau_i + \Delta} (V(\tau_i + \Delta) - V(\tau)) d\tau &\geq (1/M) \int_{\tau_i}^{\tau_i + \Delta} (V(\tau_i + \Delta) - V(\tau)) |v(\tau)| d\tau \\ &= (-1/2M)(V(\tau_i + \Delta) - V(\tau))^2 |_{\tau_i}^{\tau_i + \Delta} \\ &= (V(\tau_i + \Delta))^2 / 2M = \|v(\cdot)|_{[\tau_i, \tau_i + \Delta]}\|_{L_1}^2 / 2M. \end{aligned} \tag{6.14}$$

Deriving a similar estimate for $\int_{\tau_i - \Delta}^{\tau_i} |\tau - \tau_i| |v(\tau)| d\tau$ and repeating it for all $i \in \{1, \ldots, m\}$ we may conclude that for some $c_2 > 0$,

$$\left\langle \zeta_0, q_0 \circ \sum_{i=1}^m \int_{\tau_i - \Delta}^{\tau_i + \Delta} (X_\tau - X_i) v(\tau) d\tau \right\rangle \leq -c_2 \|v(\cdot)\|_{L_1}^2. \tag{6.15}$$

Obviously $\|v^s(\cdot)\|_{L_1} = |v| = \sum_{i=1}^m |\int_{\tau_i - \Delta}^{\tau_i + \Delta} v(\tau) d\tau| \leq \|v(\cdot)\|_{L_1}$. Since $w(\cdot) = v(\cdot) - v^s(\cdot)$, we conclude that

$$\|w(\cdot)\|_{L_1} \leq \|v(\cdot)\|_{L_1} + \|v^s(\cdot)\|_{L_1} \leq 2 \|v(\cdot)\|_{L_1}.$$

Therefore $\|v(\cdot)\|_{L_1}^2 \geq (|v|^2 + \|w(\cdot)\|_{L_1}^2)/5$, and we may change the estimate (6.15) to

$$\left\langle \zeta_0, q_0 \circ \sum_{i=1}^{m} \int_{\tau_i-\Delta}^{\tau_i+\Delta} (X_\tau - X_i)v(\tau)d\tau \right\rangle \leq -c_3(|v|^2 + \|w(\cdot)\|_{L_1}^2). \tag{6.16}$$

The projections of the first and the third summands in the right-hand side of (6.9) on the horizontal hyperplane admit *upper* estimates $\beta\|\bar{u}(\cdot)\|_{L_1}$ and $k\Delta(|v|+\|w(\cdot)\|_{L_1})$ correspondingly.

We shall prove now that $|\mathcal{V}^2(u(\cdot))| = O(\|u(\cdot)\|_{L_1}^2)$. To this end let us introduce the notation $P_t = \overrightarrow{\exp} \int_0^t X_\tau u(\tau)d\tau$ and recall that $\frac{d}{dt}P_t = P_t \circ X_t u(t)$, $P_0 = I$, and $\Phi_T(u(\cdot)) = q_0 \circ P_T$. Then

$$\Phi_T(u(\cdot)) - \mathcal{V}^1(u(\cdot)) = q_0 \circ \int_0^T (P_t - I)X_t u(t)dt = q_0 \circ \int_0^T \int_0^t P_\tau \circ X_\tau u(\tau)d\tau \circ X_t u(t)dt,$$

and the required estimate is now obvious.

This implies an upper estimate $|\mathcal{V}_h^2(u(\cdot))| \leq C(\|\bar{u}(\cdot)\|_{L_1}^2 + |v|^2 + \|w(\cdot)\|_{L_1}^2)$. Assuming the neighborhood $\mathcal{U}^\Delta$ to be such that $\max(\|u(\cdot)\|_{L_1}, |v|, \Delta^{-1}\|w(\cdot)\|_{L_1}) \leq \varepsilon < b_3/2C$ we may change this estimate to

$$|\mathcal{V}_h^2(u(\cdot))| \leq C\varepsilon(\|\bar{u}(\cdot)\|_{L_1} + |v| + \Delta\|w(\cdot)\|_{L_1}). \tag{6.17}$$

The estimates for the summands of the right-hand side of (6.9) imply a lower estimate for the vector $\mathcal{V}^1(u(\cdot))$:

$$\begin{aligned} &|\mathcal{V}^1(u(\cdot))| \\ &\geq a_1\|\bar{u}(\cdot)\|_{L_1} + c_3(|v|^2 + \|w(\cdot)\|_{L_1}^2) + \max(0, b_3|v| - \beta_3\|\bar{u}(\cdot)\|_{L_1} - k\Delta\|w(\cdot)\|_{L_1}). \end{aligned}$$

By virtue of the upper estimate (6.17) a similar lower estimate (possibly with different constants $b_3, \beta_3, k > 0$) holds for $|\mathcal{V}^1(u(\cdot)) + \mathcal{V}_h^2(u(\cdot))|$:

$$\begin{aligned} &|\mathcal{V}^1(u(\cdot)) + \mathcal{V}_h^2(u(\cdot))| \geq a_1\|\bar{u}(\cdot)\|_{L_1} + c_3(|v|^2 + \|w(\cdot)\|_{L_1}^2) \\ &+ \max(0, b_3|v| - \beta_3\|\bar{u}(\cdot)\|_{L_1} - k\Delta\|w(\cdot)\|_{L_1}) \\ &\geq a_1\|\bar{u}(\cdot)\|_{L_1} + c_3\|w(\cdot)\|_{L_1}^2 + \max(0, b_3|v| - \beta_3\|\bar{u}(\cdot)\|_{L_1} - k\Delta\|w(\cdot)\|_{L_1}). \end{aligned} \tag{6.18}$$

To establish the estimate (6.4) let us assume first that

$$b_3|v|/2 \geq \beta_3\|\bar{u}(\cdot)\|_{L_1} + k\Delta\|w\|_{L_1}.$$

Then $\max(0, b_3|v| - \beta_3\|\bar{u}(\cdot)\|_{L_1} - k\Delta\|w\|_{L_1}) \geq b_3|v|/2$ and we come to a lower estimate

$$|\mathcal{V}^1(u(\cdot)) + \mathcal{V}_h^2(u(\cdot))| \geq a_1\|\bar{u}\|_{L_1} + b_3|v|/2 + c_3\|w\|_{L_1}^2.$$

Evidently $b_3|v|/2 \geq \gamma\Delta^{-2}|v|^2$ for all small enough $|v| > 0$, and we derive the lower estimate (6.4).

Assume now, on the contrary, that

$$b_3|v|/2 \leq \beta_3\|\bar{u}(\cdot)\|_{L_1} + k\Delta\|w\|_{L_1}. \tag{6.19}$$

If $\beta_3\|\bar{u}(\cdot)\|_{L_1} \geq k\Delta\|w\|_{L_1}$, then $b_3|v|/2 \leq 2\beta_3\|\bar{u}(\cdot)\|_{L_1}$, and we derive

$$a_1\|\bar{u}(\cdot)\|_{L_1} \geq \frac{a_1}{2}\|\bar{u}(\cdot)\|_{L_1} + a_1k\Delta\|w\|_{L_1}/4\beta_3 + a_1b_3|v|/16\beta_3.$$

The sum of the last two terms is bounded below by $\gamma\Delta^{-2}|v|^2+\mu\|w\|_{L_1}^2$ for sufficiently small $|v|, \|w\|_{L_1} > 0$.

If $\beta_3\|\bar u(\cdot)\|_{L_1} \le k\Delta\|w\|_{L_1}$, then (6.19) implies $b_3|v| \le 4k\Delta\|w\|_{L_1}$ and hence in (6.18)

$$c_3\|w\|_{L_1}^2 \ge \frac{c_3}{2}\|w\|_{L_1}^2 + \frac{c_3 b_3^2}{32k^2}\Delta^{-2}|v|^2. \qquad \square$$

Proof of Lemma 6.3. As has been already established,

$$\Phi_T(u(\cdot)) - \mathcal{V}^1(u(\cdot)) = q_0 \circ \int_0^T (P_t - I)X_t u(t)dt = q_0 \circ \int_0^T \int_0^t P_\tau X_\tau u(\tau)d\tau \circ X_t u(t)dt.$$

Substituting $P_\tau = I + (P_\tau - I)$ into the last expression we transform it into

$$\begin{aligned}\Phi_T(u(\cdot)) - \mathcal{V}^1(u(\cdot)) &= q_0 \circ \int_0^T \int_0^t X_\tau u(\tau)d\tau \circ X_t u(t)dt \\ &+ q_0 \circ \int_0^T \int_0^t (P_\tau - I)X_\tau u(\tau)d\tau \circ X_t u(t)dt. \end{aligned} \tag{6.20}$$

Since $P_\tau - I = \int_0^\tau X_\xi u(\xi)d\xi$, then (see [7]) $\|P_\tau - I\|_{s,K} \le c(s,K)\|u(\cdot)\|_{L_1}$ and therefore the second summand at the right-hand side of (6.20) admits an estimate $O(\|u(\cdot)\|_{L_1}^3)$ which implies the upper estimate (6.5), if $\|u(\cdot)\|_{L_1}$ is sufficiently small. Obviously the same estimate holds for the vertical component of this summand.

To estimate the term $q_0 \circ \int_0^T \int_0^t X_\tau u(\tau)d\tau \circ X_t u(t)dt$ we represent it as a sum

$$\begin{aligned} q_0 \circ \Bigg(&\int_0^T \int_0^t X_\tau u(\tau)d\tau \circ X_t \bar u(t)dt + \sum_{i=1}^m \int_{\tau_i-\Delta}^{\tau_i+\Delta} \int_0^t X_\tau u(\tau)d\tau \circ X_t(v_i/2\Delta)dt \\ &+ \sum_{i=1}^m \int_{\tau_i-\Delta}^{\tau_i+\Delta} \int_0^t X_\tau u(\tau)d\tau \circ X_t w(t)dt \Bigg). \end{aligned} \tag{6.21}$$

The first summand of the right-hand side admits an upper estimate $c\|u(\cdot)\|_{L_1}\|\bar u(\cdot)\|_{L_1}$, which is majorized by $A\Delta\|\bar u(\cdot)\|_{L_1}$, if $\|u(\cdot)\|_{L_1}$ is small enough. The second summand admits an upper estimate

$$c(|v|^2 + \|\bar u(\cdot)\|_{L_1}^2 + \|w(\cdot)\|_{L_1}|v|) \le A(\Delta^{-1}|v|^2 + \Delta\|w(\cdot)\|_{L_1}^2 + \|\bar u(\cdot)\|_{L_1}^2),$$

which implies the estimate (6.5) if $\|u(\cdot)\|_{L_1}$ is small enough.

Now we shall transform the third summand of (6.21) by representing it as a sum:

$$\begin{aligned} &\sum_{i=1}^m \int_{\tau_i-\Delta}^{\tau_i+\Delta} \int_0^t X_\tau \bar u(\tau)d\tau \circ X_t w(t)dt + \sum_{i=1}^m \int_{\tau_i-\Delta}^{\tau_i+\Delta} \int_0^t X_\tau v^s(\tau)d\tau \circ X_t w(t)dt \\ &+ \sum_{i=1}^m \int_{\tau_i-\Delta}^{\tau_i+\Delta} \int_0^t X_\tau w(\tau)d\tau \circ X_t w(t)dt. \end{aligned} \tag{6.22}$$

The first summand admits an upper estimate $c\|\bar u(\cdot)\|_{L_1}\|w(\cdot)\|_{L_1}$, which can be substituted by $A\Delta\|\bar u(\cdot)\|_{L_1}$, if $\|u(\cdot)\|_{L_1}$ is small enough. The second summand admits an upper estimate

$$c|v|\|w(\cdot)\|_{L_1} \le (c/2)(\Delta^{-1}|v|^2 + \Delta\|w(\cdot)\|_{L_1}^2).$$

To estimate the third summand let us represent it as

$$\sum_{i=1}^{m}\int_{\tau_i-\Delta}^{\tau_i+\Delta} X_t w(t)\circ\left(\sum_{j=1}^{i-1}\int_{\tau_j-\Delta}^{\tau_j+\Delta} X_\tau w(\tau)d\tau+\int_{\tau_i-\Delta}^{t} X_\tau w(\tau)d\tau\right)dt.$$

Taking into account that $\int_{\tau_j-\Delta}^{\tau_j+\Delta} w(\tau)d\tau=0$ $(j=0,\dots,m)$, we may rewrite it as

$$\sum_{i=1}^{m}\int_{\tau_i-\Delta}^{\tau_i+\Delta} X_t w(t)\circ\left(\sum_{j=1}^{i-1}\int_{\tau_j-\Delta}^{\tau_j+\Delta} (X_\tau-X_j) w(\tau)d\tau+\int_{\tau_i-\Delta}^{t} X_\tau w(\tau)d\tau\right)dt.$$

Since for any i,j,

$$\Delta^{-1}\left|\int_{\tau_i-\Delta}^{\tau_i+\Delta} X_t w(t)\circ\int_{\tau_j-\Delta}^{\tau_j+\Delta}(X_\tau-X_j)w(\tau)d\tau\,dt\right|=O(\|w(\cdot)\|_{L_1}^2),$$

then we only have to estimate $\int_{\tau_i-\Delta}^{\tau_i+\Delta} X_t w(t)\circ\int_{\tau_i-\Delta}^{t} X_\tau w(\tau)d\tau dt$ $(i=1,\dots,m)$.

Substituting X_i in place of X_t and X_τ $(t,\tau\in[\tau_i-\Delta,\tau_i+\Delta])$ in this integral, we change its value by $O(1)\Delta\|w(\cdot)\|_{L_1}^2$. Integrating by parts we derive

$$\int_{\tau_i-\Delta}^{\tau_i+\Delta} X_i w(t)\circ X_i\int_{\tau_i-\Delta}^{t} w(\tau)d\tau dt$$
$$=X_i\int_{\tau_i-\Delta}^{\tau_i+\Delta} w(t)dt\circ X_i\int_{\tau_i-\Delta}^{\tau_i+\Delta} w(\tau)d\tau-\int_{\tau_i-\Delta}^{\tau_i+\Delta} X_i\int_{\tau_i-\Delta}^{t} w(\tau)d\tau\circ X_i w(t)dt.$$

The first summand vanishes and hence

$$\int_{\tau_i-\Delta}^{\tau_i+\Delta} X_i w(t)\circ X_i\int_{\tau_i-\Delta}^{t} w(\tau)d\tau dt=\frac{1}{2}\int_{\tau_i-\Delta}^{\tau_i+\Delta}\left[X_i w(t), X_i\int_{\tau_i-\Delta}^{t} w(\tau)d\tau\right]dt.$$

The projection of the last term on ζ_0 vanishes, provided that the second-order horizontality condition for switchings holds. Therefore

$$\Delta^{-1}\left\langle\zeta_0,\int_{\tau_i-\Delta}^{\tau_i+\Delta} X_t w(t)\circ\int_{\tau_i-\Delta}^{t} X_\tau w(\tau)d\tau dt\right\rangle=O(\|w(\cdot)\|_{L_1}^2).\qquad\square$$

7. Second-order sufficient optimality conditions for bang-bang Pontryagin extremals. It has been established (Theorem 6.1) that the triviality of the kernel of the first variation at switching points of the bang-bang Pontryagin extremal $(\tilde u(\cdot),\tilde q(\cdot),\tilde\zeta(\cdot))$ implies optimality of the extremal control $\tilde u(\cdot)$. In this section we study what happens if the first variation at the switching points possesses a nontrivial kernel. We establish in this case the second-order sufficient optimality condition (Theorem 7.1) involving the second variation at switching points. Regrettably this condition is not as sharp as one could wish. Below we shall provide some comments and also a *sharper* form of the sufficient condition for the case where the strong generecity assumption holds (Theorem 7.2).

THEOREM 7.1 (second-order sufficient optimality condition for bang-bang extremals). *Let the strong transversality condition* (4.11)*, the weak genericity assumption*

and the second-order horizontality condition for switchings (6.2) *hold for the bang-bang Pontryagin extremal* $(\tilde{u}(\cdot), \tilde{q}(\cdot), \tilde{\zeta}(\cdot))$. *If the second variation at the switching points* τ_j $(j = 1, \dots, m)$ *of the Pontryagin extremal, i.e., the quadratic form*

$$Q(v) = \sum_{i=1}^{m} \left\langle \zeta_0, q_0 \circ \left[\sum_{j<i} X_j v_j, X_i v_i \right] \right\rangle, \tag{7.1}$$

is nonpositive (negative semidefinite) on the kernel of the first variation at switching points

$$\left\{ v = (v_1, \dots, v_m) \in \oplus_{i=1}^{m} V_i \mid \sum_{i=1}^{m} q_0 \circ X_i v_i = 0 \right\}, \tag{7.2}$$

then $\tilde{u}(\cdot)$ *is optimal for the problem* (1.1)–(1.3). □

As one can see, the formulation of the *sufficient* condition is somewhat unusual, since it involves negative semidefiniteness of the second variation, while usually second-order sufficient optimality conditions involve strict definiteness of the second variation. We should also emphasize that negative semidefiniteness of the second variation at the switching points *is not necessary* for the optimality of $\tilde{u}(\cdot)$. The reason is that the first and the second variations at switching points have been constructed on the basis of Dirac measures (located at swiching points) used as generalized variations of $\tilde{u}(\cdot)$. Were it possible to approximate these Dirac measures by ordinary controls, this construction would be quite adequate. Since the admissible controls are bounded we cannot arrange such approximations, and this is why asking for negative definiteness of the second variation at switching points means asking too much; i.e., it provides stronger optimality than the one we investigate. Technically speaking it results in the following phenomenon: provided that the weak genericity assumption holds, there is some term of the first variation, which has the same order of smallness as the second variation. Due to the Pontryagin maximum principle the vertical component of this term takes negative values for any admissible variation of $\tilde{u}(\cdot)$. Therefore a more adequate sufficient condition would be the negative definiteness of the sum of this part of the first variation with the second variation at switching points. Unfortunately this part of the first variation does not admit reasonable representation. The case where one can achieve more progress is the one where the strong genericity assumption holds. In this case the faces V_i of switching are one-dimensional edges and the above-mentioned term of the first variation admits a rather simple quadratic upper estimate which, when added to the second variation, provides us with a sharper sufficient optimality condition. The second-order horizontality condition for switchings holds in this case automatically.

To introduce this additional term let us consider the unit directing vectors ℓ_i $(i = 1, \dots, m)$ of the switching edges V_i $(i = 1, \dots, m)$; $\ell_i = (\eta_i^+ - \eta_i^-)/|\eta_i^+ - \eta_i^-|$, where $\eta_i^- = \tilde{u}(\tau_i - 0), \eta_i^+ = \tilde{u}(\tau_i + 0)$. Let $M_i = |\eta_i^+ - \eta_i^-|$ $(i = 1, \dots, m)$ be the lengths of the switching edges. The variables v_i $(i = 1, \dots, m)$ involved in the expression for the second variation are in this case *scalar*. Let us introduce the quadratic form[1] $\frac{1}{2}\sum_{i=1}^{m} \langle \zeta_0, \dot{X}_i \ell_i \rangle v_i^2 / 2M_i$ (by virtue of the Pontryagin maximum principle $\langle \zeta_0, \dot{X}_i \ell_i \rangle > 0,\ i = 1, \dots, m$). Subtracting this quadratic form from the second variation at the

[1] Under the strong genericity assumption $\dot{X}_\tau \ell_i$ is continuous at the switching point τ_i.

switching points we obtain the quadratic form

$$-\frac{1}{2}\sum_{i=1}^{m}\langle\zeta_0, q_0\circ\dot{X}_i\ell_i\rangle v_i^2/2M_i+\sum_{i=1}^{m}\left\langle\zeta_0, q_0\circ\left[\sum_{j<i}X_jv_j, X_iv_i\right]\right\rangle \tag{7.3}$$

with the domain (7.2); obviously its values are less than the values of the quadratic form (7.1)–(7.2).

THEOREM 7.2 (second-order sufficient optimality condition under the strong genericity assumption). *Let the strong transversality condition* (4.11) *and the strong genericity assumption hold for the bang-bang Pontryagin extremal* $(\tilde{u}(\cdot),\tilde{q}(\cdot),\tilde{\zeta}(\cdot))$. *If the quadratic form* (7.3) *with the domain* (7.2) *is negative definite, then* $\tilde{u}(\cdot)$ *is optimal for the problem* (1.1)–(1.3). □

Proof of Theorem 7.1. We are going to verify the conditions of the auxiliary lemma on optimality and to this end will derive estimates similar to (6.4)–(6.5).

First we slightly modify the coordinatization $u(\cdot)\mapsto(\bar{u}(\cdot),v,w(\cdot))$, splitting $\oplus_{i=1}^{m}V_i$ into the sum of the subspace (7.2) and a complementary linear subspace. Therefore any $v\in\oplus_{i=1}^{m}V_i$ can now be represented as $v=v_0+v^\sharp$ with v_0 coordinatizing the subspace (7.2) and $v^\sharp$ coordinatizing the complementary subspace. The new coordinatization is $u(\cdot)\mapsto(\bar{u}(\cdot),v^\sharp,v^0,w(\cdot))$. The formula (6.9) changes to

$$\begin{aligned}
&q_0\circ\int_0^T X_\tau u(\tau)d\tau\\
&=q_0\circ\left(\int_0^T X_\tau\bar{u}(\tau)d\tau+\sum_{i=1}^{m}\left(\int_{\tau_i-\Delta}^{\tau_i+\Delta}X_\tau d\tau\right)v_i/2\Delta+\sum_{i=1}^{m}\int_{\tau_i-\Delta}^{\tau_i+\Delta}X_\tau w(\tau)d\tau\right)\\
&=q_0\circ\left(\int_0^T X_\tau\bar{u}(\tau)d\tau+\sum_{i=1}^{m}X_iv_i^\sharp+\sum_{i=1}^{m}\int_{\tau_i-\Delta}^{\tau_i+\Delta}(X_\tau-X_i)(v_i/2\Delta+w(\tau))d\tau\right).
\end{aligned}\tag{7.4}$$

The estimate (6.11) changes to

$$\left|\sum_{i=1}^{m}q_0\circ X_iv_i^\sharp\right|\geq b_3|v^\sharp|. \tag{7.5}$$

The estimate (6.18) changes to

$$\begin{aligned}
&|\mathcal{V}^1(u(\cdot))+\mathcal{V}_h^2(u(\cdot))|\geq a_1\|\bar{u}(\cdot)\|_{L_1}+c_3(|v^\sharp|^2+|v^0|^2+\|w(\cdot)\|_{L_1}^2)\\
&+\max(0,b_3|v^\sharp|-\beta_3\|\bar{u}(\cdot)\|_{L_1}-k\Delta\|w(\cdot)\|_{L_1}-k\Delta|v^0|)\geq a_1\|\bar{u}(\cdot)\|_{L_1}\\
&+c_3(|v^0|^2+\|w(\cdot)\|_{L_1}^2)+\max(0,b_3|v^\sharp|-\beta_3\|\bar{u}(\cdot)\|_{L_1}-k\Delta\|w(\cdot)\|_{L_1}-k\Delta|v^0|),
\end{aligned}\tag{7.6}$$

from which one derives as in the previous section

$$|\mathcal{V}^1(u(\cdot))+\mathcal{V}_h^2(u(\cdot))|\geq\alpha(\|\bar{u}(\cdot)\|_{L_1}+\Delta^{-2}|v^\sharp|^2+|v^0|^2+\|w(\cdot)\|_{L_1}^2). \tag{7.7}$$

For the estimate of the vertical component of the remainder term we formulate the following lemma, which is the analogue of Lemma 6.3.

LEMMA 7.3. *Provided that the second-order horizontality condition for switchings holds, there exists a positive constant* A *and for any sufficiently small* $\Delta>0$ *there exists a neighborhood* $\mathcal{U}^\Delta$ *of the origin in the* L_1*-metric, such that for any admissible*

variation $u(\cdot) \in \mathcal{U}^\Delta \cap L_\infty^r[0,T]$ *and all possible coordinatizations the inequality*

$$\langle \zeta_0, \mathcal{V}^2(u(\cdot))\rangle \le \frac{1}{2}\sum_{i=1}^m \left\langle \zeta_0, q_0 \circ \left[\sum_{j<i} X_j v_j^0, X_i^0 v_i^0\right]\right\rangle$$
$$+A(\Delta\|\bar u(\cdot)\|_{L_1} + \Delta^{-1}|v^\sharp|^2 + \Delta(|v^0|^2 + \|w(\cdot)\|_{L_1}^2)) \tag{7.8}$$

holds. □

Providing the proof of the lemma below we finish now the proof of Theorem 7.1.

If the second variation (7.1)–(7.2) at the switching points of the extremal is non-positive, then the estimate (7.8) implies the estimate

$$\langle \zeta_0, \mathcal{V}^2(u(\cdot))\rangle \le A(\Delta\|\bar u(\cdot)\|_{L_1} + \Delta^{-1}|v^\sharp|^2 + \Delta(|v^0|^2 + \|w(\cdot)\|_{L_1}^2)). \tag{7.9}$$

The conditions of the auxiliary lemma on optimality are then derived from the estimates (7.7)–(7.9) in the same way as in the proof of Theorem 6.1. Theorem 7.1 is proved. □

Proof of Theorem 7.2. To prove the theorem we shall slightly modify the estimates (7.6), (7.7), and (7.9). First, we shall add to $\mathcal{V}^1 + \mathcal{V}_h^2$ the vector-valued quadratic form

$$\mathcal{V}_v^2(v^0) = \frac{1}{2}\left(\sum_{i=1}^m \left\langle \zeta_0, q_0 \circ \left[\sum_{j<i} X_j v_j^0, X_i v_i^0\right]\right\rangle\right) z.$$

Evidently the projection of this form onto the covector ζ_0 coincides with the second variation at the switching points. The proof of the theorem can be completed by invoking the following technical lemma.

LEMMA 7.4. *If the strong genericity assumption holds for the bang-bang Pontryagin extremal* $(\tilde u(\cdot), \tilde q(\cdot), \tilde\zeta(\cdot))$*, and the quadratic form* (7.3) *is negative definite on its domain* (7.2)*, then the values of* $\mathcal{V}^1(u(\cdot)) + \mathcal{V}_h^2(u(\cdot)) + \mathcal{V}_v^2(v^0)$ *belong to the lower semispace and there exists both a positive constant* α *and for any sufficiently small* $\Delta > 0$ *a neighborhood* $\mathcal{U}^\Delta$ *of the origin in the* L_1*-norm such that for all admissible variations* $u(\cdot) \in \mathcal{U}^\Delta \cap L_\infty^r[0,T]$ *and all possible coordinatizations the inequality*

$$|\mathcal{V}^1(u(\cdot)) + \mathcal{V}_h^2(u(\cdot)) + \mathcal{V}_v^2(v^0)| \ge \alpha(\|\bar u(\cdot)\|_{L_1} + |v^0|^2 + \Delta^{-2}|v^\sharp|^2 + \|w(\cdot)\|_{L_1}^2) \tag{7.10}$$

holds. □

Postponing the proof of this lemma we finish the proof of the Theorem 7.2. It follows from (7.8) that

$$|\mathcal{V}_v^2(u(\cdot)) - \mathcal{V}_v^2(v^0)| \le A(\Delta\|\bar u(\cdot)\|_{L_1} + \Delta^{-1}|v^\sharp|^2 + \Delta(|v^0|^2 + \|w(\cdot)\|_{L_1}^2)). \tag{7.11}$$

The conditions of the auxiliary lemma on optimality are to be derived from the estimates (7.10)–(7.11) in the same way as in the proof of the Theorem 6.1. Theorem 7.2 is proved. □

Proof of Lemma 7.3. As in the proof of the Lemma 6.5 we may restrict our consideration to the vertical component of the term $q_0 \circ \int_0^T \int_0^t X_\tau u(\tau)d\tau \circ X_t u(t)dt$ of the Volterra expansion. First we represent this term as a sum:

$$\int_0^T \int_0^t X_\tau u(\tau)d\tau \circ X_t \bar u(t)dt + \sum_{i=1}^m \int_{\tau_i-\Delta}^{\tau_i+\Delta} \int_0^t X_\tau u(\tau)d\tau \circ X_t(w(t) + v_i/2\Delta)dt.$$

The first summand of the right-hand side admits an upper estimate $c\|u(\cdot)\|_{L_1}\|\bar u(\cdot)\|_{L_1}$, which is majorized by $A\Delta\|\bar u(\cdot)\|_{L_1}$, provided that $\|u(\cdot)\|_{L_1}$ is small enough. Substituting $\bar u(t)+v^s(t)+w(t)$ in place of $u(t)$ in the second summand we transform it into

$$\sum_{i=1}^m\int_{\tau_i-\Delta}^{\tau_i+\Delta}\int_0^t X_\tau\bar u(\tau)d\tau\circ X_t(w(t)+v_i/2\Delta)dt$$
$$+\sum_{i=1}^m\int_{\tau_i-\Delta}^{\tau_i+\Delta}\int_0^t X_\tau(v^s(\tau)+w(\tau))d\tau\circ X_t(w(t)+v_i/2\Delta)dt.$$

Again the first summand admits an upper estimate $c\|u(\cdot)\|_{L_1}\|\bar u(\cdot)\|_{L_1}$, majorized by $A\Delta\|\bar u(\cdot)\|_{L_1}$, provided that $\|u(\cdot)\|_{L_1}$ is small enough. The second summand can be represented as

$$\sum_{i=1}^m\int_{\tau_i-\Delta}^{\tau_i+\Delta}\sum_{j=1}^{i-1}\int_{\tau_j-\Delta}^{\tau_j+\Delta}X_\tau(w(\tau)+v_j/2\Delta)d\tau\circ X_t(w(t)+v_i/2\Delta)dt$$
$$+\sum_{i=1}^m\int_{\tau_i-\Delta}^{\tau_i+\Delta}\int_{\tau_i-\Delta}^t X_\tau(w(\tau)+v_i/2\Delta)d\tau\circ X_t(w(t)+v_i/2\Delta)dt.$$

Changing X_τ to X_i on an interval $[\tau_i-\Delta,\tau_i+\Delta]$ in this last formula we change it by $O(1)\Delta(|v^0|^2+|v^\sharp|^2+\|w(\cdot)\|_{L_1}^2)$, obtaining

$$\sum_{i=1}^m\left(\sum_{j=1}^{i-1}X_jv_j\circ X_iv_i+\frac12X_iv_i\circ X_iv_i\right)$$
$$+\sum_{i=1}^m\int_{\tau_i-\Delta}^{\tau_i+\Delta}X_i\int_{\tau_i-\Delta}^t w(\tau)d\tau\circ X_i(w(t)+v_i/2\Delta)dt \tag{7.12}$$
$$+\int_{\tau_i-\Delta}^{\tau_i+\Delta}X_i\int_{\tau_i-\Delta}^t(v_i/2\Delta)d\tau\circ X_iw(t)dt.$$

Splitting $v=(v_1,\dots,v_m)$ into the sum $v^\sharp+v^0$, where $v^\sharp=(v_1^\sharp,\dots,v_m^\sharp)$, $v^0=(v_1^0,\dots,v_m^0)$, we represent the first summand as

$$\sum_{i=1}^m\sum_{j=1}^{i-1}X_jv_j^0\circ X_iv_i^0+O(1)(|v^\sharp||v^0|+|v^\sharp|^2).$$

Obviously $O(1)(|v^\sharp||v^0|+|v^\sharp|^2)$ is $O(1)(\Delta|v^0|^2+\Delta^{-1}|v^\sharp|^2)$, as $(|v^\sharp||v^0|+|v^\sharp|^2)\to0$.

Since $\sum_{j=1}^m q_0\circ X_jv_j^0=0$, then

$$0=\sum_{i=1}^m\sum_{j=1}^m q_0\circ X_jv_j^0\circ X_iv_i^0$$
$$=\sum_{i=1}^m\left(\left(\sum_{j<i}+\frac12\sum_{j=i}^i\right)q_0\circ X_jv_j^0\circ X_iv_i^0\right)+\sum_{i=1}^m\left(\left(\sum_{j<i}+\frac12\sum_{i=j}^j\right)q_0\circ X_iv_i^0\circ X_jv_j^0\right)$$

and hence

$$\sum_{i=1}^m\left(\sum_{j=1}^{i-1}X_jv_j^0\circ X_iv_i^0+\frac12X_iv_i^0\circ X_iv_i^0\right)=\frac12\sum_{i=1}^m\sum_{j<i}q_0\circ[X_jv_j^0,X_iv_i^0].$$

Let us estimate other terms of (7.12), i.e., the sum

$$\sum_{i=1}^{m}\int_{\tau_i-\Delta}^{\tau_i+\Delta}X_i\int_{\tau_i-\Delta}^{t}w(\tau)d\tau\circ X_iw(t)dt+\sum_{i=1}^{m}\int_{\tau_i-\Delta}^{\tau_i+\Delta}X_i\int_{\tau_i-\Delta}^{t}w(\tau)d\tau\circ X_iv_i/2\Delta dt$$
$$+\int_{\tau_i-\Delta}^{\tau_i+\Delta}X_i\int_{\tau_i-\Delta}^{t}(v_i/2\Delta)d\tau\circ X_iw(t)dt.$$

Integration by parts gives us

$$\int_{\tau_i-\Delta}^{\tau_i+\Delta}X_i\int_{\tau_i-\Delta}^{t}w(\tau)d\tau\circ X_iw(t)dt$$
$$=\int_{\tau_i-\Delta}^{\tau_i+\Delta}X_i\int_{\tau_i-\Delta}^{\tau_i+\Delta}w(\tau)d\tau\circ X_iw(t)dt-\int_{\tau_i-\Delta}^{\tau_i+\Delta}X_iw(t)\circ X_i\int_{\tau_i-\Delta}^{t}w(\tau)d\tau dt.$$

Since the first summand of this last formula vanishes, we derive

$$\int_{\tau_i-\Delta}^{\tau_i+\Delta}X_i\int_{\tau_i-\Delta}^{t}w(\tau)d\tau\circ X_iw(t)dt=\frac{1}{2}\int_{\tau_i-\Delta}^{\tau_i+\Delta}\left[X_i\int_{\tau_i-\Delta}^{t}w(\tau)d\tau,X_iw(t)\right]dt;$$

its projection onto ζ_0 vanishes by virtue of the second-order horizontality condition for switchings.

Similarly

$$\int_{\tau_i-\Delta}^{\tau_i+\Delta}X_i\int_{\tau_i-\Delta}^{t}w(\tau)d\tau\circ X_i(v_i/2\Delta)dt$$
$$=\int_{\tau_i-\Delta}^{\tau_i+\Delta}X_i\int_{\tau_i-\Delta}^{\tau_i+\Delta}w(\tau)d\tau\circ X_i(v_i/2\Delta)dt-\int_{\tau_i-\Delta}^{\tau_i+\Delta}X_iw(t)\circ X_i\int_{\tau_i-\Delta}^{t}(v_i/2\Delta)d\tau dt.$$

The first summand in the right-hand side vanishes, and we obtain

$$\int_{\tau_i-\Delta}^{\tau_i+\Delta}X_i\int_{\tau_i-\Delta}^{t}w(\tau)d\tau\circ X_i(v_i/2\Delta)dt+\int_{\tau_i-\Delta}^{\tau_i+\Delta}X_i\int_{\tau_i-\Delta}^{t}(v_i/2\Delta)d\tau\circ X_iw(t)dt$$
$$=\int_{\tau_i-\Delta}^{\tau_i+\Delta}\left[X_i\int_{\tau_i-\Delta}^{t}(v_i/2\Delta)d\tau,X_iw(t)\right]dt.$$

The projection of this last expression onto ζ_0 vanishes by virtue of the second-order horizontality condition for switchings, and therefore the Lemma 7.3 is proved. □

Proof of Lemma 7.4. The estimate (7.10) is derived from the estimate

$$|\mathcal{V}^1(u(\cdot))+\mathcal{V}_h^2(u(\cdot))+\mathcal{V}_v^2(v^0)|\geq a_1\|\bar{u}(\cdot)\|_{L_1}+c_3(|v^\sharp|^2+|v^0|^2+\|w(\cdot)\|_{L_1}^2)$$
$$+\max(0,b_3|v^\sharp|-\beta_3\|\bar{u}(\cdot)\|_{L_1}-k\Delta\|w(\cdot)\|_{L_1}-k\Delta|v^0|), \tag{7.13}$$

which is a modification of the estimate (7.6). To establish (7.13) and to prove that $\mathcal{V}^1(u(\cdot))+\mathcal{V}_h^2(u(\cdot))+\mathcal{V}_v^2(v^0)$ belongs to the lower semispace we return to the estimates (6.12) and note that under the strong genericity assumption $K_i^-=R_+\ell_i$, $K_i^+=R_-\ell_i$ $\forall i=1,\dots,m$, and there exists a constant $B>0$ such that

$$|\langle\zeta_0,q_0\circ((X_\tau-X_i)v-\dot{X}_i\ell_i(\tau-\tau_i)|v|)\rangle|\leq B\Delta|\tau-\tau_i||v| \tag{7.14}$$

$\forall(\tau,v)\in([\tau_i-\Delta,\tau_i)\times K_i^-)\cup((\tau_i,\tau_i+\Delta]\times K_i^+)$.

Repeating the computation (6.14) we derive

$$\sum_{i=1}^{m}\langle\zeta_0, q_0\circ\dot X_i\ell_i\rangle\left(\int_{\tau_i-\Delta}^{\tau_i}(\tau-\tau_i)|v(\tau)|d\tau-\int_{\tau_i}^{\tau_i+\Delta}(\tau-\tau_i)|v(\tau)|d\tau\right)$$

$$\leq-\sum_{i=1}^{m}\langle\zeta_0, q_0\circ\dot X_i\ell_i\rangle\left(\left(\int_{\tau_i-\Delta}^{\tau_i}v(\tau)d\tau\right)^2+\left(\int_{\tau_i}^{\tau_i+\Delta}v(\tau)d\tau\right)^2\right)/2M_i. \tag{7.15}$$

Denoting $v_i^-=\int_{\tau_i-\Delta}^{\tau_i}|v(\tau)|d\tau,\ v_i^+=-\int_{\tau_i}^{\tau_i+\Delta}|v(\tau)|d\tau$ we note that

$$\langle\zeta_0, q_0\circ\dot X_i\ell_i\rangle>0,\ v_i^-\geq 0,\ v_i^+\leq 0,\ \text{and } v_i=v_i^-+v_i^+.$$

To derive an upper estimate for the negative definite quadratic form (7.15) in terms of v_i we have to solve an elementary extremal problem

$$((v_i^-)^2+(v_i^+)^2)/2M_i\to\min,\ v_i^-+v_i^+=v_i, v_i^-\geq 0,\ v_i^+\leq 0.$$

Using the Kuhn–Tucker theorem we obtain the value $v_i^2/2M_i$ for the minimum. Therefore the negative definite quadratic form $-\sum_{i=1}^m\langle\zeta_0, q_0\circ\dot X_i\ell_i\rangle v_i^2/2M$ provides the upper estimate for (7.15). By virtue of (7.14) the negative definite quadratic form

$$-(1-B\Delta)\sum_{i=1}^{m}\langle\zeta_0, q_0\circ\dot X_i\ell_i\rangle v_i^2/2M$$

provides the upper estimate for $\langle\zeta_0, q_0\circ\sum_{i=1}^m\int_{\tau_i-\Delta}^{\tau_i+\Delta}(X_\tau-X_i)v(\tau)d\tau\rangle$.

Extracting the terms which are quadratic in v_0 we represent it as

$$-(1-B\Delta)\left(\sum_{i=1}^{m}\langle\zeta_0, q_0\circ\dot X_i\ell_i\rangle(v_i^0)^2/2M_i+\sum_{i=1}^{m}\langle\zeta_0, q_0\circ\dot X_i\ell_i\rangle(2v_i^\sharp v_i^0+(v_i^\sharp)^2)\right).$$

The second summand admits an estimate

$$\left|\sum_{i=1}^{m}\langle\zeta_0, q_0\circ\dot X_i\ell_i\rangle(2v_i^\sharp v_i^0+(v_i^\sharp)^2)\right|=O(1)(\Delta^{-1}|v^\sharp|^2+\Delta|v^0|^2).$$

Since

$$-\frac{1}{2}\sum_{i=1}^{m}\langle\zeta_0, q_0\circ\dot X_i\ell_i\rangle(v_i^0)^2/2M_i+\sum_{i=1}^{m}\left\langle\zeta_0, q_0\circ\left[\sum_{j<i}X_jv_j^0, X_iv_i^0\right]\right\rangle$$

is a negative definite quadratic form in v^0, then for sufficiently small $\Delta>0$ the quadratic form

$$-\frac{1}{2}(1-B\Delta)\sum_{i=1}^{m}\langle\zeta_0, q_0\circ\dot X_i\ell_i\rangle(v_i^0)^2/2M_i+\sum_{i=1}^{m}\left\langle\zeta_0, q_0\circ\left[\sum_{j<i}X_jv_j^0, X_iv_i^0\right]\right\rangle$$

admits an upper estimate $-\gamma|v^0|^2$ and we derive that $\mathcal{V}^1(u(\cdot))+\mathcal{V}_h^2(u(\cdot))+\mathcal{V}_v^2(v^0)$ belongs to the lower semispace and satisfies the estimate (7.13), which implies the estimate (7.10). □

Acknowledgments. Part of the results of this paper originated in joint research with A. A. Agrachev. The author is grateful to R. V. Gamkrelidze, H. W. Knobloch, H. J. Sussmann, and P. Zezza for helpful discussions and to K. A. Grasse for numerous suggestions on the improvement of the text of the manuscript.

REFERENCES

[1] A. A. Agrachev and R. V. Gamkrelidze, *Exponential representation of flows and chronological calculus*, Mat. Sb., 107 (1978), pp. 467–532 (in Russian). English translation in Math. USSR-Sb., 35 (1979), pp. 727–785.

[2] A. A. Agrachev and R. V. Gamkrelidze, *Symplectic geometry and optimal control*, in Nonlinear Controllability and Optimal Control, H. J. Sussmann, ed., Marcel Dekker, New York, 1990, pp. 263–277.

[3] A. A. Agrachev, R. V. Gamkrelidze, and A. V. Sarychev, *Local invariants of smooth control systems*, Acta Appl. Math., 14 (1989), pp. 191–237.

[4] A. Bressan, *A high-order test for optimality of bang-bang controls*, SIAM J. Control Optim., 23 (1985), pp. 38–48.

[5] R. V. Gamkrelidze, *Principles of Optimal Control Theory*, Plenum Press, New York, 1978.

[6] B. S. Goh, *Necessary conditions for singular extremals involving multiple control variables*, SIAM J. Control, 4 (1966), pp. 716–731.

[7] L. S. Pontryagin, V. G. Boltyanskii, R. V. Gamkrelidze, and E. F. Mischenko, *The Mathematical Theory of Optimal Processes,* Pergamon Press, Oxford, 1964.

[8] A. V. Sarychev, *Index of the second variation of control system*, Mat. Sb., 113 (1980), pp. 464–486 (in Russian). English translation in Math. USSR-Sb., 41 (1982), pp. 383–401.

[9] A. V. Sarychev, *Integral representation for the trajectories of control system with generalized right-hand side*, Differentsialnye Uravnenia, 24 (1988), pp. 1551–1564 (in Russian). English translation in Differential Equations, 24 (1988), pp. 1021–1031.

[10] A. V. Sarychev, *Nonlinear systems with impulsive and generalized function controls*, in Nonlinear Synthesis, C. I. Byrnes and A. B. Kurzhansky, eds., Birkhauser, Boston, MA, 1991, pp. 244–257.

[11] A. V. Sarychev, *On optimality of generalized (impulsive) controls in time-optimal problem*, Differentsialnye Uravnenia, 27 (1991), pp. 788–801 (in Russian). English translation in Differential Equations, 27 (1991), pp. 539–550.

[12] A. V. Sarychev, *Morse index and sufficient optimality conditions for bang-bang Pontryagin extremals*, in System Modelling and Optimization, Lecture Notes in Control and Information Sci. 180, P. Kall, ed., Springer-Verlag, Berlin, 1992, pp. 440–448.

[13] A. V. Sarychev, *On Legendre-Jacobi-Morse-Type Theory of Second Variation for Optimal Control Problems*, Report 382, Schwerpunktprogramm der Deutschen Forschungsgemeinschaft "Anwendungsbezogene Optimierung und Steuerung," Würzburg, 1992.

[14] A. V. Sarychev, *Sufficient optimality conditions for Pontryagin extremals,* Systems Control Lett., 19 (1992), pp. 451–460.

[15] H. Schaettler, *On the local structure of time-optimal bang-bang trajectories*, SIAM J. Control Optim., 26 (1988), pp. 186–204.

[16] H. J. Sussmann, *Structure of optimal trajectories for single-input systems in the plane: The C^∞ nonsingular case*, SIAM J. Control Optim., 25 (1987), pp. 433–465.

[17] H. J. Sussmann, *Envelopes, conjugate points and optimal bang-bang extremals*, in Proc. 1985 Paris Conf. on Nonlinear Systems, M. Fliess and M. Hazewinkel, eds., Reidel Publishers, Dordrecht, the Netherlands, 1987.

SIAM J. CONTROL OPTIM.
Vol. 35, No. 1, pp. 341–361, January 1997

019

A CONVERGENT ALGORITHM FOR THE OUTPUT COVARIANCE CONSTRAINT CONTROL PROBLEM*

G. ZHU†, M. A. ROTEA‡, AND R. SKELTON§

Abstract. This paper considers the optimal control problem of minimizing control effort subject to multiple performance constraints on output covariance matrices Y_i of the form $Y_i \leq \overline{Y}_i$, where $\overline{Y}_i$ is given. The contributions of this paper are a set of conditions that characterize global optimality, and an iterative algorithm for finding a solution to the optimality conditions. This iterative algorithm is completely described up to a user-specified parameter. We show that, under suitable assumptions on problem data, the iterative algorithm converges to a solution of the optimality conditions, provided that this parameter is properly chosen. Both discrete- and continuous-time problems are considered.

Key words. covariance control, convergent algorithm

AMS subject classifications. 49K15, 93C05, 93C55, 93C60

PII. S0363012994263974

1. Introduction.

Consider the following linear time-invariant system:

$$\begin{aligned} \dot{x}_p(t) &= A_p x_p(t) + B_p u(t) + D_p w_p(t), \\ y_p(t) &= C_p x_p(t), \\ z(t) &= M_p x_p(t) + v(t), \end{aligned} \tag{1.1}$$

where x_p is the state, u the control, w_p represents process noise, and v is the measurement noise. The vector y_p contains all variables whose dynamic responses are of interest. The vector z is a vector of noisy measurements.

Suppose that we apply to the plant (1.1) a full state feedback stabilizing control law of the form

$$u(t) = G x_p(t) \tag{1.2}$$

or a strictly proper output feedback stabilizing control law given by

$$\begin{aligned} \dot{x}_c(t) &= A_c x_c(t) + F z(t), \\ u(t) &= G x_c(t). \end{aligned} \tag{1.3}$$

Then the resulting closed-loop system is

$$\begin{aligned} \dot{x}(t) &= Ax(t) + Dw(t), \\ y(t) &= \begin{bmatrix} y_p(t) \\ u(t) \end{bmatrix} = \begin{bmatrix} C_y \\ C_u \end{bmatrix} x(t) = Cx(t), \end{aligned} \tag{1.4}$$

*Received by the editors March 2, 1994; accepted for publication (in revised form) November 27, 1995. This research was supported in part by National Science Foundation grants ECS-91-08493 and ECS-93-58288 and NASA grant NAG8-220.

http://www.siam.org/journals/sicon/35-1/26397.html

†Space Systems Control Laboratory, Purdue University, West Lafayette, IN 47907. Present address: Cummins Engine Company, Inc., 1900 McKinley Ave, MC 50197, Columbus, IN 47201 (g.g.zhu@ctc.cummins.com).

‡School of Aeronautics and Astronautics, 1282 Grissom Hall, Purdue University, West Lafayette, IN 47907 (rotea@ecn.purdue.edu).

§Space Systems Control Laboratory, 1293 Potter Engineering Center, Purdue University, West Lafayette, IN 47907 (skelton@ecn.purdue.edu).

where for the state feedback case we have $x = x_p$ and $w = w_p$, while for the output feedback case we have $x = [x_p^T \ x_c^T]^T$ and $w = [w_p^T \ v^T]^T$. Moreover, formulas for A, C, and D are easy to obtain from (1.1) and (1.2) or (1.3).

Consider the closed-loop system (1.4). Let W_p and V denote positive definite symmetric matrices with dimensions equal to the process noise w_p and measurement vector z, respectively. Define $W = W_p$, if the state feedback controller (1.2) is used in (1.4) or $W =$ block diag $[W_p, \ V]$ if (1.3) is used in (1.4). Let X denote the closed-loop controllability Gramian from the (weighted) disturbance input $W^{-1/2}w$. Since A is stable, X satisfies

$$0 = AX + XA^T + DWD^T. \tag{1.5}$$

Partition the performance output y_p in (1.4) into $y_p := [y_1^T, y_2^T, \ldots, y_m^T]^T$, where $y_i = C_i x \in \mathcal{R}^{m_i}$ for $i = 1, 2, \ldots, m$. In this paper we are interested in finding controllers of the form (1.2) or (1.3) that minimize the (weighted) control energy *trace* $RC_uXC_u^T$ with $R > 0$, and satisfy the constraints

$$Y_i = C_iXC_i^T \le \overline{Y}_i, \quad i = 1, 2, \ldots, m, \tag{1.6}$$

where $\overline{Y}_i > 0$ $(i = 1, 2, \ldots, m)$ are given and X solves (1.5). This problem, which we call the the output covariance constraint (OCC) problem, is defined as follows.

The OCC Problem. Find a static state feedback or full-order dynamic output feedback controller for system (1.1) to minimize the OCC cost

$$J_{OCC} = \text{trace } RC_uXC_u^T, \qquad R > 0, \tag{1.7}$$

subject to (1.5) and (1.6). □

The OCC problem may be given several interesting interpretations. For instance, assume first that w_p and v are uncorrelated zero-mean white noises with intensity matrices $W_p > 0$ and $V > 0$. That is, let $\mathcal{E}$ be an expectation operator, and

$$\begin{aligned} \mathcal{E}[w_p(t)] = 0, \qquad & \mathcal{E}[w_p(t)w_p^T(t-\tau)] = W_p\delta(\tau), \\ \mathcal{E}[v(t)] = 0, \qquad & \mathcal{E}[v(t)v^T(t-\tau)] = V\delta(\tau). \end{aligned} \tag{1.8}$$

Letting $\mathcal{E}_\infty[\cdot] := \lim_{t\to\infty} \mathcal{E}[\cdot]$ and $W = W_p$ for the case of state feedback or $W =$ block diag $[W_p, \ V]$ for the output feedback case, it is easy to see that the OCC is the problem of minimizing $\mathcal{E}_\infty u^T Ru$ subject to the OCCs $Y_i := \mathcal{E}_\infty y_i(t)y_i^T(t) \le \overline{Y}_i$. As is well known, these constraints may be interpreted as constraints on the variance of the performance variables or lower bounds on the residence time (in a given ball around the origin of the output space) of the performance variables [10].

The OCC problem may also be interpreted from a deterministic point of view. To see this, define the $\mathcal{L}_\infty$ and $\mathcal{L}_2$ norms

$$\begin{aligned} \| \ y_i\|_\infty^2 &:= \sup_{t\ge 0} \ y_i^T(t)y_i(t), \\ \| \ w\|_2^2 &:= \int_0^\infty w^T(t)w(t)dt, \end{aligned} \tag{1.9}$$

and define the (weighted) $\mathcal{L}_2$ disturbance set

$$\mathcal{W} := \{w : R \to R^{n_w} \text{ and } \| \ W^{-1/2}w\|_2^2 \le 1\}, \tag{1.10}$$

where $W > 0$ is a real symmetric matrix. Then, for any $w \in \mathcal{W}$, we have [17, 18]

$$\| \ y_i\|_\infty^2 \le \overline{\sigma}[Y_i], \qquad i = 1, 2, \ldots, m, \tag{1.11}$$

and

$$\|u_i\|_\infty^2 \le [C_u X C_u^T]_{ii}, \qquad i = 1, 2, \ldots, n_u, \tag{1.12}$$

where n_u is the dimension of u. (Here, $\overline{\sigma}[\cdot]$ denotes the maximum singular value and $[\,]_{ii}$ is the ith diagonal entry.) Moreover, [17, 18] show that the bounds in (1.11) and (1.12) are the least upper bounds that hold for an arbitrary signal $w \in \mathcal{W}$.

Thus, if we define $\overline{Y}_i := I_{m_i}\epsilon_i^2$ in (1.6) and $R :=$ diag $[r_1, r_2, \ldots, r_{n_u}]$ in (1.7), the OCC problem is the problem of minimizing the (weighted) sum of worst-case peak values on the control signals given by

$$J_{OCC} = \sum_{i=1}^{n_u} r_i \left\{ \sup_{w \in \mathcal{W}} \|u_i\|_\infty^2 \right\} \tag{1.13}$$

subject to constraints on the worst-case peak values of the performance variables of the form

$$\sup_{w \in \mathcal{W}} \|y_i\|_\infty^2 \le \epsilon_i^2, \qquad i = 1, 2, \ldots, m. \tag{1.14}$$

This interpretation is important in applications where hard constraints on responses or actuator signals cannot be ignored, such as space telescope pointing and machine tool control.

Control problems related to the OCC problem defined here have been considered before by several authors. See, for example, [6, 9, 5, 1, 3, 15, 16] for work in multiobjective optimal control with quadratic cost functionals, [13, 14, 4, 19] for the so-called variance constraint control problems, and [12] for the so-called generalized $\mathcal{H}_2$ control problem.

In the above references, one may find two different approaches for solving this class of optimal control problems: the approach based on solving the optimality conditions corresponding to the optimization problem at hand [4, 16, 19] and the approach based on reducing the given problem to a finite dimensional convex optimization problem [1, 3, 12].

In this paper, we follow the approach initiated in [4, 19]. Here, we consider a more general and realistic problem, i.e., the OCC problem, than the one studied in [4, 16, 19], and provide an iterative algorithm for solving the optimality conditions corresponding to this problem. Our main contribution is in the algorithm itself. This iterative algorithm is completely described up to a user-specified parameter. We show that the algorithm converges to a solution of the optimality conditions (assuming that one exists), provided that the user-specified parameter is properly chosen. Both discrete- and continuous-time problems are considered.

The paper is organized as follows. Section 2 provides optimality conditions for the continuous-time OCC problem in the case of state feedback. These conditions comprise one algebraic Riccati equation and one Lyapunov equation. The Riccati equation has a forcing term depending on a matrix Q (which represents the Kuhn–Tucker multipliers) that must be determined. An algorithm for finding this matrix Q is given, and its convergence analyzed. Section 2 concludes with the extension of the state feedback results to the output feedback case. Section 3 is the discrete-time version of section 2. An example is presented in section 4 to illustrate the performance of the algorithm. Section 5 gives the conclusions of this work.

The notation used in this paper is fairly standard. Given the continuous-time algebraic Riccati equation

$$0 = A_p^T K + K A_p - K B_p R^{-1} B_p^T K + C_p^T Q C_p,$$

we say that K is the stabilizing solution if $K = K^T$ satisfies the Riccati equation and $A_p - B_p R^{-1} B_p^T K$ has all eigenvalues in the open left half plane. Similarly, given the discrete-time algebraic Riccati equation

$$K = A_p^T K A_p - A_p^T K B_p (R + B^T K B_p^T)^{-1} B_p^T K A_p + C_p^T Q C_p,$$

we say that K is the stabilizing solution if $K = K^T$ satisfies the Riccati equation and $A_p - B_p(R + B_p K B_p^T)^{-1} B_p^T K A_p$ has all eigenvalues in the open unit disk. Note that when the continuous (or discrete) stabilizing solution exists it is unique. Moreover, if $Q = Q^T \geq 0$, the stabilizing solution is positive semidefinite.

2. The OCC algorithm for continuous systems.

2.1. The OCC algorithm for state feedback. In this section we consider the case of state feedback. With the state feedback controller (1.2) the closed-loop system matrices in (1.4) are given by

$$A = A_p + B_p G, \quad D = D_p, \quad C_y = C_p, \quad C_u = G. \tag{2.1}$$

The following theorem provides conditions for optimality in the state feedback case.

THEOREM 2.1. *Suppose there exists a matrix*

$$Q^* = \text{ block diag } [Q_1^*, Q_2^*, \ldots, Q_m^*] \geq 0, \quad Q_i^* = Q_i^{*T} \in R^{m_i \times m_i}, \quad i = 1, 2, \ldots, m, \tag{2.2}$$

such that the algebraic Riccati equation

$$0 = A_p^T K + K A_p - K B_p R^{-1} B_p^T K + C_p^T Q^* C_p \tag{2.3}$$

has the (unique) stabilizing solution K^*. Define

$$G^* = -R^{-1} B_p^T K^*, \tag{2.4}$$

and let X^* denote the unique solution of the Lyapunov equation

$$0 = (A_p + B_p G^*) X + X (A_p + B_p G^*)^T + D_p W_p D_p^T, \tag{2.5}$$

and define $Y_i = C_i X^* C_i^T$ $(i = 1, 2, \ldots, m)$. Then if

$$0 = (Y_i - \overline{Y}_i) Q_i^* \text{ and } Y_i \leq \overline{Y}_i \tag{2.6}$$

for all $i = 1, 2, \ldots, m$, we have that G^* given by (2.4) *is an optimal solution to the OCC problem defined in* (1.7).

Proof. Let Q^* be given by (2.1) and define the following LQ problem:

$$\min_{(G,X)} J(G, X) = \text{ trace } RGXG^T + \sum_{i=1}^{m} \text{ trace } (C_i X C_i^T - \overline{Y}_i) Q_i^* \tag{2.7}$$

subject to $A_p + B_p G$ stable and

$$0 = (A_p + B_p G) X + X (A_p + B_p G)^T + D_p W_p D_p^T. \tag{2.8}$$

Using a simple completion of square argument, it is easy to see from (2.3), (2.4), and (2.5) that (G^*, X^*) solves (2.7).

Now let G denote a feasible controller (arbitrary but fixed) for the OCC problem. That is, $(A_p + B_pG)$ is stable and $C_iXC_i^T \leq \overline{Y}_i$ (for all $i = 1, 2, \ldots, m$), where X is the closed-loop Gramian corresponding to G. From the previous paragraph, we get that

$$\begin{aligned} J_{OCC}(G^*, X^*) = &\text{ trace } RG^*X^*(G^*)^T + \sum_{i=1}^{m} \text{ trace } (C_iX^*C_i^T - \overline{Y}_i)Q_i^* \\ \leq &\text{ trace } RGXG^T + \sum_{i=1}^{m} \text{ trace } (C_iXC_i^T - \overline{Y}_i)Q_i^* \\ \leq &\text{ trace } RGXG^T. \end{aligned} \tag{2.9}$$

Using the fact that $0 = (C_iX^*C_i^T - \overline{Y}_i)Q_i^*$, from (2.9) we obtain

$$\text{trace } RG^*X^*(G^*)^T \leq \text{ trace } RGXG^T. \tag{2.10}$$

This last inequality, together with the fact that G^* is also feasible for the OCC problem because $C_iX^*C_i^T \leq \overline{Y}_i$ (for all $i = 1, 2, \ldots, m$), implies that G^* is a solution to the OCC problem. □

From (2.3) and (2.4), it follows that the solution of the OCC problem with static state feedback is an LQ controller with a special choice of output-weighting matrix Q. Therefore, our algorithm for solving the conditions in Theorem 2.1 needs only to iterate on Q.

Before giving the algorithm we would like to mention that the existence of Q^* satisfying the conditions of Theorem 2.1 is necessary in certain cases. For example, from Theorem 5.8 of [5], it follows that, when the constraints in (1.6) are scalar and (for example) the pairs (C_1, A_p), ..., (C_m, A_p) do not have imaginary axis unobservable modes, then a diagonal Q^* exists if a solution to the OCC problem exists. See also [3]. The case of block diagonal matrices Q does not seem to appear in the published literature. It should be noted that the emphasis of the present paper is an algorithm for computing Q^* (and thus a controller that solves the OCC problem) under the assumption that a matrix Q^* satisfying the conditions of Theorem 2.1 exists. This algorithm is given next.

To give this algorithm we need to introduce the following operator. Let M denote a real symmetric matrix, and suppose that

$$M = [U_1\ U_2] \text{ block diag } [E_p,\ E_n]\ [U_1\ U_2]^T \tag{2.11}$$

is the (real) Schur decomposition of M, where E_p and E_n are diagonal matrices containing the strictly positive and nonpositive eigenvalues of M in decreasing order, respectively, and $[U_1\ U_2]$ is an orthogonal matrix. Define

$$\mathcal{P}[M] = \begin{cases} 0 & \text{if } M \leq 0, \\ U_1E_pU_1^T & \text{otherwise.} \end{cases} \tag{2.12}$$

Note that if M is a symmetric matrix with block diagonal structure, the operator $\mathcal{P}[\cdot]$ preserves the block structure; i.e., $\mathcal{P}[M]$ has the block structure of M.

The following algorithm for solving the conditions in Theorem 2.1 is the main contribution of this paper.

THE OCC ALGORITHM.

(1) Given A_p, B_p, D_p, C_p, W_p, R, $\overline{Y}^b =$ block diag $[\overline{Y}_1, \overline{Y}_2, \ldots, \overline{Y}_m]$, an initial point $Q(0) =$ block diag $[Q_1(0), Q_2(0), \ldots, Q_m(0)] > 0$, and constants $\alpha > 0$, $0 < \beta < 1$, let $j = 0$ and go to 2).

(2) Compute $K(j) \geq 0$ and $G(j)$ by solving

$$\begin{aligned} 0 &= A_p^T K(j) + K(j) A_p - K(j) B_p R^{-1} B_p^T K(j) + C_p^T Q(j) C_p, \\ G(j) &= -R^{-1} B_p^T K(j). \end{aligned} \tag{2.13}$$

(3) Compute $X(j)$ by solving

$$0 = [A_p + B_p G(j)] X(j) + X(j) [A_p + B_p G(j)]^T + D_p W_p D_p^T. \tag{2.14}$$

(4) Set $Y_i(j) = C_i X(j) C_i^T$ for $i = 1, 2, \ldots, m$.

(5) Let $Y^b(j) =$ block diag $[Y_1(Q(j)), Y_2(Q(j)), \ldots, Y_m(Q(j))]$, and

$$Q(j+1) = \beta Q(j) + (1-\beta) \mathcal{P}[Q(j) + \alpha \{Y^b(j) - \overline{Y}^b\}] , \tag{2.15}$$

$j = j + 1$, go to 2). □

Several stopping criteria may be used to guarantee that the OCC algorithm terminates in a finite number of steps. In this paper, we propose stopping the algorithm whenever the first equation in condition (2.6) is satisfied to a given numerical accuracy. This can be achieved by checking if the inequality

$$\sum_{i=1}^{m} ||(Y_i(j) - \overline{Y}_i) Q_i(j)|| < \epsilon \tag{2.16}$$

holds, where $\epsilon > 0$ is the specified tolerance. Inequality (2.16) must be tested after step 4). If (2.16) holds, we stop the algorithm and declare $G(j)$, $Q_1(j)$, $Q_2(j)$, and $Q_m(j)$ to be a numerical solution to the OCC problem; if (2.16) does not hold, the algorithm continues.

The rest of this section is devoted to showing that, under the assumption

(A1) (A_p, B_p) is stabilizable and A_p has no eigenvalues on the imaginary axis,

if there exists Q^* satisfying the conditions in Theorem 2.1, then the OCC algorithm will find it, provided that α is properly chosen. More specifically, under the assumptions mentioned, we will show that the sequence of matrices $\{Q(j)\}_{j=0}^{\infty}$ generated by the OCC algorithm (see (2.15)) has a limit $\hat{Q}$ which satisfies all the conditions of Theorem 2.1. Thus, the OCC algorithm converges to a globally optimal solution to the OCC problem. Note that the existence of the limit $\hat{Q}$ implies that, given any $\epsilon > 0$, there exists an integer j such that inequality (2.16) holds.

Note that the OCC algorithm is well posed in the sense that the unique positive semidefinite solution $K(j)$ to the Riccati equation (2.13) and the solution $X(j)$ to the Lyapunov equation (2.14) exist at each iteration. This follows from assumption (A1) and the fact that, at each iteration, $Q(j) \geq 0$. As is well known [7], since (A_p, B_p) is stabilizable and the pair $(Q(j)^{1/2} C_p, A_p)$ has no imaginary axis unobservable modes, $K(j) \geq 0$ exists, is unique, and renders $A_p - B_p R^{-1} B_p^T K(j)$ asymptotically stable.

To establish our results we need to introduce the following operators:

$$\begin{aligned} \mathcal{T}_\beta[Q] &:= \beta Q + (1-\beta) \mathcal{T}[Q], \\ \mathcal{T}[Q] &:= \mathcal{P}[Q + \alpha \{Y^b(Q) - \overline{Y}^b\}], \end{aligned} \tag{2.17}$$

where $\alpha > 0$ and $\beta \in (0,1)$ are parameters of the OCC algorithm. Note that, with this notation, the sequence of matrices $Q(j)$ generated by the OCC algorithm is $\{\mathcal{T}_\beta^j[Q(0)]\}_{j=0}^\infty$, where $Q(0) \geq 0$ is block diagonal.

THEOREM 2.2. *Consider the OCC algorithm, and suppose that $\alpha > 0$, $0 < \beta < 1$, and assumption* (A1) *holds. Suppose that the algorithm converges; that is, the sequence $\{\mathcal{T}_\beta^j[Q(0)]\}_{j=0}^\infty$ converges to Q^*. Then $Q^* :=$ block diag $[Q_1^*, Q_2^*, \ldots, Q_m^*]$ satisfies the sufficient conditions in Theorem* 2.1 *for optimality. In other words, if the OCC algorithm converges, the resulting controller $u{=}-R^{-1}B_p^T K^* x$, where K^* solves* (2.3)*, is a global optimal solution to the given OCC problem, where Q^* is the limit of the convergent sequence $\{\mathcal{T}_\beta^j[Q(0)]\}_{j=0}^\infty$.*

The proof of Theorem 2.2 requires the following lemma.

LEMMA 2.3. *For any symmetric matrices $M = M^T$ and $N = N^T$ of the same dimensions, the following statements hold:*

1. *$\mathcal{P}[M+N] = M$ if and only if $M \geq 0$, $N \leq 0$, and $MN = 0$.*
2. *$\|\mathcal{P}[M] - \mathcal{P}[N]\| \leq \|M - N\|$, where $\|\cdot\|$ denotes the Frobenius norm.*

Proof. First, we shall show the necessity part of 1. The property $M \geq 0$ is a direct consequence of the definition of $\mathcal{P}[\cdot]$ in (2.12). Next, we show that $N \leq 0$ and $MN = 0$. Let $M + N$ have the Schur decomposition

$$\begin{aligned} M + N &= [U_1\ U_2]\ \text{block diag}\ [E_p\ ,\ E_n][U_1\ U_2]^T \\ &= U_1 E_p U_1^T + U_2 E_n U_2^T, \end{aligned} \tag{2.18}$$

where $E_p > 0$ and $E_n \leq 0$. Thus, from (2.12) and $\mathcal{P}[M+N] = M$ we obtain

$$\mathcal{P}[M+N] = U_1 E_p U_1^T = M. \tag{2.19}$$

Subtracting this last equation from (2.18) we obtain

$$N = U_2 E_n U_2^T \leq 0. \tag{2.20}$$

Since $U_1^T U_2 = 0$, from (2.19) and (2.20) it follows that

$$MN = U_1 E_p U_1^T U_2 E_n U_2^T = 0.$$

Second, we shall show the sufficiency part of 1. Let $M \geq 0$ and $N \leq 0$ be given and suppose that $MN = 0$. Note that if either M or N is zero, the sufficiency of property 1) is trivial. Now suppose that $M \neq 0$ and $N \neq 0$. The real Schur decompositions of M and N are

$$M = U_1 E_p U_1^T, \qquad N = U_2 E_n U_2^T,$$

where $E_p > 0$, $E_n < 0$, $U_1^T U_1 = I$, $U_2^T U_2 = I$, and $U_1^T U_2 = 0$. Then

$$\mathcal{P}[M+N] = \mathcal{P}[U_1 E_p U_1^T + U_2 E_n U_2^T] = U_1 E_p U_1^T = M. \tag{2.21}$$

Finally, we show property 2. Let M and N have the following Schur decompositions:

$$\begin{aligned} M &= [U_1\ U_2]\ \text{block diag}\ [E_p\ ,\ E_n][U_1\ U_2]^T, \quad E_p > 0, \quad E_n \leq 0, \\ N &= [\hat{U}_1\ \hat{U}_2]\ \text{block diag}\ [\hat{E}_p\ ,\ \hat{E}_n][\hat{U}_1\ \hat{U}_2]^T, \quad \hat{E}_p > 0, \quad \hat{E}_n \leq 0. \end{aligned} \tag{2.22}$$

Let

$$\begin{aligned} M^+ &:= \mathcal{P}[M] = U_1 E_p U_1^T > 0, \\ M^- &:= U_2 E_n U_2^T \le 0, \\ N^+ &:= \mathcal{P}[N] = \hat{U}_1 \hat{E}_p \hat{U}_1^T > 0, \\ N^- &:= \hat{U}_2 \hat{E}_n \hat{U}_2^T \le 0. \end{aligned}$$

Note that $M = M^+ + M^-$, $N = N^+ + N^-$, $M^+M^- = 0$, and $N^+N^- = 0$. Then

$$\begin{aligned} \|M - N\|^2 &= \|(M^+ - N^+) + (M^- - N^-)\|^2 \\ &= \|M^+ - N^+\|^2 + \|M^- - N^-\|^2 \\ &\quad -2\text{trace}M^-N^+ - 2\text{trace}M^+N^-. \end{aligned} \tag{2.23}$$

Since $-\text{trace}M^-N^+ \ge 0$ and $-\text{trace}M^+N^- \ge 0$, we obtain

$$\|M - N\|^2 \ge \|M^+ - N^+\|^2 = \|\mathcal{P}[M] - \mathcal{P}[N]\|^2, \tag{2.24}$$

which completes the proof. □

The following lemma, essentially due to [2], is also required for the proof of Theorem 2.2.

LEMMA 2.4. *Consider the plant defined in* (1.1) *and suppose that assumption* (A1) *holds. Let K denote the unique stabilizing solution to the Riccati equation* (2.3) *with $Q = Q^T \ge 0$. Then $K(Q)$ is a real analytic function of $Q = Q^T \ge 0$.*

Proof of Theorem 2.2. By Lemma 2.4, the state feedback control gain $G(Q)$ in (2.4) is a continuous function of $Q = Q^T \ge 0$. Hence, the block output covariance matrix $Y^b(Q)$ is continuous with respect to $Q = Q^T \ge 0$. Since the operator $\mathcal{P}[\cdot]$ is continuous, we obtain that $\mathcal{T}[\cdot]$ and $\mathcal{T}_\beta[\cdot]$ are well defined and continuous for any $Q = Q^T \ge 0$. Suppose that $\{\mathcal{T}_\beta^j[Q(0)]\}_{j=0}^\infty$ converges to Q^*, i.e.,

$$\lim_{j\to\infty} \mathcal{T}_\beta^j[Q(0)] = Q^*. \tag{2.25}$$

Since $\beta \in (0,1)$ and $\mathcal{P}[\cdot]$ preserves the block structure, we may conclude that, for each j, $\mathcal{T}_\beta^j[Q(0)]$ has the correct block diagonal structure and is positive semidefinite. Thus, $Q^* = \text{block diag}\,[Q_1^*, Q_2^*, \ldots, Q_m^*]$ and $Q^* \ge 0$.

From the continuity properties of $\mathcal{T}_\beta$, we obtain

$$\mathcal{T}_\beta[Q^*] = \mathcal{T}_\beta\left\{\lim_{j\to\infty} \mathcal{T}_\beta^j[Q(0)]\right\} = \lim_{j\to\infty} \mathcal{T}_\beta^{j+1}[Q(0)] = Q^*. \tag{2.26}$$

That is, Q^* is a fixed point of $\mathcal{T}_\beta[\cdot]$. Since $\beta \ne 1$, from (2.17), we get

$$Q^* = \mathcal{T}[Q^*] = \mathcal{P}[Q^* + \alpha\{Y^b(Q^*) - \overline{Y}^b\}]\,. \tag{2.27}$$

Let $M = Q^*$ and $N = \alpha[Y^b(Q^*) - \overline{Y}^b]$. From Lemma 2.3, we conclude that

$$\alpha[Y^b(Q^*) - \overline{Y}^b] \le 0 \;\text{ and }\; \alpha Q^*[Y^b(Q^*) - \overline{Y}^b] = 0\,.$$

Since $\alpha > 0$, the above inequalities imply that Q^* satisfies (2.6). Hence, Q^* satisfies the conditions in Theorem 2.1. This completes the proof. □

The following result shows that there is always a choice for the parameter α in the OCC algorithm that will guarantee its convergence, provided that the conditions in Theorem 2.1 admit one solution.

THEOREM 2.5. *Suppose that assumption* (A1) *holds. Assume also that there exists* Q^* *satisfying the conditions in Theorem* 2.1. *Then, given any* $Q(0) \geq 0 \in \mathcal{R}^{n_y \times n_y}$ *with the appropriate block diagonal structure, there exists an* $\alpha^* > 0$ *such that if* $0 < \alpha \leq \alpha^*$, *the sequence* $\{\mathcal{T}_\beta^n[Q(0)]\}_{n=0}^\infty$ *will converge to some* $\hat{Q} \geq 0$ *satisfying the conditions in Theorem* 2.1. *That is, the OCC algorithm will converge to a global optimal solution of the given OCC problem.*

In order to prove Theorem 2.5, we need a few intermediate results and definitions. Let $Q = Q^T \geq 0$ be given and K denote the (unique) stabilizing solution to

$$0 = A_p^T K + K A_p - K B_p R^{-1} B_p^T K + C_p^T Q C_p. \tag{2.28}$$

Then, with the state feedback gain $G = -R^{-1} B_p^T K$, the lth output covariance of the closed-loop system Y_l $(l = 1, 2, \cdots, m)$ is given by

$$Y_l = C_l X C_l^T,$$

where X is the unique solution to

$$0 = (A_p + B_p G)X + X(A_p + B_p G)^T + D_p W_p D_p^T. \tag{2.29}$$

Now, let

$$Q = \text{block diag}[Q_1, Q_2, \ldots, Q_m], \qquad Q_l := [q_{ij}^l] \in \mathcal{R}^{m_l \times m_l}, \tag{2.30}$$

and

$$Y^b = \text{block diag}[Y_1, Y_2, \ldots, Y_m]\ . \tag{2.31}$$

Below, we compute the derivative of Y^b with respect to the weighting matrix Q given in (2.30). We do this using vector notation. Let Q be given by (2.30) and define the operator *svec* by

$$svec[Q] = \begin{bmatrix} q^1 \\ q^2 \\ \vdots \\ q^m \end{bmatrix} \in \mathcal{R}^n\ , \tag{2.32}$$

where

$$q^i := \sqrt{2}\ \left[\frac{q_{11}^i}{\sqrt{2}}, q_{12}^i, \ldots, q_{1m_i}^i, \frac{q_{22}^i}{\sqrt{2}}, q_{23}^i, \ldots, q_{2m_i}^i, \ldots, \frac{q_{m_i m_i}^i}{\sqrt{2}}\right]^T\ . \tag{2.33}$$

Note also that the operator *svec* defined in (2.32) preserves the Frobenius norm; i.e., if Q is given by (2.30), we have

$$||Q|| = ||svec[Q]||\ . \tag{2.34}$$

Moreover, $svec[\cdot]$ is a linear operator.

Let

$$y = svec[Y^b] \,, \tag{2.35}$$

where Y^b is given by (2.31). Define also the symmetric matrix

$$E_i = svec^{-1}[e_i] \,, \tag{2.36}$$

where $e_i \in \mathcal{R}^n$ has a one in the ith row and zeros elsewhere, and $svec^{-1}$ is the inverse of the operator *svec*.

LEMMA 2.6. *Consider the system defined in* (1.1)*, and suppose that assumption* (A1) *holds. Let* $Q = Q^T \geq 0$ *be given by* (2.30)*, and define* $q = svec[Q]$*. Let* y *be given by* (2.35)*. Then the partial derivative of* $y \in \mathcal{R}^n$ *with respect to* $q \in \mathcal{R}^n$ *is*

$$\frac{\partial y}{\partial q} = -[H_{ij}] = -[2\text{trace}(P_i B_p R^{-1} B_p^T P_j X)], \qquad i,j = 1,2,\ldots,n, \tag{2.37}$$

where P_i *is the unique solution to*

$$0 = P_i(A_p + B_p G) + (A_p + B_p G)^T P_i + C_p^T E_i C_p \tag{2.38}$$

with E_i *given by* (2.36)*. Moreover, if* $Q = Q^T \geq 0$*, the matrix-valued function* $H(Q) = [H_{ij}]$ *is continuous and it satisfies* $H(Q) \geq 0$*.*

Proof. Let y_i denote the ith component of y. From the definition of the operator *svec* (see, for example, (2.32)) it follows that

$$y_i = \text{trace}(E_i C_p X C_p^T). \tag{2.39}$$

Using the Lyapunov equations (2.29) and (2.38) it follows from (2.39) that

$$y_i = \text{trace}(P_i D_p W_p D_p^T), \tag{2.40}$$

where P_i is the solution to (2.38). Hence, from (2.40), we get

$$\frac{\partial y_i}{\partial q_j} = \text{trace}(P_{ij} D_p W_p D_p^T), \tag{2.41}$$

where q_j is the jth component of q and $P_{ij} = \frac{\partial P_i}{\partial q_j}$.

Now to generate P_{ij}, differentiate equation (2.38) with respect to q_j to obtain

$$\begin{aligned} 0 &= P_{ij}(A_p + B_p G) + (A_p + B_p G)^T P_{ij} \\ &\quad - P_i B_p R^{-1} B_p^T \frac{\partial K}{\partial q_j} - \frac{\partial K}{\partial q_j} B_p R^{-1} B_p^T P_i. \end{aligned} \tag{2.42}$$

From the Riccati equation (2.28) and the Lyapunov equation (2.38) we get

$$P_j = \frac{\partial K}{\partial q_j},$$

where P_j solves (2.38) with the "E-matrix" equal to E_j. Hence, from (2.42), we obtain

$$\begin{aligned} P_{ij} = -\int_0^\infty \exp[(A_p + B_p G)^T t][P_i B_p R^{-1} B_p^T P_j \\ + P_j B_p R^{-1} B_p^T P_i] \exp[(A_p + B_p G)t] dt. \end{aligned} \tag{2.43}$$

Finally, from (2.29), (2.41), and (2.43) we obtain

$$\frac{\partial y}{\partial q} = -[H_{ij}] = -[2\text{trace}(P_i B_p R^{-1} B_p^T P_j X)] \,, \tag{2.44}$$

which gives (2.37).

The continuity of $H(Q)$ follows from the fact that, on the set of positive semidefinite matrices Q, the matrix-valued functions P_i, P_j, and X are all continuous. Note also that

$$\begin{aligned} H_{ij} &= \text{trace}\,(P_i B_p R^{-1} B_p^T P_j X) \\ &= \langle X^{1/2} P_i B_p R^{-1/2}, X^{1/2} P_j B_p R^{-1/2} \rangle, \end{aligned}$$

where $\langle M, N \rangle = \text{trace} MN^T$ is the standard inner product on the space of matrices $\mathcal{R}^{n_x \times n_u}$, where n_x and n_u are dimensions of the plant states and controls. Thus, H is of the form

$$H = [\langle X^{1/2} P_i B_p R^{-1/2}, X^{1/2} P_j B_p R^{-1/2} \rangle], \tag{2.45}$$

which shows that $H \geq 0$. □

The following results may be found in [11]; see Propositions 3.2.3 and 12.3.7.

LEMMA 2.7. *Assume that $\mathcal{F} : \mathcal{R}^n \to \mathcal{R}^n$ is Fréchet differentiable on a convex set $\mathcal{D}'_0 \subset \mathcal{R}^n$. Then for any x and $y \in \mathcal{D}'_0$,*

$$\|\mathcal{F}(y) - \mathcal{F}(x)\| \leq \sup_{0 \leq t \leq 1} \overline{\sigma}\{\mathcal{F}'[x + t(y - x)]\} \|x - y\| \,, \tag{2.46}$$

where $\mathcal{F}'(\cdot)$ denotes the Fréchet derivative of $\mathcal{F}(\cdot)$ and $\overline{\sigma}[\cdot]$ denotes the maximum singular value of $[\cdot]$.

LEMMA 2.8. *Suppose that $\mathcal{T} : \mathcal{R}^{n_y \times n_y} \to \mathcal{R}^{n_y \times n_y}$ is nonexpansive on the closed, convex set $\mathcal{D}_0$. That is, for any $x, y \in \mathcal{D}_0$, we have*

$$\|\mathcal{T}(y) - \mathcal{T}(x)\| \leq \|y - x\| \,. \tag{2.47}$$

Assume, further, that $\mathcal{T}\mathcal{D}_0 \subset \mathcal{D}_0$ and that $\mathcal{D}_0$ contains a fixed point of $\mathcal{T}$. Then for any $\beta \in (0, 1)$ and $x^0 \in \mathcal{D}_0$ the iteration

$$x^{k+1} = \beta x^k + (1 - \beta)\mathcal{T}(x^k), \qquad k = 0, 1, \ldots, \tag{2.48}$$

converges to a fixed point of $\mathcal{T}$ in $\mathcal{D}_0$.

Proof of Theorem 2.5. The proof of Theorem 2.5 consists of two steps. First, we show the nonexpansive property of operator $\mathcal{T}$ defined in (2.17). By assumption, there exists Q^* satisfying the conditions in Theorem 2.1. Define a subset of $\mathcal{R}^{n_y \times n_y}$ as follows:

$$\begin{aligned} \mathcal{D}_0 := \{ & Q \geq 0 \in \mathcal{R}^{n_y \times n_y} : Q = \text{block diag}[Q_1, Q_2, \ldots, Q_m] \\ & \text{and } \|Q - Q^*\| \leq \|Q(0) - Q^*\| \}, \end{aligned} \tag{2.49}$$

where n_y is the dimension of y_p, and $Q(0)$ is the initial output-weighting matrix for the OCC algorithm. It is obvious that the set $\mathcal{D}_0$ is compact (i.e., closed and bounded) and convex. Let $\mathcal{D}'_0$ be a set defined by

$$\mathcal{D}'_0 := \{ q = svec[Q] \in \mathcal{R}^n : Q \in \mathcal{D}_0 \} \,. \tag{2.50}$$

It is clear that $\mathcal{D'}_0$ is convex, because $svec[\cdot]$ is a linear operator and $\mathcal{D}_0$ is convex. Let $q \in \mathcal{D}'_0$ and define $y(q) = svec[Y^b(Q)]$, and

$$\mathcal{F}[q] = q + \alpha y(q). \tag{2.51}$$

Note that $\mathcal{F}[\cdot]$ is well defined and Fréchet differentiable, with respect to q, in $\mathcal{D'}_0$. In fact, from Lemma 2.6, it follows that the Frechet derivative of $\mathcal{F}[\cdot]$ is

$$\mathcal{F}'[q] = I - \alpha H(q), \tag{2.52}$$

where $H(q)$ is defined in (2.37). (Here, we think of H as a function of $q = svec[Q]$ instead of a function Q.)

Now, let Q^ν and Q^μ in $\mathcal{D}_0$ be given. Define $q^\nu = svec[Q^\nu]$ and $q^\mu = svec[Q^\mu]$. Then, since $svec[\cdot]$ preserves the Frobenius norm, we have

$$\begin{aligned} \|\mathcal{T}[Q^\nu] - \mathcal{T}[Q^\mu]\| &= \|\mathcal{P}[Q^\nu - \alpha\{Y^b(Q^\nu) - \overline{Y}^b\}] - \mathcal{P}[Q^\mu - \alpha\{Y^b(Q^\mu) - \overline{Y}^b\}]\| \\ &\le \|Q^\nu - Q^\mu - \alpha[Y^b(Q^\nu) - Y^b(Q^\mu)]\| \\ &= \|q^\nu - q^\mu - \alpha[y^\nu - y^\mu]\|, \end{aligned} \tag{2.53}$$

where $y^\nu := svec[Y^b(Q^\nu)]$ and $y^\mu := svec[Y^b(Q^\mu)]$. Since q^ν and q^μ belong to $\mathcal{D'}_0$, using Lemma 2.7, we have

$$\begin{aligned} \|q^\nu - q^\mu - \alpha[y^\nu - y^\mu]\| &= \|\mathcal{F}[q^\nu] - \mathcal{F}[q^\mu]\| \\ &\le \sup_{0\le t\le 1} \overline{\sigma}[\mathcal{F}'[tq^\nu + (1-t)q^\mu]]\|q^\nu - q^\mu\| \\ &= \sup_{0\le t\le 1} \overline{\sigma}[I - \alpha H\{[tq^\nu + (1-t)q^\mu]\}]\|q^\nu - q^\mu\|. \end{aligned} \tag{2.54}$$

Since H is a continuous function over the compact set $\mathcal{D'}_0$, there exists an $\alpha^* > 0$ such that for any $q^\nu \in \mathcal{D'}_0$, $q^\mu \in \mathcal{D}'_0$, and $0 \le t \le 1$, we have

$$\overline{\sigma}\{H[tq^\nu + (1-t)q^\mu]\} \le 2/\alpha^*. \tag{2.55}$$

Thus, since for any q^ν and $q^\mu \in \mathcal{D'}_0$ and any $t \in [0,1]$, $H[tq^\nu + (1-t)q^\mu] \ge 0$, we have

$$\sup_{0\le t\le 1} \overline{\sigma}\{I - \alpha H[tq^\nu + (1-t)q^\mu]\} \le 1 \tag{2.56}$$

for any $\alpha \le \alpha^*$. Therefore, using (2.53) and (2.54), for any $\alpha \le \alpha^*$ we obtain

$$\|\mathcal{T}[Q^\nu] - \mathcal{T}[Q^\mu]\| \le \|q^\nu - q^\mu\| = \|Q^\nu - Q^\mu\|. \tag{2.57}$$

Hence, for any $\alpha \le \alpha^*$, the operator $\mathcal{T}$ is nonexpansive on $\mathcal{D}_0$. Replacing Q^μ by Q^* proves that for any $Q \in \mathcal{D}_0$

$$\|\mathcal{T}[Q] - \mathcal{T}[Q^*]\| \le \|Q - Q^*\| \le \|Q(0) - Q^*\|. \tag{2.58}$$

Now, using Lemma 2.3 and the fact that Q^* satisfies the conditions of Theorem 2.1, we conclude that $\mathcal{T}[Q^*] = Q^*$. This equation and (2.58) imply

$$\|\mathcal{T}[Q] - Q^*\| \le \|Q(0) - Q^*\|; \tag{2.59}$$

therefore, $\mathcal{T}[Q] \in \mathcal{D}_0$.

Second, we shall show the convergence of the OCC algorithm, that is, the convergence of the sequence $\{\mathcal{T}_\beta^j[Q(0)]\}_{j=0}^\infty$. Since $\mathcal{T}\mathcal{D}_0 \subset \mathcal{D}_0$, $\mathcal{D}_0$ is convex and contains a fixed point of $\mathcal{T}$, from Lemma 2.8 we obtain that the sequence $\{\mathcal{T}_\beta^j[Q(0)]\}_{j=0}^\infty$ generated by the iteration

$$Q(j+1) = \mathcal{T}_\beta[Q(j)] = \beta Q(j) + (1-\beta)\mathcal{T}[Q(j)] \tag{2.60}$$

converges to a fixed point of $\mathcal{T}$ in $\mathcal{D}_0$—say, $\hat{Q}$. The fact that $\hat{Q}$ satisfies the sufficient conditions in Theorem 2.1 is the direct consequence of Theorem 2.2. □

2.2. The OCC algorithm for full-order dynamic feedback. The extension of the state feedback case to the full-order dynamic feedback case is straightforward. In fact, the state feedback OCC algorithm can be applied to solve the full-order dynamic feedback OCC problem. Here, for system (1.1), we assume that assumption (A1) holds and that

(A2) (M_p, A_p) is detectable.

As is well known [7], under assumption (A2), there exists a unique matrix $\tilde{X}$ that satisfies the Riccati equation

$$0 = A_p\tilde{X} + \tilde{X}A_p^T - \tilde{X}M_p^T V^{-1} M_p \tilde{X} + D_p W_p D_p^T \tag{2.61}$$

and $A_p - \tilde{X}M_p^T V^{-1}$ is asymptotically stable. Moreover, $\tilde{X} \geq 0$. With this matrix $\tilde{X}$, we define

$$F = \tilde{X}M_p^T V^{-1}\ . \tag{2.62}$$

THEOREM 2.9. *Consider the plant defined in* (1.1). *Let* $\tilde{X}$ *and* F *denote the matrices in* (2.61) *and* (2.62). *Suppose that there exists a matrix*

$$Q^* = \text{ block diag}[Q_1^*, Q_2^*, \ldots, Q_m^*] \geq 0, \quad Q_i^* = {Q_i^*}^T \in \mathcal{R}^{m_i \times m_i}, \quad i = 1, 2, \ldots, m, \tag{2.63}$$

such that the algebraic Riccati equation

$$0 = A_p^T K + K A_p - K B_p R^{-1} B_p^T K + C_p^T Q^* C_p \tag{2.64}$$

has the (unique) stabilizing solution K^*. *Define*

$$G = -R^{-1}B_p^T K^*, \tag{2.65}$$

and let X^* *denote the unique solution to the Lyapunov equation*

$$0 = (A_p + B_p G)X + X(A_p + B_p G)^T + FVF^T, \tag{2.66}$$

and define $Y_i = C_i(\tilde{X} + X^*)C_i^T$ $(i = 1, 2, \ldots, m)$. *Then if*

$$0 = (Y_i - \overline{Y}_i)Q_i^* \text{ and } Y_i \leq \overline{Y}_i \tag{2.67}$$

for all $i = 1, 2, \ldots, m$, *the dynamic controller*

$$\begin{aligned} \dot{x}_c(t) &= (A_p + B_p G - FM_p)x_c(t) + Fz(t), \\ u(t) &= Gx_c(t) \end{aligned} \tag{2.68}$$

is an optimal solution to the OCC problem defined in (1.7).

A proof of this theorem may be obtained by combining Theorem 2.1 in this paper and Lemma 4.2 and Theorem 4.1 in [12]. The result in [12] shows how to reduce the OCC problem (and other $\mathcal{H}_2$-like problems) with output feedback to an equivalent problem with state feedback.

Note that the matrices $\tilde{X}$ and F in (2.61) and (2.62) do not depend on the weighting matrix Q^*. To find a matrix Q^* satisfying the conditions in Theorem 2.9, we can use the OCC algorithm given in section 2. This requires that we replace D_p, W_p, and $\overline{Y}_i$ in the OCC algorithm F, V, and $\overline{Y}_i - C_i\tilde{X}C_i^T$, respectively.

3. Discrete-time version. The discrete-time version of the OCC problem is very much like the continuous-time case. Here, we give the definition of the OCC problem and the main results.

Consider the following discrete system:

$$\begin{aligned} x_p(k+1) &= A_p x_p(k) + B_p u(k) + D_p w_p(k), \\ y_p(k) &= C_p x_p(k), \\ z(k) &= M_p x_p(k) + v(k). \end{aligned} \tag{3.1}$$

Suppose that we apply to the plant (3.1) a full state feedback stabilizing control, i.e.,

$$u(k) = Gx(k), \tag{3.2}$$

or a strictly proper stabilizing control

$$\begin{aligned} x_c(k+1) &= A_c x_c(k) + Fz(k), \\ u(k) &= Gx_c(k). \end{aligned} \tag{3.3}$$

Then the closed-loop system has the following form:

$$\begin{aligned} x(k+1) &= Ax(k) + Dw(k), \\ y(k) &= \begin{bmatrix} y_p(k) \\ u(k) \end{bmatrix} = \begin{bmatrix} C_y \\ C_u \end{bmatrix} x(k) = Cx(k), \end{aligned} \tag{3.4}$$

where the definitions of matrices A, B, and C and vectors x, w, and y are as in the continuous-time case.

As in section 1, let $W_p > 0$ and $V > 0$ denote symmetric matrices with dimensions equal to the dimensions of w_p and z, respectively. Define $W = W_p$ if (3.2) is used in (3.4) or $W =$ block diag$[W_p,\ V]$ if (3.3) is used. Let X denote the closed-loop controllability Gramian from the input $W^{-1/2}w$. Since A is stable, X is given by

$$X = AXA^T + DWD^T. \tag{3.5}$$

As in the continuous-time case, we seek a solution to the following optimal control problem.

The Discrete-Time OCC Problem. Find a state feedback stabilizing controller (3.2) or a strictly proper output feedback stabilizing controller (3.3) for the system (3.1) to minimize the OCC cost

$$J_{OCC} = \text{trace}RC_uXC_u^T, \qquad R > 0, \tag{3.6}$$

subject to

$$Y_i = C_iXC_i^T \le \overline{Y}_i, \qquad i = 1, 2, \ldots, m, \tag{3.7}$$

where X is given by (3.5). □

The discrete-time OCC problem has interpretations similar to the ones of the continuous-time case. For example, the discrete-time OCC problem may be interpreted as the problem of minimizing a weighted sum of the worst-case peak values of the control signals u_i subject to constraints on the worst-case peak values of the response y_i, when the disturbance w is unknown but has bounded energy. This is because, as in the continuous-time case, discrete-time gains from ℓ_2 to ℓ_∞ may also be computed using controllability Gramians [18].

3.1. State feedback case. In this section we consider the case of state feedback. The following theorem provides conditions for global optimality. Its proof is similar to that of Theorem 2.1 and is omitted.

THEOREM 3.1. *Suppose there exists a matrix*

$$Q^* = \text{ block diag}[Q_1^*, Q_2^*, \dots, Q_m^*] \geq 0, \quad Q_i^* = {Q_i^*}^T \in R^{m_i \times m_i}, \qquad i = 1, 2, \dots, m, \tag{3.8}$$

such that the algebraic Riccati equation

$$K = A_p^T K A_p - A_p^T K B_p (R + B_p^T K B_p)^{-1} B_p^T K A_p + C_p^T Q^* C_p \tag{3.9}$$

has the (unique) stabilizing solution K^*. Define

$$G^* = -(R + B_p^T K^* B_p)^{-1} B_p^T K^* A_p, \tag{3.10}$$

let X^* denote the unique solution of the Lyapunov equation

$$X = (A_p + B_p G^*) X (A_p + B_p G^*)^T + D_p W_p D_p^T, \tag{3.11}$$

and define $Y_i = C_i X^* C_i^T$ $(i = 1, 2, \dots, m)$. Then, if

$$0 = (Y_i - \overline{Y}_i) Q_i^* \text{ and } Y_i \leq \overline{Y}_i \tag{3.12}$$

for all $i = 1, 2, \dots, m$, we have that G^* given by (3.10) *is an optimal solution to the OCC problem defined in* (3.6).

The following algorithm may be used to find a matrix Q^* and consequently a matrix G^* for the OCC problem satisfying the conditions in Theorem 3.1.

THE DISCRETE-TIME OCC ALGORITHM.

(1) Given A_p, B_p, D_p, C_p, W_p, R, $\overline{Y}_i = \text{block diag}[\overline{Y}_1, \overline{Y}_2, \dots, \overline{Y}_m]$, an initial point $Q(0) = \text{block diag}[Q_1(0), Q_2(0), \dots, Q_m(0)] > 0$, and constants $\alpha > 0$, $0 < \beta < 1$, let $j = 0$ and go to 2).

(2) Compute $K(j) \geq 0$ and $G(j)$ by solving

$$\begin{aligned} K(j) &= A_p^T K(j) A_p + C_p^T Q(j) C_p \\ &\quad - A_p^T K(j) B_p [R + B_p^T K(j) B_p]^{-1} B_p^T K(j) A_p, \\ G(j) &= -[R + B_p^T K(j) B_p]^{-1} B_p^T K(j) A_p. \end{aligned} \tag{3.13}$$

(3) Compute $X(j)$ by solving

$$X(j) = [A_p + B_p G(j)] X(j) [A_p + B_p G(j)]^T + D_p W_p D_p^T. \tag{3.14}$$

(4) Set $Y_i(j) = C_i X(j) C_i^T$ for $i = 1, 2, \dots, m$.

(5) Let $Y^b(j) = \text{block diag}[Y_1(Q(j)), Y_2(Q(j)), \ldots, Y_m(Q(j))]$, and

$$Q(j+1) = \beta Q(j) + (1-\beta)\mathcal{P}[Q(j) + \alpha\{Y^b(j) - \overline{Y}^b\}], \tag{3.15}$$

$j = j+1$, go to 2).

In (3.15), the operator $\mathcal{P}[\cdot]$ is as defined in (2.12).

The same stop criterion as the continuous-time case is proposed for the discrete-time OCC algorithm. That is, (2.16) needs to be tested after step 4) for a given tolerance $\epsilon > 0$. If (2.16) holds, $G(j)$, $Q_1(j)$, $Q_2(j)$, and $Q_m(j)$ are numerical solutions for the given OCC problem; otherwise, the algorithm continues.

In the rest of this section, we will assume the following:

(A3) (A_p, B_p) is stabilizable and A_p has no eigenvalues on the unit circle.

Note that the discrete-time OCC algorithm is well posed in the sense that the unique positive semidefinite solution $K(j)$ to the Riccati equation (3.13) and the solution $X(j)$ to the Lyapunov equation (3.14) exist at each iteration. This follows from assumption (A3) and the fact that, at each iteration, $Q(j) \geq 0$. As is well known [8], since (A_p, B_p) is stabilizable and the pair $(Q(j)^{1/2}C_p, A_p)$ has no unobservable modes on the unit circle, $K(j) \geq 0$ exists, is unique, and renders $A_p - B_p(R + B_p^T K(j) B_p)^{-1} B_p^T K(j) A_p$ asymptotically stable.

A close examination of the proofs of the continuous-time results given in Theorems 2.2 and 2.5 reveals that the convergence property of the continuous-time algorithm follows from

(i) the properties of the operator $\mathcal{P}[\cdot]$ given in Lemma 2.3,

(ii) the properties of the stabilizing solution to the continuous-time Riccati equation given in Lemma 2.4,

(iii) the formula for the derivatives of the output covariance matrices $Y_1(Q)$, $Y_2(Q)$, ..., $Y_m(Q)$, with respect to Q, given in Lemma 2.6.

Certainly, property (i) above holds in the discrete-time case because the operator $\mathcal{P}[\cdot]$ is the same. Also, it is relatively easy to show that, under assumption (A3), property (ii) extends to the discrete-time setting. Finally, property (iii) above also holds in the discrete-time case, provided that the Lyapunov equation (2.38) is replaced by its discrete-time counterpart and the matrices R and X in (2.37) are replaced by $R + B_p^T K B_p$ and $X - D_p W_p D_p^T$, respectively. Thus, we may now conclude the following result.

THEOREM 3.2. *Suppose that the assumption* (A3) *holds. Assume also that there exists Q^* satisfying the conditions in Theorem* 3.1. *Then, given any $Q(0) \geq 0 \in \mathcal{R}^{n_y \times n_y}$ with the appropriate block diagonal structure, there exists an $\alpha^* > 0$ such that if $0 < \alpha \leq \alpha^*$, the sequence $\{\mathcal{T}_\beta^j[Q(0)]\}_{j=0}^\infty$ will converge to some $\hat{Q} \geq 0$ satisfying the conditions in Theorem* 3.1. *That is, the discrete-time OCC algorithm will converge to a global optimal solution of the given OCC problem.*

3.2. Full-order dynamic feedback. As in the continuous-time case, the discrete-time state feedback results can be readily extended to solve the discrete-time OCC problem with output feedback.

Consider the system (3.1) and suppose that

(A4) (M_p, A_p) is detectable.

Then there exists a unique matrix $\tilde{X}$ that satisfies the Riccati equation

$$\tilde{X} = A_p \tilde{X} A_p^T - A_p \tilde{X} M_p^T (V + M_p \tilde{X} M_p^T)^{-1} M_p \tilde{X} A_p^T + D_p W_p D_p^T \tag{3.16}$$

and $A_p - A_p\tilde{X}M_p^T(V + M_p\tilde{X}M_p^T)^{-1}M_p$ is asymptotically stable; see, for example, [8]. Moreover, $\tilde{X} \geq 0$. With this matrix $\tilde{X}$, we define

$$F = A_p\tilde{X}M_p^T(V + M_p\tilde{X}M_p^T)^{-1}. \tag{3.17}$$

The next result gives a solution to the OCC problem with strictly proper output feedback controllers. The proof follows the continuous-time case and is omitted.

THEOREM 3.3. *Consider the plant defined in* (3.1). *Let* $\tilde{X}$ *and* F *denote the matrices in* (3.16) *and* (3.17). *Suppose that there exists a matrix*

$$Q^* = \text{ block diag}[Q_1^*, Q_2^*, \ldots, Q_m^*] \geq 0, \quad Q_i^* = Q_i^{*T} \in \mathcal{R}^{m_i \times m_i}, \quad i = 1, 2, \ldots, m, \tag{3.18}$$

such that the algebraic Riccati equation

$$K = A_p^T K A_p - A_p^T K B_p(R + B_p^T K B_p)^{-1} B_p^T K A_p + C_p^T Q^* C_p \tag{3.19}$$

has the (unique) stabilizing solution K^*. *Define*

$$G = -(R + B_p^T K^* B_p)^{-1} B_p^T K^* A_p, \tag{3.20}$$

let X^* *denote the unique solution to the Lyapunov equation*

$$X = (A_p + B_pG)X(A_p + B_pG)^T + F(V + M_p\tilde{X}M_p^T)F^T, \tag{3.21}$$

and define $Y_i = C_i(\tilde{X} + X^*)C_i^T$ $(i = 1, 2, \ldots, m)$. *Then, if*

$$0 = (Y_i - \overline{Y}_i)Q_i^* \text{ and } Y_i \leq \overline{Y}_i \tag{3.22}$$

for all $i = 1, 2, \ldots, m$, *the dynamic controller*

$$\begin{aligned} x_c(k+1) &= (A_p + B_pG - FM_p)x_c(k) + Fz(k), \\ u(k) &= Gx_c(k) \end{aligned} \tag{3.23}$$

is an optimal solution to the OCC problem defined in (3.6).

Note that, as in the continuous-time case, the computation of $\tilde{X}$ and F are independent of the selection of the output-weighting matrix Q. Hence, we can apply the discrete-time OCC algorithm with state feedback to solve the discrete-time full-order output feedback OCC problem under the assumption that the optimal solutions are strictly proper. This requires that in the algorithm given in section 3.1, we replace D_p, W_p, and $\overline{Y}_i$ with F, $V + M_p\tilde{X}M_p^T$, and $\overline{Y}_i - C_i\tilde{X}C_i^T$, respectively.

4. An example. We consider the continuous-time OCC problem defined in (1.7) for the plant (1.1) with the following system matrices:

$$A_p = \begin{bmatrix} 0 & 1 & 0 \\ -1 & -0.1 & 1 \\ 0 & 0 & -10 \end{bmatrix}, \quad B_p = \begin{bmatrix} 0 \\ 0 \\ 1 \end{bmatrix}, \quad D_p = \begin{bmatrix} 0 \\ 0 \\ 1 \end{bmatrix}, \tag{4.1a}$$

$$M_p = \begin{bmatrix} 1 & 1 & 0 \end{bmatrix}, \tag{4.1b}$$

$$C_p = \begin{bmatrix} 1 & 0.5 & 0 \\ 0 & 0 & 0.5 \\ 1 & 1 & 0 \end{bmatrix}. \tag{4.1c}$$

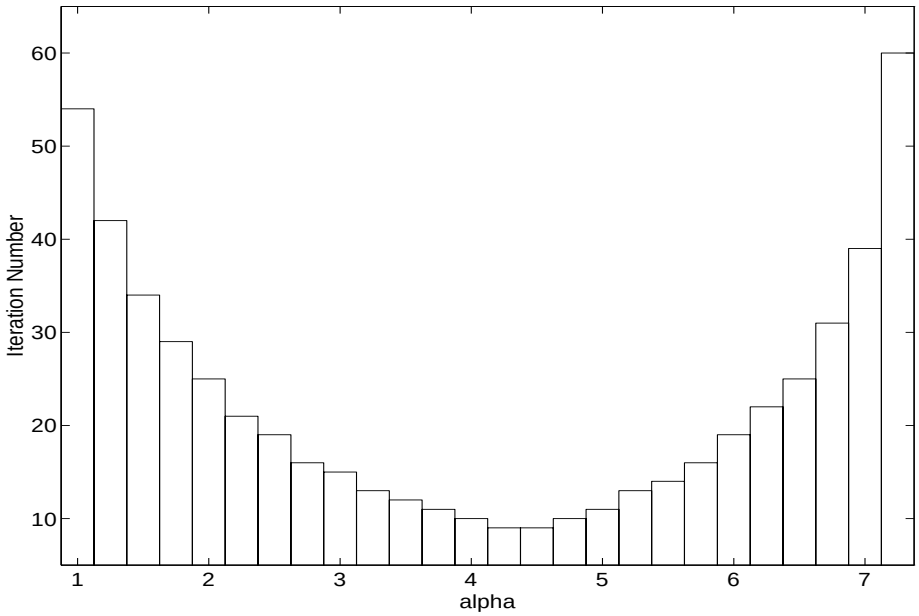


FIG. 4.1. *Iteration number versus* α.

Both the process noise w_p and the measurement noise v are scalar variables, while the performance variable y_p has three components. The weighting matrices required to define the OCC problem (1.7) are taken to be

$$W_p = 1, \quad V = 0.01, \quad \text{and} \quad R = 1. \tag{4.2}$$

Below, we consider two different OCC problems. These two problems differ in the grouping of the performance variables y_i used to define the constraints (1.6). For each problem, we consider both state feedback and dynamic output feedback.

4.1. Problem 1. Here, the OCC problem has the performance constraints

$$Y_1 \leq 0.035, \quad Y_2 \leq 0.050, \quad Y_3 \leq 0.050, \tag{4.3}$$

where Y_1, Y_2, and Y_3 denote the output covariance (1×1) matrices introduced in (1.6), corresponding to the first, second, and third performance variables, respectively. Note that this OCC problem can be also solved by the ellipsoid algorithms given in [1, 3, 12] or the quadratically convergent algorithms given in [16].

First, we consider the case of state feedback. We use the algorithm described in section 2.1 with the following parameters:

$$Q(0) = I_3, \quad \beta = 0.1, \quad \epsilon = 10^{-6}. \tag{4.4}$$

To assess the effect of the user-specified parameter α, we ran the the algorithm with $1.0 \leq \alpha \leq 7.25$. Figure 4.1 shows the number of iterations required to meet the stopping criteria of the algorithm versus α. Clearly, as α approaches 1 or 7.25, the iteration number increases. From Figure 4.1, it follows that there exists an optimal α which uses the least number of iterations. Finding such an optimal α in terms of the system and specification matrices remains an open problem.

TABLE 4.1
Solution to Problem 1 with $\alpha = 4.5$.

State feedback design					
Iteration	Constraints		Optimal cost		
number	Spec. ($\overline{Y}_i$)	Actual (Y_i)	J_{OCC}	Q_i	G^T
9	0.0350	0.0314	0.0234	0.0000	0.0237
	0.0500	0.0123		0.0000	−0.9522
	0.0500	0.0500		1.4268	−0.0948
Output feedback design					
Iteration	Constraints		Optimal cost		
number	Spec. ($\overline{Y}_i$)	Actual (Y_i)	J_{OCC}	Q_i	G^T
22	0.0350	0.0314	0.0340	0.0000	0.0193
	0.0500	0.0126		0.0000	−1.3839
	0.0500	0.0500		2.3765	−0.1374

Table 4.1 shows the results of running the algorithm with $\alpha = 4.5$. Both state and output feedback cases are computed. In the state feedback case, G denotes the state feedback gain. In the output feedback case, G denotes the controller output matrix; see (2.65) and (2.68). The controller input matrix F is precomputed according to (2.62). In this case we have

$$F = [\ 0.4412\ \ 0.7633\ \ 0.4796\ ]^T. \tag{4.5}$$

From Table 4.1, we can see that both controllers are feasible, since Y_i satisfies the bound $Y_i \leq \overline{Y}_i$. The only active constraint is the third one, i.e., $Y_3 = \overline{Y}_3$; hence, the corresponding output weight Q_3 is nonzero. As expected, the optimal cost J_{OCC} with output feedback is bigger than that with state feedback.

4.2. Problem 2. Now, the OCC problem has the performance constraints

$$Y_1 \leq 0.035, \qquad Y_2 \leq 0.050 \times I_2, \tag{4.6}$$

where Y_1 denotes the (1×1) output covariance matrix corresponding to the first performance output and Y_2 denotes the (2×2) output covariance matrix of the second and third performance outputs grouped together.

Table 4.2 shows the results of running the algorithm with $\alpha = 30$ for both state and output feedback cases. The other parameters required by the algorithm are those in (4.4). For the output feedback case the input gain matrix F of the controller given in (4.5).

From Table 4.2, we can see that both controllers are feasible. As expected, the optimal cost J_{OCC} with output feedback is bigger than that with state feedback. Also, note that the constraint on the second output group $Y_2 \leq 0.05 \times I_2$ is sufficient for the output covariance constraints of Problem 1 in (4.3), that is, $Y_2 \leq 0.05$ and $Y_3 \leq 0.05$. As expected, the costs of Problem 2 for both state and output feedback cases are bigger than those of Problem 1.

5. Conclusion. In this paper we have considered the so-called output covariance constraint (OCC) control problem. This is the problem of minimizing control effort subject to matrix inequality constraints on several closed-loop covariance matrices. Optimality conditions for characterizing a global solution are provided. In the state feedback case, these conditions comprise one algebraic Riccati equation, one Lyapunov equation, and a coupling condition. The unknown in this system of equations is a

TABLE 4.2
Solution to Problem 2 with $\alpha = 30$.

State feedback design					
Iteration number	Constraints		Optimal cost		
	Spec. ($\overline{Y}_i$)	Actual (Y_i)	J_{OCC}	Q_i	G^T
31	0.0300 $0.050\times I_2$	0.0313 $\begin{bmatrix} 0.0123 & 0.0014 \\ 0.0014 & 0.0499 \end{bmatrix}$	0.0235	0.0000 $\begin{bmatrix} 0.0019 & 0.0527 \\ 0.0527 & 1.4277 \end{bmatrix}$	0.0212 −0.9542 −0.0950
Output feedback design					
Iteration number	Constraints		Optimal cost		
	Spec. ($\overline{Y}_i$)	Actual (Y_i)	J_{OCC}	Q_i	G^T
65	0.0350 $0.050\times I_2$	0.0314 $\begin{bmatrix} 0.0126 & 0.0014 \\ 0.0014 & 0.0499 \end{bmatrix}$	0.0341	0.0000 $\begin{bmatrix} 0.0035 & 0.0919 \\ 0.0919 & 2.3809 \end{bmatrix}$	0.0149 −1.3878 −0.1379

matrix Q which may be interpreted as a matrix of Kuhn–Tucker multipliers. We have given an iterative algorithm to find such a matrix Q. Under the assumption that the optimality conditions admit a solution Q, we have shown that the iterative algorithm converges to one such solution, provided that the step size parameter α is properly chosen. Using the separation property of a closed-loop covariance matrix given in [12], we have shown how to modify the state feedback algorithm to solve the OCC problem with output feedback. Both discrete- and continuous-time problems have been solved. Finally, an example is presented to demonstrate the applicability of our results.

REFERENCES

[1] S. P. BOYD AND C. H. BARRATT, *Linear Controller Design: Limits of Performance*, Prentice-Hall, Englewood Cliffs, NJ, 1991.

[2] D. F. DELCHAMPS, *A note on the analyticity of the Riccati metric*, in Algebraic and Geometric Methods in Linear Systems Theory, Lectures in Applied Mathematics, Vol. 18, AMS, Providence, RI, 1980.

[3] A. M. EUDARIC, *Ellipsoid Method for Multiobjective Control with Quadratic Performance Measures*, Master thesis, Purdue University, 1992.

[4] C. HSIEH, R. E. SKELTON, AND F. M. DAMRA, *Minimum energy controllers with inequality constraints on output variances*, Optimal Control Appl. Methods, 10 (1989), pp. 347–366.

[5] P. P. KHARGONEKAR AND M. A. ROTEA, *Multiple objective optimal control of linear systems: The quadratic norm case*, IEEE Trans. Automat. Control, 36 (1991) pp. 14–24.

[6] N. T. KOUSSOULAS AND C. T. LEONDES, *The multiple linear quadratic Gaussian problem*, Internat. J. Control, 43 (1986), pp. 337–349.

[7] V. KUCERA, *A contribution to matrix quadratic equations*, IEEE Trans. Automat. Control, 17 (1972), pp. 344–347.

[8] V. KUCERA, *The discrete Riccati equation of optimal control*, Kybernetika, 8 (1972), pp. 430–447.

[9] P. M. MAKILA, *On multiple criteria stationary linear quadratic control*, IEEE Trans. Automat. Control, 34 (1989), pp. 1311–1313.

[10] S. MEERKOV AND T. RUNOLFSSON, *Output residence time control*, IEEE Trans. Automat. Control, 34 (1989), pp. 1171–1176.

[11] J. M. ORTEGA AND W. C. RHEINBOLDT, *Iterative Solution of Nonlinear Equations in Several Variables*, Academic Press, New York, 1970.

[12] M. A. ROTEA, *The generalized H_2 control problem*, Automatica J. IFAC, 29 (1993), pp. 373–385.

[13] R. E. SKELTON AND M. DELORENZO, *Space structure control design by variance assignment*, J. Guidance Control Dynamics, 8 (1985), pp. 454–462.

[14] H. T. Toivonen, *Variance constrained self-tuning control*, Automatica J. IFAC, 19 (1983), pp. 415–418.
[15] H. T. Toivonen, *A multiobjective linear quadratic Gaussian control problem*, IEEE Trans. Automat. Control, 29 (1984), pp. 279–280.
[16] H. T. Toivonen and P. M. Makila, *Computer-aided design procedure for multiobjective LQG control problems*, Internat. J. Control, 49 (1989), pp. 655–666.
[17] D. A. Wilson, *Convolution and Hankel operator norms for linear systems*, IEEE Trans. Automat. Control, 34 (1989), pp. 94–97.
[18] G. Zhu, M. Corless, and R. Skelton, *Robustness properties of covariance controllers*, in Proce. of Allerton Conf., Monticello, IL, September 1989.
[19] G. Zhu and R. E. Skelton, *Mixed L_2 and L_∞ problems by weight selection in quadratic optimal control*, Internat. J. Control, 53 (1991), pp. 1161–1176.

SIAM J. CONTROL OPTIM.
Vol. 35, No. 2, pp. 363–383, March 1997

001

CENTRALIZED AND DECENTRALIZED SUPERVISORY CONTROL OF NONDETERMINISTIC SYSTEMS UNDER PARTIAL OBSERVATION*

RATNESH KUMAR† AND MARK A. SHAYMAN‡

Abstract. In this paper we extend our earlier work on supervisory control of nondeterministic systems using prioritized synchronization as the mechanism of control and trajectory model as the modeling formalism by considering design of supervisors under partial observation. We introduce the notion of *observation-compatible* systems and show that prioritized synchronous composition (PSC) of observation-compatible systems can be used as a mechanism of control of nondeterministic systems under partial observation in presence of driven events. Necessary and sufficient conditions that depend on the trajectory model as opposed to the language model of the plant are obtained for the existence of *centralized* as well as *decentralized* supervision. Our work on centralized control shows that the results of the traditional supervisory control can be "extended" to the above setting, provided that the supervisor is deterministic and the observation mask is projection type. On the other hand, our work on decentralized control is based on a new relation between controllability, observability, co-observability, and PSC that we derive in this paper.

Key words. discrete event systems, supervisory control, partial observation, nondeterministic automata, driven events, prioritized synchronization, trajectory models, controllability, observability, co-observability

AMS subject classifications. 68Q75, 93B25, 93C83

PII. S0363012994272903

1. Introduction.

Discrete event systems (DESs) are systems which involve quantities that take a discrete set of values and which evolve according to occurrence of certain discrete qualitative changes, called *events*, such as arrival of a customer in a queue, termination of an algorithm in a computer program, loss of a message packet in a communication network, breakdown of a machine in a manufacturing system, and the like. The theory of supervisory control of DESs was introduced by Ramadge and Wonham [26, 27] for designing controllers so that the controlled system satisfies certain desired *qualitative* constraints, such as that a buffer in a manufacturing system should never overflow, a message sequence in a communication network must be received in the same order as it was transmitted, and so on.

Such qualitative behavior of a *deterministic*[1] DES can be described by the set of all possible event *traces*, called a *language* model, that the system can execute starting from its initial state. However, due to partial observation and/or unmodeled dynamics, it is too restrictive to require determinism of a system. If a DES is nondeterministic, then its language model may not adequately describe its qualita-

*Received by the editors August 17, 1994; accepted for publication (in revised form) December 8, 1995. This research was supported in part by the Center for Robotics and Manufacturing, University of Kentucky, and National Science Foundation grants CDR-8803012, EEC-94-02384, ECS-9312587, and ECS-9409712. A preliminary version of this paper appeared as *Supervisory control of nondeterministic systems under partial observation*, in Proc. 1994 IEEE Conf. Decision and Control, Orlando, FL, December 1994, pp. 3649–3654.

http://www.siam.org/journals/sicon/35-2/27290.html

†Department of Electrical Engineering, University of Kentucky, Lexington, KY 40506-0046 (kumar@engr.uky.edu).

‡Department of Electrical Engineering and Institute of Systems Research, University of Maryland, College Park, MD 20742 (shayman@src.umd.edu).

[1] A DES is said to be deterministic if, given the current state and an event that occurs in that state, the next state is uniquely determined.

tive behavior, and more detailed models are needed. Several models such as *failures model* [10], *refusal-trace model* [25], *ready-trace model* [1], *bisimulation model* [23, 24], etc. have been proposed in the literature for representing qualitative behavior of nondeterministic DESs. A nice comparative study of such modeling formalisms can be found in [2, 31]. As a designer, it is desirable to choose the *least detailed* modeling formalism that is adequate for the design task at hand. As is argued below, this is the reason for us to choose the *trajectory model* proposed by Heymann [8], also known as *refusal-trace model*, for representing nondeterministic DESs.

Most of the prior work on supervisory control of DESs, such as [26, 16, 4], essentially use *strict synchronous composition* (SSC) of *plant* DES and *supervisor* DES as the mechanism of control. In SSC of systems, it is required that the common events must occur synchronously. This is restrictive, because due to nondeterminism the plant state is not uniquely known after the execution of a certain observed trace, and the set of executable events in each such state may differ. If we require strict synchronization, then the supervisor is restricted to enable those events that are executable in each of those states, which imposes a severe restriction on the supervisor. Moreover, there is no a priori reason for a supervisor to synchronously execute all the *uncontrollable* events that the plant can execute. Similarly, it is restrictive to require that the plant synchronously executes the so-called *forcible* [7] or *command* [4] or *driven* [8] events that are initiated by the supervisor. The motivating example in [30, section 2, Example 5] describes a nondeterministic plant that can be controlled only when the requirement of strict synchronization is relaxed.

In this paper we study the control of qualitative behavior of nondeterministic DESs using *prioritized synchronous composition* (PSC) as the mechanism of control. PSC was originally proposed by Heymann [8, 9] and was later applied to supervisory control in the deterministic setting by Balemi [3] and in the nondeterministic setting by Shayman and Kumar [30]. PSC is a generalization of the SSC. The parallel operator considered by Inan [12, 13], an extension of the parallel operator defined in [14, 15], can be viewed as a generalization of PSC when applied to the so-called *improper* systems. However, while studying supervisory control only proper systems are considered; consequently the resulting operation is that of strict synchronization.

In PSC each system is associated with a certain priority set of events, and for an event to occur in the composition of a pair of systems operating in prioritized synchrony, each system having the priority over the event must participate. So if an event belongs to the common priority set, then it occurs synchronously. On the other hand, if a certain event belongs to the priority set of a single system, then it can occur asynchronously without the participation of the second system. However, the second system will participate whenever possible; such synchronization is called *broadcast synchronization*. Thus PSC does not impose the unnecessarily restrictive requirement of SSC that common events must always occur in synchrony. For supervisory control, the priority set of a plant consists of the uncontrollable and the controllable events, while the priority set of a supervisor consists of the controllable and the driven events. Since controllable events are in the priority sets of plant as well as supervisor, they always occur in synchrony in the controlled system, whereas the uncontrollable and the driven events may occur asynchronously.

Heymann showed via an example [8, Example 7] that if PSC is *admitted* as a mechanism of interconnection, then a modeling formalism which is more detailed than the failures model (and consequently, more detailed than the language model) is needed to adequately describe the behavior of nondeterministic DESs. For this reason,

Heymann proposed the modeling formalism called *trajectory model*. A trajectory model consists of *generated* and *recognized* trajectories, also called *refusal-traces*, of a system. A refusal-trace is a sequence of alternating refusal sets and events, where a refusal set consists of those events that the system "refuses" to execute when offered at a certain execution point. The trajectory model is quite similar to the refusal-testing model of Phillips [25] but differs in its treatment of *silent* or epsilon transitions.

In our previous work [30, 19] we showed that the trajectory model can be used for adequately describing behaviors of nondeterministic DESs that may be interconnected using PSC, and also that the operation of PSC is associative. Since an event that belongs to the priority set of a single system can occur asynchronously if we *augment* the other system by adding *self-loops* on such events, then the operation of PSC can be reduced to the operation of SSC, provided that the priority sets of the two systems exhaust the entire event set. Under this condition, we proved in [30, 19] that the PSC of a pair of systems is equivalent to SSC of appropriately augmented systems. In particular, if the plant is augmented with driven events and the supervisor is augmented with uncontrollable events, then the PSC of plant and supervisor is equivalent to SSC of augmented plant and augmented supervisor. Using these results we obtained necessary and sufficient conditions for the existence of a supervisor so that the language of the controlled plant equals a desired language.

In this paper we extend our earlier work on supervisory control of nondeterministic systems using prioritized synchronization as the mechanism of control and trajectory model as the modeling formalism by considering design of supervisors under partial observation. Partial observation in the setting of supervisory control arises due to lack of sufficient number of sensors. As in the work of Lin and Wonham [21], we use a projection function, also called an *observation mask*, to represent such partial observation. A supervisor under partial observation must take identical control action following indistinguishable traces. We call this property of a supervisor *observation-compatibility*, which captures physically realizable supervisors. Such supervisors make control decisions based on *only* the observed event trace of the system and do not require any "special" internal knowledge of the system.

We define the notion of observation-compatibility of a trajectory model and prove that this property is preserved under augmentation whenever the system is deterministic. Using this result we obtain a necessary and sufficient condition for the existence of an observation-compatible supervisor so that the language of the plant operating in prioritized synchrony with the supervisor equals the desired one. This result is then applied to obtain a supervisor which achieves mutually exclusive usage of a shared channel in a communication system. We also obtain conditions for the existence of *nonblocking* supervisors [27, 5].

Finally, we study the problem of decentralized supervision [29, 20, 22, 6, 32]. Decentralized supervision is inevitable when the plant is physically distributed for example as in communication networks and manufacturing systems. A supervisor is installed at each location of the "subplant." In such a situation, a supervisor is able to control a certain set of events, called *local* events, and is able to observe a partial set of events. The problem of decentralized supervision requires design of supervisors that are observation-compatible with respect to their own observation function, and control events in their own priority sets. This problem is naturally formulated in our framework. We show that the condition of controllability together with the condition of *co-observability* is necessary and sufficient for decentralized supervision. Our constructive proof is novel and is based on a nice relationship between control-

lability, observability, co-observability, and PSC that we derive in this paper. These conditions, however, are significantly different from the standard ones [21, 29], as they depend on the trajectory model (rather than language model) of the plant.

The remainder of the paper is organized as follows: In section 2 we introduce the relevant notation. In section 3, we define the notion of observation-compatibility and study some of its properties. In section 4 we study the supervisory control problem under partial observation in the proposed framework and apply it to achievement of mutually exclusive usage of a shared communication channel in a communication system. In section 5 we study the problem of decentralized supervision. Finally, section 6 concludes the work presented here.

2. Notation and preliminaries. Given a finite event set Σ, Σ^* is used to denote the collection of all *traces*, i.e., finite sequences of events, including the zero-length sequence, denoted ϵ. A subset of Σ^* is called a language. Symbols H, K, etc. are used to denote languages. For a language $K \subseteq \Sigma^*$, the notation $\mathrm{pr}(K) \subseteq \Sigma^*$, called the *prefix-closure* of K, is the set of all prefixes of traces from K. K is said to be prefix-closed if $K = \mathrm{pr}(K)$.

The set $2^\Sigma(\Sigma \times 2^\Sigma)^*$ is used to denote the collection of all *refusal-traces*, i.e., finite sequences of alternating *refusals* and events [9, 30] of the type

$$\Sigma_0(\sigma_1, \Sigma_1) \dots (\sigma_n, \Sigma_n),$$

where $n \in \mathcal{N}$. The sequence $\sigma_1 \dots \sigma_n \in \Sigma^*$ is the trace, and for each $i \leq n$, $\Sigma_i \subseteq \Sigma$ is a set of events refused (if offered) at the indicated point. Symbols P, Q, R, S, etc., are used to denote sets of refusal-traces. Refusal-traces are also referred to as *trajectories*.

Given $e \in 2^\Sigma(\Sigma \times 2^\Sigma)^*$, we use $|e|$ to denote the length of e, and for each $k \leq |e|$, $\Sigma_k(e) \subseteq \Sigma$ is used to denote the kth refusal in e and $\sigma_k(e) \in \Sigma$ is used to denote the kth event in e, i.e.,

$$e = \Sigma_0(e)(\sigma_1(e), \Sigma_1(e)) \dots (\sigma_k(e), \Sigma_k(e)) \dots (\sigma_{|e|}(e), \Sigma_{|e|}(e)).$$

The *trace* of e, denoted $\mathrm{tr}(e) \in \Sigma^*$, is defined as $\mathrm{tr}(e) := \sigma_1(e) \dots \sigma_{|e|}(e)$. Given a set of refusal-traces $P \subseteq 2^\Sigma(\Sigma \times 2^\Sigma)^*$, we use $L(P) := \mathrm{tr}(P)$ to denote its set of traces.

If $f \in 2^\Sigma(\Sigma \times 2^\Sigma)^*$ is another refusal-trace such that $|f| \leq |e|$ and for each $k \leq |f|$, $\Sigma_k(f) = \Sigma_k(e)$ and $\sigma_k(f) = \sigma_k(e)$, then f is said to be a prefix of e, denoted by $f \leq e$. For each $k \leq |e|$, the notation $e^k \leq e$ is used to denote the prefix of length k of e. The prefix-closure of e, denoted $\mathrm{pr}(e) \subseteq 2^\Sigma(\Sigma \times 2^\Sigma)^*$, is the set of all prefixes of e. If $f \in 2^\Sigma(\Sigma \times 2^\Sigma)^*$ is such that $|f| = |e|$ and for each $k \leq |f|$, $\Sigma_k(f) \subseteq \Sigma_k(e)$ and $\sigma_k(f) = \sigma_k(e)$, then f is said to be *dominated* by e, denoted by $f \sqsubseteq e$. The *dominance-closure* of e, denoted $\mathrm{dom}(e) \subseteq 2^\Sigma(\Sigma \times 2^\Sigma)^*$, is the set of all refusal-traces dominated by e.

Symbols $\mathcal{P}, \mathcal{Q}, \mathcal{R}$, etc., are used to denote nondeterministic state machines (NSMs) (with ϵ-moves). Let the 5-tuple

$$\mathcal{P} := (X_\mathcal{P}, \Sigma, \delta_\mathcal{P}, x^0_\mathcal{P}, X^m_\mathcal{P})$$

represent a DES modeled as an NSM, where $X_\mathcal{P}$ is the state set, Σ is the finite event set, $\delta_\mathcal{P} : X_\mathcal{P} \times (\Sigma \cup \{\epsilon\}) \to 2^{X_\mathcal{P}}$ denotes the nondeterministic transition function,[2] $x^0_\mathcal{P} \in X_\mathcal{P}$ is the initial state, and $X^m_\mathcal{P} \subseteq X_\mathcal{P}$ is the set of accepting or marked states.

[2] ϵ represents both an *internal* or *unobservable* event and an *internal* or *nondeterministic* choice [10, 23].

A triple $(x_1, \sigma, x_2) \in X_{\mathcal{P}} \times (\Sigma \cup \{\epsilon\}) \times X_{\mathcal{P}}$ is said to be a transition if $x_2 \in \delta_{\mathcal{P}}(x_1, \sigma)$. A transition (x_1, ϵ, x_2) is referred to as a *silent* or *hidden* transition. We assume that the plant cannot undergo an unbounded sequence of silent transitions.

The ϵ-*closure* of $x \in X_{\mathcal{P}}$, denoted $\epsilon^*_{\mathcal{P}}(x) \subseteq X_{\mathcal{P}}$, is defined inductively as

$$x \in \epsilon^*_{\mathcal{P}}(x) \text{ and } x' \in \epsilon^*_{\mathcal{P}}(x) \Rightarrow \delta_{\mathcal{P}}(x', \epsilon) \subseteq \epsilon^*_{\mathcal{P}}(x),$$

and the set of *refusal events* at $x \in X_{\mathcal{P}}$, denoted $\Re_{\mathcal{P}}(x) \subseteq \Sigma$, is defined as

$$\Re_{\mathcal{P}}(x) := \{\sigma \in \Sigma \mid \delta_{\mathcal{P}}(x', \sigma) = \emptyset, \forall x' \in \epsilon^*_{\mathcal{P}}(x)\}.$$

In other words, given $x \in X_{\mathcal{P}}$, $\epsilon^*_{\mathcal{P}}(x)$ is the set of states that can be reached from x on zero or more ϵ-moves, and $\Re_{\mathcal{P}}(x)$ is the set of events that are undefined at each state in the ϵ-closure of x. Using the definitions of the ϵ-closure and refusal maps, the transition function $\delta_{\mathcal{P}} : X_{\mathcal{P}} \times (\Sigma \cup \{\epsilon\}) \to 2^{X_{\mathcal{P}}}$ is extended (i) to the set of *traces*, denoted $\delta^*_{\mathcal{P}} : X_{\mathcal{P}} \times \Sigma^* \to 2^{X_{\mathcal{P}}}$, which is defined in the usual way [11], and (ii) to the set of *refusal-traces*, denoted $\delta^T_{\mathcal{P}} : X \times (2^{\Sigma}(\Sigma \times 2^{\Sigma})^*) \to 2^{X_{\mathcal{P}}}$, which is defined inductively as

$$\forall x \in X_{\mathcal{P}} : \begin{cases} \forall \Sigma' \subseteq \Sigma : \delta^T_{\mathcal{P}}(x, \Sigma') := \{x' \in \epsilon^*_{\mathcal{P}}(x) \mid \Sigma' \subseteq \Re_{\mathcal{P}}(x')\}, \\ \forall e \in 2^{\Sigma}(\Sigma \times 2^{\Sigma})^*,\ \sigma \in \Sigma, \Sigma' \subseteq \Sigma : \\ \delta^T_{\mathcal{P}}(x, e(\sigma, \Sigma')) := \{x' \in \epsilon^*_{\mathcal{P}}(\delta_{\mathcal{P}}(\delta^T_{\mathcal{P}}(x, e), \sigma)) \mid \Sigma' \subseteq \Re_{\mathcal{P}}(x')\}. \end{cases}$$

These maps are then used to obtain the language models and the trajectory models of $\mathcal{P}$ as follows:

$$L(\mathcal{P}) := \{s \in \Sigma^* \mid \delta^*_{\mathcal{P}}(x^0_{\mathcal{P}}, s) \neq \emptyset\},\ \ L^m(\mathcal{P}) := \{s \in L(\mathcal{P}) \mid \delta^*_{\mathcal{P}}(x^0_{\mathcal{P}}, s) \cap X^m_{\mathcal{P}} \neq \emptyset\},$$

$$T(\mathcal{P}) := \{e \in 2^{\Sigma}(\Sigma \times 2^{\Sigma})^* \mid \delta^T_{\mathcal{P}}(x^0_{\mathcal{P}}, e) \neq \emptyset\}, T^m(\mathcal{P}) := \{e \in T(\mathcal{P}) \mid \delta^T_{\mathcal{P}}(x^0_{\mathcal{P}}, e) \cap X^m_{\mathcal{P}} \neq \emptyset\}.$$

$L(\mathcal{P}), L^m(\mathcal{P}), T(\mathcal{P})$, and $T^m(\mathcal{P})$ are called the *generated language, recognized language, generated trajectory set, and recognized trajectory set,* respectively, of $\mathcal{P}$. The pairs $(L^m(\mathcal{P}), L(\mathcal{P}))$ and $(T^m(\mathcal{P}), T(\mathcal{P}))$ are called the language model and the trajectory model, respectively, of $\mathcal{P}$. Two language models, (K^m_1, K_1) and (K^m_2, K_2), are said to be equal, written $(K^m_1, K_1) = (K^m_2, K_2)$, if $K^m_1 = K^m_2$ and $K_1 = K_2$; equality of two trajectory models is defined analogously.

Given a trajectory model, the trace map can be used to obtain the associated language model. On the other hand, given a language model (K^m, K), the *trajectory* map, $\mathrm{tr}j_K : K \to 2^{\Sigma}(\Sigma \times 2^{\Sigma})^*$ can be used to obtain the *deterministic* trajectory model[3] having the language model (K^m, K), which is defined as follows:

$$\mathrm{tr}j_K(s) := \Sigma_0(s)(\sigma_1(s), \Sigma_1(s)) \ldots (\sigma_{|s|}(s), \Sigma_{|s|}(s)) \in 2^{\Sigma}(\Sigma \times 2^{\Sigma})^*, \text{ where}$$

$$\Sigma_k(s) := \{\sigma \in \Sigma \mid s^k\sigma \notin K\} \quad \forall k \leq |s|.$$

Define $(\det^m(K^m, K),\ \det(K)) := (\mathrm{dom}(\mathrm{tr}j_K(K^m)),\ \mathrm{dom}(\mathrm{tr}j_K(K)))$. Then it is shown in [19, Proposition 1] that it is the unique deterministic trajectory model that has the language model (K^m, K).

[3] A trajectory model (P^m, P) is said to be deterministic if there exists a deterministic state machine $\mathcal{P}$ such that $(T^m(\mathcal{P}), T(P)) = (P^m, P)$.

In [8, 9, 30, 19] PSC of systems is used as the mechanism of control. In this setting, associated with each system is a priority set of events, which endows the system with the ability to prevent the occurrence of events belonging to its priority set; a system must participate in the execution of an event belonging to its priority set for that event to occur in the PSC with other system(s). Letting $\mathcal{P}\ {}_A\|_B\ \mathcal{Q}$ denote the PSC of NSMs $\mathcal{P}$ and $\mathcal{Q}$ with priority sets $A, B \subseteq \Sigma$, respectively, and $T^m(\mathcal{P})\ {}_A\|_B\ T^m(\mathcal{Q})$, $T(\mathcal{P})\ {}_A\|_B\ T(\mathcal{Q})$ denote the PSC of corresponding trajectory models, it was proven in [30, Theorem 2] and [19, Theorem 2] that

$$T^m(\mathcal{P})\ {}_A\|_B\ T^m(\mathcal{Q}) = T^m(\mathcal{P}\ {}_A\|_B\ \mathcal{Q}); \quad T(\mathcal{P})\ {}_A\|_B\ T(\mathcal{Q}) = T(\mathcal{P}\ {}_A\|_B\ \mathcal{Q}).$$

Various properties of PSC of trajectory models were studied in [30, 19]. In particular, associativity of PSC was proven [19, Proposition 2, Corollary 6], a language intersection result for the case when $A = B = \Sigma$ was obtained [19, Corollary 4, Corollary 5], and the notion of *augmentation* and its properties were studied.

We recall from [30, 19] that the *augmentation* of an NSM $\mathcal{P}$ by an event set $D \subset \Sigma$ is the NSM $\mathcal{P}^D := \mathcal{P}\ {}_\emptyset\|_\emptyset\ \mathcal{D}$, where $\mathcal{D}$ denotes the deterministic state machine with one state, which is marked, and has self-loops labeled by every event in D. Thus the augmented NSM $\mathcal{P}^D$ can also be obtained by adding self-loops on each state of $\mathcal{P}$ on those events in D that are refused at that state, i.e., $\mathcal{P}^D := (X_\mathcal{P}, \Sigma, \delta_{\mathcal{P}^D}, x^0_\mathcal{P}, X_m)$, where the transition function is defined as

$$\forall x \in X_\mathcal{P}, \sigma \in \Sigma : \delta_{\mathcal{P}^D}(x,\sigma) := \begin{cases} \{x\} & \text{if } \sigma \in D \cap \Re_\mathcal{P}(x), \\ \delta_\mathcal{P}(x,\sigma) & \text{otherwise.} \end{cases}$$

Refer to Example 1 for illustration. Since the trajectory model of $\mathcal{D}$ is $(\det(D^*), \det(D^*))$, the augmented trajectory model is given by

$$((T^m(\mathcal{P}))^D, (T(\mathcal{P}))^D) := (T^m(\mathcal{P}^D), T(\mathcal{P}^D)) = (T^m(\mathcal{P})\ {}_\emptyset\|_\emptyset\ \det(D^*), T(\mathcal{P})\ {}_\emptyset\|_\emptyset\ \det(D^*)).$$

It was shown in [30, Proposition 4] and [19, Proposition 3] that whenever the priority sets of a given pair of systems exhaust the entire event set, then the operation of PSC can be reduced to that of SSC of appropriately augmented systems. Specifically, given a pair of trajectory models (P^m, P) and (Q^m, Q) with priority sets $A, B \subseteq \Sigma$, respectively, if $A \cup B = \Sigma$, then

$$P^m\ {}_A\|_B\ Q^m = (P^m)^{\Sigma - A}\ {}_\Sigma\|_\Sigma\ (Q^m)^{\Sigma - B}; \quad P\ {}_A\|_B\ Q = P^{\Sigma - A}\ {}_\Sigma\|_\Sigma\ Q^{\Sigma - B}.$$

Consequently we have the following identities:

$$L(P^m\ {}_A\|_B\ Q^m) = L((P^m)^{\Sigma-A}) \cap L((Q^m)^{\Sigma-B}); \quad L(P\ {}_A\|_B\ Q) = L(P^{\Sigma-A}) \cap L(Q^{\Sigma-B}).$$

Thus the technique of augmentation is useful in studying the behavior of a pair of systems operating in prioritized synchrony if their priority sets jointly exhaust the entire event set. In particular, we can apply the technique of augmentation in supervisory control, as the event set Σ can be written as the union of the priority set of plant, which is the set of uncontrollable and controllable events, and the priority set of supervisor, which is the set of controllable and driven events.

3. Observation-compatible systems. In many control designs, it is not possible to completely observe the behavior of the uncontrolled plant due to lack of sufficient number of sensors. Thus, certain events executed by the uncontrolled plant

may be *unobservable*. In the setting of supervisory control, an observation mask—a projection map defined from the set of events to the set of observable events—is used to describe such partial observation. In such situations it is natural to require that the control actions taken by a supervisor following indistinguishable traces be identical. We call this property of a supervisor *observation-compatibility*. In this section, we formally define the notion of observation-compatibility of the trajectory model of a nondeterministic DES and study some of its properties.

Let $\Sigma^o \subseteq \Sigma$ be the set of *observable* events, i.e., the events that can be sensed by a supervisor. A projection function $M : \Sigma \to \Sigma^o \cup \{\epsilon\}$, called an *observation mask* [21, 6], is used to represent such a partial observation; it is defined as

$$\forall \sigma \in \Sigma : M(\sigma) := \begin{cases} \sigma & \text{if } \sigma \in \Sigma^o, \\ \epsilon & \text{otherwise.} \end{cases}$$

Note that we assume that the observation mask is a projection function.

Recall from [26] that a language $K \subseteq \Sigma^*$ is said to be controllable with respect to a given prefix-closed language H and the set of uncontrollable events $\Sigma - B$, called $(H, \Sigma - B)$-*controllable*, if

$$\mathrm{pr}(K)(\Sigma - B) \cap H \subseteq \mathrm{pr}(K),$$

i.e., if the extension of a certain prefix of K by an uncontrollable event results in a trace of H, then this extended trace should also be a prefix of K. Also, recall from [21] that K is said to be observable with respect to H and a given observation mask $M(\cdot)$, called (H, M)-*observable*, if

$$\forall s, t \in \mathrm{pr}(K), \sigma \in \Sigma : \quad M(s) = M(t), \quad s\sigma \in \mathrm{pr}(K), \quad t\sigma \in H \Rightarrow t\sigma \in \mathrm{pr}(K).$$

In other words, K is said to be (H, M)-observable if given an indistinguishable pair of traces in $\mathrm{pr}(K)$, the pair of traces resulting from appending a common event to the given pair has identical membership in $\mathrm{pr}(K)$ whenever they have identical membership in H. It was shown in [21] that the observability of prefix-closed languages is preserved under intersection so that the *infimal prefix-closed and observable superlanguage* of a given language exists. Using the above notion of observability we next define the concept of observation-compatibility.

DEFINITION 1. *Given a trajectory model (S^m, S) and an observation mask $M(\cdot)$, (S^m, S) is said to be* observation compatible with respect to $M(\cdot)$ *or simply* M-compatible *if*

$$\forall s, t \in L(S), \sigma \in \Sigma : \quad M(s) = M(t), \quad s\sigma \in L(S) \Rightarrow t\sigma \in L(S).$$

An NSM is said to be M-compatible if its associated trajectory model is M-compatible. Thus a trajectory model is M-compatible if and only if its generated language is (Σ^*, M)-observable. Note that the property of *observation-compatibility* captures physically realizable supervisors. Such supervisors make control decisions based on *only* the observed event trace of the system and do not require any "special" internal knowledge of the system. Next we show that M-compatibility of a *deterministic* trajectory model is preserved under augmentation. We first need to establish two lemmas.

DEFINITION 2. *Given a nonempty prefix-closed language K, the* projection of Σ^* onto K *is defined inductively by*

$$\pi_K(\epsilon) := \epsilon; \quad \forall s \in \Sigma^*, \sigma \in \Sigma : \pi_K(s\sigma) := \begin{cases} \pi_K(s)\sigma & \text{if } \pi_K(s)\sigma \in K, \\ \pi_K(s) & \text{otherwise.} \end{cases}$$

When the choice of K is clear, we use the abbreviated notation s' for $\pi_K(s)$.

The first lemma asserts that the state reached by the execution of a certain trace in an augmented deterministic state machine is the same as that reached in the unaugmented state machine by the execution of the trace projected onto its language.

LEMMA 1. *Let $\mathcal{P} := (X_{\mathcal{P}}, \Sigma, \delta_{\mathcal{P}}, x^0_{\mathcal{P}}, X^m_{\mathcal{P}})$ be a deterministic state machine and $D \subseteq \Sigma$. Then for each $s \in L(\mathcal{P}^D)$, $\delta^*_{\mathcal{P}^D}(x^0_{\mathcal{P}}, s) = \delta^*_{\mathcal{P}}(x^0_{\mathcal{P}}, \pi_{L(\mathcal{P})}(s))$.*

Proof. We use induction on length of s for proving the assertion. For notational simplicity, define $\pi_{L(\mathcal{P})}(s) := s'$. If $|s| = 0$, then $s = s' = \epsilon$. Hence $\delta^*_{\mathcal{P}^D}(x^0_{\mathcal{P}}, s) = \delta^*_{\mathcal{P}}(x^0_{\mathcal{P}}, s') = x^0_{\mathcal{P}}$, since $\mathcal{P}$, and thus $\mathcal{P}^D$, are deterministic. Thus the base step trivially holds. In order to prove the induction step, suppose $s = \bar{s}\sigma$, where $\sigma \in \Sigma$. Define $\bar{s}' := \pi_{L(\mathcal{P})}(\bar{s})$. Then it follows from induction hypothesis that $\delta^*_{\mathcal{P}^D}(x^0_{\mathcal{P}}, \bar{s}) = \delta^*_{\mathcal{P}}(x^0_{\mathcal{P}}, \bar{s}') := x_{\bar{s}}$. If $\sigma \notin \Re_{\mathcal{P}}(x_{\bar{s}})$, then $\delta^*_{\mathcal{P}^D}(x^0_{\mathcal{P}}, s) = \delta^*_{\mathcal{P}}(x_{\bar{s}}, \sigma) = \delta^*_{\mathcal{P}}(x^0_{\mathcal{P}}, s')$. On the other hand, if $\sigma \in D \cap \Re_{\mathcal{P}}(x_{\bar{s}})$, then $s' = \bar{s}'$, and $\delta^*_{\mathcal{P}^D}(x^0_{\mathcal{P}}, s) = x_{\bar{s}}$, so $\delta^*_{\mathcal{P}}(x^0_{\mathcal{P}}, s') = \delta^*_{\mathcal{P}}(x^0_{\mathcal{P}}, \bar{s}') = x_{\bar{s}} = \delta^*_{\mathcal{P}^D}(x^0_{\mathcal{P}}, s)$. This proves the induction step and completes the proof. ∎

The next lemma asserts that if a certain language is (Σ^*, M)-observable, then the indistinguishability of a pair of traces implies indistinguishability of their projections onto the language.

LEMMA 2. *Consider an observation mask $M(\cdot)$, and a nonempty prefix-closed language $K \subseteq \Sigma^*$. If K is (Σ^*, M)-observable, then*

$$\forall s, t \in \Sigma^* : M(s) = M(t) \Rightarrow M(\pi_K(s)) = M(\pi_K(t)).$$

Proof. For notational simplicity, define $s' := \pi_K(s)$ and $t' := \pi_K(t)$. We prove the assertion by induction on $|s| + |t|$. For the base step, if $|s| = 0$ or $|t| = 0$, then $M(s) = M(t) = \epsilon$, so $M(s') = M(t') = \epsilon$. For the induction step, consider $s = \bar{s}\sigma_s$ and $t = \bar{t}\sigma_t$ with $\bar{s}, \bar{t} \in \Sigma^*$ and $\sigma_s, \sigma_t \in \Sigma$. Define $\bar{s}' := \pi_K(\bar{s})$ and $\bar{t}' := \pi_K(\bar{t})$. We have three possibilities: (i) $M(\sigma_s) = \epsilon$, which implies that $M(\bar{s}) = M(t)$. Then, $M(s') = M(\bar{s}') = M(t')$, where the first equality follows trivially from the unobservability of σ_s and the second equality follows by induction hypothesis. (ii) $M(\sigma_t) = \epsilon$. Then it follows from symmetry and case (i) above that $M(s') = M(t')$. (iii) $M(\sigma_s) \neq \epsilon$, $M(\sigma_t) \neq \epsilon$, which implies that $\sigma_s = \sigma_t := \sigma$ and $M(\bar{s}) = M(\bar{t})$. By the induction hypothesis, $M(\bar{s}') = M(\bar{t}')$. Since K is (Σ^*, M)-observable, either $\bar{s}'\sigma, \bar{t}'\sigma \in K$ or $\bar{s}'\sigma, \bar{t}'\sigma \notin K$. In the first case, $M(s') = M(\bar{s}'\sigma) = M(\bar{s}')\sigma = M(\bar{t}')\sigma = M(\bar{t}'\sigma) = M(t')$. In the second case, $M(s') = M(\bar{s}') = M(\bar{t}') = M(t')$. ∎

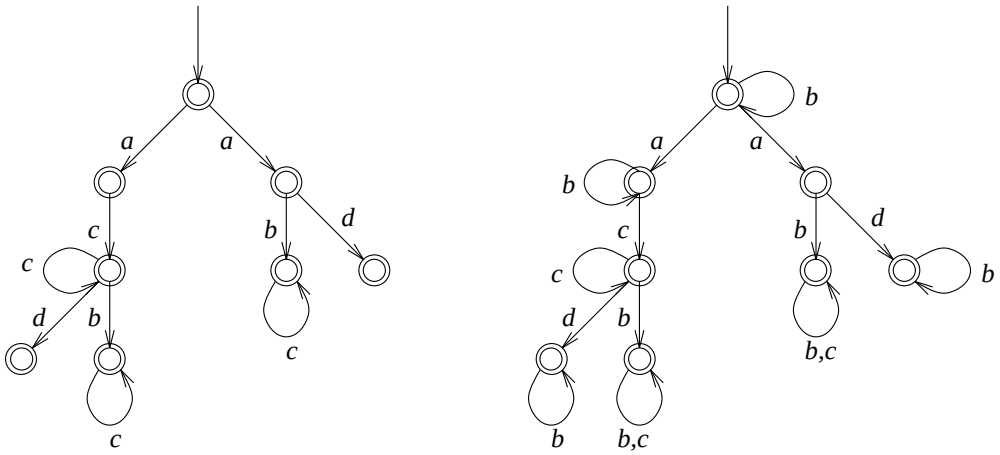
The results of Lemma 1 and 2 are now used to prove that the observation-compatibility of a deterministic system is preserved under augmentation.

THEOREM 1. *Let (S^m, S) be a deterministic trajectory model, $M(\cdot)$ be an observation mask, and $D \subseteq \Sigma$. Suppose that (S^m, S) is M-compatible. Then $((S^m)^D, S^D)$ is also M-compatible (and deterministic).*

Proof. It suffices to show that $L(S^D)$ is (Σ^*, M)-observable. Pick $s, t \in L(S^D), \sigma \in \Sigma$ such that $M(s) = M(t)$ and $s\sigma \in L(S^D)$. Then we need to show that $t\sigma \in L(S^D)$. Since (S^m, S) is a deterministic trajectory model, there exists a deterministic state machine $\mathcal{S} := (X_{\mathcal{S}}, \Sigma, \delta_{\mathcal{S}}, x^0_{\mathcal{S}}, X^m_{\mathcal{S}})$ with trajectory model (S^m, S). Then $((S^m)^D, S^D) = (T^m(\mathcal{S}^D), T(\mathcal{S}^D))$. Define $s' := \pi_{L(\mathcal{S})}(s)$ and $t' := \pi_{L(\mathcal{S})}(t)$. Then it follows from Lemma 1 that

$$\delta^*_{\mathcal{S}^D}(x^0_{\mathcal{S}}, s) = \delta^*_{\mathcal{S}}(x^0_{\mathcal{S}}, s'); \quad \delta^*_{\mathcal{S}^D}(x^0_{\mathcal{S}}, t) = \delta^*_{\mathcal{S}}(x^0_{\mathcal{S}}, t'). \tag{1}$$

Also, since $M(s) = M(t)$, it follows from Lemma 2 that $M(s') = M(t')$. Hence if $s'\sigma \in L(S)$, then it follows from (Σ^*, M)-observability of $L(S)$ that $t'\sigma \in L(S)$. Hence



(a) NSM P (b) Augmented NSM $P^{\{b\}}$

FIG. 1. *Diagram illustrating Example* 1.

(1) implies $t\sigma \in L(S^D)$. On the other hand, if $s'\sigma \notin L(S)$, then $\sigma \in D$, so $t\sigma \in L(S^D)$ trivially. □

We show via the following example that the requirement of determinism cannot be relaxed in Theorem 1.

Example 1. In order to see that the determinism is a necessary condition for Theorem 1 to hold, consider the NSM $\mathcal{P}$ shown in Figure 1(a) with $\Sigma = \{a, b, c, d\}$ and $M(\cdot)$ such that $M(a) = a, M(b) = b, M(c) = \epsilon, M(d) = d$. Then (i) $ac^* \in L(\mathcal{P})$, and each trace in ac^* has identical mask value. It can be checked that the set of events enabled after each trace in ac^* equals $\{b, c, d\}$. (ii) $ac^*bc^* \in L(\mathcal{P})$, and each trace in ac^*bc^* has identical mask value. It can also be checked that the set of events enabled after each trace in ac^*bc^* equals $\{c\}$. (iii) Finally, $ac^*d \in L(\mathcal{P})$, and each trace in ac^*d has identical mask value. One can verify that no event is enabled after each such trace. Thus $L(\mathcal{P})$ is (Σ^*, M)-observable, so the associated trajectory model $(T^m(\mathcal{P}), T(\mathcal{P}))$ is M-compatible.

The augmented NSM $\mathcal{P}^{\{b\}}$ is shown in Figure 1(b). Then $ab, abc \in L(\mathcal{P}^{\{b\}})$ with $M(ab) = M(abc)$. However, the set of events enabled after ab equals $\{b, c\}$, whereas the set of events enabled after abc equals $\{b, c, d\}$. Thus $L(\mathcal{P}^{\{b\}})$ is not (Σ^*, M)-observable, and so the associated trajectory model $(T^m(\mathcal{P}^{\{b\}}), T(\mathcal{P}^{\{b\}})$ is not M-compatible. □

4. Centralized control under partial observation. In a previous paper [30], we showed that PSC can be used as a mechanism of control under the restriction that all controllable events are observable to the supervisor. We show in this section that PSC can be used as a mechanism of control without imposing this restriction on the observation mask. As discussed in the previous section, whenever the observations of a supervisor are filtered through a mask, the supervisor must be observation-compatible with respect to its observation mask; i.e., a supervisor under partial observation must satisfy the constraint that following each pair of traces that look alike under the observation mask, it must take identical control action.

Prior to establishing the main result of this section, we prove the following preliminary result.

LEMMA 3. *Let $H \subseteq \Sigma^*$ be prefix-closed, $K \subseteq H$, and $M(\cdot)$ be an observation mask. If $K^M \subseteq \Sigma^*$ denotes the infimal prefix-closed and (Σ^*, M)-observable superlanguage of K, then $K^M \cap H$ equals the infimal prefix-closed and (H, M)-observable superlanguage of K.*

Proof. For simplicity of notation define $K' := K^M \cap H$. Let $\hat{K} \subseteq \Sigma^*$ denote the infimal prefix-closed and (H, M)-observable superlanguage of K. We need to show that $K' = \hat{K}$. In order to show that $\hat{K} \subseteq K'$, it suffices to show that K' is a prefix-closed (H, M)-observable superlanguage of K. Since $K \subseteq H$, it follows that $K' = K^M \cap H$ is a superlanguage of K. Also, from the fact that prefix-closure is preserved under intersection, it follows that K' is prefix-closed. Finally, since K^M is (Σ^*, M)-observable, clearly, it is (H, M)-observable. Then it follows from the fact that observability of prefix-closed languages is preserved under intersection [21, 28] that $K' = K^M \cap H$ is also (H, M)-observable.

It remains to show that $K' \subseteq \hat{K}$. Suppose for contradiction that $\hat{K}$ is a proper sublanguage of K'. Then there exists $s \in \hat{K}$ and $\sigma \in \Sigma$ such that $s\sigma \in K' - \hat{K}$. Since $K' \subseteq K^M$, it follows that $s\sigma \in K^M$. Also, since $K \subseteq \hat{K} \subset K' \subseteq K^M$, and K^M is the infimal prefix-closed and (Σ^*, M)-observable superlanguage of K, it follows that K^M is also the infimal prefix-closed and (Σ^*, M)-observable superlanguage of $\hat{K}$. Finally, since $s \in \hat{K}$, $s\sigma \notin \hat{K}$, and $s\sigma \in K^M$, it follows from the fact that K^M is the infimal prefix-closed and (Σ^*, M)-observable superlanguage of $\hat{K}$ that there exists $t \in \hat{K}$ such that $M(t) = M(s)$ and $t\sigma \in \hat{K}$. We also have that $s \in \hat{K}$, and $s\sigma \in K' - \hat{K} \subseteq H - \hat{K}$. Thus we arrive at a contradiction to the fact that $\hat{K}$ is (H, M)-observable. ∎

The following corollary is immediate from Lemma 3.

COROLLARY 1. *Let $H \subseteq \Sigma^*$ be prefix-closed, $M(\cdot)$ be an observation function, and $K \subseteq H$ be prefix-closed and (H, M)-observable. If $K^M \subseteq \Sigma^*$ denotes the infimal prefix-closed and (Σ^*, M)-observable superlanguage of K, then $K^M \cap H = K$.*

Recall from [19] that a supervisor with trajectory model (S^m, S) is said to be *nonmarking* if $S^m = S$. In the following theorem we obtain a necessary and sufficient condition for the existence of a nonmarking and observation-compatible deterministic supervisor. We need the following result from [30, Remark 11]: Given a plant trajectory model (P^m, P) with priority set A, if a language K satisfies the controllability condition of Theorem 2 below and H is any prefix-closed language satisfying $L(P^{\Sigma-A}) \cap H = K$, then the nonmarking deterministic supervisor $(S, S) := (\det(H), \det(H))$ with priority set B such that $A \cup B = \Sigma$ yields K as the closed-loop behavior $L(P\ {}_A\|_B\ S)$.

THEOREM 2. *Let (P^m, P) be the trajectory model of a plant, $A, B \subseteq \Sigma$, with $A \cup B = \Sigma$; $M(\cdot)$ be an observation mask; and $K \subseteq L(P^{\Sigma-A})$ be a nonempty language. Then there exists a deterministic, nonmarking, and M-compatible supervisor with trajectory model (S, S) such that $L(P\ {}_A\|_B\ S) = K$ if and only if all of the following conditions are met:*

Prefix-closure: $\mathrm{pr}(K) = K$,

Controllability: $\mathrm{pr}(K)(\Sigma - B) \cap L(P^{\Sigma-A}) \subseteq \mathrm{pr}(K)$,

Observability: $\forall s, t \in \mathrm{pr}(K), \sigma \in \Sigma : M(s) = M(t), s\sigma \in \mathrm{pr}(K), t\sigma \in L(P^{\Sigma-A}) \Rightarrow t\sigma \in \mathrm{pr}(K)$.

In this case S can be chosen to be $\det(K^M)$, where K^M is the infimal prefix-closed and (Σ^*, M)-observable superlanguage of K.

Proof. In order to see the sufficiency part, consider the supervisor with $S := \det(K^M)$. Then $L(S) = K^M$, so that S is M-compatible. Also, it follows from Corollary 1 that $K^M \cap L(P^{\Sigma-A}) = K$. Using [30, Remark 11], we obtain $L(P\ {}_A\|_B\ S) = K$.

In order to see the necessity part, suppose that (S, S) is the trajectory model of a deterministic nonmarking and M-compatible supervisor such that $L(P\ {}_A\|_B\ S) = K$. Then it follows from the necessity part of [30, Theorem 4] that K is prefix-closed and controllable. It remains to show that K is $(L(P^{\Sigma-A}), M)$-observable. Since $K = L(P\ {}_A\|_B\ S) = L(P^{\Sigma-A}) \cap L(S^{\Sigma-B})$, it suffices to show that $L(S^{\Sigma-B})$ is (Σ^*, M)-observable. This follows from the fact that (S, S) is a deterministic and M-compatible trajectory model, and as shown in Theorem 1, M-compatibility of deterministic trajectory models is preserved under augmentation. □

Remark 1. In contrast to the standard controllability and observability condition of the Ramadge–Wonham setting, the conditions of Theorem 2 refer to the language of the *augmented* plant. This language depends on the *trajectory model* of the plant and in general cannot be deduced from the language model of the plant. Readers are referred to [30, Remark 9, Example 3] for further elaboration on this point.

Also, since the necessity part of Theorem 2 uses the result of Theorem 1, it follows from Example 3 that the necessity part of Theorem 2 may not hold if the supervisor is not required to be deterministic. In a recent paper Inan has studied the design of nondeterministic supervisors under partial observation [13], where he has introduced the notion of *co-closure* (a condition weaker than controllability and observability combined) and has proved its necessity and sufficiency.

Finally, it may seem that the result of Theorem 2 is an immediate consequence of our prior work on nondeterministic systems and the standard supervisory control results. However, this is not true as it is not clear at the outset whether our results on nondeterministic systems under *complete* observations will immediately "carry over" to the case of *partial* observations (with appropriate extensions as in the standard supervisory control). In fact the result of Theorem 2 fails to hold if more general *nonprojection* type observation masks are considered. This is because the observation-compatibility of a deterministic system is not preserved under augmentation if the observation mask is no longer the projection type. To see this consider an observation mask that identifies the only events a and b of a deterministic system which executes the event a in its initial state and deadlocks. Clearly, the system is observation-compatible. However, its augmentation with the event b has a self-loop on b in both its states. So, in the augmented system a as well as b can occur after the occurrence of the initial b, whereas only b can occur after the occurrence of the initial a, which violates the observation-compatibility since a and b are indistinguishable.

We next apply the result of Theorem 2 to the design of a supervisor that achieves mutually exclusive usage of a shared communication channel in a communication system.

Example 2. Consider the nondeterministic plant P depicted in Figure 2(a). In this example, the plant represents a partial model of a multiuser communication system. Only the portions of the model needed to illustrate the main result are included. The communication system has two channels. The first user can transmit messages using either channel and switches between the channels in a manner that is unmodeled and hence nondeterministic. The second user can transmit only on channel 2. The event a represents the commencement of transmission by user 1 and results in a nondeterministic transition to one of two successor states depending on which channel is used. The event b represents the commencement of transmission by user 2. Both the commencement events are controllable but are unobservable to the supervisor to be constructed. If both users are able to transmit their messages without collision, then an uncontrollable completion event c occurs which returns the plant to its initial state.

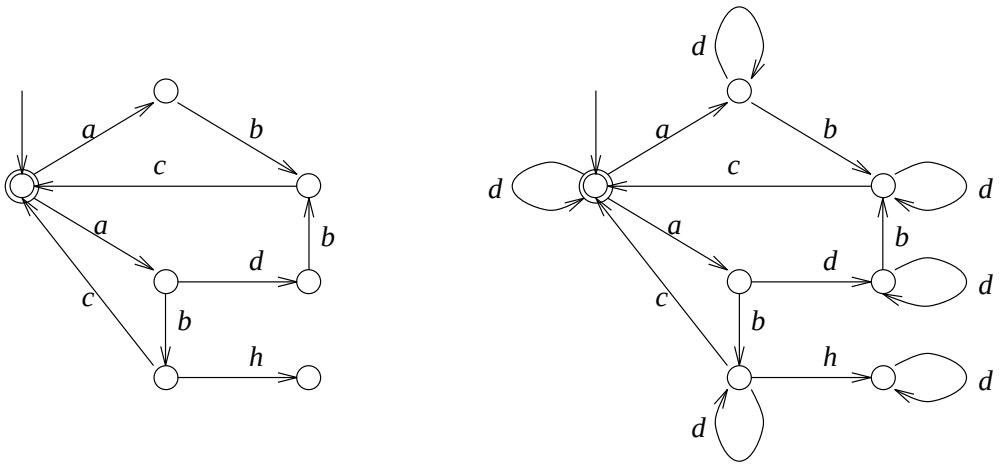


FIG. 2. *Plant P and augmented plant $P^{\Sigma-A} = P^{\{d\}}$.*

In order to avoid collision of messages, user 1 may receive a signal that causes it to vacate channel 2, provided that it has in fact chosen channel 2. This is represented by the event d. It is a driven event because it must be initiated by a supervisor and is executed synchronously by the plant only if able to do so, i.e., only if user 1 is transmitting on channel 2. If user 1 has been transmitting on channel 2 and user 2 commences transmission without it being preceded by d, then there are two possibilities: If user 1 has happened to finish before user 2 starts, then b is followed by the completion event c; otherwise b is followed by the collision event h, an uncontrollable event.

Thus, in this example,

$$\Sigma = \{a, b, c, d, h\}, \quad A = \{a, b, c, h\}, \quad B = \{a, b, d\},$$

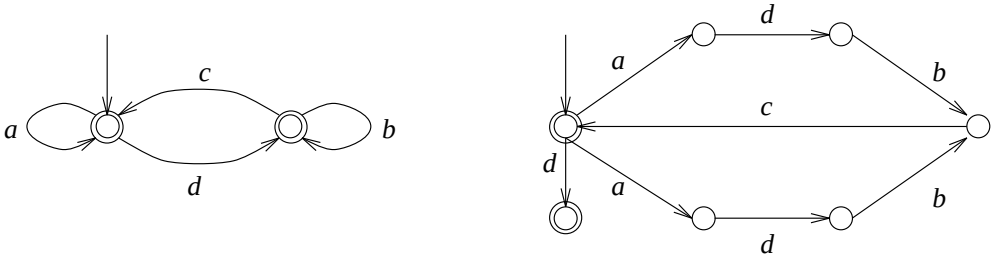
since a and b are controllable, c and h are uncontrollable, and d is a driven event. Note that a and b are the only events that are unobservable to the supervisor to be constructed. The basic performance specification is that a collision-free service should be provided. This can be represented by the prefix-closed sublanguage of the augmented plant (shown in Figure 2(b)) given by

$$K_0 := \{s \in L(P^{\Sigma-A}) \mid s \text{ does not contain } h\} = \text{ pr}[(d^*ad^*bd^*c)^*].$$

However, since user 1 cannot vacate channel 2 unless it is using it, it is reasonable to consider the desired behavior to be the sublanguage of K_0 consisting of those traces that do not contain any occurrence of d that is not immediately preceded by a. This is given by

$$K_1 := \text{ pr}[(abc + adbc)^*].$$

Since the uncontrollable event h can occur following the trace $ab \in K_1$, it is not controllable. The supremal prefix-closed and controllable sublanguage of K_1 is given



(a) Supervisor 1 (b) Closed-loop system 1

FIG. 3. *Supervisor S_1 and closed-loop system $P\ {}_A\|_B\ S_1$.*

by

$$K_1^{\uparrow} = \ \mathrm{pr}[(adbc)^*].$$

However, this is not $L(P^{\Sigma-A}, M)$-observable. In fact since $\epsilon, a \in K_1^{\uparrow}$ with $M(\epsilon) = M(a)$ and $d \in L(P^{\Sigma-A}) - K_1^{\uparrow}$, it follows that any prefix-closed sublanguage of $K_1^{\uparrow}$ that is $(L(P^{\Sigma-A}), M)$-observable cannot contain ad. Thus, a prefix-closed $(L(P^{\Sigma-A}), M)$-observable sublanguage of $K_1^{\uparrow}$ is contained in $\mathrm{pr}(a)$. By Theorem 2, it follows any M-compatible supervisor that results in a closed-loop generated language contained in the specification language K_1 gives a closed-loop generated language contained in $\mathrm{pr}(a)$. This is clearly unsatisfactory.

Thus, we must relax the specification given by K_1 keeping in mind that the constraint given by K_0 must be satisfied. The infimal prefix-closed and $(L(P^{\Sigma-A}), M)$-observable superlanguage of $K_1^{\uparrow}$ is $\mathrm{pr}[(adbc)^*d]$, which is a sublanguage of K_0. Since $\mathrm{pr}[(adbc)^*d]$ is also controllable, and since its infimal prefix-closed and (Σ^*, M)-observable superlanguage is $\mathrm{pr}[(a^*db^*c)^*d]$, it follows from Theorem 2 that the nonmarking supervisor

$$S_1 := \ \det[(pr(a^*db^*c)^*d)] = \ \det[(pr(a^*db^*c)^*)]$$

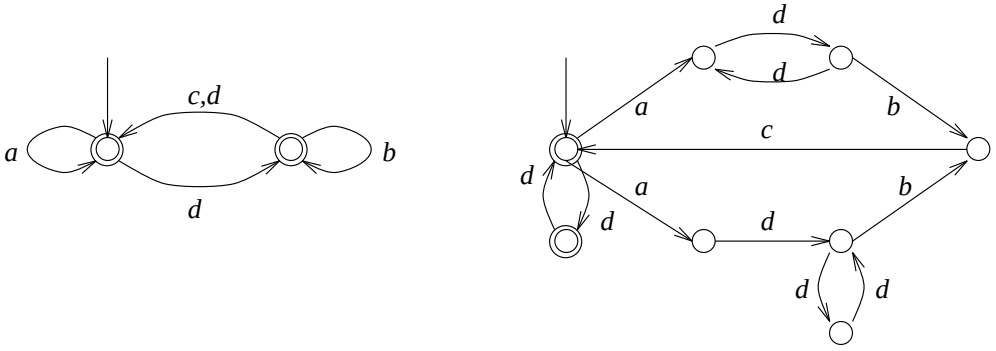
depicted in Figure 3(a) is M-compatible and yields $\mathrm{pr}[(adbc)^*d]$ as the closed-loop generated language. The closed-loop system is shown in Figure 3(b).

The supervisor implements the following simple control strategy: Initially it allows only user 1 to transmit. Before enabling transmission by user 2, it signals user 1 to vacate channel 2. This command is synchronously executed in the plant only when user 1 is transmitting on channel 2; otherwise, it is "refused" by the plant and occurs asynchronously in the supervisor. The supervisor then allows user 2 to communicate and returns to its initial state when the completion event c occurs. The ability of the plant to refuse a driven event initiated by the supervisor is essential to our control and is available because of the PSC-based control design. (Such a feature is certainly unavailable in an SSC-based control design.)

This design is not entirely satisfactory since, as can be seen from Figure 3(b), the closed-loop system deadlocks following the execution of any trace in $(adbc)^*d$.[4] This is because we did not require that the closed-loop behavior be *live* [17].[5] So the next

[4]Note that although the closed-loop system is nonblocking in the sense that the prefix-closure of the recognized refusal-traces is the same as the generated refusal-traces, it may deadlock.

[5]Informally, a language is said to be live if each of its trace has an extension in the language.



(a) Supervisor 2 (b) Closed-loop system 2

FIG. 4. *Supervisor S_2 and closed-loop system $P\ {}_A\|_B\ S_2$.*

alternative is to consider a live superlanguage of the "nonlive" language $\mathrm{pr}[(adbc)^*d]$ that is also controllable and observable and is contained in K_0. Although controllability and observability of prefix-closed languages are preserved under intersection, liveness is not. Similarly, although controllability and liveness of languages is preserved under union, observability is not. Hence, no unique solution can be identified. So a "semiautomatic" design involving some human reasoning is unavoidable.

With a little insight into the problem, it is easy to see that a simple modification of the supervisor in which a transition is added to permit the supervisor to return to its initial state by execution of d achieves liveness of the closed-loop behavior. The new supervisor, denoted S_2, and the resulting closed-loop system are shown in Figure 4. The closed-loop system can no longer deadlock. The language of the closed-loop system equals $\mathrm{pr}[(dd + ad(dd)^*bd^*c)^*]$, which is a sublanguage of K_0 as desired.

Note that both S_1 and S_2 do not change their state when either a or b occur, showing that they are compatible with the unobservability of these events.

We conclude this section by extending the result of Theorem 2 to obtain conditions for the existence of *nonblocking* supervisors. Recall from [19, Definition 6] that given a plant (P^m, P) with priority set A, a supervisor (S^m, S) with priority set B is said to be *language model nonblocking* if $\mathrm{pr}(L(P^m\ {}_A\|_B\ S^m)) = L(P\ {}_A\|_B\ S)$; it is said to be *trajectory model nonblocking* if $\mathrm{pr}(P^m\ {}_A\|_B\ S^m) = P\ {}_A\|_B\ S$. In the following corollary we provide a necessary and sufficient condition for the existence of an observation-compatible and language model nonblocking supervisor.

COROLLARY 2. *Let (P^m, P) be the trajectory model of a plant, $A, B \subseteq \Sigma$, with $A \cup B = \Sigma$; $M(\cdot)$ be an observation mask; and $K^m \subseteq L((P^m)^{\Sigma - A})$ be a nonempty language. Then there exists a deterministic, nonmarking, language model nonblocking, and M-compatible supervisor with trajectory model (S, S) such that $L(P^m\ {}_A\|_B\ S^m) = K^m$ if and only if all of the following conditions are met:*

Relative-closure: $\mathrm{pr}(K^m) \cap L((P^m)^{\Sigma - A}) = K^m$,

Controllability: $\mathrm{pr}(K^m)(\Sigma - B) \cap L(P^{\Sigma - A}) \subseteq \mathrm{pr}(K^m)$,

Observability: $\forall s, t \in \mathrm{pr}(K^m), \sigma \in \Sigma : M(s) = M(t), s\sigma \in \mathrm{pr}(K^m), t\sigma \in L(P^{\Sigma - A}) \Rightarrow t\sigma \in \mathrm{pr}(K^m)$.

In this case S can be chosen to be $\det((K^m)^M)$, *where $(K^m)^M$ denotes the infimal prefix-closed and (Σ^*, M)-observable superlanguage of K^m.*

Proof. First consider sufficiency. Since $\mathrm{pr}(K^m)$ is nonempty, prefix-closed, controllable, and $(L(P^{\Sigma - A}), M)$-observable, it follows from the sufficiency part of

Theorem 2 that the nonmarking supervisor with $S := \det((\mathrm{pr}(K^m))^M) = \det((K^m)^M)$ is M-compatible, and $L(P\ {}_A\|_B\ S) = \mathrm{pr}(K^m)$. Hence, using the relative closure condition we obtain the following series of equalities:

$$\begin{aligned} K^m &= \mathrm{pr}(K^m) \cap L((P^m)^{\Sigma-A}) \\ &= L(P\ {}_A\|_B\ S) \cap L((P^m)^{\Sigma-A}) \\ &= [L(P^{\Sigma-A}) \cap L(S^{\Sigma-B})] \cap L((P^m)^{\Sigma-A}) \\ &= L((P^m)^{\Sigma-A}) \cap L(S^{\Sigma-B}) \\ &= L(P^m\ {}_A\|_B\ S). \end{aligned}$$

Since $\mathrm{pr}(K^m) = L(P\ {}_A\|_B\ S)$ and $K^m = L(P^m\ {}_A\|_B\ S)$, the supervisor is language model nonblocking.

The necessity part follows from the necessity parts of Theorem 2 and [19, Theorem 5]. □

The result of Corollary 2 can be extended to obtain a necessary and sufficient condition for the existence of an observation-compatible and trajectory model nonblocking supervisor. We need the following result from [19, Proposition 4]: Given a plant (P^m, P) with priority set A and a nonempty language $K^m \subseteq L((P^m)^{\Sigma-A})$, if there exists a deterministic, nonmarking, and language model nonblocking supervisor (S, S) with priority set B such that $A \cup B = \Sigma$ and $L(P^m\ {}_A\|_B\ S) = K^m$, then

$$P^m\ {}_A\|_B\ \det(\mathrm{pr}(K^m)) = P^m\ {}_A\|_B\ S; \quad P\ {}_A\|_B\ \det(\mathrm{pr}(K^m)) = P\ {}_A\|_B\ S.$$

COROLLARY 3. *Let (P^m, P) be the trajectory model of a plant, $A, B \subseteq \Sigma$, with $A \cup B = \Sigma$; $M(\cdot)$ be an observation mask; and $K^m \subseteq L((P^m)^{\Sigma-A})$ be a nonempty language. Then there exists a deterministic, nonmarking, trajectory model nonblocking, and M-compatible supervisor with trajectory model (S, S) such that $L(P^m\ {}_A\|_B\ S^m) = K^m$ if and only if all of the following conditions are met:*

Relative-closure: $\mathrm{pr}(K^m) \cap L((P^m)^{\Sigma-A}) = K^m$,

Controllability: $\mathrm{pr}(K^m)(\Sigma - B) \cap L(P^{\Sigma-A}) \subseteq \mathrm{pr}(K^m)$,

Observability: $\forall s, t \in \mathrm{pr}(K^m), \sigma \in \Sigma : M(s) = M(t),\ s\sigma \in \mathrm{pr}(K^m), t\sigma \in L(P^{\Sigma-A}) \Rightarrow t\sigma \in \mathrm{pr}(K^m)$,

Trajectory-closure: $P\ {}_A\|_B \det(\mathrm{pr}(K^m)) = \mathrm{pr}[P^m\ {}_A\|_B \det(\mathrm{pr}(K^m))]$.

In this case S can be chosen to be $\det((K^m)^M)$, where $(K^m)^M$ denotes the infimal prefix-closed and (Σ^*, M)-observable superlanguage of K^m.

Proof. The necessity part follows from the necessity part of Corollary 2 and that of [19, Theorem 6]; the sufficiency part follows from the sufficiency part of Corollary 2 and [19, Proposition 4]. □

5. Decentralized control. So far we have restricted our attention to the problem of *centralized* control under partial observation. However, in many applications, such as manufacturing systems, communication networks, and so on, the plant is physically distributed and it is desirable to have decentralized controllers [6, 20, 22, 29, 32], where each controller is able to control a certain set of events and is able to observe certain other events. The problem of decentralized control can be studied quite elegantly in our PSC-based approach.

Without any loss of generality we consider the case of "two-decentralization"; i.e., given a discrete event plant P with priority set A we consider synthesis of two supervisors S_1 and S_2 with priority sets B_1 and B_2, respectively, which are compatible with their own observation masks $M_1(\cdot)$ and $M_2(\cdot)$, respectively, such that the

controlled plant $P\ {}_A\|_{B_1\cup B_2}\ (S_1\ {}_{B_1}\|_{B_2}\ S_2)$ satisfies a desired behavior constraint. The priority set of supervisor $S_i (i = 1, 2)$ is B_i, and its observations are filtered through the mask function $M_i(\cdot)$. Thus the events in the set $A \cap B_i$ are the controllable events for S_i, those in the set $A - B_i$ are the uncontrollable events for S_i, and finally those in $B_i - A$ are the driven events for S_i. Also, S_i must be compatible with $M_i(\cdot)$; i.e., its generated language must be (Σ^*, M_i)-observable. Since an event must belong to at least one of the priority sets we have that $A \cup B_1 \cup B_2 = \Sigma$.

For notational simplicity we define $B := B_1 \cup B_2$ and $S := S_1\ {}_{B_1}\|_{B_2}\ S_2$. Since the events in the set $A - B$ are in the priority set of neither of the supervisors, these represent the uncontrollable events. Thus for decentralized supervision it is expected that the desired behavior be controllable with respect to these uncontrollable events. The remaining events are in the priority set(s) of one or both of the supervisors; however, their enablement/disablement must satisfy the restriction that results from the partial observability of the supervisors. This is captured by the following condition of co-observability, which is similar to that given by Rudie and Wonham [29].

DEFINITION 3. *Given the priority sets B_1 and B_2 of two supervisors, and their observation masks $M_1(\cdot)$ and $M_2(\cdot)$, respectively, a language $K \subseteq \Sigma^*$ is said to be* co-observable *with respect to a prefix-closed language $H \subseteq \Sigma^*$, called* (H, B_1, B_2, M_1, M_2)–co-observable*, if*

$\forall s_1, s_2, t \in \mathrm{pr}(K), \sigma \in B_1 \cup B_2$*:*

(1) $[\sigma \in B_1 - B_2, M_1(s_1) = M_1(t), s_1\sigma \in \mathrm{pr}(K), t\sigma \in H] \Rightarrow [t\sigma \in \mathrm{pr}(K)]$

(2) $[\sigma \in B_2 - B_1, M_2(s_2) = M_2(t), s_2\sigma \in \mathrm{pr}(K), t\sigma \in H] \Rightarrow [t\sigma \in \mathrm{pr}(K)]$

(3) $[\sigma \in B_1 \cap B_2, M_1(s_1) = M_1(t), M_2(s_2) = M_2(t), s_1\sigma, s_2\sigma \in \mathrm{pr}(K), t\sigma \in H] \Rightarrow [t\sigma \in \mathrm{pr}(K)]$.

Thus if an event belongs solely to priority set of one of the supervisors and it is enabled following a trace, then it must be enabled following any other trace that is indistinguishable to that supervisor (provided it can occur in the plant). On the other hand, if the event belongs to the common priority set of the supervisors and it can occur in the plant following a trace which is indistinguishable from a certain trace to the first supervisor, and from another trace to the second supervisor, and the event is enabled following these latter pair of traces, then the event must also be enabled following the former trace. It is clear that K is co-observable if and only if $\mathrm{pr}(K)$ is co-observable. Also, as is the case with observability, co-observability of prefix-closed languages is preserved under intersection [29]; consequently, the infimal prefix-closed and co-observable superlanguage of a given language exists.

We show below that controllability together with co-observability is necessary and sufficient for decentralized supervision. It is clear that observability with respect to each of the masks implies co-observability. Thus a weaker condition than observability with respect to each of the masks is needed for decentralized supervision; this is because the events in the common priority set of the two supervisors can be disabled by either of them. However, if the common priority set is empty, then under the condition of controllability, co-observability is equivalent to observability with respect to each of the masks.

We saw above that the operation of PSC of a pair of systems can be reduced to that of SSC when the priority sets of the two systems exhaust the entire event set. We next prove that this is also the case when more than two systems are involved. We need the following lemma.

LEMMA 4. *Consider NSMs $\mathcal{S}_1$ and $\mathcal{S}_2$ with priority sets B_1 and B_2, respectively. Then*

$$(\mathcal{S}_1\ {}_{B_1}\|_{B_2}\ \mathcal{S}_2)^{\Sigma - B} = \mathcal{S}_1^{\Sigma - B_1}\ {}_{\Sigma}\|_{\Sigma}\ \mathcal{S}_2^{\Sigma - B_2},$$

where $B := B_1 \cup B_2$.

Proof. The above lemma follows from the following series of equalities:

$$\begin{aligned}
\mathcal{S}_1^{\Sigma-B_1}\ {}_{\Sigma}\|_{\Sigma}\ \mathcal{S}_2^{\Sigma-B_2} &= (\mathcal{S}_1^{B-B_1})^{\Sigma-B}\ {}_{\Sigma}\|_{\Sigma}\ (\mathcal{S}_2^{B-B_2})^{\Sigma-B} \\
&= [\mathcal{S}_1^{B-B_1}{}_B\|_{\Sigma-B}\ \det((\Sigma-B)^*)]_{\Sigma}\|_{\Sigma} \\
&\quad \times[\mathcal{S}_2^{B-B_2}{}_B\|_{\Sigma-B}\ \det((\Sigma-B)^*)] \\
&= [\mathcal{S}_1^{B-B_1}{}_B\|_B\mathcal{S}_2^{B-B_2}]\ {}_B\|_{\Sigma-B} \\
&\quad \times[\det((\Sigma-B)^*)\ {}_{\Sigma-B}\|_{\Sigma-B}\ \det((\Sigma-B)^*)] \\
&= [\mathcal{S}_1^{B-B_1}\ {}_B\|_B\ \mathcal{S}_2^{B-B_2}]\ {}_B\|_{\Sigma-B}\ \det((\Sigma-B)^*) \\
&= [\mathcal{S}_1\ {}_{B_1}\|_{B_2}\ \mathcal{S}_2]^{\Sigma-B},
\end{aligned}$$

where the first, second, and final equalities follow from definition of augmentation and the third equality follows from associativity of PSC. ∎

The following corollary is immediate from the above lemma.

COROLLARY 4. *Consider NSMs $\mathcal{P}$, $\mathcal{S}_1$, and $\mathcal{S}_2$ with priority sets A, B_1, and B_2, respectively, such that $A \cup B_1 \cup B_2 = \Sigma$. Then*

$$\mathcal{P}\ {}_A\|_B\ \mathcal{S} = P^{\Sigma-A}\ {}_{\Sigma}\|_{\Sigma}\ [\mathcal{S}_1^{\Sigma-B_1}\ {}_{\Sigma}\|_{\Sigma}\ \mathcal{S}_2^{\Sigma-B_2}],$$

where $B := B_1 \cup B_2$ and $\mathcal{S} := \mathcal{S}_1\ {}_{B_1}\|_{B_2}\ \mathcal{S}_2$.

Proof. Since $A \cup B = \Sigma$, it follows from a PSC property [9, 30, 19] that

$$\mathcal{P}\ {}_A\|_B\ \mathcal{S} = \mathcal{P}^{\Sigma-A}\ {}_{\Sigma}\|_{\Sigma}\ \mathcal{S}^{\Sigma-B}.$$

Thus the result follows from Lemma 4. ∎

Remark 2. Corollary 4 shows that the operation of PSC of two or more systems can be reduced to that of SSC whenever the priority sets of all the systems jointly exhaust the entire event set. It also follows that under the hypothesis of Corollary 4

$$L(\mathcal{P}\ {}_A\|_B\ \mathcal{S}) = L(\mathcal{P}^{\Sigma-A}) \cap L(\mathcal{S}_1^{\Sigma-B_1}) \cap L(\mathcal{S}_2^{\Sigma-B_2}). \tag{2}$$

Next we establish a relationship between controllability, observability, co-observability, and PSC. In the following lemma we prove that if supervisors $\mathcal{S}_1$ and $\mathcal{S}_2$ are M_1-compatible and M_2-compatible, respectively, and both generate $(\Sigma^*, \Sigma - B)$ controllable languages, then the language of $\mathcal{S}_1\ {}_{B_1}\|_{B_2}\ \mathcal{S}_2$ is $(\Sigma^*, \Sigma - B)$-controllable and $(\Sigma^*, B_1, B_2, M_1, M_2)$–co-observable.

LEMMA 5. *Consider deterministic state machines $\mathcal{S}_1$ and $\mathcal{S}_2$ with priority sets B_1 and B_2, respectively. Suppose that $\mathcal{S}_1$ and $\mathcal{S}_2$ are observation-compatible with respect to masks M_1 and M_2, respectively, and $L(\mathcal{S}_1)$ as well as $L(\mathcal{S}_2)$ are $(\Sigma^*, \Sigma - B)$ controllable, where $B := B_1 \cup B_2$. Then $L(\mathcal{S})$, where $\mathcal{S} := \mathcal{S}_1\ {}_{B_1}\|_{B_2}\ \mathcal{S}_2$, is $(\Sigma^*, B_1, B_2, M_1, M_2)$–co-observable and $(\Sigma^*, \Sigma - B)$ controllable.*

Proof. In order to see co-observability, pick $s_1, s_2, t \in L(\mathcal{S})$, and $\sigma \in B$. Since $\mathcal{S}_1$ and $\mathcal{S}_2$ are deterministic, $\mathcal{S}$ is also deterministic. Let $(x^1_{s_1}, x^2_{s_1})$, $(x^1_{s_2}, x^2_{s_2})$, and (x^1_t, x^2_t) denote the states reached in $\mathcal{S}$ after execution of s_1, s_2, and t, respectively, where the first coordinate denotes the state reached in $\mathcal{S}_1$ and the second coordinate denotes the state reached in $\mathcal{S}_2$.

In order to prove co-observability of $L(\mathcal{S})$ we must consider the three different cases of the definition of co-observability. First suppose $\sigma \in B_1 - B_2$, $M_1(s_1) = M_1(t)$, and $s_1\sigma \in L(\mathcal{S})$; we need to show that $t\sigma \in L(\mathcal{S})$. Since $\sigma \in B_1 - B_2$ and $s_1\sigma \in L(\mathcal{S})$, it follows that σ is defined at the state $x^1_{s_1}$ of $\mathcal{S}_1$. Then using the result of Lemma 2 and the fact that $\mathcal{S}_1$ is M_1-compatible, we obtain that σ is also defined at the state x^1_t of $\mathcal{S}_1$, which implies that $t\sigma \in L(\mathcal{S})$. It can be argued in a similar manner that the second and third cases of the definition of co-observability also hold.

In order to see controllability, consider $s \in L(\mathcal{S})$ and $\sigma \in \Sigma - B$. Let (x_s^1, x_s^2) be the state reached in $\mathcal{S}$ by execution of s. Then it follows from the controllability of $L(\mathcal{S}_1)$ that σ is defined at state x_s^1 of $\mathcal{S}_1$. Hence $s\sigma \in L(\mathcal{S})$. □

Given a language K, we use K^{BM_i} $(i = 1, 2)$ to denote the infimal prefix-closed, $(\Sigma^*, \Sigma - B)$-controllable, and (Σ^*, M_i)-observable superlanguage of K, which exists as the controllability and observability of prefix-closed languages is preserved under intersection. The notation $K^{BM_{12}}$ is used to denote the infimal prefix-closed, $(\Sigma^*, \Sigma - B)$-controllable, and $(\Sigma^*, M_1, M_2, B_1, B_2)$–co-observable superlanguage of K. The result of Lemma 5 can be used to show that if $\mathcal{S}_1$ generates K^{BM_1} and $\mathcal{S}_2$ generates K^{BM_2}, then $\mathcal{S}$ generates $K^{BM_{12}}$. This we state in the following theorem.

THEOREM 3. *Let $M_1, M_2, B_1, B_2, K^{BM_1}, K^{BM_2}$, and $K^{BM_{12}}$ be as defined above. Suppose that $\mathcal{S}_1$ and $\mathcal{S}_2$ are deterministic state machines with $L(\mathcal{S}_1) = K^{BM_1}$ and $L(\mathcal{S}_2) = K^{BM_2}$. Then $L(\mathcal{S}) = K^{BM_{12}}$, where $\mathcal{S} := \mathcal{S}_1 \,{}_{B_1}\|_{B_2}\, \mathcal{S}_2$ and $B := B_1 \cup B_2$.*

Proof. Since $K \subseteq L(\mathcal{S}_1)$ and $K \subseteq L(\mathcal{S}_2)$, it follows that $K \subseteq L(\mathcal{S})$. Also, it follows from Lemma 5 that $L(\mathcal{S})$ is controllable and co-observable. Thus $L(\mathcal{S})$ is a prefix-closed, controllable, and co-observable superlanguage of K. Hence we have that $K^{BM_{12}} \subseteq L(\mathcal{S})$. In order to see the reverse containment, it suffices to show that non–zero-length strings of $L(\mathcal{S})$ are also in $K^{BM_{12}}$, as the zero-length string ϵ does belong to $K^{BM_{12}}$. Thus we need to show that for any string $t \in K^{BM_{12}}$ and an event σ such that $t\sigma \in L(\mathcal{S})$, $t\sigma \in K^{BM_{12}}$. If $\sigma \in \Sigma - B$, then it follows from the prefix-closure and $(\Sigma^*, \Sigma - B)$-controllability of $K^{BM_{12}}$ that $t\sigma \in K^{BM_{12}}$. On the other hand, if $\sigma \in B$, then we show that the following hold:

(1) $[\sigma \in B_1 - B_2] \Rightarrow [\exists s_1 : M_1(s_1) = M_1(t), s_1\sigma \in \mathrm{pr}(K)]$,

(2) $[\sigma \in B_2 - B_1] \Rightarrow [\exists s_2 : M_2(s_2) = M_2(t), s_2\sigma \in \mathrm{pr}(K)]$,

(3) $[\sigma \in B_1 \cap B_2] \Rightarrow [\exists s_1, s_2 : M_1(s_1) = M_1(t), M_2(s_2) = M_2(t), s_1\sigma, s_2\sigma \in \mathrm{pr}(K)]$,

as this together with $(\Sigma^*, B_1, B_2, M_1, M_2)$–co-observability of $K^{BM_{12}}$ clearly implies that $t\sigma \in K^{BM_{12}}$.

We prove this using induction on length of t. We only prove that the case (1) holds, as the proof for the other two cases is similar. In order to see the base step, set $t = \epsilon$ (note that we do have $\epsilon \in K^{BM_{12}}$) and pick $\sigma \in B_1 - B_2$. Since $t\sigma = \sigma \in L(\mathcal{S})$ and $\sigma \in B_1 - B_2$, it follows from construction of $\mathcal{S}$ that $\sigma \in L(\mathcal{S}_1) = K^{BM_1}$. Since K^{BM_1} is the infimal prefix-closed, $(\Sigma^*, \Sigma - B)$-controllable, and (Σ^*, M_1)-observable superlanguage of K, and σ is not an uncontrollable event, it follows that there exists a string s_1 such that $s_1\sigma \in \mathrm{pr}(K)$ and $M_1(s_1) = M_1(t) = \epsilon$. The other two cases of the base step can be proved analogously.

In order to see the induction step set $t = \bar{t}\bar{\sigma}$ and pick $\sigma \in B_1 - B_2$. Suppose $\bar{\sigma} \in B_1 - B_2$. Then it follows from induction hypothesis that there exists $\bar{s}_1$ such that $s_1' := \bar{s}_1\bar{\sigma} \in \mathrm{pr}(K)$ and $M_1(\bar{s}_1) = M_1(\bar{t})$. Let (x_t^1, x_t^2) and $(x_{s_1'}^1, x_{s_1'}^2)$ denote the states reached in $\mathcal{S}$ after execution of t and s_1', respectively, where the first coordinate denotes the state reached in $\mathcal{S}_1$ and the second coordinate denotes the state reached in $\mathcal{S}_2$. Since $\sigma \in B_1 - B_2$, we have that σ is defined at state x_t^1. Hence it follows from Lemma 2 and M_1-compatibility of $\mathcal{S}_1$ that σ is also defined at state $x_{s_1'}^1$, which implies that $s_1'\sigma \in L(\mathcal{S})$. Since $s_1' \in \mathrm{pr}(K) \subseteq L(\mathcal{S}_1)$ and $\sigma \in B_1 - B_2$, we must have $s_1'\sigma \in L(\mathcal{S}_1) = K^{BM_1}$. Since K^{BM_1} is the infimal prefix-closed controllable and observable superlanguage of K, and σ is not an uncontrollable event, this implies that there exists s_1 such that $s_1\sigma \in \mathrm{pr}(K)$ and $M_1(s_1) = M_1(s_1')$. Thus s_1 is the desired string, as $M_1(s_1) = M_1(s_1') = M_1(t)$. □

Using the results derived in this section, we are now ready to present a necessary and sufficient condition for decentralized supervision.

THEOREM 4. *Consider A, B_1, B_2, M_1, and M_2 as defined above with $A \cup B_1 \cup B_2 = \Sigma$. Let (P^m, P) be the trajectory model of a plant and $K \subseteq L(P^{\Sigma-A})$ be a nonempty language. Then there exist deterministic, nonmarking, and M_1-compatible supervisor with trajectory model (S_1, S_1) and M_2-compatible supervisor with trajectory model (S_2, S_2) such that $L(P\ {}_A\|_B\ S) = K$, where $S := S_1\ {}_{B_1}\|_{B_2}\ S_2$ and $B := B_1 \cup B_2$ if and only if all of the following hold:*

Prefix-closure: $\mathrm{pr}(K) = K$,

Controllability: $\mathrm{pr}(K)(\Sigma - B) \cap L(P^{\Sigma-A}) \subseteq \mathrm{pr}(K)$,

Co-observability: K is $(L(P^{\Sigma-A}), B_1, B_2, M_1, M_2)$–co-observable.

In this case S_i $(i = 1, 2)$ can be chosen to be $\det(K^{BM_i})$, where K^{BM_i} is the infimal prefix-closed, $(\Sigma^*, \Sigma - B)$-controllable, and (Σ^*, M_i)-observable superlanguage of K.

Proof. We begin by proving the necessity. Prefix-closure and controllability conditions follow from the necessity part of [30, Theorem 4]. We need to show that the co-observability condition also holds. It follows from hypothesis and Corollary 4 that $K = L(P\ {}_A\|_B\ S) = L(P^{\Sigma-A}) \cap L(S_1^{\Sigma-B_1}) \cap L(S_2^{\Sigma-B_2})$. Hence it suffices to show that $H := L(S_1^{\Sigma-B_1}) \cap L(S_2^{\Sigma-B_2})$ is $(\Sigma^*, B_1, B_2, M_1, M_2)$–co-observable. Pick $s_1, s_2, t \in H$ and $\sigma \in \Sigma$. We must consider the three different cases of the definition of co-observability. We consider only the first case, as the other cases can be proven in a similar manner. Suppose $\sigma \in B_1 - B_2$, $s_1\sigma \in H$, and $M_1(s_1) = M_1(t)$. We need to show that $t\sigma \in H$. Since $t \in L(S_2^{\Sigma-B_2})$ and $\sigma \in B_1 - B_2 \subseteq \Sigma - B_2$, $t\sigma \in L(S_2^{\Sigma-B_2})$ trivially. It remains to show that $t\sigma \in L(S_1^{\Sigma-B_1})$. This follows from the fact that $S_1^{\Sigma-B_1}$ is M_1-compatible (as S_1 is M_1-compatible and deterministic, and observation-compatibility of deterministic systems is preserved under augmentation). This completes the proof of the necessity part.

In order to see the sufficiency part, select $S_1 := \det(K^{BM_1})$ and $S_2 := \det(K^{BM_2})$. Then S_1 and S_2 are deterministic, S_1 is M_1-compatible, and S_2 is M_2-compatible. It remains to show that the controlled plant language equals K. From Theorem 3 we have that $L(S) = K^{BM_{12}}$, where $K^{BM_{12}}$ is the infimal prefix-closed, $(\Sigma^*, \Sigma - B)$-controllable, and $(\Sigma^*, B_1, B_2, M_1, M_2)$–co-observable superlanguage of K. Using arguments similar to those in Lemma 3 we can readily conclude that $L(P^{\Sigma-A}) \cap L(S)$ is the infimal prefix-closed, $(L(P^{\Sigma-A}), \Sigma - B)$-controllable, and $(L(P^{\Sigma-A}), B_1, B_2, M_1, M_2)$–co-observable superlanguage of K. Hence it follows from the prefix-closure, controllability, and co-observability conditions that

$$L(P^{\Sigma-A}) \cap L(S) = K. \tag{3}$$

We need to show that we also have the following equality: $H := L(P^{\Sigma-A}) \cap L(S^{\Sigma-B}) = K$. This follows from (3) and the fact that K is controllable as is shown next.

Since $L(S) \subseteq L(S^{\Sigma-B})$, clearly $K \subseteq H$. Suppose for contradiction that there exists a string s such that $s \in H - K$. Let s be a minimal-length such string. Since $\epsilon \in K$, we have $s \neq \epsilon$, which implies $s = \bar{s}\sigma$, where $\bar{s} \in K$ and $\sigma \in \Sigma$. Since $\bar{s} \in K$ and $\bar{s}\sigma \notin K$, it must be the case that $\sigma \in \Sigma - B$. This is contradictory to the fact that K is controllable, as we have $\bar{s} \in K$, $\sigma \in \Sigma - B$, $\bar{s}\sigma \in H$, which implies $\bar{s}\sigma \in L(P^{\Sigma-A})$; however, $\bar{s}\sigma \notin K$. This completes the proof. ∎

Remark 3. Note that the conditions of controllability and co-observability in Theorem 4 are with regard to the language of the *augmented* plant, which depends on the trajectory model of the plant and cannot be inferred from the language model of the plant. Also, as is the case of the necessity part of Theorem 2, the necessity part of Theorem 4 may not hold if the supervisors are nondeterministic.

Finally, the result of Theorem 4 can be easily extended to obtain conditions for either language model nonblocking or trajectory model nonblocking supervisors. In

fact arguments similar to those given in Corollaries 2 and 3 can be used to show that language model nonblocking supervision would require the condition of relative-closure instead of that of prefix-closure, and a trajectory model nonblocking supervision would require the additional trajectory-closure condition.

6. Conclusion. In this paper we have extended our earlier work on supervisory control of nondeterministic systems using prioritized synchronization as the mechanism of control and trajectory model as the modeling formalism to control under partial observation. The notion of observation-compatibility of trajectory models has been introduced, and necessary and sufficient conditions for the existence of observation-compatible supervisors have been obtained for centralized as well as decentralized supervision. Although these conditions are similar to the standard conditions of controllability, observability, and co-observability found in literature, they are different, as they depend on the trajectory model as opposed to the language model of the plant. Also, our work demonstrates the suitability of PSC-based supervisor design for nondeterministic systems under centralized as well as decentralized setting. These results have been applied to the design of a supervisor that achieves a mutually exclusive usage of a communication channel in a communication system.

REFERENCES

[1] J. C. M. Baeten, J. A. Bergstra, and J. W. Klop, *Ready-trace semantics for concrete process algebra with the priority operator*, Comput. J., 30 (1987), pp. 498–506.
[2] J. C. M. Baeten and W. P. Weijland, *Process Algebra*, Cambridge University Press, Cambridge, UK, 1990.
[3] S. Balemi, *Input/output discrete event processes and communication delays*, Discrete Event Dynamical Systems: Theory and Applications, 4 (1994), pp. 41–85.
[4] S. Balemi, G. J. Hoffmann, P. Gyugyi, H. Wong-Toi, and G. F. Franklin, *Supervisory control of a rapid thermal multiprocessor*, IEEE Trans. Automat. Control, 38 (1993), pp. 1040–1059.
[5] E. Chen and S. Lafortune, *Dealing with blocking in supervisory control of discrete event systems*, IEEE Trans. Automat. Control, 36 (1991), pp. 724–735.
[6] R. Cieslak, C. Desclaux, A. Fawaz, and P. Varaiya, *Supervisory control of discrete event processes with partial observation*, IEEE Trans. Automat. Control, 33 (1988), pp. 249–260.
[7] C. H. Golaszewski and P. J. Ramadge, *Control of discrete event processes with forced events*, in Proc. 26th IEEE Conf. Decision and Control, Los Angeles, CA, 1987, pp. 247–251.
[8] M. Heymann, *Concurrency and discrete event control*, IEEE Control Systems Magazine, 10 (1990), pp. 103–112.
[9] M. Heymann and G. Meyer, *Algebra of Discrete Event Processes*, Technical report NASA 102848, NASA Ames Research Center, Moffett Field, CA, June 1991.
[10] C. A. R. Hoare, *Communicating Sequential Processes*, Prentice-Hall, Englewood Cliffs, NJ, 1985.
[11] J. E. Hopcroft and J. D. Ullman, *Introduction to Automata Theory, Languages and Computation*, Addison-Wesley, Reading, MA, 1979.
[12] K. Inan, *An algebraic approach to supervisory control*, Math. Control Signals Systems, 5 (1992), pp. 151–164.
[13] K. Inan, *Nondeterministic supervision under partial observations*, in 11th Internat. Conf. on Analysis and Optimization of Systems, Discrete Event Systems, Proc. Conf. Held in Sophia-Antipolis, June 15–17, 1994, Lecture Notes in Control and Inform. Sci. 199, G. Cohen and J.-P. Quadrat, eds., Springer-Verlag, New York, 1994, pp. 39–48.
[14] K. Inan and P. Varaiya, *Finitely recursive process models for discrete event systems*, IEEE Trans. Automat. Control, 33 (1988), pp. 626–639.
[15] K. Inan and P. Varaiya, *Algebras of discrete event models*, Proc. IEEE, 77 (1989), pp. 24–38.
[16] R. Kumar, V. K. Garg, and S. I. Marcus, *On controllability and normality of discrete event dynamical systems*, Systems Control Lett., 17 (1991), pp. 157–168.
[17] R. Kumar, V. K. Garg, and S. I. Marcus, *On supervisory control of sequential behaviors*, IEEE Trans. Automat. Control, 37 (1992), pp. 1978–1985.

[18] R. Kumar and M. A. Shayman, *Supervisory control of nondeterministic systems under partial observation*, in Proc. 1994 IEEE Conf. Decision and Control, Orlando, FL, December 1994, pp. 3649–3654.
[19] R. Kumar and M. A. Shayman, *Non-blocking supervisory control of nondeterministic systems via prioritized synchronization*, IEEE Trans. Automat. Control, 41 (1995), pp. 1160–1175.
[20] F. Lin and W. M. Wonham, *Decentralized supervisory control of discrete event systems*, Inform. Sci., 44 (1988), pp. 199–224.
[21] F. Lin and W. M. Wonham, *On observability of discrete-event systems*, Inform. Sci., 44 (1988), pp. 173–198.
[22] F. Lin and W. M. Wonham, *Decentralized control and coordination of discrete-event systems with partial observation*, IEEE Trans. Automat. Control, 35 (1990), pp. 1330–1337.
[23] R. Milner, *A Calculus of Communicating Systems*, Springer-Verlag, New York, 1980.
[24] R. Milner, *Communication and Concurrency*, Prentice-Hall, New York, 1989.
[25] I. Phillips, *Refusal testing*, Theoret. Comput. Sci., 50 (1987), pp. 241–284.
[26] P. J. Ramadge and W. M. Wonham, *Supervisory control of a class of discrete event processes*, SIAM J. Control Optim., 25 (1987), pp. 206–230.
[27] P. J. Ramadge and W. M. Wonham, *The control of discrete event systems*, Proc. IEEE: Special Issue on Discrete Event Systems, 77 (1989), pp. 81–98.
[28] K. Rudie and W. M. Wonham, *The infimal prefix closed and observable superlanguage of a given language*, Systems Control Lett., 15 (1990), pp. 361–371.
[29] K. Rudie and W. M. Wonham, *Think globally, act locally: Decentralized supervisory control*, IEEE Trans. Automat. Control, 37 (1992), pp. 1692–1708.
[30] M. Shayman and R. Kumar, *Supervisory control of nondeterministic systems with driven events via prioritized synchronization and trajectory models*, SIAM J. Control Optim., 33 (1995), pp. 469–497.
[31] R. J. van Glabbeek, *Comparative Concurrency Semantics, with Refinement of Actions*, Ph.D. thesis, Free University of Amsterdam, 1990.
[32] Y. Willner and M. Heymann, *Supervisory control of concurrent discrete-event systems*, Internat. J. Control, 54 (1991), pp. 1143–1169.

SIAM J. CONTROL OPTIM.
Vol. 35, No. 2, pp. 384–398, March 1997

002

ON CONTROLLABILITY CONCEPTION FOR STOCHASTIC SYSTEMS*

AGAMIRZA E. BASHIROV† AND KERIM R. KERIMOV‡

Abstract. The controllability notions for partially observed stochastic systems are defined. Their relation with complete and approximate controllabilities is shown. In particular, it is proven that the approximate controllability condition is necessary and the complete controllability condition is sufficient for the partially observed linear Gaussian control system to attain the arbitrarily small neighborhood of each point in the state space with probability arbitrarily closely to one.

Key words. controllability, stochastic system, separation principle

AMS subject classifications. Primary, 93B, 93E; Secondary, 60G, 49N

PII. S0363012994260970

1. Introduction. A quite complete theory of the controllability for the deterministic linear systems exists (see, for example, Curtain and Pritchard [1]). At the same time there have been only several attempts to introduce a controllability notion for stochastic systems, which do not agree in general with either complete or approximate controllabilities when a deterministic system is considered as a stochastic system with zero noise, and to obtain the conditions of its contents. For example, the stochastic ε-controllability with probability p, defined in Sunahara et al. [2], cannot be reduced to the known controllability notions for the deterministic systems.

In Bashirov and Hajiyev [3, 4] the approximate and complete controllability notions for the deterministic systems and the stochastic controllability from [2] were combined. In this paper, using the approach from [3, 4]—that is, based on the separation principle—we introduce the controllability notions for partially observed stochastic systems, show their relation with complete and approximate controllabilities, and study them for the linear systems. In particular, in this paper it is proven that the approximate controllability condition is necessary and the complete controllability condition is sufficient for the partially observed linear Gaussian control system to attain the arbitrarily small neighborhood of each point in the state space with probability arbitrarily closely to one.

2. Notations. In this paper X and Y are the real separable Hilbert spaces. R^n denotes the n-dimensional real Euclidean space. The closure of the set D is denoted by $\overline{D}$.

The space of all linear bounded operators on X to Y is denoted by $\mathcal{L}(X,Y)$. The brief notation $\mathcal{L}(X)=\mathcal{L}(X,X)$ is used as well. A^* denotes the adjoint to the operator A. The trace of the operator A is denoted by $\mathrm{tr}A$. If $A\in\mathcal{L}(X)$ is self-adjoint and $\langle h,Ah\rangle\geq 0$ (respectively, $\langle h,Ah\rangle\geq c\|h\|^2$, where $c=\text{const.}>0$) for all $h\in X$,

*Received by the editors January 3, 1994; accepted for publication (in revised form) December 12, 1995. A version of this paper was presented at the 35th IEEE Conference on Decision and Control, Kobe, Japan, December 11–13, 1996.

http://www.siam.org/journals/sicon/35-2/26097.html

†Department of Mathematics, Eastern Mediterranean University, Gazi Magusa, Mersin 10, Turkey, and Institute of Cybernetics, Azerbaijan Academy of Sciences, F. Agayev St. 9, Baku 370141, Azerbaijan (rza@mozart.as.emu.edu.tr).

‡Institute of Cybernetics, Azerbaijan Academy of Sciences, F. Agayev St. 9, Baku 370141, Azerbaijan.

then we write $A \geq 0$ (respectively, $A > 0$), where $\langle \cdot\,, \cdot \rangle$ is an inner product and $\| \cdot \|$ is a norm.

Always it is supposed that $(\Omega, \mathcal{F}, \mathbf{P})$ is a complete probability space, and two time moments are given. The initial time moment is identified with zero. The terminal moment is denoted by T. The notation $\mathbf{T} = [0, T]$ is used for the finite time interval on which all random and nonrandom processes will be considered. $L_2(\mathbf{T}, X)$ (respectively, $L_2(\Omega, X)$) denotes the space of equivalence classes of all functions on $\mathbf{T}$ (respectively, Ω) to X that are Lebesgue measurable (respectively, $\mathcal{F}$-measurable) and square integrable with respect to the Lebesgue measure (respectively, measure $\mathbf{P}$).

The notation $\Delta = \{(t, s) : \ 0 \leq s \leq t \leq T\}$ is used for the triangular set over $\mathbf{T}$. $B_2(\Delta, \mathcal{L}(X, Y))$ denotes the class of all $\mathcal{L}(X, Y)$-valued functions on Δ that are strongly measurable and square integrable with respect to the Lebesgue measure on Δ (see Curtain and Ichikawa [5]).

All integrals of the abstract functions are in Bochner sense. For the expectation and the conditional expectation the notations $\mathbf{E}$ and $\mathbf{E}(\cdot\,|\,\cdot)$ are used, respectively. $\operatorname{cov}(x, y)$ is the covariance operator of the random variables x and y. The brief notation $\operatorname{cov} x = \operatorname{cov}(x, x)$ is used as well. The integrals of the operator valued functions (except the stochastic integrals) are in strong Bochner sense.

3. Main definitions. Consider a control system on $\mathbf{T}$. Let x_t^u be its (random or not) state value at time $t \in \mathbf{T}$ corresponding to the control u taken from the set of the admissible controls U. If the considered control system is stochastic, then by $\mathcal{F}^u$ we denote the smallest σ-algebra generated by the observations on the time interval $\mathbf{T}$ corresponding to the control u. Suppose that X is the state space. Introduce the following sets:

$$D = \{x_T^u : u \in U\}, \tag{1}$$

$$S(\varepsilon, p) = \{h \in X : \exists u \in U \ \mathbf{P}\{\|\mathbf{E}(x_T^u|\mathcal{F}^u) - h\|^2 > \varepsilon\} \leq 1 - p\}, \tag{2}$$

$$C(\varepsilon, p) = \{h \in X : \exists u \in U \ h = \mathbf{E}x_T^u, \ \mathbf{P}\{\|\mathbf{E}(x_T^u|\mathcal{F}^u) - h\|^2 > \varepsilon\} \leq 1 - p\}, \tag{3}$$

where $0 \leq \varepsilon < \infty$ and $0 \leq p \leq 1$ are the parameters.

DEFINITION 1. *If $D = X$ (respectively, $\overline{D} = X$), then the corresponding deterministic control system will be called* D^c-controllable (*respectively,* D^a-controllable).

DEFINITION 2. *If $0 \in S(\varepsilon, p)$, then the corresponding stochastic control system will be called* $S^0_{\varepsilon,p}$-controllable.

DEFINITION 3. *If $S(\varepsilon, p) = X$ (respectively, $\overline{S(\varepsilon, p)} = X$), then the corresponding stochastic control system will be called* $S^c_{\varepsilon,p}$-controllable (*respectively,* $S^a_{\varepsilon,p}$-controllable).

It is clear that D^c- and D^a-controllabilities are the well-known complete and approximate controllabilities for the deterministic systems, respectively. The $S^0_{\varepsilon,p}$-controllability is a generalization of the ε-controllability with probability p, defined in Sunahara et al. [2], to the partially observed stochastic systems.

The geometric interpretation of the $S^c_{\varepsilon,p}$ (respectively, $S^a_{\varepsilon,p}$)-controllability is as follows. If a stochastic system with the initial (random or not) state x_0 is $S^c_{\varepsilon,p}$ (respectively, $S^a_{\varepsilon,p}$)-controllable, then with probability not less than p it can pass from x_0 for the time T into the $\sqrt{\varepsilon}$-neighborhood of the arbitrary point of the state space

(respectively, the set that is dense in the state space). The $S^0_{\varepsilon,p}$-controllability means that the hitting probability into the $\sqrt{\varepsilon}$-neighborhood of zero is not less than p. The smaller ε is and the larger p is for the stochastic system, the more controllable it is; i.e., it is possible to hit into the smaller neighborhood with higher probability. In particular, if a D^c (respectively, D^a)-controllable deterministic system is considered as stochastic with zero noise, then this system is $S^c_{0,1}$ (respectively, $S^a_{0,1}$)-controllable (with parameters $\varepsilon = 0$ and $p = 1$). At the same time it is clear that all stochastic systems are $S^c_{\varepsilon,p}$- and $S^a_{\varepsilon,p}$-controllable with $\varepsilon \geq 0$ and $p = 0$ or $\varepsilon = \infty$ and $0 \leq p \leq 1$ if we admit ∞ as a value for ε.

It is also to be noted that if a given stochastic system is $S^c_{\varepsilon,p}$ (respectively, $S^a_{\varepsilon,p}$)-controllable, then it is also $S^c_{\varepsilon_1,p_1}$ (respectively, $S^a_{\varepsilon_1,p_1}$)-controllable, where $\varepsilon \leq \varepsilon_1 < \infty$ and $0 \leq p_1 \leq p$. So, if ε is given for some system, it becomes important to find the largest value of p with which the system is $S^c_{\varepsilon,p}$ (respectively, $S^a_{\varepsilon,p}$)-controllable. Similarly, if there is given p, then it is worth finding the smallest ε.

We also introduce the following stronger controllability notions.

DEFINITION 4. *If $C(\varepsilon,p) = X$ (respectively, $\overline{C(\varepsilon,p)} = X$), then the corresponding stochastic control system will be called* $C^c_{\varepsilon,p}$-controllable (*respectively,* $C^a_{\varepsilon,p}$-controllable).

It is clear that $C^c_{\varepsilon,p}$ (respectively, $C^a_{\varepsilon,p}$)-controllability implies $S^c_{\varepsilon,p}$ (respectively, $S^a_{\varepsilon,p}$) -controllability, but the converses are not true in general. The geometric interpretation of the $C^c_{\varepsilon,p}$- and $C^a_{\varepsilon,p}$-controllabilities differs from that of $S^c_{\varepsilon,p}$- and $S^a_{\varepsilon,p}$-controllabilities since among the controls, with the help of which the $\sqrt{\varepsilon}$-neighborhood of any point h is achieved, there exists one with property that the expectation of the target state, corresponding to this control, coincides with h.

To introduce the next controllability notion we need the following lemma.

LEMMA 1. *A stochastic system is $S^a_{\varepsilon,p}$-controllable for all $\varepsilon > 0$ and $0 \leq p < 1$ if and only if it is $S^c_{\varepsilon,p}$-controllable for all $\varepsilon > 0$ and $0 \leq p < 1$.*

Proof. The sufficiency is obvious. Let us prove the necessity. Suppose that a given stochastic control system is $S^a_{\varepsilon,p}$-controllable for all $\varepsilon > 0$ and $0 \leq p < 1$. Let $S(\varepsilon,p)$ be the set (2) corresponding to this system. We have $\overline{S(\varepsilon,p)} = X$ for all $\varepsilon > 0$ and $0 \leq p < 1$, where X is the state space. We should show that the stronger condition $S(\varepsilon,p) = X$ for all $\varepsilon > 0$ and $0 \leq p < 1$ holds. Fix arbitrary $\varepsilon_0 > 0$, $0 \leq p_0 < 1$ and $h \in X$. Since $\overline{S(\varepsilon,p)} = X$ for all $\varepsilon > 0$ and $0 \leq p < 1$, there is $h_0 \in S(\varepsilon_0/4, p_0)$ such that $\|h_0 - h\|^2 \leq \varepsilon_0/4$. At the same time, since $h_0 \in S(\varepsilon_0/4, p_0)$, there exists $u \in U$ with

$$\mathbf{P}\{\|\mathbf{E}(x^u_T|\mathcal{F}^u) - h_0\|^2 > \varepsilon_0/4\} \leq 1 - p_0.$$

Hence, for this $u \in U$, we have

$$\mathbf{P}\{\|\mathbf{E}(x^u_T|\mathcal{F}^u) - h\|^2 > \varepsilon_0\} \leq \mathbf{P}\{\|\mathbf{E}(x^u_T|\mathcal{F}^u) - h_0\| + \|h_0 - h\| > \sqrt{\varepsilon_0}\}$$

$$\leq \mathbf{P}\{\|\mathbf{E}(x^u_T|\mathcal{F}^u) - h_0\| + \sqrt{\varepsilon_0}/2 > \sqrt{\varepsilon_0}\} = \mathbf{P}\{\|\mathbf{E}(x^u_T|\mathcal{F}^u) - h_0\|^2 > \varepsilon_0/4\} \leq 1 - p_0.$$

Thus, $h \in S(\varepsilon_0, p_0)$. Since $\varepsilon_0 > 0$, $0 \leq p_0 < 1$ and $h \in X$ are arbitrary, we have $S(\varepsilon,p) = X$ for all $\varepsilon > 0$ and $0 \leq p < 1$. The lemma is proven.

DEFINITION 5. *A stochastic control system will be called* S-controllable *if it is $S^c_{\varepsilon,p}$-controllable (or, equivalently, $S^a_{\varepsilon,p}$-controllable) for all $\varepsilon > 0$ and $0 \leq p < 1$.*

Obviously, S-controllability is independent on parameters ε and p. Moreover, by Lemma 1, the complete and approximate versions of this controllability are equivalent.

S-controllability is the main object of our consideration. The geometric interpretation of S-controllability is as follows: an S-controllable stochastic control system can attain the arbitrarily small neighborhood of each point in the state space with probability arbitrarily closely to one.

Finally, it is to be noted that in the previously introduced controllability notions the abbreviations D, S, and C mean deterministic, stochastic, and combined, respectively.

4. Separation theorem. In this section we shall prove that in the case of linear systems the $C^c_{\varepsilon,p}$ (respectively, $C^a_{\varepsilon,p}$)-controllability is equivalent to the $S^0_{\varepsilon,p}$- and D^c (respectively, D^a)-controllabilities combined.

Suppose that A is the infinitesimal generator of a strongly continuous semigroup $\mathcal{U}$; $B \in \mathcal{L}(Y,X)$; $C \in \mathcal{L}(X,R^k)$; $f \in L_2(\mathbf{T},X)$; $x_0 \in L_2(\Omega,X)$ is a Gaussian random variable with $\mathrm{cov}x_0 = P_0$; m and n are X- and R^k-valued Wiener processes, respectively; $n_0 = 0$; $m_0 = 0$; $\mathbf{E}n_t = 0$; $\mathbf{E}m_t = 0$; $\mathrm{cov}n_t = It$; I is the unit $(k \times k)$-matrix, $\mathrm{cov}m_t = Mt$; M is a nuclear operator on X; and x_0, n, m are mutually independent. Consider the linear partially observed stochastic system

$$\left\{ \begin{array}{ll} dx^u_t = (Ax^u_t + Bu_t + f_t)dt + dm_t, & 0 < t \le T, \quad x^u_0 = x_0, \\ d\xi^u_t = Cx^u_t dt + dn_t, & 0 < t \le T, \quad \xi^u_0 = 0, \end{array} \right. \tag{4}$$

where x, u, and ξ are the state, control, and observation processes. Under a set U of the admissible controls we consider the set of all controls in the linear form

$$u_t = \bar{u}_t + \int_0^t K_{t,s} d\xi^u_s, \tag{5}$$

where $K \in B_2(\Delta, \mathcal{L}(R^k,Y))$, $\bar{u} \in L_2(\mathbf{T},Y)$.

Note that under the above and some additional conditions the random function x, defined by

$$x_t = \mathcal{U}_t x_0 + \int_0^t \mathcal{U}_{t-s}(B_s u_s + f_s)ds + \int_0^t \mathcal{U}_{t-s} dm_s, \qquad 0 \le t \le T,$$

satisfies the equation in (4) and stands for its unique solution (see Curtain and Pritchard [1]). On the other hand, the function x_t, $0 \le t \le T$, is well defined even if these additional conditions do not hold. According to the theory of the differential equations in Banach spaces the function x_t, $0 \le t \le T$, is called the mild solution of the equation in (4) that becomes the solution (in the ordinary sense) when the abovementioned additional conditions hold. Below under the solution of the equation in (4) we shall keep in mind its mild solution.

One can associate two systems with the system (4). The first of them is the deterministic system

$$\frac{d}{dt} y^v_t = Ay^v_t + Bv_t + f_t, \qquad 0 < t \le T, \ y^v_0 = y_0 = \mathbf{E}x_0, \tag{6}$$

where v is a control from $V = \{v : v_t = \mathbf{E}u_t, \ u \in U\}$. The second one is the partially observed stochastic system

$$\left\{ \begin{array}{ll} dz^w_t = (Az^w_t + Bw_t)dt + dm_t, & 0 < t \le T, \quad z^w_0 = z_0 = x_0 - \mathbf{E}x_0, \\ d\eta^w_t = Cz^w_t dt + dn_t, & 0 < t \le T, \quad \eta^w_0 = 0, \end{array} \right. \tag{7}$$

where w is a control from $W = \{w : w_t = u_t - \mathbf{E}u_t, \ u \in U\}$.

Note that under the solution of the equations in (6) and (7) we shall also mean their mild solution.

LEMMA 2. *Under the above conditions and notation the following equalities hold:*
(a) $V = L_2(\mathbf{T}, Y)$;
(b) $W = \{w : w_t = \int_0^t K_{t,s} d\eta_s^w,\ K \in B_2(\Delta, \mathcal{L}(R^k, Y))\}$.

Proof. Item (a) can be proven by easy verification. To prove item (b) let $w \in W$. Then there exists such $u \in U$ that $w = u - \mathbf{E}u$. Let u be in the form (5) with $K \in B_2(\Delta, \mathcal{L}(R^k, Y))$ and $\bar{u} \in L_2(\mathbf{T}, Y)$. Then

$$\begin{aligned} w_t = u_t - \mathbf{E}u_t &= \int_0^t K_{t,s} C(x_s^u - \mathbf{E}x_s^u) ds + \int_0^t K_{t,s} dn_s \\ &= \int_0^t K_{t,s} C z_s^w ds + \int_0^t K_{t,s} dn_s = \int_0^t K_{t,s} d\eta_s^w. \end{aligned}$$

On the other hand, if

$$w_t = \int_0^t K_{t,s} d\eta_s^w, \tag{8}$$

then for u, which has the representation (5), we have $w = u - \mathbf{E}u$. So item (b) is true.

LEMMA 3. *Under the above conditions and notation the equality $U = V + W$ holds, where $+$ is the sign of the sum of the sets.*

Proof. Suppose $v \in V$, $w \in W$, and w has the form (8). Let $u = v + w$. Then

$$\begin{aligned} u_t &= v_t + \int_0^t K_{t,s} C z_s^w ds + \int_0^t K_{t,s} dn_s \\ &= v_t - \int_0^t K_{t,s} C y_s^v ds + \int_0^t K_{t,s} C x_s^u ds + \int_0^t K_{t,s} dn_s. \end{aligned}$$

Denote

$$\bar{u}_t = v_t - \int_0^t K_{t,s} C y_s^v ds. \tag{9}$$

Then u has the form of (5) with $\bar{u}$ as in (9), i.e., $u \in U$. On the other hand, each element of U can be shown as a sum of some elements taken from V and W. So $U = V + W$.

LEMMA 4. *Under the above conditions and notation, if $u = v + w$, $v \in V$, $w \in W$, then the σ-algebras $\mathcal{F}^{u,\xi}$ and $\mathcal{F}^{w,\eta}$, generated by ξ_s^u, $0 \le s \le T$, and η_s^w, $0 \le s \le T$, respectively, are equal.*

Proof. Using $u = v + w$ with $v = \mathbf{E}u$ it is easy to show that

$$\xi_t^u = \eta_t^w + C \int_0^t y_s^v ds, \qquad 0 \le t \le T. \tag{10}$$

Since the second term in the right-hand side of (10) is nonrandom, we conclude that ξ_s^u, $0 \le s \le T$, and η_s^w, $0 \le s \le T$, generate the same σ-algebra.

THEOREM 1. *Under the above conditions and notation the system* (4) *is $C^c_{\varepsilon,p}$ (respectively, $C^a_{\varepsilon,p}$)-controllable if and only if the system* (6) *is D^c (respectively, D^a)-controllable and the system* (7) *is $S^0_{\varepsilon,p}$-controllable.*

Proof. Let $C(\varepsilon, p)$ be the set (3) corresponding to the system (4). Similarly, let D be the set (1) corresponding to the system (6). Suppose that the system (4) is

$C^c_{\varepsilon,p}$ (respectively, $C^a_{\varepsilon,p}$)-controllable. Then, from the inclusion $C(\varepsilon,p)\subset D$, it follows that the system (6) is D^c (respectively, D^a)-controllable. Let $h\in C(\varepsilon,p)$. Then there exists $u\in U$ such that $h=\mathbf{E}x^u_T$ and

$$\mathbf{P}\{\|\mathbf{E}(x^u_T|\mathcal{F}^{u,\xi})-h\|^2>\varepsilon\}\le 1-p.$$

Consider $w=u-\mathbf{E}u\in W$. By Lemma 4, we have $\mathcal{F}^{u,\xi}=\mathcal{F}^{w,\eta}$. Therefore,

$$\mathbf{P}\{\|\mathbf{E}(z^w_T|\mathcal{F}^{w,\eta})\|^2>\varepsilon\}=\mathbf{P}\{\|\mathbf{E}(x^u_T|\mathcal{F}^{u,\xi})-\mathbf{E}x^u_T\|^2>\varepsilon\}\le 1-p;$$

i.e., the system (7) is $S^0_{\varepsilon,p}$-controllable. So the necessity is proven. To prove the sufficiency let $h\in D$. Then there exists $v\in V$ such that $h=y^v_T$. From the $S^0_{\varepsilon,p}$-controllability of the system (7), we conclude that there exists $w\in W$ with

$$\mathbf{P}\{\|\mathbf{E}(z^w_T|\mathcal{F}^{w,\eta})\|^2>\varepsilon\}\le 1-p.$$

Consider $u=v+w$. By Lemma 3, $u\in U=V+W$. Moreover,

$$\mathbf{P}\{\|\mathbf{E}(x^u_T|\mathcal{F}^{u,\xi})-h\|^2>\varepsilon\}=\mathbf{P}\{\|\mathbf{E}(z^w_T|\mathcal{F}^{w,\eta})\|^2>\varepsilon\}\le 1-p,$$

i.e., $h\in C(\varepsilon,p)$. Therefore, $D\subset C(\varepsilon,p)$. As $D=X$ (respectively, $\overline{D}=X$), then $C(\varepsilon,p)=X$ (respectively, $\overline{C(\varepsilon,p)}=X$). Thus, the system (4) is $C^c_{\varepsilon,p}$ (respectively, $C^a_{\varepsilon,p}$)-controllable.

Theorem 1 separates the study of the $C^c_{\varepsilon,p}$ (respectively, $C^a_{\varepsilon,p}$)-controllability of the general system (4) into the study of the D^c (respectively, D^a)-controllability and the $S^0_{\varepsilon,p}$-controllability of the systems (6) and (7), respectively. The D^c- and D^a-controllabilities of the linear system (6) on the set $L_2(\mathbf{T},Y)$ of admissible controls are investigated in a number of papers. Therefore, we mention only the following results from Curtain and Pritchard [1, pp. 56, 60] that will be used later.

THEOREM 2. *Under the above conditions and notation the following statements hold:*

(a) *the system* (6) *is D^c-controllable if and only if*

$$\int_t^T \mathcal{U}_{T-s}BB^*\mathcal{U}^*_{T-s}ds>0,\qquad 0\le t<T;$$

(b) *the system* (6) *is D^a-controllable if and only if $B^*\mathcal{U}^*_t x=0$ implies $x=0$ for all $t\in\mathbf{T}$.*

5. Sufficient condition for $S^0_{\varepsilon,p}$-controllability. Consider the system (7) under the above conditions and notation. Let Q^i and P be the solutions of the following operator Riccati equations, respectively:

$$\frac{d}{dt}Q_t+Q_tA+A^*Q_t-iQ_tBB^*Q_t=0,\qquad 0\le t<T,\ Q_T=I,\ i=1,2,\dots, \tag{11}$$

$$\frac{d}{dt}P_t-AP_t-P_tA^*-M+P_tC^*CP_t=0,\qquad 0<t\le T,\ P_0=\operatorname{cov}z_0, \tag{12}$$

where I is the identity operator. Note that under the solution of (11) we mean the operator-valued function Q that for all $x,y\in D(A)$ satisfies

$$\frac{d}{dt}\langle Q_tx,y\rangle+\langle Q_tAx,y\rangle+\langle Q_tx,Ay\rangle-i\langle Q_tBB^*Q_tx,y\rangle=0,\qquad 0\le t<T,$$

where $D(A)$ is the domain of A. The same sense is applied to the solution of (12). It follows from Curtain and Pritchard [1] that these equations have the unique strongly continuous solutions Q^i and P with $Q_t^i \geq 0$ and $P_t \geq 0$ for all $t \in \mathbf{T}$.

LEMMA 5. *Under the above conditions and notation the equality*

$$\inf_W \mathbf{E}\|\mathbf{E}(z_T^w|\mathcal{F}^{w,\eta})\|^2 = \lim_{i\to\infty} \int_0^T \text{tr } CP_sQ_s^iP_sC^*ds \tag{13}$$

holds, where Q^i and P are the solutions of (11) *and* (12)*, respectively, and there exists a finite limit in the right-hand side of* (13)*.*

Proof. First we shall prove the existence of the finite limit. Consider the family of the stochastic optimal control problems on W with the state and observation systems defined by (7) and the functional

$$J^i(w) = \mathbf{E}\left(\|z_T^w\|^2 + i^{-1}\int_0^T \|w_t\|^2 dt\right), \qquad i = 1, 2, \dots,$$

to be minimized. It is known (see Curtain and Ichikawa [5]) that there exists the unique optimal control $w^i \in W$ in the considered optimal control problem and

$$J^i(w^i) = \text{tr}P_T + \int_0^T \text{tr } CP_sQ_s^iP_sC^*ds.$$

Since P_T is the covariance of the error $z_T^w - \mathbf{E}(z_T^w|\mathcal{F}^{w,\eta})$ (see Curtain [6]) independently on w,

$$\mathbf{E}\|z_T^w\|^2 - \mathbf{E}\|\mathbf{E}(z_T^w|\mathcal{F}^{w,\eta})\|^2 = \mathbf{E}\|z_T^w - \mathbf{E}(z_T^w|\mathcal{F}^{w,\eta})\|^2 = \text{tr}P_T.$$

Therefore, if we denote

$$\tilde{J}^i(w) = \mathbf{E}\left(\|\mathbf{E}(z_T^w|\mathcal{F}^{w,\eta})\|^2 + i^{-1}\int_0^T \|w_t\|^2 dt\right),$$

then

$$\tilde{J}^i(w^i) = J^i(w^i) - \text{tr}P_T = \int_0^T \text{tr } CP_sQ_s^iP_sC^*ds. \tag{14}$$

Let us show that $\tilde{J}^i(w^i)$ does not increase as $i \to \infty$. Let $j \geq i$. Then

$$\begin{aligned}
\tilde{J}^j(w^j) &= \mathbf{E}\left(\|z_T^{w^j}\|^2 + j^{-1}\int_0^T \|w_t^j\|^2 dt\right) - \text{tr}P_T \\
&\leq \mathbf{E}\left(\|z_T^{w^i}\|^2 + j^{-1}\int_0^T \|w_t^i\|^2 dt\right) - \text{tr}P_T \\
&\leq \mathbf{E}\left(\|z_T^{w^i}\|^2 + i^{-1}\int_0^T \|w_t^i\|^2 dt\right) - \text{tr}P_T = \tilde{J}^i(w^i).
\end{aligned}$$

We conclude that $\tilde{J}^i(w^i)$, $i = 1, 2, \dots$, is a nonnegative and nonincreasing sequence. Therefore, there exists a finite limit of $\tilde{J}^i(w^i)$ as $i \to \infty$. From (14), it follows that there exists a finite limit in the right-hand side of (13). Now let us show that (13) is true. Indeed

$$\inf_W \mathbf{E}\|\mathbf{E}(z_T^w|\mathcal{F}^{w,\eta})\|^2 \leq \tilde{J}^i(w^i) \leq \mathbf{E}\left(\|\mathbf{E}(z_T^{\tilde{w}^r}|\mathcal{F}^{\tilde{w}^r,\eta})\|^2 + i^{-1}\int_0^T \|\tilde{w}_t^r\|^2 dt\right), \tag{15}$$

where $\tilde{w}^r$, $r = 1, 2, \ldots$, is a minimizing sequence of the functional

$$J_0(w) = \mathbf{E}\|\mathbf{E}(z_T^w|\mathcal{F}^{w,\eta})\|^2. \tag{16}$$

Consequently, taking the limit in (15) as $i \to \infty$ and $r \to \infty$, we obtain the equality (13). The lemma is proven.

Denote

$$a = \lim_{i\to\infty} \int_0^T \operatorname{tr} CP_sQ_s^iP_sC^*ds. \tag{17}$$

THEOREM 3. *Under the above conditions and notation the system* (7) *is* $S^0_{\varepsilon,p}$*-controllable if*

$$a < \varepsilon(1-p), \tag{18}$$

where a *is defined by* (17).

Proof. By Lemma 5, we have

$$\inf_W \mathbf{E}\|\mathbf{E}(z_T^w|\mathcal{F}^{w,\eta})\|^2 = a < \varepsilon(1-p).$$

Therefore, there exists $w^0 \in W$ such that

$$\mathbf{E}\|\mathbf{E}(z_T^{w^0}|\mathcal{F}^{w^0,\eta})\|^2 < \varepsilon(1-p).$$

Using Chebyshev's inequality, we obtain

$$\mathbf{P}\{\|\mathbf{E}(z_T^{w^0}|\mathcal{F}^{w^0,\eta})\|^2 > \varepsilon\} \le \frac{1}{\varepsilon}\mathbf{E}\|\mathbf{E}(z_T^{w^0}|\mathcal{F}^{w^0,\eta})\|^2 \le 1-p.$$

Hence, the theorem is proven.

It should be noted that the condition (18) that is the sufficient condition for $S^0_{\varepsilon,p}$-controllability is not necessary in general. In view of this we present the following arguments. Define the following functions for a given system:

$$\varphi_p = \inf \Phi_p,\ \Phi_p = \{\varepsilon :\ \text{the system is } S^0_{\varepsilon,p}\text{-controllable}\}, \tag{19}$$

$$\psi_\varepsilon = \sup \Psi_\varepsilon,\ \Psi_\varepsilon = \{p :\ \text{the system is } S^0_{\varepsilon,p}\text{-controllable}\}. \tag{20}$$

Obviously, φ and ψ are the nondecreasing functions and $\varphi_0 = 0$, $\lim_{\varepsilon\to\infty} \psi_\varepsilon = 1$. It follows from the definitions that the necessary and sufficient condition for the system to be $S^0_{\varepsilon,p}$-controllable is

$$\begin{cases} \varphi_p < \varepsilon & \text{if } \inf \Phi_p \text{ is not achieved,} \\ \varphi_p \le \varepsilon & \text{if } \inf \Phi_p \text{ is achieved,} \end{cases} \tag{21}$$

which can be written in the following equivalent form:

$$\begin{cases} \psi_\varepsilon > p & \text{if } \sup \Psi_\varepsilon \text{ is not achieved,} \\ \psi_\varepsilon \ge p & \text{if } \sup \Psi_\varepsilon \text{ is achieved.} \end{cases} \tag{22}$$

Using (18), define the functions

$$\tilde{\varphi}_p = \begin{cases} a(1-p)^{-1}, & 0 \le p < 1, \\ \infty, & p = 1, \end{cases} \qquad \tilde{\psi}_\varepsilon = \begin{cases} 1 - a\varepsilon^{-1}, & a < \varepsilon < \infty, \\ 0, & 0 \le \varepsilon \le a. \end{cases}$$

By (21), (22), and Theorem 3, it follows that

$$\varphi_p \le \tilde{\varphi}_p,\ 0 \le p \le 1,\ \text{and}\ \psi_\varepsilon \ge \tilde{\psi}_\varepsilon,\ 0 \le \varepsilon < \infty;$$

i.e., in the case of the system (7) the functions $\tilde{\varphi}$ and $\tilde{\psi}$, defined with the help of (18), give only approximations of the functions φ and ψ and may not be equal to them. In the case $\varphi_p < \tilde{\varphi}_p$ or $\psi_\varepsilon > \tilde{\psi}_\varepsilon$ the condition (18) cannot be necessary for $S^0_{\varepsilon,p}$-controllability. But it turns out that the condition (18), being sufficient for $S^0_{\varepsilon,p}$-controllability of the system (7), is weaker than the D^c-controllability of the system (6). We shall prove this result in the next section.

6. Necessary and sufficient condition for $C^c_{\varepsilon,p}$-controllability. Using the special form of the Riccati equation (11), we can present its solution in the following explicit form.

LEMMA 6. *Under the above conditions and notation the Riccati equation* (11) *has a solution in the form*

$$Q^i_t = \mathcal{U}^*_{T-t}\left(I + i\int_t^T \mathcal{U}_{T-s}BB^*\mathcal{U}^*_{T-s}ds\right)^{-1}\mathcal{U}_{T-t}, \qquad 0 \le t \le T,\ i = 1, 2, \dots . \tag{23}$$

Proof. First note that the right-hand side of (23) is a composition of three operators, each depending on time t. The first and third of them satisfy

$$\frac{d}{dt}\mathcal{U}_{T-t}x = -A\mathcal{U}_{T-t}x = -\mathcal{U}_{T-t}Ax, \qquad x \in D(A). \tag{24}$$

But the middle one is the inverse of the operator that is strongly differentiable in t. According to the rule for the derivative of the inverse operator, we have

$$\frac{d}{dt}\left(I + i\int_t^T \mathcal{U}_{T-s}BB^*\mathcal{U}^*_{T-s}ds\right)^{-1}$$

$$= i\left(I + i\int_t^T \mathcal{U}_{T-s}BB^*\mathcal{U}^*_{T-s}ds\right)^{-1}\mathcal{U}_{T-t}BB^*\mathcal{U}^*_{T-t}\left(I + i\int_t^T \mathcal{U}_{T-s}BB^*\mathcal{U}^*_{T-s}ds\right)^{-1}. \tag{25}$$

Thus, for all $x, y \in D(A)$, from (23), we have

$$\begin{aligned}
\frac{d}{dt}\langle Q^i_t x, y\rangle &= \frac{d}{dt}\left\langle \left(I + i\int_t^T \mathcal{U}_{T-s}BB^*\mathcal{U}^*_{T-s}ds\right)^{-1}\mathcal{U}_{T-t}x, \mathcal{U}_{T-t}y\right\rangle \\
&= \left\langle \left(\frac{d}{dt}\left(I + i\int_t^T \mathcal{U}_{T-s}BB^*\mathcal{U}^*_{T-s}ds\right)^{-1}\right)\mathcal{U}_{T-t}x \right. \\
&\quad + \left(I + i\int_t^T \mathcal{U}_{T-s}BB^*\mathcal{U}^*_{T-s}ds\right)^{-1}\frac{d}{dt}\mathcal{U}_{T-t}x, \mathcal{U}_{T-t}y\Big\rangle \\
&\quad \left. + \left\langle \left(I + i\int_t^T \mathcal{U}_{T-s}BB^*\mathcal{U}^*_{T-s}ds\right)^{-1}\mathcal{U}_{T-t}x, \frac{d}{dt}\mathcal{U}_{T-t}y\right\rangle\right.
\end{aligned} \tag{26}$$

Substituting (24) and (25) in (26) and using (23), one can easily show that Q^i, defined by (23), is a solution of (11).

LEMMA 7. *Under the above conditions and notation the D^c-controllability condition for the system* (6) *from Theorem* 2(a) *implies*

$$a = \lim_{i\to\infty} \int_0^T \operatorname{tr}\, CP_sQ_s^iP_sC^*ds = 0.$$

Proof. By D^c-controllability condition, for all $0 \le t < T$ we have

$$\left\langle x, \left(I + i\int_t^T \mathcal{U}_{T-s}BB^*\mathcal{U}_{T-s}^*ds\right)x\right\rangle$$

$$= \|x\|^2 + i\left\langle x, \int_t^T \mathcal{U}_{T-s}BB^*\mathcal{U}_{T-s}^*xds\right\rangle \ge (1+ib_t)\|x\|^2,$$

where $b_t > 0$ for all $0 \le t < T$. Therefore,

$$\left\|\left(I + i\int_t^T \mathcal{U}_{T-s}BB^*\mathcal{U}_{T-s}^*ds\right)^{-1}\right\| \le (1+ib_t)^{-1},$$

and by (23),

$$\|Q_t^i\| \le \|\mathcal{U}_{T-t}\|^2(1+ib_t)^{-1}.$$

The last inequality implies

$$\|Q_t^i\| \le \|\mathcal{U}_{T-t}\|^2,\ i = 1,2,\dots,\ 0\le t\le T; \qquad \|Q_t^i\| \to 0,\ i\to\infty,\ 0\le t<T. \tag{27}$$

Thus, applying majorized convergence theorem, we obtain

$$a = \lim_{i\to\infty}\int_0^T \operatorname{tr}\, CP_sQ_s^iP_sC^*ds \le \lim_{i\to\infty}\int_0^T (\operatorname{tr}P_s)^2\|C\|^2\|Q_s^i\|ds$$

$$= \int_0^T (\operatorname{tr}P_s)^2\|C\|^2 \lim_{i\to\infty}\|Q_s^i\|ds = 0.$$

THEOREM 4. *Under the above conditions and notation the system* (4) *is $C^c_{\varepsilon,p}$-controllable for all $\varepsilon > 0$ and $0 \le p < 1$ if and only if the system* (6) *is D^c-controllable.*

Proof. The necessity follows from Theorem 1. Suppose that the system (6) is D^c-controllable. Using Theorem 2(a), we obtain that the condition of Lemma 7 holds. Therefore, $a = 0$, and, applying Theorem 3, we get the $S^0_{\varepsilon,p}$-controllability for all ε and p satisfying $\varepsilon(1-p) > 0$ for the system (7). Note that the condition $\varepsilon(1-p) > 0$ includes all pairs (ε, p) with $\varepsilon > 0$ and $0 \le p < 1$. Finally, by Theorem 1, we obtain $C^c_{\varepsilon,p}$-controllability of the system (4) for all $\varepsilon > 0$ and $0 \le p < 1$. The theorem is proven.

Remark. One may ask, is the analogue of Theorem 4 true in the approximate controllability case; i.e., is $C^a_{\varepsilon,p}$-controllability of the system (4) for all $\varepsilon > 0$ and $0 \le p < 1$ equivalent to D^a-controllability of the system (6)? The necessity part again holds in view of Theorem 1. Problems arise in proving the sufficiency. The sufficiency part of Theorem 4 is based on Lemma 7, in the proof of which the uniform operator convergence (27) under the D^c-controllability condition from Theorem 2(a)

was used. There is an example (see section 7) constructed by the reviewer of this article which shows that, if the exact controllability condition in Lemma 7 is replaced with the approximate controllability condition (see Theorem 2(b)), then the uniform operator convergence (27) is not true. Nevertheless, it might be possible to prove the convergence $Q_t^i \to 0,\ i \to \infty$, in the strong (or weak) operator topology. This problem needs further investigation.

7. Necessary and sufficient conditions for *S*-controllability. Now we can consider the main controllability notion for the stochastic systems that was defined in Definition 5.

LEMMA 8. *Under the above conditions and notation let w be the random process defined by* (8)*, where $K \in B_2(\Delta, \mathcal{L}(R^k, Y))$ and η^w is defined by* (7)*. Then there exists $M \in B_2(\Delta, \mathcal{L}(R^k, Y))$ such that*

$$w_t = \int_0^t M_{t,s} d\eta_s^0, \tag{28}$$

where η^0 is the observation process of the system (7) *corresponding to the zero-control. Conversely, if w is defined by* (28) *with $M \in B_2(\Delta, \mathcal{L}(R^k, Y))$, then there exists $K \in B_2(\Delta, \mathcal{L}(R^k, Y))$ such that w has the representation* (8)*.*

Proof. The direct statement is proven in Curtain [6]. The converse will be proven in the same way as in Curtain [6]. Suppose w has the form (28). It is easy to observe that there exists the following relation between η^w and η^0:

$$d\eta_s^0 = d\eta_s^w - \int_0^s C\mathcal{U}_{s-r} B w_r dr ds. \tag{29}$$

Substituting (29) in (28) we have

$$w_t = \int_0^t M_{t,s} d\eta_s^w - \int_0^t \int_0^s M_{t,s} C\mathcal{U}_{s-r} B w_r dr ds. \tag{30}$$

Equation (30) is a Volterra integral equation with respect to w which has the kernel

$$L_{t,r} = -\int_r^t M_{t,s} C\mathcal{U}_{s-r} B ds.$$

It is known that (30) has a solution in the form

$$w_t = \int_0^t M_{t,s} d\eta_s^w + \int_0^t N_{t,s} \int_0^s M_{s,r} d\eta_r^w ds$$

for some $N \in B_2(\Delta, \mathcal{L}(Y))$. Applying the stochastic analogue of the Fubini theorem (see Curtain and Pritchard [1]) we obtain that w has the form (8) with

$$K_{t,s} = M_{t,s} + \int_s^t N_{t,r} M_{r,s} dr.$$

So $w \in W$. The lemma is proven.

LEMMA 9. *Under the above conditions and notation the set U of admissible controls of the system* (4) *is convex.*

Proof. Suppose $u^1, u^2 \in U$ and $\lambda_1 > 0, \lambda_2 > 0$ with $\lambda_1 + \lambda_2 = 1$. Consider $u = \lambda_1 u^1 + \lambda_2 u^2$. We have

$$u_t^i = v_t^i + \int_0^t M_{t,s}^i d\eta_s^0, \qquad i = 1, 2,$$

for some $v^1, v^2 \in L_2(\mathbf{T}, Y)$ and $M^1, M^2 \in B_2(\Delta, \mathcal{L}(R^k, Y))$. Denoting $v = \lambda_1 v^1 + \lambda_2 v^2$ and $M = \lambda_1 M^1 + \lambda_2 M^2$, we have

$$u_t = v_t + \int_0^t M_{t,s} d\eta_s^0$$

for $v \in L_2(\mathbf{T}, Y)$ and $M \in B_2(\Delta, \mathcal{L}(R^k, Y))$. It means that $u \in U$ and, therefore, U is convex. The lemma is proven.

THEOREM 5. *Under the above conditions and notation,*

(a) D^c*-controllability of the system* (6) *is sufficient for the* S*-controllability of the system* (4)*;*

(b) D^a*-controllability of the system* (6) *is necessary for the* S*-controllability of the system* (4)*.*

Proof. By Theorem 4, we obtain that the D^c-controllability of the system (6) implies $C^c_{\varepsilon,p}$-controllability of the system (4) for all $\varepsilon > 0$ and $0 \le p < 1$. Since $C(\varepsilon, p) \subset S(\varepsilon, p)$ for an arbitrary system, where $C(\varepsilon, p)$ and $S(\varepsilon, p)$ are defined by (3) and (2), then $C(\varepsilon, p) = X$ implies $S(\varepsilon, p) = X$. Therefore, the system (4) is $S^c_{\varepsilon,p}$-controllable for all $\varepsilon > 0$ and $0 \le p < 1$, which means the S-controllability of the system (4). This proves item (a). Let us prove item (b). Suppose that the system (4) is S-controllable, i.e., $S^a_{\varepsilon,p}$-controllable for all $\varepsilon > 0$ and $0 \le p < 1$. To prove the D^a-controllability of the system (6) let us consider arbitrary $h \in X$. We shall show that there exists a sequence $\{\tilde{u}^n\}$ in U such that $\|\mathbf{E}x_T^{\tilde{u}^n} - h\| \to 0$ as $n \to \infty$, where x_t^u is the state of the system (4) at time t corresponding to control $u \in U$. Consider the sequences $\{\varepsilon_n\}$ and $\{p_n\}$ with $\varepsilon_n > 0$, $0 \le p_n < 1$ and $\varepsilon_n \to 0$, $p_n \to 1$ as $n \to \infty$. Then from $S^a_{\varepsilon_n,p_n}$-controllability of the system (4) we obtain the existence of the sequence $\{u^n\}$ in U such that

$$\mathbf{P}\{\|\mathbf{E}(x_T^{u^n} | \mathcal{F}^{u^n,\xi}) - h\|^2 > \varepsilon_n\} \le 1 - p_n. \tag{31}$$

Inequality (31) implies the convergence in probability of $\|\mathbf{E}(x_T^{u^n} | \mathcal{F}^{u^n,\xi}) - h\|$ to zero. Indeed, for any $\varepsilon > 0$ we can find a number N such that $0 < \varepsilon_n < \varepsilon^2$ for all $n > N$. Therefore, for $n > N$ we have

$$\mathbf{P}\{\|\mathbf{E}(x_T^{u^n} | \mathcal{F}^{u^n,\xi}) - h\| > \varepsilon\} \le \mathbf{P}\{\|\mathbf{E}(x_T^{u^n} | \mathcal{F}^{u^n,\xi}) - h\|^2 > \varepsilon_n\} \le 1 - p_n \to 0, \qquad n \to \infty.$$

Hence, $\mathbf{E}(x_T^{u^n} | \mathcal{F}^{u^n,\xi})$ converges to h in probability. Since $\mathbf{E}(x_T^{u^n} | \mathcal{F}^{u^n,\xi})$ is a Gaussian random variable for all n, the characteristic functions of these random variables have the form (see Vakhania [7])

$$\chi_n(x) = \exp\left(i\langle m_n, x\rangle - \frac{1}{2}\langle \Lambda_n x, x\rangle\right), \qquad x \in X, \tag{32}$$

where $m_n = \mathbf{E}(\mathbf{E}(x_T^{u^n} | \mathcal{F}^{u^n,\xi})) = \mathbf{E}x_T^{u^n}$ and $\Lambda_n = \mathrm{cov}\mathbf{E}(x_T^{u^n} | \mathcal{F}^{u^n,\xi})$. Also, the vector $h \in X$ is considered as a degenerate Gaussian random variable with characteristic function

$$\chi(x) = \exp\Big(i\langle h, x\rangle\Big), \qquad x \in X. \tag{33}$$

The convergence of $\mathbf{E}(x_T^{u^n}|\mathcal{F}^{u^n,\xi})$ to h in probability implies $\chi_n(x) \to \chi(x)$ for all $x \in X$. The last convergence is possible when

$$\langle m_n, x\rangle = \langle \mathbf{E}x_T^{u^n}, x\rangle \to \langle h, x\rangle, \quad \langle \Lambda_n x, x\rangle \to 0, \quad n \to \infty.$$

First of these convergences means the convergence of $\mathbf{E}x_T^{u^n}$ to h in the weak topology of the Hilbert space X. By Mazur's theorem (see Balakrishnan [8]) we can construct a sequence

$$h_n = \sum_{i=1}^{n} c_i^n \mathbf{E}x_T^{u^i}, \quad c_i^n \geq 0, \quad \sum_{i=1}^{n} c_i^n = 1, \quad i = 1, 2, \dots, n, \quad n = 1, 2, \dots,$$

of convex combinations of $\mathbf{E}x_T^{u^n}$ such that h_n converges to h in the strong topology of X. Denote $\tilde{u}_n = \sum_{i=1}^{n} c_i^n u^i$, $n = 1, 2, \dots$. By Lemma 9, $\tilde{u}_n \in U$ for all n. Moreover, in view of the affineness of the system (4), $h_n = \mathbf{E}x_T^{\tilde{u}^n}$. In terms of the system (6) it means that, for the sequence of controls $\tilde{v}_n = \mathbf{E}\tilde{u}_n$ in V, the sequence of vectors $h_n = \mathbf{E}x_T^{\tilde{u}^n} = y_T^{\tilde{v}^n}$ converges to h in the strong topology of X. Since h is an arbitrary point of X, we conclude that the set D defined by (1) for the system (6) is dense in X. The theorem is proven.

8. Discussion. One can call the stochastic system (4) D^c (respectively, D^a)-controllable if the associated deterministic system (6) has the same property. Theorem 5 indicates the following implications between D^c-, D^a-, and S-controllability properties for the system (4):

$$D^c \;\Rightarrow\; S \;\Rightarrow\; D^a, \tag{34}$$

in which S-controllability has the middle position between the stronger D^c and weaker D^a-controllabilities. The following example shows that in general the reverse implication $D^c \Leftarrow S$ does not hold. Note that this example was constructed by the reviewer of this article to show that if D^c-controllability in Lemma 7 is replaced with the D^a-controllability condition, then the uniform operator convergence (27) does not hold.

Let $X = Y = l_2$ (the Hilbert space of numerical sequences $\{x_n\}$ satisfying $\sum_{n=1}^{\infty} x_n^2 < \infty$ with scalar product $\langle\{x_n\}, \{y_n\}\rangle = \sum_{n=1}^{\infty} x_n y_n$), $\mathcal{U}_t \equiv I$ (the identity operator) and

$$B = \begin{pmatrix} 1 & 0 & 0 & \cdots \\ 0 & \frac{1}{2} & 0 & \cdots \\ 0 & 0 & \frac{1}{3} & \cdots \\ \cdot & \cdot & \cdot & \cdots \end{pmatrix}.$$

Consider the basis $e_1 = (1, 0, 0, \dots)$, $e_2 = (0, 1, 0, \dots)$, $e_3 = (0, 0, 1, \dots), \dots$ in l_2. Since

$$\sum_{n=1}^{\infty} \langle Be_n, Be_n\rangle = \sum_{n=1}^{\infty} \frac{1}{n^2} < \infty,$$

then B is a Hilbert–Schmidt operator on l_2 (the notation is $B \in \mathcal{L}_2(l_2)$) and, therefore, $B \in \mathcal{L}(l_2)$. Obviously, $B = B^*$. Also, $B^*\mathcal{U}_t^* x = 0$ implies $Bx = 0$ and, hence, $x = 0$. So the system (4) with $\mathcal{U}$ and B defined as above is D^a-controllable (see Theorem

2(b)). But it is not D^c-controllable since $\|B^2 e_n\|_{l_2} = n^{-2} \to 0$ as $n \to \infty$, and, therefore, the operator

$$\int_t^T \mathcal{U}_{T-s} B B^* \mathcal{U}_{T-s}^* ds = (T-t)B^2$$

is not positive (see Theorem 2(a)).

Let us show that the system (4) with $\mathcal{U}$ and B defined as above is S-controllable. Consider Q^i defined by (23). For above $\mathcal{U}$ and B, we have

$$Q_t^i x = (I + i(T-t)B^2)^{-1} x = \left\{ \frac{n^2 x_n}{n^2 + i(T-t)} \right\}, \qquad x = \{x_n\} \in l_2.$$

Therefore,

$$\|Q_t^i x\|^2 = \sum_{i=1}^{\infty} \frac{n^4 x_n^2}{(n^2 + i(T-t))^2}.$$

The last series is majorized by the convergent series $\sum_{n=1}^{\infty} x_n^2$. Hence, for $0 \le t < T$, we have

$$\lim_{i\to\infty} \|Q_t^i x\|^2 = \lim_{i\to\infty} \sum_{i=1}^{\infty} \frac{n^4 x_n^2}{(n^2 + i(T-t))^2} = \sum_{i=1}^{\infty} \lim_{i\to\infty} \frac{n^4 x_n^2}{(n^2 + i(T-t))^2} = 0.$$

We obtain that $Q_t^i \to 0$ as $i \to \infty$ for all $0 \le t < T$ in the strong operator topology. Now consider a defined by (17):

$$a = \lim_{i\to\infty} \int_0^T \operatorname{tr}\, C P_s Q_s^i P_s C^* ds = \lim_{i\to\infty} \int_0^T \sum_{i=1}^{\infty} \langle Q_s^i P_s C^* e_n, P_s C^* e_n \rangle ds. \tag{35}$$

In (35) we can change the places of the limit, integration, and summation operations since $Q_s^i \le I$ for all i and for all $0 \le s \le T$. Therefore,

$$\sum_{i=1}^{\infty} \langle Q_s^i P_s C^* e_n, P_s C^* e_n \rangle \le \sum_{i=1}^{\infty} \|P_s C^* e_n\|^2 = \operatorname{tr}\, C P_s P_s C^* \le (\operatorname{tr} P_s)^2 \|C\|^2 < \infty.$$

Hence, from (35), by strong operator convergence $Q_t^i \to 0$ as $i \to \infty$, for all $0 \le t < T$, we obtain

$$a = \int_0^T \sum_{i=1}^{\infty} \langle \lim_{i\to\infty} Q_s^i P_s C^* e_n, P_s C^* e_n \rangle ds = 0.$$

Finally, having $a = 0$ and following the proof of Theorem 4 (sufficiency part) and Theorem 5(a), the S-controllability of the system (4) can be obtained.

This example indicates that in (34) there is the middle case of $a = 0$ between D^c- and S-controllability conditions; i.e., (34) could be completed as

$$D^c \;\Rightarrow\; a = 0 \;\Rightarrow\; S \;\Rightarrow\; D^a.$$

For completeness a counterexample showing the nonvalidity of the implication $D^a \Rightarrow S$ would be also presented. We do not present such an example for two reasons. First, it is not easy to construct such an example even if it exists. Second, the authors tend to think that in the case of the system (4) the equivalence $S \Leftrightarrow D^a$ takes place. Of course, this subject needs further study.

Acknowledgments. The authors would like to thank the reviewers of this paper for their useful recommendations. In particular, the results of section 6 were predicted by them.

REFERENCES

[1] R. F. Curtain and A. J. Pritchard, *Infinite Dimensional Linear Systems Theory*, Lecture Notes on Control and Inform. Sci. 8, Springer-Verlag, New York, 1978.
[2] Y. Sunahara, T. Kabeuchi, Y. Asada, S. Aihara, and K. Kishino, *On stochastic controllability for nonlinear systems*, IEEE Trans. Automat. Control, 19 (1974), pp. 49–54.
[3] A. E. Bashirov and R. R. Hajiyev, *On controllability of the partially observed stochastic systems*, 1, Izv. Akad. Nauk Azerbaidzhan. SSR Ser. Fiz.-Tekhn.-Mat. Nauk, 4 (1983), pp. 109–114 (in Russian).
[4] A. E. Bashirov and R. R. Hajiyev, *On controllability of the partially observed stochastic systems*, 2, Izv. Akad. Nauk Azerbaidzhan. SSR Ser. Fiz.-Tekhn.-Mat. Nauk, 5 (1984), pp. 99–103 (in Russian).
[5] R. F. Curtain and A. Ichikawa, *The separation principle for stochastic evolution equations*, SIAM J. Control Optim., 15 (1977), pp. 367–383.
[6] R. F. Curtain, *Estimation theory for abstract evolution equations excited by general white noise processes*, SIAM J. Control Optim., 14 (1976), pp. 1124–1150.
[7] N. N. Vakhania, *Probability Distributions on Linear Spaces*, Elsevier–North-Holland, New York, Oxford, 1981.
[8] A. V. Balakrishnan, *Applied Functional Analysis*, Springer-Verlag, New York, 1976.

SIAM J. CONTROL OPTIM.
Vol. 35, No. 2, pp. 399–434, March 1997

003

DETERMINISTIC EXIT TIME CONTROL PROBLEMS WITH DISCONTINUOUS EXIT COSTS*

ALAIN-PHILIPPE BLANC†

Abstract. We study deterministic exit time control problems with discontinuous exit costs. When the exit cost φ is upper semicontinuous and there is an outer field on the boundary, we show that all the value functions have the same lower semicontinuous envelope which is the unique lower semicontinuous viscosity solution of the associated Dirichlet problem. We also prove uniqueness results for the generalized Dirichlet problem for first-order Hamilton–Jacobi equations with convex Hamiltonians and with discontinuous boundary conditions, under some nondegeneracy conditions on the Hamiltonians on the boundary.

Key words. exit time problems, Hamilton–Jacobi equations, Dirichlet problems, discontinuous data, viscosity solutions

AMS subject classifications. 49L05, 49L20, 49L25, 35B37, 35F30, 35R05

PII. S0363012994267340

1. Introduction.

This paper is concerned with deterministic exit time control problems with discontinuous exit costs. We want to characterize the value functions of this type of control problems as the unique viscosity solutions of the corresponding Hamilton–Jacobi–Bellman problem. We obtain, under some suitable conditions, the uniqueness of the lower semicontinuous (l.s.c.) envelope of the value function.

In order to be more specific, we now describe the optimal control problem. Let Ω be a smooth bounded domain of $\mathbb{R}^N$. We consider a system whose state is given by the solution of the ordinary differential equation

$$\begin{cases} dy_x(t) &= b(y_x(t), \alpha(t))dt, \\ y_x(0) &= x \in \overline{\Omega}, \end{cases}$$

where b is a Lipschitz continuous function from $\overline{\Omega} \times \mathcal{A}$ into $\mathbb{R}^N$ and $\alpha(.) \in L^\infty(\mathbb{R}^+, \mathcal{A})$ is the control; $\mathcal{A}$, the control space, is a compact metric space. We denote by τ the first exit time of the trajectory y_x from Ω, i.e.,

$$\tau \;=\; \inf\{t \geq 0,\; y_x(t) \notin \Omega\}.$$

In this framework, it is well known that the value function can be defined in several ways: the first value function that we want to consider is given by

$$u(x) \;=\; \inf_{\alpha(.) \in L^\infty(\mathbb{R}^+, \mathcal{A})} \left\{ \int_0^\tau f(y_x(t), \alpha(t)) e^{-\lambda t} dt + \varphi(y_x(\tau)) e^{-\lambda \tau} \right\},$$

where f is a given continuous real-valued function and λ is a positive constant, the discount factor. Precise assumptions on f and b are detailed in the first part. The main point is that we assume only that φ, the exit cost, is a bounded function on $\partial\Omega$ defined pointwise; in particular, it may present discontinuities.

*Received by the editors May 9, 1994; accepted for publication (in revised form) December 12, 1995.

http://www.siam.org/journals/sicon/35-2/26734.html

†Faculté des Sciences et Techniques, Université de Tours, Parc de Grandmont, 37200 Tours, France (blanc@univ-tours.fr).

It is well known that when the value function is continuous, it is a viscosity solution of the corresponding Hamilton–Jacobi–Bellman equation

$$H(x,u,Du) \;=\; 0 \quad \text{in } \Omega,$$

where

$$H(x,t,p) \;=\; \sup_{\alpha\in\mathcal{A}}\{-b(x,\alpha)\cdot p + \lambda t - f(x,\alpha)\}. \tag{1}$$

This is a consequence of the so-called dynamic programming principle (cf. [17]). We recall that the notion of viscosity solutions was introduced in [11] by M. G. Crandall and P. L. Lions (see also [10]). We refer the reader to [12, 17], where the applications of viscosity solutions to deterministic and stochastic control problems are described.

But the value function u may be discontinuous even when the exit cost φ is continuous. Therefore, we have to use the notion of discontinuous viscosity solution, which was introduced by H. Ishii [13, 14] and requires the concepts of l.s.c. and upper semicontinuous (u.s.c.) envelopes of functions. In all the following discussion, ξ_* (resp., ξ^*) will denote the l.s.c. (resp., u.s.c.) envelope of the locally bounded function ξ, which is defined by

$$\xi_*(x) \;=\; \liminf_{y\to x}\, \xi(y) \quad (\text{resp.,} \quad \xi^*(x) \;=\; \limsup_{y\to x}\, \xi(y)).$$

G. Barles and B. Perthame [5] first considered the connections of such discontinuous solutions with optimal control problems.

Moreover, the concept of viscosity solutions is used to identify the appropriate boundary conditions satisfied by the value function. In the case of a continuous exit cost, H. Ishii [15] and G. Barles and B. Perthame [6] show the connections between the above control problem and the following Hamilton–Jacobi–type problem:

$$\begin{cases} H(x,u,Du) \;=\; 0 & \text{in } \Omega, \\ \min\{H(x,u,Du), u-\varphi\} \;\le\; 0 & \text{on } \partial\Omega, \\ \max\{H(x,u,Du), u-\varphi\} \;\ge\; 0 & \text{on } \partial\Omega. \end{cases} \tag{2}$$

When the function φ is discontinuous, the definition of viscosity solutions deals with the l.s.c. and the u.s.c. envelope of φ. For illustration, we say that the bounded function u is a viscosity subsolution of (2) on the boundary $\partial\Omega$ if

$$\text{for every function } \phi\in C^1(\overline{\Omega}), \quad \text{if } x\in\partial\Omega \;\text{ is a maximum point of } u^*-\phi, \quad \text{we have}$$
$$H(x,u^*(x),D\phi(x)) \;\le\; 0 \qquad \text{or} \qquad u^*(x) \;\le\; \varphi^*(x).$$

Our first aim is to prove that the value function u is a viscosity solution of (2).

Then we have to look at "uniqueness" (or characterization) properties for the viscosity solution u. In general, the discontinuous viscosity solution is not unique. One reason is that the Hamilton–Jacobi–Bellman equation is the same for the control problem and for the relaxed control problem. But the value function can be different for these two problems. Moreover, it is quite natural to consider different stopping times on the boundary: for example, the first exit time from the closed set $\overline{\Omega}$, i.e.,

$$\bar{\tau} \;=\; \inf\{t\ge 0,\; y_x(t)\notin\overline{\Omega}\}.$$

In this paper, since the exit cost φ is discontinuous, an additional choice is to use φ_* or φ^*, instead of taking φ, in the definition of u.

We first prove the existence of a minimum and of a maximum solution, which are given respectively by

$$u_-(x) \;=\; \inf\left\{\int_0^\theta \int_{\mathcal{A}} f(\hat{y}_x(t),\alpha)e^{-\lambda t}d\mu_t(\alpha)dt + \varphi_*(\hat{y}_x(\theta))e^{-\lambda\theta},\ \hat{\tau}\le\theta\le\hat{\bar{\tau}},\right.$$
$$\left.\hat{y}_x(\theta)\in\partial\Omega,\ \text{and } \mu\in L^\infty(\mathbb{R}^+,P(\mathcal{A}))\right\},$$

and

$$u^+(x) \;=\; \inf_{\alpha(.)\in L^\infty(\mathbb{R}^+,\mathcal{A})}\left\{\sup\left\{\int_0^\theta f(y_x(t),\alpha(t))e^{-\lambda t}dt + \varphi^*(y_x(\theta))e^{-\lambda\theta},\ \tau\le\theta\le\bar{\tau},\right.\right.$$
$$\left.\left.\text{and}\quad y_x(\theta)\in\partial\Omega\right\}\right\}.$$

It is clear from these definitions that one has

$$u_- \;\le\; u \;\le\; u^+ \quad \text{in } \Omega.$$

But these functions may be very different, in particular because of the trajectories which remain on $\overline{\Omega}$ but which touch the boundary several times ($\tau\neq\bar{\tau}$).

Nevertheless, in the case when the function φ is continuous, G. Barles and B. Perthame [7] showed that, essentially, if there exist outer and inner fields at each point of $\partial\Omega$, the value function u is continuous and is the unique solution of (2); hence, we get

$$u_- \;=\; u \;=\; u^+ \quad \text{in } \Omega. \tag{3}$$

The underlying idea is that, with these outer and inner fields, one can control a trajectory close to the boundary to make it stay in Ω or to leave $\overline{\Omega}$.

An other aim of this article is to give a similar uniqueness result when φ is discontinuous. But since the value functions may be discontinuous, we first have to explain how to interpret uniqueness of the solution in this case. We want the equalities (3) to hold again in a weaker sense, namely,

$$u_- \;=\; u_* \;=\; (u^+)_* \quad \text{in } \Omega,$$

since u_- is l.s.c. on $\overline{\Omega}$. Therefore uniqueness means in this context that all the solutions have the same l.s.c. envelope.

In the case when the exit cost φ satisfies

$$(\varphi^*)_* \;=\; \varphi_* \quad \text{on } \partial\Omega, \tag{4}$$

and when there is an outer field on the boundary, we show that

$$u_- \;=\; (u^+)_* \quad \text{in } \Omega$$

by working directly on the control formulas. And therefore, all the solutions of (2) have the same l.s.c. envelope since u_- and u^+ are, respectively, the minimum and the maximum solution of (2).

But we want also an uniqueness result without the condition (4) on φ since, for control problems, it is quite natural to consider exit costs φ which are only l.s.c. In fact, one of our motivations for this work is the exit problem from a part Γ of the boundary $\partial\Omega$. If, for instance, Γ is reduced to a point $\{x_0\}$, then φ is equal to 0 at x_0 and 1 elsewhere; therefore it does not satisfy (4). To solve this difficulty, we adapt PDE arguments introduced by E. N. Barron and R. Jensen [8, 9] for convex Hamiltonians.

Under a nondegeneracy condition on the Hamiltonian which we interpret as the existence of an outer field on the boundary for control problems, we prove that there exists a unique l.s.c. viscosity solution in the sense of Barron and Jensen in Ω which is a classical viscosity supersolution on the boundary and which satisfies, for every $x \in \partial\Omega$,

$$\liminf_{y\to x, y\in\Omega} u(y) \;\leq\; \varphi_*(x). \tag{5}$$

We emphasize that this uniqueness result is obtained by PDE arguments.

The application of this result to the control problem is the following: since P. Soravia [19] showed that all the value functions are viscosity solutions in the sense of Barron and Jensen, we have the uniqueness of the l.s.c. envelope of the value functions which satisfy the condition (5). Besides, when the condition (4) holds, we recover the uniqueness result obtained by working directly on the representation formulas of the value functions.

We also refer the reader interested in the discontinuous viscosity solution approach to Dirichlet problems to the work of A. I. Subbotin [20], M. Bardi and P. Soravia [1], P. Soravia [19], and G. Barles [3]. These authors consider mainly continuous Dirichlet function on the boundary. In the case of discontinuous data, the pioneering paper of Barron and Jensen [8] was concerned with the Cauchy problem in $\mathbb{R}^N$. G. Barles [2] extended their ideas to stationary optimal stopping time in $\mathbb{R}^N$. To the best of our knowledge, the exit time problems with discontinuous exit costs have not yet been considered in the literature with such a generality.

This paper is organized as follows: the first section is devoted to the study of the exit time problems and its connections with (2). We also introduce the condition on the behavior of the controlled vector field at the boundary to obtain an uniqueness result. In the second section, we describe the new idea for discontinuous viscosity solutions for convex Hamiltonians and prove uniqueness results in the case when φ is discontinuous by PDE arguments. The third section is devoted to applications of the uniqueness results. The two first parts are nearly independent.

2. The exit time control problems with discontinuous exit costs. We recall that Ω is a smooth bounded domain of $\mathbb{R}^N$ and that the state of the system is described by the solution of

$$\begin{cases} dy_x(t) &= b(y_x(t), \alpha(t))dt, \\ y_x(0) &= x \in \overline{\Omega} \end{cases}$$

or, in the case of relaxed controls, by

$$\begin{cases} d\hat{y}_x(t) &= \displaystyle\int_{\mathcal{A}} b(\hat{y}_x(t), \alpha) d\mu_t(\alpha)dt, \\ \hat{y}_x(0) &= x \in \overline{\Omega}, \end{cases}$$

where $\alpha(.) \in L^\infty(\mathbb{R}^+, \mathcal{A})$ and $\mu_. \in L^\infty(\mathbb{R}^+, P(\mathcal{A}))$; $\mathcal{A}$ is a compact metric space and $P(\mathcal{A})$ is the set of probability measures on $\mathcal{A}$. The function b is continuous from $\overline{\Omega}\times\mathcal{A}$

into $\mathbb{R}^N$ and satisfies

$$\text{(6)} \qquad \begin{cases} |b(x,\alpha)| \le C, \quad \forall x \in \overline{\Omega},\ \forall \alpha \in \mathcal{A}; \\ |b(x,\alpha) - b(y,\alpha)| \le C|x-y|, \quad \forall x,y \in \overline{\Omega},\ \forall \alpha \in \mathcal{A}. \end{cases}$$

These assumptions imply the existence and the uniqueness of the solution y_x for all $t > 0$.

In the same way as for τ and $\bar{\tau}$, we define $\hat{\tau}$ and $\hat{\bar{\tau}}$ for a relaxed trajectory $\hat{y}_x$ respectively by

$$\hat{\tau} = \inf\{t \ge 0,\ \hat{y}_x(t) \notin \Omega\} \quad \text{and} \quad \hat{\bar{\tau}} = \inf\{t \ge 0,\ \hat{y}_x(t) \notin \overline{\Omega}\}.$$

The associated cost functions are

$$J(x,\alpha,\theta,\varphi) = \int_0^\theta f(y_x(t),\alpha(t))e^{-\lambda t}dt + \varphi(y_x(\theta))e^{-\lambda\theta}$$

and

$$\hat{J}(x,\mu,\theta,\varphi) = \int_0^\theta \int_{\mathcal{A}} f(\hat{y}_x(t),\alpha)e^{-\lambda t}d\mu_t(\alpha)dt + \varphi(\hat{y}_x(\theta))e^{-\lambda\theta},$$

where λ is some positive constant and f is a continuous function from $\overline{\Omega} \times \mathcal{A}$ into $\mathbb{R}$ satisfying

$$\text{(7)} \qquad \begin{cases} |f(x,\alpha)| \le C \quad \forall x \in \overline{\Omega},\ \forall \alpha \in \mathcal{A}; \\ |f(x,\alpha) - f(y,\alpha)| \le C|x-y| \quad \forall x,y \in \overline{\Omega},\ \forall \alpha \in \mathcal{A}. \end{cases}$$

We use particularly the following three value functions:

$$u_-(x) = \inf_{\mu \in L^\infty(\mathbb{R}^+, P(\mathcal{A}))} \{\hat{J}(x,\mu,\theta,\varphi_*),\ \hat{\tau} \le \theta \le \hat{\bar{\tau}} \text{ and } \hat{y}_x(\theta) \in \partial\Omega\},$$

which is l.s.c. on $\overline{\Omega}$;

$$u^+(x) = \inf_{\alpha(.) \in L^\infty(\mathbb{R}^+, \mathcal{A})} \{\sup\{J(x,\alpha,\theta,\varphi^*),\ \tau \le \theta \le \bar{\tau} \text{ and } y_x(\theta) \in \partial\Omega\}\},$$

which is u.s.c. on $\overline{\Omega}$; and the value function already introduced in the introduction,

$$u[\varphi](x) = \inf_{\alpha(.) \in L^\infty(\mathbb{R}^+, \mathcal{A})} \{J(x,\alpha,\tau,\varphi)\}.$$

These three types of value functions have already been studied in [6, 13] when the function φ is continuous. In the case of a discontinuous exit cost, we introduce different exit cost functions, namely, φ_* and φ^*. Our main results are the following. First, we prove that the function u_- is the minimum supersolution of (2) with the Hamiltonian given by (1) and the function u^+ is the maximum subsolution. We show that the value functions $u[\varphi_*]$, $u[\varphi]$, and $u[\varphi^*]$ are viscosity solutions of (2), and an example shows that all these value functions may be very different.

In order to obtain the uniqueness of the l.s.c. envelope of the value functions, we are going to prove that

$$(u^+)_* = u_- \quad \text{on } \overline{\Omega}.$$

Therefore we have to show the following series of equalities:

$$(u^+)_* \;=\; (u[\varphi^*])_* \;=\; (u[\varphi_*])_* \;=\; u_- \quad \text{on } \overline{\Omega}.$$

The first equality is proven without additional assumptions; it already holds in the continuous case (see [6]). For the second, since we minimize, "interesting values" may be lost by taking the exit cost φ^* instead of φ_*: for instance, if φ is equal to 0 at one point and 1 elsewhere, then φ^* is equal to 1 everywhere. To avoid this difficulty in this section, we consider only "regular" exit costs, i.e., satisfying the property (4). Finally, the difficulty with the trajectories which touch the boundary several times is solved by assuming that there exists an outer field at every point on the boundary or that the boundary is such that there are only inner fields on $\partial\Omega$.

This section is divided in three subsections. In the first, we study the value functions u_- and u^+. The second is devoted to the properties of the function $u[\varphi]$. And, in the last one, we prove the uniqueness result.

2.1. The value functions u_- and u^+. In this subsection, we characterize u_- and u^+ as the minimal and maximal solutions of (2). We also prove the connection between u^+ and $u[\varphi^*]$.

THEOREM 2.1. *We assume that φ is a bounded function defined pointwise, the constant λ is positive, and the assumptions* (6) *and* (7) *hold.*

1. *The function u_- is l.s.c. and the function u^+ is u.s.c. on $\overline{\Omega}$.*
2. *The functions u^+ and u_- are viscosity solutions of* (2).
3. *The functions u^+ and u_- are, respectively, the maximal subsolution and the minimal supersolution of* (2).

Proof. We first consider the case of u^+. We introduce a nonincreasing sequence $(\varphi^n)_n$ of continuous functions such that

$$\inf_n \varphi^n \;=\; \varphi^*.$$

We note $u^+[\varphi^n]$, the value function defined by the same formula as u^+ except that the exit cost function φ^* is replaced by φ^n (in particular, $u^+[\varphi^*]$ is equal to u^+). Then we need the following lemma.

LEMMA 2.2. *We have*

$$\inf_n u^+[\varphi^n] \;=\; u^+[\varphi^*] \quad \textit{on } \overline{\Omega}.$$

Proof. It is clear enough that

$$u^+[\varphi^n] \;\geq\; u^+[\varphi^{n+1}] \;\geq\; u^+[\varphi^*] \quad \text{on } \overline{\Omega}, \tag{8}$$

and thus $\inf_n u^+[\varphi^n] \geq u^+[\varphi^*]$ on $\overline{\Omega}$.

To prove the opposite inequality, we use the definition of $u^+[\varphi^n]$. For any control $\alpha(.)$ and any point $x \in \overline{\Omega}$, we have

$$u^+[\varphi^n](x) \;\leq\; \sup\{J(x,\alpha,\theta,\varphi^n), \tau \leq \theta \leq \bar{\tau} \text{ and } y_x(\theta) \in \partial\Omega\}.$$

Now, for a fixed control $\alpha(.)$, we pick a sequence $(\theta_n)_n$ such that

$$u^+[\varphi^n](x) \;\leq\; \frac{1}{n} + \int_0^{\theta_n} f(y_x(t),\alpha(t))e^{-\lambda t}dt + \varphi^n(y_x(\theta_n))e^{-\lambda\theta_n}. \tag{9}$$

First case. If the sequence $(\theta_n)_n$ is bounded, considering a subsequence if necessary, we may assume that $\theta_n \to \bar{\theta}$. Taking the limit superior as $n \to \infty$ in the inequality (9), we get

$$\limsup_n\ u^+[\varphi^n](x)\ \le\ \int_0^{\bar{\theta}} f(y_x(t),\alpha(t))e^{-\lambda t}dt + \limsup_n\ \varphi^n(y_x(\theta_n))e^{-\lambda\bar{\theta}}. \tag{10}$$

But since $y_x(\theta_n) \to y_x(\bar{\theta})$, since the function φ^n is continuous and since $\inf_n\ \varphi^n = \varphi^*$, we easily deduce

$$\limsup_n\ \varphi^n(y_x(\theta_n))\ \le\ \varphi^*(y_x(\bar{\theta})).$$

Moreover, using (8), we obtain

$$\limsup_n\ u^+[\varphi^n](x)\ =\ \inf_n\ u^+[\varphi^n](x).$$

Combining these two results with (10), we get

$$\inf_n\ u^+[\varphi^n](x)\ \le\ \int_0^{\bar{\theta}} f(y_x(t),\alpha(t))e^{-\lambda t}dt + \varphi^*(y_x(\bar{\theta}))e^{-\lambda\bar{\theta}}\ =\ J(x,\alpha,\bar{\theta},\varphi^*),$$

and taking the supremum in $\bar{\theta}$ in the right-hand side, we have

$$\inf_n\ u^+[\varphi^n](x)\ \le\ \sup\{J(x,\alpha,\theta,\varphi^*), \tau \le \theta \le \bar{\tau},\ \text{and}\ y_x(\theta) \in \partial\Omega\}. \tag{11}$$

Second case. If the sequence $(\theta_n)_n$ is not bounded, then there exists a subsequence, still denoted $(\theta_n)_n$, such that $\theta_n \to +\infty$. The inequality (9) implies

$$\begin{aligned} u^+[\varphi^n](x) &\le \frac{1}{n} + \int_0^{\theta_n} f(y_x(t),\alpha(t))e^{-\lambda t}dt + \varphi^*(y_x(\theta_n))e^{-\lambda\theta_n} \\ &\quad + \varphi^n(y_x(\theta_n))e^{-\lambda\theta_n} - \varphi^*(y_x(\theta_n))e^{-\lambda\theta_n} \\ &\le \frac{1}{n} + J(x,\alpha,\theta_n,\varphi^*) + (\varphi^n(y_x(\theta_n)) - \varphi^*(y_x(\theta_n)))e^{-\lambda\theta_n}. \end{aligned}$$

Since φ^n and φ^* are bounded, the last term tends to zero and we conclude as before.

Hence, finally, since the inequality (11) holds for any control $\alpha(.)$ and for any point $x \in \overline{\Omega}$, we conclude

$$\inf_n\ u^+[\varphi^n](x)\ \le\ u^+[\varphi^*](x). \qquad \square$$

Now we deduce the properties of u^+ from this lemma. By using the results of [6], the function $u^+[\varphi^n]$ is u.s.c. Then the function u^+ is u.s.c. too since u^+ is equal to the limit of the nonincreasing sequence of function $u^+[\varphi^n]$. Moreover, again by the results of [6], $u^+[\varphi^n]$ is the maximal subsolution (and solution) of

$$\begin{cases} H(x,u,Du) &= 0 \quad \text{in } \Omega, \\ \min\{H(x,u,Du), u-\varphi^n\} &\le 0 \quad \text{on } \partial\Omega, \\ \max\{H(x,u,Du), u-\varphi^n\} &\ge 0 \quad \text{on } \partial\Omega. \end{cases} \tag{12}$$

Then, using standard stability results (cf. [5]), $u^+[\varphi^*]$ is a subsolution of (2) because, for every $x \in \overline{\Omega}$, we have

$$\limsup_{n\to\infty, y\to x} u^+[\varphi^n](y) \;=\; \inf_n u^+[\varphi^n](x) \;=\; u^+[\varphi^*](x)$$

and

$$\limsup_{n\to\infty, y\to x} \varphi^n(y) \;=\; \inf_n \varphi^n(x) \;=\; \varphi^*(x).$$

It is also a supersolution of (2) since we have, for every $x \in \overline{\Omega}$,

$$\liminf_{n\to\infty, y\to x} u^+[\varphi^n](y) \;=\; \liminf_{y\to x} u^+[\varphi^*](y) \;=\; (u^+[\varphi^*])_*(x)$$

and since

$$\begin{aligned}\liminf_{n\to\infty, y\to x} \varphi^n(y) \;&=\; (\inf_n \varphi^n)_*(x) = (\varphi^*)_*(x)\\ &\geq \varphi_*(x).\end{aligned}$$

Moreover, if w is a subsolution of (2), w is also a subsolution of (12) since $\varphi^n \geq \varphi^*$ on $\partial\Omega$. This implies, for every n, that

$$w \;\leq\; u^+[\varphi^n] \quad \text{on } \overline{\Omega},$$

because $u^+[\varphi^n]$ is the maximal subsolution of (12). Therefore, by Lemma 2.2, taking the infimum over n yields

$$w \;\leq\; u^+[\varphi^*] \quad \text{on } \overline{\Omega},$$

which implies that the function $u^+[\varphi^*]$ is a maximal subsolution of (2).

For the function u_-, we proceed exactly as for the value function u^+ . Let $(\varphi_n)_n$ be a nondecreasing sequence of continuous functions such that

$$\sup_n \varphi_n \;=\; \varphi_*.$$

Again we introduce the notation $u_-[\varphi_n]$ to denote the value function defined as u_- but with the exit cost φ_n instead of φ_*. Then we consider the following lemma, which corresponds to Lemma 2.2 for u^+.

LEMMA 2.3. *We have*

$$\sup_n u^+[\varphi_n] \;=\; u_-[\varphi_*] \quad \textit{on } \overline{\Omega}.$$

With this lemma, we conclude the proof by deducing the properties of $u_-[\varphi_*]$ from the properties of $u_-[\varphi_n]$ as before. □

Proof of Lemma 2.3. It is clear enough that

$$\sup_n u_-[\varphi_n] \;\leq\; u_-[\varphi_*] \quad \text{on } \overline{\Omega}.$$

It remains to prove the opposite inequality. For every $x \in \overline{\Omega}$, we consider a minimizing sequence $(\mu_n, \theta_n)_n$ for $u_-[\varphi_n]$ such that

$$u_-[\varphi_n](x) + \frac{1}{n} \;\geq\; \int_0^{\theta_n} \int_{\mathcal{A}} f(\hat{y}_x^n(t), \alpha) e^{-\lambda t} d\mu_n(\alpha) dt + \varphi_n(\hat{y}_x^n(\theta_n)) e^{-\lambda\theta_n}. \tag{13}$$

First case. If the sequence $(\theta_n)_n$ is bounded, considering a subsequence if necessary, we may assume that there exists a stopping time $\bar{\theta}$ such that $\theta_n \to \bar{\theta}$ as $n \to \infty$. Moreover, using classical arguments relying on the compactness of relaxed controls, we may also assume that $\mu_n \to \bar{\mu}$ weakly in $L^\infty(\mathbb{R}^+, P(\mathcal{A}))$ for some relaxed control $\bar{\mu}$. Let $\hat{y}_x$ be the relaxed trajectory associated with $\bar{\mu}$. Since $\hat{y}^n_x$ converges locally uniformly to $\hat{y}_x$, we can check that $\hat{y}_x(\bar{\theta}) \in \partial\Omega$ and $\hat{y}^n_x(\theta_n) \to \hat{y}_x(\bar{\theta})$.

Then, again using the local uniform convergence of $\hat{y}^n_x$ to $\hat{y}_x$ together with the fact that f is Lipschitz continuous, we replace the trajectory $\hat{y}^n_x$ by $\hat{y}_x$ in the integral of (13), i.e.,

$$u_-[\varphi_n](x) + \frac{1}{n} \;\geq\; \int_0^{\bar{\theta}} \int_{\mathcal{A}} f(\hat{y}_x(t), \alpha) e^{-\lambda t} d\bar{\mu}(\alpha) dt + \varphi_n(\hat{y}^n_x(\theta_n)) e^{-\lambda\bar{\theta}} + \varepsilon_n,$$

with a sequence of numbers ε_n such that $\varepsilon_n \to 0$ when $n \to \infty$. From the above inequality, we deduce

$$u_-[\varphi_n](x) + \frac{1}{n} \;\geq\; u_-[\varphi_*](x) + (\varphi_n(\hat{y}^n_x(\theta_n)) - \varphi_*(\hat{y}_x(\bar{\theta}))) e^{-\lambda\bar{\theta}} + \varepsilon_n.$$

Then, taking the limit inferior, we get

$$\liminf_n\, u_-[\varphi_n](x) \;\geq\; u_-[\varphi_*](x) + \liminf_n \{\varphi_n(\hat{y}^n_x(\theta_n)) - \varphi_*(\hat{y}_x(\bar{\theta}))\} e^{-\lambda\bar{\theta}}. \tag{14}$$

But since $\hat{y}^n_x(\theta_n) \to \hat{y}_x(\bar{\theta})$, since the function φ_n is continuous, and since $\sup_n \varphi_n = \varphi_*$, we easily deduce

$$\liminf_n\, \varphi_n(\hat{y}^n_x(\theta_n)) \;\geq\; \varphi_*(\hat{y}_x(\bar{\theta}));$$

thus, combining this with (14), we obtain

$$\begin{aligned} \liminf_n\, u_-[\varphi_n](x) &= \sup_n\, u_-[\varphi_n](x) \\ &\geq u_-[\varphi_*](x). \end{aligned}$$

Second case. If the sequence $(\theta_n)_n$ is not bounded, we may assume without loss of generality that $\theta_n \to \infty$, and then the inequality (13) implies

$$u_-[\varphi_n](x) + \frac{1}{n} \;\geq\; u_-[\varphi_*](x) + (\varphi_n(\hat{y}^n_x(\theta_n)) - \varphi_*(\hat{y}^n_x(\theta_n))) e^{-\lambda\theta_n}.$$

Since φ_* is bounded and since $(\varphi_n)_n$ is uniformly bounded, letting $n \to \infty$, we have the desired result. □

The following theorem shows that u^+ and $u[\varphi^*]$ are not very different.

THEOREM 2.4. *Under the same assumptions of Theorem* 2.1*, we have*

$$(u^+)_* \;=\; (u[\varphi^*])_* \qquad \text{in } \Omega.$$

Proof. We first remark that the proof is inspired from the corresponding one in [6].

It is clear that $(u^+)_* \geq (u[\varphi^*])_*$ in Ω since $u^+ \geq u[\varphi^*]$ on $\overline{\Omega}$.

It remains to prove the opposite inequality. For any point x of Ω, there exists a sequence $(x_n)_n$ of points of Ω such that $x_n \to x$ and

$$\lim_n u[\varphi^*](x_n) \;=\; (u[\varphi^*])_*(x).$$

For each x_n, we consider a control $\alpha_n(.)$ such that

$$u[\varphi^*](x_n) + \frac{1}{n} \ \geq \ J(x_n, \alpha_n, \tau_n, \varphi^*). \tag{15}$$

Then we need the following lemma.

LEMMA 2.5. *For every $x \in \Omega$ and $\alpha \in L^\infty(\mathbb{R}^+, \mathcal{A})$, there exists a sequence $(x_p)_p$ of points of Ω such that*

$$\liminf_p u^+(x_p) \ \leq \ J(x, \alpha, \tau, \varphi^*),$$

where τ is the first exit time of the trajectory y_x associated with the control α.

For every n, we apply this lemma to the point x_n and to the control α_n. Thus there exists a sequence $(x_n^p)_p$ such that $x_n^p \to x_n$ as $p \to \infty$ and

$$J(x_n, \alpha_n, \tau_n, \varphi^*) \ \geq \ u^+(x_n^p) - \frac{1}{p}.$$

Combining this with (15), we pass to the limit, and by a diagonal procedure, we get

$$\begin{aligned}(u[\varphi^*])_*(x) &= \lim_n u[\varphi^*](x_n) \\ &\geq (u^+)_*(x),\end{aligned}$$

which is the desired result. □

Proof of Lemma 2.5. If $\tau = \infty$, then $u^+(x) \leq J(x, \alpha, \tau, \varphi^*)$, and therefore it suffices to take $x_p := x$ for all p.

If $\tau \neq \infty$, we consider the map $Y_\tau : z \mapsto y_z(\tau)$, which is an homeomorphism from a neighborhood of x onto some neighborhood of $y_x(\tau)$. We introduce the domain D defined by $D := Y^{-1}(\mathcal{B}(y_x(\tau), \varepsilon) \cap \overline{\Omega}^c)$, where $\mathcal{B}(z, r)$ is the open ball centered at z and of radius r and $\overline{\Omega}^c$ is the complementary set of $\overline{\Omega}$ in $\mathbb{R}^N$. For ε small enough, the domain D is a nonempty open subset of $\mathbb{R}^N$ and the point x is in its closure $\overline{D}$.

Moreover, since the function b is Lipschitz continuous, classical ODE estimates yield

$$|y_z(s) - y_x(s)| \ \leq \ |z - x| e^{Cs} \quad \text{for every } s \in [0, \tau].$$

And, for every p, since we have

$$\inf \left\{ \operatorname{dist}(y_x(s), \partial\Omega), 0 \leq s \leq \tau - \frac{1}{p} \right\} \ = \ \delta \ > \ 0,$$

we consider the constant η equal to

$$\eta \ := \ \max \left\{ \frac{1}{p}, \delta \right\} e^{-C\tau},$$

and therefore we have, for $|z - x| < \eta$,

$$|y_z(s) - y_x(s)| \ < \ \frac{1}{p} \quad \text{for every } s \in [0, \tau], \tag{16}$$

and

$$y_z(s) \in \Omega \quad \text{for every } s \in \left[0, \tau - \frac{1}{p}\right].$$

Then we choose a point $x_p \in \mathcal{B}(x,\eta) \cap D$. We first remark that the first exit time τ_p of the trajectory y_{x_p} from Ω satisfies $\tau - \frac{1}{p} \le \tau_p$ since $x_p \in \mathcal{B}(x,\eta)$. The exit time $\bar{\tau}_p$ satisfies $\bar{\tau}_p \le \tau$ since $x_p \in D$. We may write

$$u^+(x_p) \ \le \ \sup\{J(x_p,\alpha,\theta,\varphi^*), \theta \in [\tau_p, \bar{\tau}_p] \text{ and } y_{x_p}(\theta) \in \partial\Omega\}.$$

Using the Lipschitz continuity of f and the inequality (16), we compute, for every $\theta \in [\tau_p, \bar{\tau}_p]$,

$$\begin{aligned} J(x_p,\alpha,\theta,\varphi^*) &\le J(x,\alpha,\tau,\varphi^*) - \varphi^*(y_x(\tau))e^{-\lambda\tau} + \varphi^*(y_{x_p}(\theta))e^{-\lambda\tau} \\ &\quad + \|\varphi\|_\infty |e^{-\lambda\theta} - e^{-\lambda\tau}| + \|f\|_\infty |\tau - \theta| + \frac{C\tau}{p} \\ &\le J(x,\alpha,\tau,\varphi^*) + \rho_{\varphi^*}(|y_x(\tau) - y_{x_p}(\theta)|) + \tilde{C}\left(|\theta - \tau| + \frac{1}{p}\right), \end{aligned} \tag{17}$$

where ρ_{φ^*} is a continuous function such that $\rho_{\varphi^*}(t) \to 0$ as $t \to 0^+$ and, for every $y \in \partial\Omega$,

$$\varphi^*(y) - \varphi^*(y_x(\tau)) \ \le \ \rho_{\varphi^*}(|y_x(\tau) - y|).$$

We pass to the limit inferior in (17), and again using (16) and the fact that

$$\tau - \frac{1}{p} \ \le \ \tau_p \ \le \ \theta \ \le \ \bar{\tau}_p \ \le \ \tau,$$

we get

$$\begin{aligned} J(x,\alpha,\tau,\varphi^*) &\ge \liminf_p J(x_p,\alpha,\theta,\varphi^*) \\ &\ge \liminf_p u^+(x_p). \end{aligned}$$

Thus the proof is complete. □

2.2. The properties of the value function $u[\varphi]$. In this section, we prove that the value functions $u[\varphi_*]$, $u[\varphi]$, and $u[\varphi^*]$ are viscosity solutions. Then we give an example which shows that these functions may be very different.

Let us begin by a result concerning the values of the solutions of (2) on the boundary.

PROPOSITION 2.6. *Let u be a bounded function from Ω into $\mathbb{R}$.*

We define

$$\check{u} \ = \ \begin{cases} u & \text{in } \Omega, \\ \varphi & \text{on } \partial\Omega, \end{cases} \qquad \text{and} \qquad \tilde{u} \ = \ \begin{cases} u & \text{in } \Omega, \\ \xi & \text{on } \partial\Omega, \end{cases}$$

where ξ is a real-valued function such that

$$\varphi_* \ \le \ \xi \ \le \ \varphi^* \quad \text{on } \partial\Omega.$$

Then $\check{u}$ is a subsolution (resp., supersolution) of (2) *if and only if $\tilde{u}$ is a subsolution (resp., supersolution) of* (2).

REMARK 2.7. *This proposition shows that the values of the solutions of* (2) *are not prescribed by the viscosity condition on the boundary. Thus the uniqueness results cannot be extended up to the boundary.*

Proof. We prove only the proposition in the subsolutions case. The other case may be obtained by the same method. Let us first remark that, for every $x \in \partial\Omega$, we have

$$\tilde{u}^*(x) \;=\; \sup\{u^*(x), \xi^*(x)\} \;\leq\; \check{u}^*(x) \;=\; \sup\{u^*(x), \varphi^*(x)\}. \tag{18}$$

We first assume that $\check{u}$ is a subsolution, and therefore we have to prove that $\tilde{u}$ is a subsolution too. Let $\phi \in C^1(\overline{\Omega})$ be a test function and $x_0 \in \partial\Omega$ be a maximum point of $\tilde{u}^* - \phi$. Changing ϕ in $\phi + (\tilde{u}^*(x_0) - \phi(x_0))$, we may assume without loss of generality that $\tilde{u}^*(x_0) = \phi(x_0)$ and thus $\tilde{u}^* \leq \phi$.

• If $\tilde{u}^*(x_0) \leq \varphi^*(x_0)$, there is nothing to prove.

• If $\tilde{u}^*(x_0) > \varphi^*(x_0)$, then by definition of $\tilde{u}$ and since we have $\xi \leq \varphi^*$ on $\partial\Omega$, this implies $u^*(x_0) > \varphi^*(x_0)$. Hence, using the inequality (18), we deduce that $\check{u}^*(x_0) = \tilde{u}^*(x_0)$.

We introduce the positive constant $\delta := \tilde{u}^*(x_0) - \varphi^*(x_0)$. Then, since $\tilde{u}^*(x_0) = \phi(x_0)$, we get

$$\varphi^*(x_0) + \frac{\delta}{2} \;=\; \phi(x_0) - \frac{\delta}{2}. \tag{19}$$

Since ϕ is continuous, there exists some constant $\varepsilon_1 > 0$ such that, for all $y \in \overline{\Omega}$ and $|x_0 - y| < \varepsilon_1$, we have

$$|\phi(x_0) - \phi(y)| \;<\; \frac{\delta}{2}.$$

And since φ^* is u.s.c., there exists some constant $\varepsilon_2 > 0$ such that, for all $y \in \partial\Omega$ and $|x_0 - y| < \varepsilon_2$, we have

$$\varphi^*(y) - \varphi^*(x_0) \;<\; \frac{\delta}{2}.$$

Hence, combining these two preceding estimates with the equality (19), we deduce

$$\varphi^*(y) \;<\; \varphi^*(x_0) + \frac{\delta}{2} \;=\; \phi(x_0) - \frac{\delta}{2} \;<\; \phi(y) \quad \text{for } y \in \partial\Omega \cap \mathcal{B}(x_0, \varepsilon_1 \wedge \varepsilon_2),$$

where $a \wedge b := \inf\{a, b\}$ for $a, b \in \mathbb{R}$. Combining this with the fact that

$$\check{u}^* \;\leq\; \phi \quad \text{for } y \in \Omega \cap \mathcal{B}(x_0, \varepsilon_1 \wedge \varepsilon_2),$$

since $u^* \leq \phi$ in Ω, we deduce that x_0 is also a maximum point of $\check{u}^* - \phi$. And thus $\tilde{u}^*$ is a subsolution.

Conversely, if $\tilde{u}$ is a subsolution, let $x_0 \in \partial\Omega$ be a maximum point of $\check{u}^* - \phi$ with $\phi \in C^1(\overline{\Omega})$. As before, we choose $\check{u}^*(x_0) = \phi(x_0)$.

• If $\check{u}^*(x_0) \leq \varphi^*(x_0)$, $\check{u}^*$ is a subsolution.

• If $\check{u}^*(x_0) > \varphi^*(x_0)$, then using the inequality (18) and the fact that $\xi^* \leq \varphi^*$ on $\partial\Omega$, we get

$$\check{u}^*(x_0) \;=\; u^*(x_0) \;=\; \tilde{u}^*(x_0).$$

Since $\check{u}^* \geq \tilde{u}^*$ on $\overline{\Omega}$, it is easy to show that $\check{u}^*$ is a subsolution. ∎

Now, we recall a general result in control theory which is the dynamic programming principle.

THEOREM 2.8. *Under the same assumptions of Theorem* 2.1*, for any* $T > 0$ *and for any* $x \in \overline{\Omega}$*, we have*

$$u[\varphi](x) = \inf_{\alpha(.)\in L^\infty(\mathbb{R}^+,\mathcal{A})} \left\{ \int_0^{\tau\wedge T} f(y_x(t),\alpha(t))e^{-\lambda t}dt \right.$$
$$\left. +\mathbb{1}_{\{\tau\le T\}}\varphi(y_x(\tau))e^{-\lambda\tau} + \mathbb{1}_{\{\tau>T\}}u[\varphi](y_x(T))e^{-\lambda T} \right\}.$$

Proof. To prove the theorem, we just remark that the continuity of the exit cost is no used in the classical proof (see [17], for instance). □

THEOREM 2.9. *Under the assumptions of Theorem* 2.1*, the value functions* $u[\varphi_*]$*,* $u[\varphi]$ *and* $u[\varphi^*]$ *are viscosity solutions of* (2).

Proof. Since the proof is inspired from the corresponding one in the continuous case, we point out only the modifications that we need. Let us denote by $\bar{\varphi}$ the functions φ_*, φ, or φ^*. We prove only that $u[\bar{\varphi}]$ is a supersolution of (2) on $\partial\Omega$. The other properties may be obtained by the same method. We consider some point $x_0 \in \partial\Omega$.

• If $(u[\bar{\varphi}])_*(x_0) \ge \varphi_*(x_0)$, we have nothing to prove.

• If $(u[\bar{\varphi}])_*(x_0) < \varphi_*(x_0)$, we have to show that the function u satisfies the equation $H(x,u,Du) = 0$ in viscosity sense.

There exists a sequence $(x_n)_n$ of points of $\overline{\Omega}$ such that $x_n \to x_0$ and

$$\lim_n u[\bar{\varphi}](x_n) \;=\; (u[\bar{\varphi}])_*(x_0).$$

We may assume without loss of generality that $x_n \in \Omega$ for all n. Indeed, if there exists a subsequence $(x_p)_p$ of $(x_n)_n$ such that $x_p \in \partial\Omega$ for all p, then, by definition of the value function $u[\bar{\varphi}]$,

$$u[\bar{\varphi}](x_p) \;=\; \bar{\varphi}(x_p).$$

And since $\bar{\varphi}$ is equal to φ_*, φ, or φ^*, passing to the limit, we obtain

$$\liminf_p \bar{\varphi}[x_p] \;\ge\; \varphi_*(x_0).$$

Hence we get a contradiction.

Now we choose a constant T such that

$$T\|f\|_\infty + \rho_{\varphi_*}(\|b\|_\infty T) + \|\varphi\|_\infty|1-e^{-\lambda T}| \;\le\; \frac{\varphi_*(x_0)-(u[\bar{\varphi}])_*(x_0)}{2}, \tag{20}$$

where ρ_{φ_*} is a nondecreasing continuous function satisfying $\rho_{\varphi_*}(t) \to 0$ as $t \to 0^+$ and, for every $y \in \partial\Omega$,

$$\varphi_*(x_0) - \varphi_*(y) \;\le\; \rho_{\varphi_*}(|x_0-y|).$$

(Such a function ρ_{φ_*} exists because φ_* is l.s.c.)

For every x_n, we introduce a control $\alpha_n(.)$ such that

$$u[\bar{\varphi}](x_n) + \frac{1}{n} \geq \int_0^{\tau_n \wedge T} f(y_{x_n}(t), \alpha_n(t)) e^{-\lambda t} dt + \mathbb{1}_{\{\tau_n \leq T\}} \bar{\varphi}(y_{x_n}(\tau_n)) e^{-\lambda \tau_n}$$

$$+ \mathbb{1}_{\{\tau_n > T\}} u[\bar{\varphi}](y_{x_n}(T)) e^{-\lambda T}. \tag{21}$$

If, for n large enough, we have $\tau_n > T$, then the inequality (21) implies

$$u[\bar{\varphi}](x_n) + \frac{1}{n} \ \geq \ \int_0^T f(y_{x_n}(t), \alpha_n(t)) e^{-\lambda t} dt + u[\bar{\varphi}](y_{x_n}(T)) e^{-\lambda T}.$$

And, since this inequality does not deal with the discontinuous exit cost $\bar{\varphi}$, we may apply classical arguments (see [3], for instance).

Otherwise, we may assume that, for all n, there exists p_n such that $\tau_{p_n} \leq T$. Then the inequality (21) implies, for the index p_n, that

$$u[\bar{\varphi}](x_{p_n}) + \frac{1}{p_n} \ \geq \ \int_0^{\tau_{p_n}} f(y_{x_{p_n}}(t), \alpha_{p_n}(t)) e^{-\lambda t} dt + \bar{\varphi}(y_{x_{p_n}}(\tau_{p_n})) e^{-\lambda \tau_{p_n}}.$$

Then we compute

$$\begin{aligned}
\varphi_*(x_0) - u[\bar{\varphi}](x_{p_n}) - \frac{1}{p_n} & \\
&\leq \int_0^{\tau_{p_n}} \|f\|_\infty dt + \varphi_*(x_0) - \bar{\varphi}(y_{x_{p_n}}(\tau_{p_n})) + \|\varphi\|_\infty |1 - e^{-\lambda \tau_{p_n}}| \\
&\leq \int_0^T \|f\|_\infty dt + \rho_{\varphi_*(x_0)}(|x_0 - y_{x_{p_n}}(\tau_{p_n})|) + \|\varphi\|_\infty |1 - e^{-\lambda T}| \\
&\leq T\|f\|_\infty + \rho_{\varphi_*(x_0)}(|x_0 - x_{p_n}| + \|b\|_\infty T) + \|\varphi\|_\infty |1 - e^{-\lambda T}|.
\end{aligned}$$

We pass to the limit, and we get a contradiction with the choice of T in (20). □

Example 2.10. We are going to show in this example that the value functions $u[\varphi_*]$, $u[\varphi]$, and $u[\varphi^*]$ may be not equal.

We take in $\mathbb{R}^2$

$$\Omega \ := \ \{(x, y) \in [-1, 1]^2, \ y \ > \ 0 \text{ or } x \ < \ 0\},$$

$$f \equiv 0, \ \lambda \ = \ 0$$

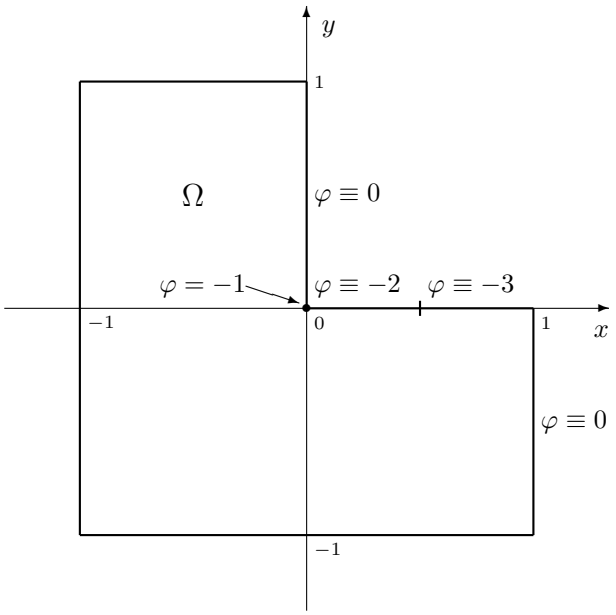
(λ is null for the sake of simplicity but this is not relevant here), and

$$b(x, \alpha) \ := \ \alpha \in \mathcal{A},$$

with $\mathcal{A} := \mathcal{B}(0, 1) \cap (\mathbb{R}^+ \times \mathbb{R}^-)$, i.e., $\{(\alpha_1, \alpha_2) \in \mathbb{R}^2, \ \alpha_1^2 + \alpha_2^2 \leq 1, \alpha_1 \geq 0, \alpha_2 \leq 0\}$.

The exit cost φ is equal to 0 except on $[0, 1] \times \{0\}$, where

$$\varphi(0, 0) \ = \ -1, \quad \varphi(x, 0) \ = \ -2 \ \text{ for } x \in \left]0, \frac{1}{2}\right[, \quad \varphi(x, 0) \ = \ -3 \ \text{ for } x \in \left[\frac{1}{2}, 1\right].$$



In the region $]-1,0[\times]0,1[$, it is easy to show that the best strategy is to reach the point $(0,0)$. To do so, we may, for example, use the control $(0,-1)$ until we touch the line $y=0$, and then we take the control $(1,0)$. Hence we compute that $u[\varphi_*] \equiv -2$, $u[\varphi] \equiv -1$, $u[\varphi^*] \equiv 0$, and $u_- \equiv -3$ in $]-1,0[\times]0,1[$.

The value functions $u[\varphi_*]$, $u[\varphi]$, and $u[\varphi^*]$ are very different because at the point $(0,0)$, the exit costs φ_*, φ, and φ^* take different values. Moreover, the functions u_- and $u[\varphi_*]$ are not equal because the trajectories which reach $[\frac{1}{2},1]\times\{0\}$ must be tangent to the boundary.

The open set Ω is not regular for the sake of simplicity. But we can easily change it. With φ continuous, a similar example was used to show that u_- and $u[\varphi]$ are not necessarily equal (cf. [6]).

This example shows that the discontinuous viscosity solutions may be very different even if we consider only their l.s.c. or u.s.c. envelopes.

2.3. Partial controllability on the boundary. We introduce now new assumptions which allow us to prove that all the value functions have the same l.s.c. envelope, namely, u_-.

We denote by $d(.)$ the distance function to the boundary $\partial\Omega$. We are given a smooth bounded domain; more precisely, we assume that

$$d \text{ is a } C^{1,1} \text{ function in the neighborhood } \mathcal{V} \text{ of } \partial\Omega. \tag{22}$$

Then we set $n(x) := -Dd(x)$ for $x \in \mathcal{V}$.

We assume that $\partial\Omega = \partial\Omega_1 \cup \partial\Omega_2$, where $\partial\Omega_1$ and $\partial\Omega_2$ are unions of connected components of $\partial\Omega$, and, at every point of $\partial\Omega_1$, there exists an outer field, i.e.,

$$\forall x \in \partial\Omega_1,\ \exists \alpha \in \mathcal{A},\ b(x,\alpha)\cdot n(x) \ \geq\ \beta \ >\ 0, \tag{23}$$

and, on $\partial\Omega_2$, there are only inner fields, i.e.,

$$\forall x \in \partial\Omega_2,\ \forall \alpha \in \mathcal{A}, \quad b(x,\alpha)\cdot n(x) \ \leq\ -\beta \ <\ 0. \tag{24}$$

Note that in (23) and (24), β can be chosen independent of x since $\partial\Omega_1$ and $\partial\Omega_2$ are compact subsets of $\mathbb{R}^N$ and since the functions b and n are continuous on $\partial\Omega$.

THEOREM 2.11. *We assume that φ is a bounded function defined pointwise satisfying* (4)*, that the constant λ is positive, and that the assumptions* (6), (7)*, and* (22)–(24) *hold.*

Then we have

$$u_- \;=\; (u^+)_* \quad \text{in } \Omega \cup \partial\Omega_1.$$

Using the representation formula, one easily build special examples where the equality may be wrong on $\partial\Omega_2$. For example, if the exit cost φ is equal to 0 on $\partial\Omega_2$ and to 1 on $\partial\Omega_1$, the value function u_- is null on the boundary $\partial\Omega_2$ but not the function u^+ since the trajectories cannot exit through $\partial\Omega_2$, i.e., $\bar{\tau} > \tau$.

Since u^+ and u_- are the maximum and minimum solutions of (2), this result characterizes the discontinuous solutions of (2) in $\Omega \cup \partial\Omega_1$. Indeed, if w is a solution of (2), then

$$u_- \;\le\; w_* \;\le\; w^* \;\le\; u^+ \quad \text{in } \Omega \cup \partial\Omega_1,$$

and therefore, taking the l.s.c. envelope of these inequalities and using Theorem 2.11 yield

$$u_- \;=\; w_* \;=\; (w^*)_* \quad \text{in } \Omega \cup \partial\Omega_1.$$

When the exit cost is continuous (cf. [6]), to get the uniqueness result, it is enough to know that the first exit time τ is equal to the "best exit time," i.e., which gives the minimal value for the value function or that $u_- = \varphi$ on $\partial\Omega$. But this is not the case with discontinuous exit cost since we deal with φ_* and φ^*.

Example 2.12. We show that the assumption (23) is necessary to prove Theorem 2.11. In fact, we take up the above Example 2.10, but we change φ to satisfy (4). Precisely,

$$\varphi(x) \;=\; \begin{cases} -2 & \text{on } [0,1]\times\{0\}, \\ \\ 0 & \text{otherwise.} \end{cases}$$

Then we can easily compute $u[\varphi_*] \equiv -2$ and $u[\varphi^*] \equiv 0$ in the region $]-1,0[\times]0,1[$.

REMARK 2.13. *The assumption* (24) *was introduced and used by H. M. Soner* [18] *for control problems with state-space constraints to prove the continuity and the uniqueness of the value function.*

Proof of Theorem 2.11. It is easy to see that the trajectories cannot exit through the boundary $\partial\Omega_2$ because of (24). And since the result holds only in $\Omega \cup \partial\Omega_1$, we do not need to take care of the value of the exit cost on the boundary $\partial\Omega_2$.

First, we prove that

$$(u^+)_* \;=\; (u[\varphi^*])_* \;=\; (u[\varphi_*])_* \;=\; u_- \quad \text{in } \Omega.$$

First equality: $(u^+)_* = (u[\varphi^*])_*$.

This is nothing but Theorem 2.4.

Second equality: $(u[\varphi^*])_* = (u[\varphi_*])_*$.

It is clear that $(u[\varphi_*])_* \le (u[\varphi^*])_*$ in the domain Ω, since $u[\varphi_*] \le u[\varphi^*]$ on $\overline{\Omega}$.

To prove the opposite inequality, let x be a point of Ω. Then there exists a sequence $(x_n)_n$ of points of Ω such that $x_n \to x$ and

$$\lim_n u[\varphi_*](x_n) \;=\; (u[\varphi_*])_*(x).$$

For each x_n, we consider a control $\alpha_n(.)$ such that

$$u[\varphi_*](x_n) + \frac{1}{n} \;\geq\; J(x_n, \alpha_n, \tau_n, \varphi_*). \tag{25}$$

We denote by $y^n_{x_n}$ the trajectory associated with α_n and by z_n its first exit point from Ω, i.e., $z_n := y^n_{x_n}(\tau_n) \in \partial\Omega_1$.

By using the assumption (4), we take a sequence $(z^p_n)_p$ of points of the boundary such that $z^p_n \to z_n$ and

$$\lim_p \varphi^*(z^p_n) \;=\; \varphi_*(z_n). \tag{26}$$

Now we need the following lemma.

LEMMA 2.14. *Let y_x be a trajectory such that its first exit time τ from Ω is positive and bounded. Then, for ε small enough and for $\tilde{z} \in \partial\Omega_1$ such that $|y_x(\tau) - \tilde{z}| < \varepsilon$, there exists a trajectory $\tilde{y}_{\tilde{x}}$ such that $\tilde{z} = \tilde{y}_{\tilde{x}}(\tilde{\tau})$ and*

$$|y_x(t) - \tilde{y}_{\tilde{x}}(t)| \;<\; D\varepsilon \quad \text{for } t \in [0, \tau]$$

with some constant D independent of ε.

For each trajectory $y^n_{x_n}$ and for p large enough, using (26), we apply Lemma 2.14 to the point z^p_n. Then, using again (26) and the Lipschitz continuity of the function f, we obtain after tedious but straightforward computations

$$J(x_n, \alpha_n, \tau_n, \varphi_*) \;\geq\; J(x^p_n, \alpha^p_n, \tau^p_n, \varphi^*) - \varepsilon^p_n,$$

with a sequence $(\varepsilon^p_n)_p$ of numbers such that $\varepsilon^p_n \to 0$ as $p \to \infty$. Combining this with (25), we get

$$\begin{aligned} u[\varphi_*](x_n) + \frac{1}{n} &\geq J(x^p_n, \alpha^p_n, \tau^p_n, \varphi^*) - \varepsilon^p_n \\ &\geq u[\varphi^*](x^p_n) - \varepsilon^p_n. \end{aligned}$$

Hence, passing to the limit, by a diagonal procedure, we obtain

$$\begin{aligned} (u[\varphi_*])_*(x) &= \lim_n u[\varphi_*](x_n) \\ &\geq (u[\varphi^*])_*(x). \end{aligned}$$

Since it is true for every $x \in \Omega$, the result is complete.

Third equality: $(u[\varphi_*])_* = u_-$.

By their definitions, $u_- \leq u[\varphi_*]$ in Ω. And since the value function u_- is l.s.c., we deduce $u_- \leq (u[\varphi_*])_*$ in Ω.

It remains to prove the opposite inequality. Let x_0 be a point of Ω. By the compactness of the set of relaxed controls, using classical arguments, there exist $\bar{\mu} \in L^\infty(\mathbb{R}^+, P(\mathcal{A}))$ and $\bar{\theta} \in [0, \infty]$ such that

$$u_-(x_0) \;=\; \hat{J}(x_0, \bar{\mu}, \bar{\theta}, \varphi_*).$$

Then we need the following lemma.

LEMMA 2.15. *Let $\hat{y}_x$ be a trajectory associated with a relaxed control μ and with a point $x \in \Omega$.*

Then, for any bounded exit time θ of $\hat{y}_x$, there exists a classical trajectory $\tilde{y}_{\tilde{x}}$ arbitrarily close to $\hat{y}_x$ and such that

$$\tilde{y}_{\tilde{x}}(\tilde{\tau}) \;=\; \hat{y}_x(\theta),$$

where $\tilde{\tau}$ is the first exit time of $\tilde{y}_{\tilde{x}}$ from Ω.

If $\bar{\theta}$ is bounded, using Lemma 2.15, there exists a sequence of points x_n and of controls $\alpha_n(.)$ such that $x_n \to x_0$ and

$$\hat{J}(x_0, \bar{\mu}, \bar{\theta}, \varphi_*) \;=\; J(x_n, \alpha_n, \tau_n, \varphi_*) - \varepsilon_n,$$

where $\varepsilon_n \to 0$ when $n \to \infty$ using the Lipschitz continuity of f. Then this implies

$$u_-(x_0) \;\geq\; u[\varphi_*](x_n) - \varepsilon_n.$$

Passing to the limit, we get

$$u_-(x_0) \;\geq\; (u[\varphi_*])_*(x_0).$$

It remains the cases when $\bar{\theta}$ is not bounded. We have several cases to consider. The easier is when the trajectory $\hat{y}_{x_0}$ stays far to the boundary, i.e.,

$$\inf_{t \,\geq\, 0} d(\hat{y}_{x_0}(t)) \;=\; \delta \;>\; 0;$$

then classical arguments imply that $\hat{y}_{x_0}$ can be approximated by classical trajectories. A connected case is when it is true but only after some finite time T, i.e.,

$$\inf_{t \,\geq\, T} d(\hat{y}_{x_0}(t)) \;=\; \delta \;>\; 0;$$

then we mix the preceding arguments with those of Lemma 2.14. The last case is when, for any $T > 0$, we get

$$\inf_{t \,\geq\, T} d(\hat{y}_{x_0}(t)) \;=\; 0.$$

Then we may assume that there exists a sequence of trajectories y_n such that their exit times τ_n tend to ∞ as $n \to \infty$, by using Lemma 2.14 for exit times of $\hat{y}_{x_0}$ large enough, or after modifying $\hat{y}_{x_0}$ when $\hat{y}_{x_0}$ is near the boundary, in order that it touches the boundary by using (23). Therefore, using the term $e^{-\lambda t}$ which tends to zero as $t \to \infty$ and the boundedness of f and φ, there exists a sequence $(\varepsilon_n)_n$ of numbers such that $\varepsilon_n \to 0$ as $n \to \infty$ and

$$\begin{aligned}\hat{J}(x_0, \bar{\mu}, \bar{\theta}, \varphi_*) &= \int_0^\infty \int_{\mathcal{A}} f(\hat{y}_{x_0}(t), \alpha) e^{-\lambda t} d\bar{\mu}(\alpha) dt \\ &\geq J(x_n, \alpha_n, \tau_n, \varphi_*) - \varepsilon_n.\end{aligned}$$

Then we conclude as before.

It remains to prove that $u_- = (u[\varphi_*])_*$ on the boundary $\partial\Omega_1$. Then we need the following lemma.

LEMMA 2.16. *The function u_- satisfies*

$$u_-(x) \;=\; \liminf_{y \to x, y \in \Omega} u_-(y) \quad \textit{for every } x \in \partial\Omega_1.$$

We postpone the proof of this lemma.

Moreover, since the function u^+ is u.s.c., for every $x \in \partial\Omega_1$, we have

$$(u^+)_*(x) \;=\; \liminf_{y\to x, y\in\Omega} u^+(y).$$

Therefore, since $u_- = (u^+)_*$ in Ω, we deduce that $u_- = (u^+)_*$ on $\partial\Omega$. □

Now we turn to the proof of the lemmas.

Proof of Lemma 2.14. In order to simplify the notation, we denote $z := y_x(\tau)$, by $\alpha(.)$ the control associated with the trajectory y_x, and by $\tilde{\alpha}(.)$ the control associated with $\tilde{y}_{\tilde{x}}$, which is the unknown.

We first remark that since the map $Y_\tau : x \mapsto y_x(\tau)$ is locally an homeomorphism, we may always find a point $\tilde{x}$ such that $y_{\tilde{x}}(\tau) = \tilde{z}$. Then a possible candidate for $\tilde{\alpha}$ may directly be α (and thus the constant D is equal to $e^{C\tau}$ by standard ODE estimates). But the difficulty is that the trajectory $y_{\tilde{x}}$ may cross the complementary set of Ω to reach $\tilde{z}$. In order to avoid this difficulty, we use the assumption (23) of a partial controllability on the boundary. Briefly speaking, we consider a backward trajectory which comes from the point $\tilde{z}$ and which goes in the interior of Ω for a short time. Afterward, the trajectory is again associated with α in order to reach a neighborhood of x.

First, using the assumption (23), we consider the control $\bar{\alpha} \in \mathcal{A}$ such that $b(z,\bar{\alpha}) \cdot n(z) > 0$. Thus we define the control $\tilde{\alpha}$ by

$$\tilde{\alpha}(t) \;=\; \begin{cases} \alpha(t) & \text{for } t \in [0,\tau[, \\ \bar{\alpha} & \text{for } t \in [\tau, \tau+\varepsilon]. \end{cases}$$

Then the trajectory $\tilde{y}_{\tilde{x}}$ is the solution of

$$\begin{cases} d\tilde{y}_{\tilde{x}}(s) &= b(\tilde{y}_{\tilde{x}}(s), \tilde{\alpha}(s))ds \quad \text{for } s \in [0, \tau+\varepsilon[, \\ \tilde{y}_{\tilde{x}}(\tau+\varepsilon) &= \tilde{z}. \end{cases}$$

Using the assumptions (22) on n and (6) on b, there exists a constant $\delta > 0$ such that

$$b(\xi,\bar{\alpha}) \cdot n(\xi) \;>\; \frac{\beta}{2} \quad \text{for } \xi \in \mathcal{B}(z,\delta) \cap \overline{\Omega}.$$

Moreover, using the boundedness of b, standard ODE estimates yield, for $t \in [\tau, \tau+\varepsilon]$,

$$\begin{aligned} |z - \tilde{y}_{\tilde{x}}(t)| &\le |z - \tilde{z}| + |\tilde{z} - \tilde{y}_{\tilde{x}}(t)| \\ &\le \varepsilon + C\varepsilon. \end{aligned}$$

Therefore, for ε small enough, the trajectory $\tilde{y}_{\tilde{x}}$ stays in $\mathcal{B}(z,\delta)$, and then we compute, for $t \in [\tau, \tau+\varepsilon]$,

$$\begin{aligned} d(\tilde{y}_{\tilde{x}}(t)) &= d(\tilde{z}) + \int_t^{\tau+\varepsilon} n(\tilde{y}_{\tilde{x}}(s)) \cdot b(\tilde{y}_{\tilde{x}}(s), \bar{\alpha}) ds \\ &\ge \frac{\beta}{2}(\tau + \varepsilon - t). \end{aligned} \tag{27}$$

Next, we prove that the trajectory $\tilde{y}_{\tilde{x}}$ does not touch the boundary during $[\tau-\eta, \tau]$, where η is a positive constant to determine. In order to simplify the notation, we set $z_\varepsilon := \tilde{y}_{\tilde{x}}(\tau)$ and $\tilde{y}_{z_\varepsilon}(.) := \tilde{y}_{\tilde{x}}(.)$. To prove the claim, it suffices to show that $d(\tilde{y}_{z_\varepsilon}(t))$

is positive for $t \in [\tau - \eta, \tau]$. Since the functions b and n are Lipschitz continuous, at least, in the neighborhood $\mathcal{V}$ of the boundary, for η small enough, we compute

$$\begin{aligned} d(\tilde{y}_{z_\varepsilon}(t)) - d(y_x(t)) &= d(z_\varepsilon) - d(z) \\ &\quad + \int_t^\tau (n(\tilde{y}_{z_\varepsilon}(s)) \cdot b(\tilde{y}_{z_\varepsilon}(s), \alpha(s)) - n(y_x(s)) \cdot b(y_x(s), \alpha(s)))ds \\ &\geq d(z_\varepsilon) - CK \int_t^\tau |\tilde{y}_{z_\varepsilon}(s) - y_x(s)|ds. \end{aligned}$$

Using standard ODE estimates, we obtain

$$\begin{aligned} d(\tilde{y}_{z_\varepsilon}(t)) - d(y_x(t)) &\geq d(z_\varepsilon) - CK \int_t^\tau |z_\varepsilon - z| e^{C(\tau - s)} ds \\ &\geq d(z_\varepsilon) - K|z_\varepsilon - z|(e^{C(\tau - t)} - 1) \\ &\geq d(z_\varepsilon) - K(|z_\varepsilon - \tilde{z}| + |\tilde{z} - z|)(e^{C\eta} - 1). \end{aligned} \tag{28}$$

We need an estimate for $|z_\varepsilon - \tilde{z}|$. But, since b is bounded and since $z_\varepsilon = \tilde{y}_{z_\varepsilon}(\tau)$, we have

$$\begin{aligned} |z_\varepsilon - \tilde{z}| &= \left| \int_\tau^{\tau + \varepsilon} b(\tilde{y}_{z_\varepsilon}(s), \bar{\alpha}) ds \right| \\ &\leq C\varepsilon. \end{aligned} \tag{29}$$

Combining this and (27) with (28), we get

$$d(\tilde{y}_{z_\varepsilon}(t)) - d(y_x(t)) \ \geq \ \frac{\beta\varepsilon}{2} - K(C\varepsilon + |\tilde{z} - z|)(e^{C\eta} - 1).$$

But, since $\tilde{z}$ is such that $|\tilde{z} - z| < \varepsilon$, this implies

$$d(\tilde{y}_{z_\varepsilon}(t)) - d(y_x(t)) \ \geq \ \varepsilon \left(\frac{\beta}{2} - K(1 + C)(e^{C\eta} - 1) \right). \tag{30}$$

Then we take η such that the right-hand side of this inequality is null, i.e.,

$$\eta \ := \ \frac{1}{C} \ln \left(1 + \frac{\beta}{2K(1 + C)} \right).$$

Hence we conclude that

$$d(\tilde{y}_{z_\varepsilon}(t)) \ \geq \ 0 \quad \text{for every } t \in [\tau - \eta, \tau],$$

and the inequality is strict because of the term $d(y_x(t))$. We may remark that the constant η depends only on the functions b and n.

Finally, since τ is the first exit time of y_x, we have

$$\inf\{d(y_x(t)), t \in [0, \tau - \eta]\} \ = \ \rho \ > \ 0.$$

We compute, for $t \in [0, \tau - \eta]$,

$$\begin{aligned} d(\tilde{y}_{z_\varepsilon}(t)) &\geq d(y_x(t)) - |(\tilde{y}_{z_\varepsilon} - y_x)(t)| \\ &\geq \rho - |z_\varepsilon - z|(e^{C\tau} - 1) \\ &\geq \rho - (C + 1)(e^{C\tau} - 1)\varepsilon. \end{aligned}$$

Therefore, for ε small enough, we obtain the result with a constant D equal to $(C+1)(e^{C\tau}-1)$. □

Proof of Lemma 2.15. First we approximate $\hat{y}_x$ by a relaxed trajectory whose first exit point is $\tilde{z} := \hat{y}_x(\theta)$. We use backward trajectories. We recall that a backward trajectory $\check{y}_z$ is a solution of the ODE

$$\begin{cases} d\check{y}_z(t) = -b(\check{y}_z(t), \alpha(t))dt, \\ \check{y}_z(0) = z. \end{cases}$$

We begin by choosing a control $\bar{\alpha} \in \mathcal{A}$ such that $b(\tilde{z}, \bar{\alpha}) \cdot n(\tilde{z}) > 0$. Then we consider the backward trajectory $\check{y}_1$ with $\check{y}_1(0) = \tilde{z}$ and associated with the relaxed control μ_1 defined by

$$\mu_1(t) \;=\; \begin{cases} \bar{\alpha} & \text{if } t \in [0, \varepsilon_1[, \\ \mu(t) & \text{elsewhere,} \end{cases}$$

with a parameter $\varepsilon_1 > 0$ to be chosen later. Using the estimate (27), which can be extended to backward relaxed trajectories, we know that the new trajectory $\check{y}_1$ does not touch the boundary during $]0, \varepsilon_1[$, for ε_1 small enough. Now, if the new trajectory $\check{y}_1$ touches the boundary at the time τ_1, we set $z_1 := \check{y}_1(\tau_1)$ and choose $\alpha_1 \in \mathcal{A}$ such that $b(z_1, \alpha_1) \cdot n(z_1) > 0$. Then we change the control μ_1 for μ_2 given by

$$\mu_2(t) \;=\; \begin{cases} \alpha_1 & \text{if } t \in [\tau_1 - \varepsilon_2, \tau_1], \\ \mu_1(t) & \text{elsewhere.} \end{cases}$$

Therefore the trajectory associated with μ_2 does not touch $\partial\Omega$, at least during $[\tau_1 - \varepsilon_2, \tau_1]$. In fact, by using the arguments of the proof of Lemma 2.14, we know that the trajectory lies in the open set Ω during the fixed time η.

Next, we do the same thing as many times as the trajectory touches the boundary. Finally, since after each modification, we know that the trajectory does not touch $\partial\Omega$ during η, we change the control only a finite number of time.

In order to be close to the trajectory $\hat{y}_x$, it suffices to choose the parameters ε_i small enough since the relaxed control of the new trajectory is equal to μ except during the times ε_i where μ is replaced by α_i.

Finally, it remains to approximate the new relaxed trajectory by a classical trajectory. But it is standard since the relaxed trajectory lies in Ω during $[\varepsilon_1, \theta]$. □

Proof of Lemma 2.16. We consider a point z of $\partial\Omega_1$. Using the assumption (23), there exists a control $\bar{\alpha} \in \mathcal{A}$ such that $b(z, \bar{\alpha}) \cdot n(z) > 0$. As in Lemma 2.14, we consider the solution $\tilde{y}_z$ of

$$\begin{cases} d\tilde{y}_z(s) &=& b(\tilde{y}_z(s), \bar{\alpha})ds \quad \text{for } s \in [0, 1[, \\ \tilde{y}_z(1) &=& z. \end{cases}$$

Then we consider the points

$$x_n \;:=\; \tilde{y}_z\left(1 - \frac{1}{n}\right)$$

and the trajectories

$$\tilde{y}_{x_n}(.) \;:=\; \tilde{y}_z\left(. + 1 - \frac{1}{n}\right).$$

It is clear enough that, for n large enough, $x_n \in \Omega$ and $x_n \to z$ as $n \to \infty$. Note also that the first exit time of $\tilde{y}_{x_n}$ is equal to $\frac{1}{n}$. Thus the cost function associated with the trajectory $\tilde{y}_{x_n}$ is

$$J\left(x_n, \bar{\alpha}, \frac{1}{n}, \varphi_*\right) \;=\; \int_0^{\frac{1}{n}} f(\tilde{y}_{x_n}(t), \bar{\alpha}) e^{-\lambda t} dt + \varphi_*(z) e^{-\lambda \frac{1}{n}}.$$

Therefore this implies

$$u_-(x_n) \;\leq\; J\left(x_n, \bar{\alpha}, \frac{1}{n}, \varphi_*\right).$$

Then we pass to the limit

$$\begin{aligned} \liminf_n \; u_-(x_n) &\leq \liminf_n \; J\left(y_n, \alpha_1, \frac{1}{n}, \varphi_*\right) \\ &\leq \varphi_*(x_0). \end{aligned}$$

Finally, by definition of the limit inferior, we deduce

$$\begin{aligned} \liminf_{y \to z, y \in \Omega} u_-(y) &\leq \liminf_n \; u_-(x_n) \\ &\leq \varphi_*(x_0), \end{aligned}$$

and the proof is complete. □

Example 2.17. We will show that the assumption (23) is necessary to prove Lemma 2.15. More precisely, we give a relaxed trajectory which cannot be approximated by classical trajectories in a closed set. To this end, we consider in $\mathbb{R}^3$ the field b given, for $x = (x_1, x_2, x_3) \in \mathbb{R}^3$, by

$$b(x, \alpha) \;=\; \begin{bmatrix} 1 \\ \alpha \\ \frac{1}{1+\alpha^2} + |x_2| \end{bmatrix}$$

with $\alpha \in \mathcal{A} := [-1, 1]$. (This field was already considered in [5].)

Around the point $O = (0, 0, 0)$, the domain Ω is reduced by two cylinders, C_1 and C_2:

$$\begin{aligned} C_1 &= \left\{(x_1, x_2, x_3) \in \mathbb{R}^3,\; (x_1 - 5)^2 + (x_3)^2 \;\leq\; 5\right\}, \\ C_2 &= \left\{(x_1, x_2, x_3) \in \mathbb{R}^3,\; (x_1 - 9)^2 + (x_3 - 7)^2 \;\leq\; 5\right\}. \end{aligned}$$

We can define a relaxed trajectory $\hat{y}_O$ satisfying $(\hat{y}_O)_3(t) = \frac{t}{2}$ which is tangent to the two cylinders. We can easily prove that classical trajectories cannot approximate the trajectory $\hat{y}_O$ when $t \geq 10$ because, since the classical trajectories are such that

$$(\dot{y}(t))_3 \;>\; \frac{1}{2} \quad \text{for } t > 0$$

(see [5]), at the time $t = 5$ they have to be above the cylinder C_1 and then they cannot pass below C_2 for $t = 10$.

3. The uniqueness results. In this part, we prove uniqueness results for Hamilton–Jacobi equations in the case when the Dirichlet boundary data φ is assumed only to be defined pointwise, by using PDE methods. We recall that uniqueness means in this context that all the solutions have the same l.s.c. envelope.

For simplicity in what follows, we assume henceforth that $H(x,t,p) = H(x,p) + t$ and that $(x,p) \to H(x,p)$ is a continuous function from $\overline{\Omega} \times \mathbb{R}^N$ into $\mathbb{R}$ which satisfies

$$H(x,p) \text{ is convex in } p \text{ for every } x \in \overline{\Omega} \tag{31}$$

and

$$\left|\frac{\partial H}{\partial p}\right| \le C, \quad \left|\frac{\partial H}{\partial x}\right| \le C(1+|p|) \quad \text{on } \overline{\Omega} \times \mathbb{R}^N \text{ for some constant } C > 0. \tag{32}$$

To explain the problems coming from the discontinuity of the Dirichlet condition, we consider the function φ being equal to 0 at some point $x_0 \in \partial\Omega$ and 1 elsewhere, as in the introduction. Then the viscosity subsolution condition on the boundary is satisfied by the function u if

$$u^* \le \varphi^* \equiv 1 \quad \text{on } \partial\Omega.$$

It is clear that this condition is not restrictive enough because the fact that $\varphi(x_0) = 0$ is not seen by this boundary condition.

Indeed, for illustration, we consider the exit time problem from the point x_0. To this end, we even assume that the controllability is complete; i.e., the field b is given by $b(x,\alpha) := \alpha$ with α in the unit ball $\mathcal{B}(0,1)$ of $\mathbb{R}^N$. Then, if we assume that the domain Ω is convex, any point $x \in \Omega$ may reach the point x_0 by following the straight line $[x, x_0]$. Finally, we take $f \equiv 1$ and $\lambda = 1$. With the notation of the introduction, the value function $u[\varphi]$ is defined, for every $x \in \overline{\Omega}$, by

$$\begin{aligned} u[\varphi](x) &= \inf_{\alpha(.) \in L^\infty(\mathbb{R}^+, \mathcal{B}(0,1))} \left\{ \int_0^\tau 1 e^{-t} dt + \varphi(y_x(\tau)) e^{-\tau} \right\} \\ &= \inf_{\alpha(.) \in L^\infty(\mathbb{R}^+, \mathcal{B}(0,1))} \{1 - e^{-\tau} + \mathbb{1}_{\partial\Omega \setminus \{x_0\}}(y_x(\tau)) e^{-\tau}\}. \end{aligned}$$

It is easy to show that the best strategy is to reach the point x_0. Then, the value function is equal to

$$u[\varphi](x) = 1 - e^{-|x - x_0|}.$$

But by the result of the first part, all the functions $u_\zeta(x) := 1 - \zeta e^{-|x-x_0|}$ with $\zeta \in [0,1]$ are also solutions of (2) since $\varphi_* \le \zeta \mathbb{1}_{\partial\Omega\setminus\{x_0\}} \le \varphi^*$. In particular, the maximal subsolution u^+, which is equal to $u[\varphi^*] \equiv 1$, appears to be a pathological solution for this exit time problem.

We consider two ways to avoid this difficulty: the first one is to consider only "regular" boundary data φ, i.e., functions satisfying the condition (4) as we did in the first section. The second one is to impose additional conditions on the solution u: since the viscosity subsolution condition on the boundary turns out to be not restrictive enough, we replace it by the condition (5), i.e., for any $x \in \partial\Omega$,

$$\liminf_{y \to x, y \in \Omega} u(y) \le \varphi_*(x),$$

and moreover, we need that the function u is a viscosity solution in the sense of Barron and Jensen inside the domain Ω. For the sake of completeness, we recall the following definition.

DEFINITION 3.1. *Let u be a bounded function. We say that u is a Barron–Jensen solution of*

$$H(x, Du) + u \;=\; 0 \quad \text{in } \Omega \tag{33}$$

if it satisfies

$$\begin{cases} \forall\, \phi \in C^1(\Omega) \text{ at each minimum point } x_0 \in \Omega \text{ of } u_* - \phi, \text{ we have} \\ H(x_0, D\phi(x_0)) + u_*(x_0) \;=\; 0. \end{cases} \tag{34}$$

In the case of continuous solutions, we have an equivalence between (34) and the definition of viscosity solutions of (33): if the function u satisfies (34), then u is a viscosity solution of (33) and vice versa.

For discontinuous solutions, the connections are far less simple (cf. [2]).

• If u is a bounded viscosity solution of (33) satisfying $(u^*)_* = u_*$ in Ω, then the property (34) holds for u_*. (In particular, it holds if u is u.s.c.)

• If u is a l.s.c. bounded function satisfying the property (34), then u is a viscosity solution of (33).

To prove the uniqueness result, we need to add nondegeneracy conditions on the Hamiltonian on the boundary. To do so, we recall that we assume that the distance function d satisfies (22). We assume that $\partial\Omega = \partial\Omega_1 \cup \partial\Omega_2$, where $\partial\Omega_1$ and $\partial\Omega_2$ are unions of connected components of $\partial\Omega$ and

$$\begin{cases} \forall x \in \partial\Omega_1,\ \forall R \;>\; 0,\ \exists C^R \;>\; 0 \text{ such that} \\ \text{if } |y - x| \;\le\; \dfrac{1}{C^R},\ y \in \overline{\Omega}, \text{ and } \lambda \;\ge\; C^R(1 + |p|),\ \lambda \in \mathbb{R}^+,\ p \in \mathbb{R}^N, \\ \text{then} \quad H(y, p - \lambda n(y)) \;\ge\; R, \end{cases} \tag{35}$$

and

$$\begin{cases} \forall x \in \partial\Omega_2,\ \forall R \;>\; 0,\ \exists C^R \;>\; 0 \text{ such that} \\ \text{if } |y - x| \;\le\; \dfrac{1}{C^R},\ y \in \overline{\Omega}, \text{ and } \lambda \;\ge\; C^R(1 + |p|),\ \lambda \in \mathbb{R}^+,\ p \in \mathbb{R}^N, \\ \text{then} \quad H(y, p - \lambda n(y)) \;\le\; -R. \end{cases} \tag{36}$$

We recall that $n(x) := -Dd(x)$ for every x in the neighborhood $\mathcal{V}$ of $\partial\Omega$. We also assume that

$$\begin{cases} \text{in a neighborhood of } \partial\Omega_2, \text{ for every } p \in \mathbb{R}^N, \\ \text{the map } \lambda \to H(x, p + \lambda n(x)) \text{ is a nondecreasing function.} \end{cases} \tag{37}$$

In the case of control problems when the Hamiltonian is given by (1), the assumption (35) is equivalent to (23), i.e., the existence of an outer field on $\partial\Omega_1$ and the assumptions (36) and (37) are equivalent to (24), which implies that there are only inner fields on $\partial\Omega_2$.

The assumption (36) is used in the comparison result of [7].

Our result is the following theorem.

THEOREM 3.2. *We assume that φ is a bounded function defined pointwise and that the assumptions* (22), (31), (32), *and* (35)–(37) *hold.*

(1) *If the condition* (4) *holds on $\partial\Omega_1$ and if the functions u and v are viscosity solutions of* (2), *then*

$$u_* \;=\; v_* \quad \text{in } \Omega.$$

(2) *If the functions u and v are Barron–Jensen solutions of* (33) *which are supersolutions of* (2) *on the boundary and which satisfy the condition* (5) *on $\partial\Omega_1$, then*

$$u_* \;=\; v_* \quad \text{in } \Omega.$$

In the first point, we recover the uniqueness result of the first section in the case of control problems. We will detail in the part on applications the adaptation of the second point. In particular, we will show that all the value functions are Barron–Jensen solutions of the associated Bellman equation.

Finally, let us come back to the example of the beginning of the section. The function u^+, which is equal to 1, and the functions u_ζ with $\zeta \in]0,1[$ are Barron–Jensen solutions of the associated Bellman equation in Ω, but they do not satisfy the condition (5) on $\partial\Omega$. Only the function $u[\varphi]$ satisfies both properties.

A viscosity solution u is a Barron–Jensen solution in particular if u satisfies the property $(u^*)_* = u_*$ in Ω which is a criterion for uniqueness in [6] or if the function u is continuous. We recall that the solution u is continuous if $H(x,p) \to \infty$ when $|p| \to \infty$.

This result is similar to that obtained in the case of optimal stopping time problems with discontinuous stopping costs (cf. [2]).

The following comparison principle is the keystone of the proof of the uniqueness results.

THEOREM 3.3. *Under the assumptions of Theorem* 3.2, *if the function u is a Barron–Jensen solution of* (33) *satisfying* (5) *on $\partial\Omega_1$ and if the function v is a bounded supersolution of* (2), *then*

$$u_* \;\le\; v_* \quad \text{in } \Omega.$$

Proof of Theorem 3.2. (1) First, we remark that all the subsolutions of equation (2) satisfy condition (5): indeed, using a classical result when the assumption (35) holds, we have

$$u^* \;\le\; \varphi^* \quad \text{on } \partial\Omega_1.$$

Taking the limit inferior, we find that

$$(u^*)_* \;\le\; (\varphi^*)_* \;=\; \varphi_* \quad \text{on } \partial\Omega_1, \tag{38}$$

since the function φ satisfies the assumption (4). Finally, since we have, for every $x \in \partial\Omega$,

$$\liminf_{y\to x, y\in\Omega} u^*(y) \;=\; (u^*)_*(x),$$

we deduce from (38) that the solution u^* satisfies, for every $x \in \partial\Omega_1$,

$$\liminf_{y\to x, y\in\Omega} u^*(y) \;\le\; \varphi_*(x).$$

Now since the function u^* is not necessarily a Barron–Jensen solution of (33), we are not able to use directly Theorem 3.3 with u. Then we introduce the maximal subsolution u^+ of equation (2). The function u^+ may be obtained by considering the following standard approximate Dirichlet problem:

$$\left\{ \begin{array}{ll} -\dfrac{1}{n}|Du^n|^2 + H(x, Du^n) + u^n = 0 & \text{in } \Omega, \\ u^n = \varphi^n & \text{on } \partial\Omega, \end{array} \right. \tag{39}$$

where $(\varphi^n)_n$ is a nonincreasing sequence of continuous functions such that

$$\inf_n \varphi^n \;=\; \varphi^*.$$

The existence of solutions of the problem (39) is a classical consequence of Perron's method (cf. [14]) since the constants

$$M_n \;:=\; \max\{\|H(x,0)\|_\infty, \|\varphi_n\|_\infty\}$$

and $-M_n$ are, respectively, a supersolution and a subsolution of this problem. Moreover, the new Hamiltonian satisfies the property

$$-\frac{1}{n}|p|^2 + H(x,p) \to -\infty \quad \text{as} \quad |p| \to \infty,$$

which corresponds to the opposite of the classical coercivity property. Nevertheless, we recover the results of [6], but for supersolutions instead of subsolutions: the supersolutions of (39) are uniformly Lipschitz continuous in Ω and we have a comparison result for this equation. Thus we get that the function u^n is continuous and unique. Thanks to the formulation of the approximate problem, a subsolution w of the equation (2) is still a subsolution of the new equation (39). Then this implies that $w \le u^n$ in Ω. Hence, the maximal subsolution u^+ is defined, for $x \in \overline{\Omega}$, by

$$u^+(x) \;:=\; \limsup_{n\to\infty, y\to x} u^n(y) \;=\; \inf_n u^n(x).$$

Therefore the function u^+ is u.s.c.; hence it satisfies the property (34) and the condition (5) by the first remark. Using the comparison principle of Theorem 3.3, we get

$$(u^+)_* \;\le\; v_* \quad \text{in } \Omega.$$

But since u^+ is the maximal subsolution, we have $u^+ \ge u^*$ in Ω. Hence

$$\begin{aligned} v_* &\ge (u^+)_* \\ &\ge (u^*)_* \;\ge\; u_* \quad \text{in } \Omega. \end{aligned}$$

Finally, we exchange the roles of the functions u and v, and the proof is complete.

(2) The second point is a direct consequence of Theorem 3.3. □

Now we turn to the proof of the comparison principle.

Proof of Theorem 3.3. We first recall that the distance function to the boundary d is assumed to be $C^{1,1}$ in the neighborhood $\mathcal{V}$ of $\partial\Omega$ by the assumption (22). We still denote by d a nondecreasing $C^{1,1}$ function on $\overline{\Omega}$ which is equal to the distance

function to $\partial\Omega$ in a neighborhood $\mathcal{V}'$ of $\partial\Omega$ included in $\mathcal{V}$. And we set $n(x) := -Dd(x)$ for $x \in \overline{\Omega}$, even if, for $x \notin \mathcal{V}'$, $n(x)$ is not necessarily unitary.

For every $\alpha > 0$, we introduce the function u^α defined, for every $(x,t) \in \Omega \times \mathbb{R}^+$, by

$$u^\alpha(x,t) \;=\; \inf_{y\in\overline{\Omega}} \{u_*(y) + e^{-Kt}\phi_\alpha(x,y) + \bar{C}(d(x) - d(y))\},$$

with

$$\phi_\alpha(x,y) \;:=\; \frac{|x-y|^4}{\alpha} + L\frac{|x-y|^3}{\alpha}(d(x)-d(y)) + M\frac{|d(x)-d(y)|^4}{\alpha},$$

where $\bar{C}$, K, L, and M are some positive constants to be chosen later.

The properties of u^α are described in the following lemma.

LEMMA 3.4. *There exist constants $\bar{C}$, K, L, and M such that, for every $T > 0$ and for α small enough, the following properties hold:*

(1) *the function u^α is Lipschitz continuous in $\Omega \times [0,T]$. (Thus we may extend it by continuity to the boundary $\partial\Omega \times [0,T]$.)*

(2) *The function u^α is a viscosity subsolution of*

$$\frac{\partial w}{\partial t} + H(x,Dw) + w - B\sqrt[4]{\alpha} \;=\; 0 \qquad \text{in } (\Omega \cup \partial\Omega_2)\times]0,T] \tag{40}$$

for some constant $B > 0$, independent of α.

(3) *We have $u^\alpha \le \varphi^\alpha$ on $\partial\Omega_1 \times [0,T]$, where the function φ^α is defined, for every $x \in \partial\Omega_1$ and $t \ge 0$, by*

$$\varphi^\alpha(x,t) \;=\; \inf_{y\in\partial\Omega_1} \left\{ \varphi_*(y) + e^{-Kt}\frac{|x-y|^4}{\alpha} \right\}.$$

We postpone the proof of the lemma.

REMARK 3.5. *The inf-convolution procedure, which usually leads a supersolution in the viscosity solution theory, allows us to obtain a subsolution following new ideas introduced by Barron and Jensen* [8] *for convex Hamiltonians. The time-dependent formulation of u^α is a technical point which permits to treat discontinuous solutions for stationary problems* (*cf.* [2]).

The terms with the distance function is a trick to obtain the classical property of the inf-convolution procedure despite the presence of the boundary, thanks to the assumptions (35) *and* (36). *The distance function to $\partial\Omega$ plays here essentially the same role as the time variable in the classical Barron–Jensen approach for the Cauchy problem.*

To conclude the proof of Theorem 3.3, we compare the functions u^α and v. Since $\varphi_* \ge \varphi^\alpha$ on Ω and since the function v_* does not depend on t, v_* is a supersolution of

$$\begin{cases} \dfrac{\partial w}{\partial t} + H(x,Dw) + w = 0 & \text{in } \Omega\times]0,T], \\ w = \varphi^\alpha & \text{on } \partial\Omega\times]0,T]. \end{cases} \tag{41}$$

Because of the Lipschitz continuity of u^α and φ^α, an easy adaptation of the comparison result of [6] for the Cauchy problem (41) (see also [17] and [18]) yields

$$u^\alpha(x,T) - v_*(x) \;\le\; e^{-T}\|(u^\alpha(.,0) - v_*(.))^+\|_\infty + B\sqrt[4]{\alpha}.$$

We first let $\alpha \to 0$ in this inequality and then $T \to \infty$. Since the functions u and v are bounded, we obtain

$$u_* \ \leq \ v_* \quad \text{in } \Omega,$$

which is the inequality that we wanted to prove. □

Now we turn to the proofs of the lemmas.

Proof of Lemma 3.4. (1) Regularity and elementary properties of u^α.

For every x, $z \in \Omega$, and $t \geq 0$, we have

$$\begin{aligned} u^\alpha(x,t) - u^\alpha(z,t) &\leq \sup_{y\in\overline{\Omega}}\{(e^{-Kt}\phi_\alpha(x,y) + \bar{C}(d(x) - d(y))) \\ &\qquad -(e^{-Kt}\phi_\alpha(z,y) + \bar{C}(d(z) - d(y)))\} \\ &\leq e^{-Kt}\sup_{y\in\overline{\Omega}}\{\phi_\alpha(x,y) - \phi_\alpha(z,y)\} + \bar{C}(d(x) - d(z)). \end{aligned}$$

By the compactness of the domain $\overline{\Omega}$ and by Lipschitz regularity of the distance function, we obtain

$$u^\alpha(x,t) - u^\alpha(z,t) \ \leq \ \frac{C}{\alpha}|x - z|.$$

Therefore the function u^α is Lipschitz continuous in the space variable. It is easy to check the same property in the time variable.

Then, by Rademacher's theorem, u^α is differentiable almost everywhere, and then, by classical result in optimization theory, if y_α is a point such that $u^\alpha(x,t) = u_*(y_\alpha) + e^{-Kt}\phi_\alpha(x,y_\alpha) + \bar{C}(d(x) - d(y_\alpha))$, we have

$$Du^\alpha(x,t) \ = \ e^{-Kt}D_x\phi_\alpha(x,y_\alpha) - \bar{C}n(x)$$

and

$$\frac{\partial u^\alpha}{\partial t}(x,t) \ = \ -Ke^{-Kt}\phi_\alpha(x,y_\alpha).$$

Let us check some elementary properties of the function u^α. First, using the fact that the functions u and v are bounded, we fix the constant $\bar{C}$ to the largest of the constants C^R which appear in the assumptions (35) and (36) for $R = \max\{\|u\|_\infty, \|v\|_\infty\}$.

LEMMA 3.6. *For L large enough and $M \gg L$ and for every $(x,t) \in \Omega \times [0,T]$, the following properties hold:*

(1) $u^\alpha(x,t) \leq u_*(x)$.

(2) *If y_α is a point such that $u^\alpha(x,t) = u_*(y_\alpha) + e^{-Kt}\phi_\alpha(x,y_\alpha) + \bar{C}(d(x) - d(y_\alpha))$, there exists a constant E independent of α such that*

$$|x - y_\alpha| \ \leq \ E\sqrt[4]{\alpha}.$$

(3) *For x in a neighborhood of the boundary, if we write $Du^\alpha(x,t) = p_\alpha - \lambda_\alpha n(x)$ and if we have $d(x) \geq d(y_\alpha)$, then*

$$\lambda_\alpha \ \geq \ \bar{C}(1 + |p_\alpha|).$$

(4) *There exists a constant $\bar{K}$ independent of α and of K such that*

$$|Du^\alpha(x,t)||x - y_\alpha| \ \leq \ \bar{K}(|x - y_\alpha| + e^{-Kt}\phi_\alpha(x,y_\alpha)).$$

We postpone the proof of this lemma.

(2) Equation satisfied by u^α.

Since the Hamiltonian is convex, in order to prove that the function u^α is a viscosity subsolution of (40) in $\Omega\times]0,T[$, it is enough to show that u^α satisfies (40) in the almost everywhere sense (cf. [17]). Let (x_0,t_0) be a point of $\Omega\times]0,T[$ where u^α is differentiable.

We consider a point $y_\alpha\in\overline{\Omega}$ such that

$$u^\alpha(x_0,t_0) \;=\; u_*(y_\alpha)+e^{-Kt_0}\phi_\alpha(x_0,y_\alpha)+\bar{C}(d(x_0)-d(y_\alpha)).$$

By the definition of u^α, the point y_α is a minimum point of $u_*-\psi$ with

$$\psi(y) \;:=\; -e^{-Kt_0}\phi_\alpha(x_0,y)-\bar{C}(d(x_0)-d(y)).$$

In order to use the equation on u, we have to show that $y_\alpha\in\Omega$. We first remark that the assumptions (22), (35), and (36) are satisfied in the neighborhood of $\partial\Omega$ included in $\mathcal{V}'$,

$$\Omega_\delta \;:=\; \{x\in\overline{\Omega} \text{ such that } \; d(x)<\delta\},$$

for some positive constant δ small enough.

• If $x_0\notin\Omega_\delta$, then $d(x_0)\geq\delta$ and, by using the second point of Lemma 3.6, we obtain

$$\begin{aligned} d(y_\alpha) &\geq d(x_0)-|x_0-y_\alpha| \\ &\geq \delta - E\sqrt[4]{\alpha}. \end{aligned}$$

Hence, for α small enough, we find that $d(y_\alpha)>0$.

• If $x_0\in\Omega_\delta$, we need the following lemma.

LEMMA 3.7. *If $x_0\in\Omega_\delta$, then we have*

$$d(x_0) \;\leq\; d(y_\alpha).$$

Using Lemma 3.7, we deduce that $d(y_\alpha)>0$ since $x_0\in\Omega$.

Hence, we get

$$H(y_\alpha,D\psi(y_\alpha))+u_*(y_\alpha) \;=\; 0, \tag{42}$$

where

$$D\psi(y_\alpha) \;=\; -e^{-Kt}D_y\phi_\alpha(x,y_\alpha)-\bar{C}n(y_\alpha).$$

We want to replace y_α by x_0 and $u_*(y_\alpha)$ by $u^\alpha(x_0,t_0)$ in the equality (42). By using the assumption (32) and the definition of u^α, we get

$$\begin{aligned} H(x_0,Du^\alpha(x_0,t_0))+u^\alpha(x_0,t_0) \leq\; & C(1+|Du^\alpha(x_0,t_0)|)|x_0-y_\alpha| \\ & +C|Du^\alpha(x_0,t_0)-D\psi(y_\alpha)| \\ & +e^{-Kt_0}\phi_\alpha(x_0,y_\alpha)+\bar{C}(d(x_0)-d(y_\alpha)). \end{aligned} \tag{43}$$

To estimate the right side of this inequality, we use Lemma 3.6 and we compute

$$\begin{aligned} |Du^\alpha(x_0,t_0)-D\psi(y_\alpha)| &\leq |\lambda_\alpha||n(x_0)-n(y_\alpha)| \\ &\leq \|Dn\|_\infty|Du^\alpha(x_0,t_0)||x_0-y_\alpha|. \end{aligned}$$

Therefore we deduce from the inequality (43) that

$$H(x_0, Du^\alpha(x_0,t_0)) + u^\alpha(x_0,t_0) \;\le\; B\sqrt[4]{\alpha} + Ke^{-Kt_0}\phi_\alpha(x_0,y_\alpha),$$

with $B := (C(1+\bar{K}+\bar{K}\|Dn\|_\infty)+\bar{C})E$ and $K := C\bar{K}(1+\|Dn\|_\infty)+1$.

It remains to introduce the time derivative of u^α, i.e.,

$$\frac{\partial u^\alpha}{\partial t}(x_0,t_0) \;=\; -Ke^{-Kt_0}\phi_\alpha(x_0,y_\alpha).$$

Then, we obtain

$$\frac{\partial u^\alpha}{\partial t}(x_0,t_0) + H(x_0, Du^\alpha(x_0,t_0)) + u^\alpha(x_0,t_0) \;\le\; B\sqrt[4]{\alpha}.$$

Hence, the function u^α satisfies the equation (40) in the almost everywhere sense.

Finally, since u^α is a viscosity subsolution of (40) in $\Omega\times]0,T[$, u^α is also a subsolution of (40) on the boundary $\Omega\times\{T\}$ by classical properties of Cauchy problems (cf. [7]).

Moreover, on the boundary $\partial\Omega_2\times]0,T]$, the assumption (37) implies the following.

LEMMA 3.8. *The function u^α is a viscosity subsolution of* (40) *on $\partial\Omega_2\times]0,T]$.*

We again postpone the proof of this lemma.

(3) Boundary properties of u^α.

By the definition of φ^α, for any $x_0\in\partial\Omega_1$ and any $t\ge 0$, there exists a point y_0 of $\partial\Omega_1$ such that

$$\varphi^\alpha(x_0,t) \;=\; \varphi_*(y_0) + e^{-Kt}\frac{|x_0-y_0|^4}{\alpha}.$$

We take a sequence $(y_n)_n$ of points of Ω such that $y_n\to y_0$ and

$$\liminf_{y\to y_0, y\in\Omega} u(y) \;=\; \lim_n u(y_n).$$

By the property (5), we know that

$$\lim_n u(y_n) \;\le\; \varphi_*(y_0).$$

Now we introduce the sequence $(x_n)_n$ defined by $x_n := x_0 - d(y_n)n(x_0)$ for every n. Since we have $d(y_n)\to 0$ as $n\to\infty$, the sequence $(x_n)_n$ converges to x_0. Note also that $d(x_n)=d(y_n)$ for n large enough.

By the definition of u^α, we may write

$$\begin{aligned}
u^\alpha(x_n,t) &\le u_*(y_n) + e^{-Kt}\phi_\alpha(x_n,y_n) + \bar{C}(d(x_n)-d(y_n))\\
&\le u_*(y_n) + e^{-Kt}\left[\frac{|x_n-y_n|^4}{\alpha} + L\frac{|x_n-y_n|^3}{\alpha}(d(x_n)-d(y_n))\right.\\
&\qquad\left. + M\frac{|d(x_n)-d(y_n)|^4}{\alpha}\right] + \bar{C}(d(x_n)-d(y_n))\\
&\le u_*(y_n) + e^{-Kt}\frac{|x_n-y_n|^4}{\alpha} \qquad \text{for } n \text{ large enough.}
\end{aligned}$$

Then, letting $n\to\infty$, we obtain

$$u^\alpha(x_0,t) \;=\; \lim_n u^\alpha(x_n,t) \;\le\; \varphi_*(y_0) + e^{-Kt}\frac{|x_0-y_0|^4}{\alpha} \;=\; \varphi^\alpha(x_0,t).$$

Since it is true for every $x_0\in\partial\Omega_1$ and $t\ge 0$, the result is proven. □

Proof of Lemma 3.6. (1) Since $\phi_\alpha(x,x)=0$ for any $x\in\overline{\Omega}$, we have $u_*\geq u^\alpha$ in $\Omega\times[0,T]$.

(2) Since the function u_* is l.s.c., for any α, there exists a point $y_\alpha\in\overline{\Omega}$ such that

$$u^\alpha(x,t) \;=\; u_*(y_\alpha)+e^{-Kt}\phi_\alpha(x,y_\alpha)+\bar{C}(d(x)-d(y_\alpha)),$$

and since $u^\alpha \;\leq\; u_*$ in $\Omega\times[0,T]$, we have

$$u_*(y_\alpha)+e^{-Kt}\phi_\alpha(x,y_\alpha)+\bar{C}(d(x)-d(y_\alpha)) \;\leq\; u_*(x).$$

Then we deduce

$$e^{-Kt}\phi_\alpha(x,y_\alpha) \;\leq\; 2\|u\|_\infty+2\bar{C}\|d\|_\infty. \tag{44}$$

By using the Young inequality

$$L\frac{|x-y|^3}{\alpha}(d(x)-d(y)) \;\leq\; \frac{3}{4}\frac{|x-y|^4}{\alpha}+\frac{1}{4}L^4\frac{|d(x)-d(y)|^4}{\alpha},$$

we get

$$\phi_\alpha(x,y_\alpha)\geq\frac{1}{4}\frac{|x-y|^4}{\alpha}+(M-\frac{1}{4}L^4)\frac{|d(y)-d(x)|^4}{\alpha} \tag{45}$$

$$\geq\frac{1}{4}\frac{|x-y|^4}{\alpha} \tag{46}$$

for $M\geq\frac{1}{2}L^4$. Combining it with (44), this implies

$$|x-y_\alpha| \;\leq\; e^{\frac{Kt}{4}}\sqrt[4]{2(\|u\|_\infty+\bar{C}\|d\|_\infty)\alpha}.$$

(3) If the function u^α is differentiable at (x,t), we may write $Du^\alpha(x,t)=p_\alpha-\lambda_\alpha n(x)$ with

$$p_\alpha \;:=\; e^{-Kt}\left[4(x-y_\alpha)\frac{|x-y_\alpha|^2}{\alpha}+3L(x-y_\alpha)\frac{|x-y_\alpha|}{\alpha}(d(x)-d(y_\alpha))\right]$$

and

$$\lambda_\alpha \;:=\; e^{-Kt}\left[L\frac{|x-y_\alpha|^3}{\alpha}+4M(d(x)-d(y_\alpha))\frac{|d(x)-d(y_\alpha)|^2}{\alpha}\right]+\bar{C}.$$

Again, we use the Young inequality to get

$$3L\frac{|x-y_\alpha|^2}{\alpha}|d(x)-d(y_\alpha)| \;\leq\; 2\frac{|x-y_\alpha|^3}{\alpha}+L^3\frac{|d(x)-d(y_\alpha)|^3}{\alpha}.$$

If we have $d(x)\geq d(y_\alpha)$, then we deduce

$$\bar{C}(1+|p_\alpha|)-\lambda_\alpha \;\leq\; e^{-Kt}\left[(6\bar{C}-L)\frac{|x-y_\alpha|^3}{\alpha}+(\bar{C}L^3-4M)\frac{|d(x)-d(y_\alpha)|^3}{\alpha}\right].$$

The right-hand side of this inequality is negative for $L\geq 6\bar{C}$ since we already know that $M\geq\frac{1}{2}L^4$.

(4) We compute

$$|Du^\alpha(x,t)||x-y_\alpha| \le e^{-Kt}\left[4\frac{|x-y_\alpha|^4}{\alpha} + 3L\frac{|x-y_\alpha|^3}{\alpha}|d(x)-d(y_\alpha)| + L\frac{|x-y_\alpha|^4}{\alpha}\right.$$

$$\left. + 4M\frac{|d(x)-d(y_\alpha)|^3}{\alpha}|x-y_\alpha|\right] + \bar{C}|x-y_\alpha|. \tag{47}$$

Then, using again Young inequalities, we obtain after straightforward computations

$$|Du^\alpha(x,t)||x-y_\alpha| \le e^{-Kt}\left[\left((4+\frac{13}{4}L+M\right)\frac{|x-y_\alpha|^4}{\alpha}\right.$$

$$\left. + \left(\frac{3}{4}L+3M\right)\frac{|d(x)-d(y_\alpha)|^4}{\alpha}\right] + \bar{C}|x-y_\alpha|.$$

Finally, we use the inequality (45) to fix the constant $\bar{K}$. □

Proof of Lemma 3.7. We argue by contradiction assuming that

$$d(x_0) \;>\; d(y_\alpha). \tag{48}$$

We first consider the case when the point y_α is in the domain Ω. Hence, we use the equation on u and we get

$$H(y_\alpha, D\psi(y_\alpha)) + u_*(y_\alpha) \;=\; 0, \tag{49}$$

where

$$D\psi(y_\alpha) \;=\; p_\alpha - \lambda_\alpha n(y_\alpha)$$

with the notation of Lemma 3.6.

But, since $d(x_0) < \delta$, we get that $y_\alpha \in \Omega_\delta$ using the inequality (48). Then, we may apply the assumptions (35) or (36) to the equality (49), and thus we get a contradiction since, by Lemma 3.6, we have

$$\lambda_\alpha \;\ge\; \bar{C}(1+|p_\alpha|).$$

It remains the case when the minimum point y_α is on the boundary $\partial\Omega$. We set

$$\chi_\varepsilon(y) \;:=\; u_*(y) - \psi(y) + |y-y_\alpha|^2 + \frac{\varepsilon}{d(y)}$$

for $y \in \Omega$ and $\varepsilon > 0$. We add the term $|.-y_\alpha|^2$ to be sure that the point y_α is a strict local minimum point of $u_*(.) - \psi(.) + |.-y_\alpha|^2$. Then there exists a sequence $(y_\varepsilon)_\varepsilon$ of local minimum points of χ_ε such that

$$\lim_{\varepsilon\to 0} y_\varepsilon \;=\; y_\alpha \quad \text{and} \quad \lim_{\varepsilon\to 0} \chi_\varepsilon(y_\varepsilon) \;=\; u_*(y_\alpha) - \psi(y_\alpha).$$

Since $y_\varepsilon \in \Omega$, we may use the equation on u:

$$H\left(y_\varepsilon, D\psi(y_\alpha) - 2(y_\varepsilon - y_\alpha) - \frac{\varepsilon}{d(y_\varepsilon)^2}n(y_\varepsilon)\right) + u_*(y_\varepsilon) \;=\; 0. \tag{50}$$

But since $d(y_\varepsilon) \to 0$ as $\varepsilon \to 0$, we have $d(y_\varepsilon) < d(x_0)$ for ε small enough. Then, using the same arguments as in preceding case, we get a contradiction with (50) since the term $|y_\varepsilon - y_\alpha|$ tends to 0 and the additional term with the normal is negative.

Hence, we have $d(x_0) \le d(y_\alpha)$. □

Proof of Lemma 3.8. Let ϕ be a C^1 function on $\overline{\Omega} \times [0,T]$ and let $(x_0,t_0) \in \partial\Omega_2 \times]0,T]$ be a maximum point of $u^\alpha - \phi$. We may assume (x_0,t_0) to be a strict local maximum point of $u^\alpha - \phi$ by changing, if needed, ϕ in $\phi + |.-x_0|^2 + |.-t_0|^2$. Following an idea of [4], for $\varepsilon > 0$ and for every $(x,t) \in \Omega \times]0,T]$, we set

$$\chi_\varepsilon(x,t) \;=\; u^\alpha(x,t) - \phi(x,t) - \frac{\varepsilon}{d(x)}.$$

As in Lemma 3.7, since the point (x_0,t_0) is a strict local maximum point of $u^\alpha - \phi$, there exists a sequence $(x_\varepsilon,t_\varepsilon)_\varepsilon$ of local maximum points of χ_ε such that

$$(x_\varepsilon,t_\varepsilon) \to (x_0,t_0) \quad \text{ and } \quad u^\alpha(x_\varepsilon,t_\varepsilon) \to u^\alpha(x_0,t_0) \quad \text{ as } \varepsilon \to 0.$$

Since $x_\varepsilon \in \Omega$, (40) holds for u^α and we get

$$\frac{\partial\phi}{\partial t}(x_\varepsilon,t_\varepsilon) + H\left(x_\varepsilon, D\phi(x_\varepsilon,t_\varepsilon) + \frac{\varepsilon}{d(x_\varepsilon)^2} n(x_\varepsilon)\right) + u^\alpha(x_\varepsilon,t_\varepsilon) - B\sqrt[4]{\alpha} \;\le\; 0. \tag{51}$$

Using the assumption (37), we get

$$H(x_\varepsilon, D\phi(x_\varepsilon,t_\varepsilon)) \;\le\; H\left(x_\varepsilon, D\phi(x_\varepsilon,t_\varepsilon) + \frac{\varepsilon}{d(x_\varepsilon)^2} n(x_\varepsilon)\right).$$

Combining this with (51), we obtain

$$\frac{\partial\phi}{\partial t}(x_\varepsilon,t_\varepsilon) + H(x_\varepsilon, D\phi(x_\varepsilon,t_\varepsilon)) + u^\alpha(x_\varepsilon,t_\varepsilon) - B\sqrt[4]{\alpha} \;\le\; 0.$$

We conclude by letting ε go to zero. □

4. Applications. In this section, we apply the uniqueness results of Theorem 3.2 to exit time control problems.

We use the notations of the first section; in particular, the Hamiltonian is given by (1).

THEOREM 4.1. *We assume that φ is a bounded function defined pointwise, that the constant λ is positive, and that the assumptions* (6), (7), *and* (22)–(24) *hold. Then*

(1) *if the property* (4) *holds for φ on $\partial\Omega_1$, the l.s.c. envelope of the solutions of* (2) *is equal to the value function u_-.*

(2) *all the Barron–Jensen solutions of* (33) *in Ω which are supersolutions of* (2) *on the boundary and which satisfy the condition* (5) *on $\partial\Omega_1$ have the same l.s.c. envelope, u_-.*

In the second point, the only value functions satisfying the condition (5) are, in general, u_- and $u[\varphi_*]$. It is clear enough in the example of the introduction where $\varphi := \mathbb{1}_{\partial\Omega\setminus\{x_0\}}$. Besides, we can directly show that $(u[\varphi_*])_* = u_-$ by Lemma 2.15.

P. Soravia [19] showed that all the value functions are Barron–Jensen solutions of (33) in Ω, using their explicit representation formulas. We have the following theorem.

THEOREM 4.2. *Assume* (6) *and* (7). *The value functions $u[\varphi_*]$, $u[\varphi]$, $u[\varphi^*]$, u_-, and u^+ are Barron–Jensen solutions of* (33) *in Ω.*

It is an open problem to know under which assumptions the other solutions of (2) are Barron–Jensen solutions.

Now we turn to the proofs.

Proof of Theorem 4.1. This result is an easy adaptation of Theorem 3.2. We just have to verify the assumptions of this theorem. The Hamiltonian satisfies (32)

because of the conditions (6) and (7) on b and f. Also, the assumptions (23) and (24) lead to (35), (36), and (37).

We prove only the claim for (35) since the other cases use the same kind of arguments. For any $x \in \partial\Omega_1$, using (23), we choose a control $\bar{\alpha}$ such that $b(x,\bar{\alpha}) \cdot n(x) \geq \beta$. Then, by definition of the Hamiltonian, for $R > 0$, $\lambda \in \mathbb{R}^+$ and $p \in \mathbb{R}^N$, and for $y \in \mathcal{V}$, we have

$$\begin{aligned} H(y, p - \lambda n(y)) - R &= \sup_{\alpha \in \mathcal{A}} \{-b(y,\alpha) \cdot (p - \lambda n(y)) - f(y,\alpha)\} - R \\ &\geq -b(y,\bar{\alpha}) \cdot (p - \lambda n(y)) - f(y,\bar{\alpha}) - R, \\ H(y, p - \lambda n(y)) - R &\geq \lambda\beta - \lambda(b(x,\bar{\alpha}) \cdot n(x) - b(y,\bar{\alpha}) \cdot n(y)) \\ &\quad - \|b\|_\infty |p| - \|f\|_\infty - R. \end{aligned}$$

Using the Lipschitz continuity of the functions b and n, we get

$$H(y, p - \lambda n(y)) - R \ \geq \ \lambda(\beta - K|x-y|) - \max\{\|b\|_\infty,\ \|f\|_\infty + R\}(1 + |p|).$$

Then, if we take

$$C^R \ := \ \frac{2}{\beta} \max\{K,\ \|b\|_\infty,\ \|f\|_\infty + R\},$$

the right side of the preceding inequality is negative and therefore the assumption (35) is satisfied.

Now, by Theorem 4.2, we already know that u_- satisfies (34). And Lemma 2.16 implies that u_- satisfies (5). Then the proof is complete. □

Proof of Theorem 4.2. We have already proved that all these value functions are supersolutions of (2) in the first part. In order to prove the opposite viscosity inequality, we need the following lemma introduced by P. Soravia [19] as the backward dynamic programming principle.

LEMMA 4.3. *Let x be a point of Ω and $\alpha(.)$ be a control, i.e., $\alpha \in L^\infty(\mathbb{R}^+, \mathcal{A})$. We consider the backward trajectory $\check{y}_x$ solution of the dynamical system*

$$\begin{cases} d\check{y}_x(t) = -b(\check{y}_x(t), \alpha(t))dt, \\ \check{y}_x(0) = x \in \overline{\Omega}, \end{cases}$$

We denote by τ_x the first exit time of the backward trajectory $\check{y}_x$ from Ω.

Then, for all T such that $0 < T < \tau_x$, we have

$$u(x) \ \geq \ -\int_0^T f(\check{y}_x(t), \alpha(t)) e^{\lambda t} dt + u(\check{y}_x(T)) e^{\lambda T}.$$

To prove Theorem 4.2, we just have to use the classical arguments as in Theorem 2.9, but using the inequality of Lemma 4.3 instead of the usual dynamic programming principle. □

Now we explain how use the uniqueness result to pass to the limit. We consider an exit time problem satisfying the assumption (23) on $\partial\Omega$ with a exit cost satisfying (5). Let $(\varphi_n)_n$ be a nondecreasing sequence of continuous functions such that

$$\sup_n \varphi_n \ = \ \varphi_* \ = \ (\varphi^*)_*.$$

We note u_n, a viscosity solution of problem (2) where φ is replaced by φ_n.

By the second part, $((u_n)^*)_*$ satisfies the property (34). Using (23) and since φ_n is continuous, we get

$$((u_n)^*)_* \ \leq \ \varphi_n \quad (\leq \ \varphi_{n+1}) \quad \text{ on } \partial\Omega.$$

Therefore $((u_n)^*)_*$ satisfies also the property (5). Then we apply Theorem 4.1 to get

$$\begin{aligned}(u_{n+1})_* &\geq ((u_n)^*)_* \\ &\geq (u_n)_* \quad \text{ in } \Omega.\end{aligned}$$

Thus, for $x \in \Omega$, we have

$$\liminf_{n\to\infty, y\to x} (u_n)_*(y) \ = \ \sup_n (u_n)_*(x).$$

Finally, by classical stability results, the function

$$\underline{u}(x) \ := \ \liminf_{n\to\infty, y\to x} (u_n)_*(y) \ = \ \sup_n (u_n)_*(x)$$

is a supersolution of (2). Moreover it easy to check that $\underline{u}$ satisfies the properties (5) and (34). Hence, by using again the preceding uniqueness result, we conclude

$$\lim_n (u_n)_* \ = \ \sup_n (u_n)_* \ = \ \underline{u} \ = \ u_- \quad \text{ in } \Omega.$$

REFERENCES

[1] M. Bardi and P. Soravia, *A comparison result for Hamilton-Jacobi equations and applications to differential games lacking controllability*, Funkcial. Ekvac., 37 (1994), pp. 19–43.

[2] G. Barles, *Discontinuous viscosity solutions of first-order Hamilton-Jacobi equations: A guided visit*, Nonlinear Anal., 20 (1993), pp. 1123–1134.

[3] G. Barles, *Solutions de viscosite des equations de Hamilton-Jacobi*, Math. Appl. 17, Springer-Verlag, Paris, 1994.

[4] G. Barles and J. Burdeau, *The Dirichlet problem for semilinear second-order degenerate elliptic equations and applications to stochastic exit time control problem*, Comm. Partial Differential Equations, 20 (1995), pp. 129–178.

[5] G. Barles and B. Perthame, *Discontinuous solutions of deterministic optimal stopping time problems*, RAIRO Modél. Math. Anal. Numér., 21 (1987), pp. 557–579.

[6] G. Barles and B. Perthame, *Exit time problems in optimal control and vanishing viscosity method*, SIAM J. Control Optim., 26 (1988), pp. 1133–1148.

[7] G. Barles and B. Perthame, *Comparison principle for Dirichlet type Hamilton-Jacobi equations and singular perturbations of degenerated elliptic equations*, Appl. Math. Optim., 21 (1988), pp. 21–44.

[8] E. N. Barron and R. Jensen, *Semicontinuous viscosity solutions of Hamilton-Jacobi equations with convex Hamiltonians*, Comm. Partial Differential Equations, 15 (1990), pp. 1713–1740.

[9] E. N. Barron and R. Jensen, *Optimal control and semicontinuous viscosity solutions*, Proc. Amer. Math. Soc., 113 (1991), pp. 49–79.

[10] M. G. Crandall, L. C. Evans, and P. L. Lions, *Some properties of viscosity solutions of Hamilton-Jacobi equations*, Trans. Amer. Math. Soc., 282 (1984), pp. 487–502.

[11] M. G. Crandall and P. L. Lions, *Viscosity solutions of Hamilton-Jacobi equations*, Trans. Amer. Math. Soc., 277 (1983), pp. 1–42.

[12] W. H. Fleming and H. M. Soner, *Controlled Markov Processes and Viscosity Solutions*, Springer-Verlag, New York, 1993.

[13] H. Ishii, *Hamilton-Jacobi equations with discontinuous Hamiltonians on arbitrary subsets*, Bull. Fac. Sci. Engrg. Chuo Univ. Ser. I Math., 28 (1985), pp. 33–77.

[14] H. Ishii, *Perron's method for Hamilton-Jacobi equations*, Duke Math. J., 55 (1987), pp. 369–384.

[15] H. Ishii, *A boundary value problem of the Dirichlet type for the Hamilton-Jacobi equations*, Ann. Scuola Norm. Sup. Pisa Cl. Sci. (4), 16 (1989), pp. 105–135.
[16] J. M. Lasry and P. L. Lions, *A remark on regularisation in Hilbert space*, Israel J. Math, 55 (1986), pp. 257–266.
[17] P. L. Lions, *Generalized Solutions of Hamilton-Jacobi Equations*, Pitman Res. Notes Math. Ser. 69, Pitman, Boston, 1982.
[18] H. M. Soner, *Optimal control with state-space constraint*, SIAM J. Control Optim., 28 (1986), pp. 552–561.
[19] P. Soravia, *Discontinuous viscosity solutions to Dirichlet problems for the Hamilton-Jacobi equations with convex Hamiltonians*, Comm. Partial Differential Equations, 18 (1993), pp. 1493–1514.
[20] A. I. Subbotin, *Discontinuous solutions of a Dirichlet type boundary value problem for first order P.D.E.*, Russian J. Numer. Anal. Math. Modelling, 8 (1993), pp. 145–164.

SIAM J. CONTROL OPTIM.
Vol. 35, No. 2, pp. 435–461, March 1997

004

NONLINEAR FILTERING REVISITED: A SPECTRAL APPROACH*

SERGEY LOTOTSKY†, REMIGIJUS MIKULEVICIUS‡, AND BORIS L. ROZOVSKII†

Abstract. The objective of this paper is to develop an approach to nonlinear filtering based on the Cameron–Martin version of Wiener chaos expansion. This approach gives rise to a new numerical scheme for nonlinear filtering. The main feature of this algorithm is that it allows one to separate the computations involving the observations from those dealing only with the system parameters and to shift the latter off-line.

Key words. Cameron–Martin development, Wick polynomials, Wiener chaos, Zakai equation

AMS subject classifications. 60G35, 60H15, 62M20

PII. S0363012993248918

1. Introduction.

Nonlinear filtering is a classic problem of applied stochastic analysis (see, e.g., Kallianpur [19], Kunita [23], Kushner [24], Liptser and Shiryayev [27], etc.). It is of notable theoretical and practical importance by itself and also as a part of control theory for partially observable stochastic systems (see, e.g., Fleming and Pardoux [11]).

In this paper we consider the filtering scheme where the signal process $x(t)$ is a Markov diffusion process and the observation process is of the form

$$y(t) = y_0 + \int_0^t h(x(s))ds + w(t),$$

where $w(t)$ is a Brownian motion independent of the process $x(t)$.

Let f be a given bounded function on $\mathbf{R}^d$ and $\hat{f}(x(t))$ be the optimal filter (the best in the mean-square estimate for $f(x(t))$ based on observations $y(s)$, $s \le t$). A fundamental result of filtering theory says that the optimal filter is given by the formula

$$\hat{f}(x(t)) = \frac{\int_{\mathbf{R}^d} f(x)u(t,x)dx}{\int_{\mathbf{R}^d} u(t,x)dx}, \tag{1.1}$$

where $u(t,x)$ is the so-called unnormalized filtering density (UFD); of course, some regularity assumptions are needed to ensure the existence of the density.

A standard way to study the UFD (analytically or numerically) is to treat it as a solution of the Zakai equation

$$du(t,x) = \mathcal{L}^* u(t,x)dt + h(x)u(t,x)dy(t), \tag{1.2}$$

where $\mathcal{L}^*$ is the formally adjoint operator to the generator of the Markov process $x(t)$ (see, e.g., Baras [2]; Benesh [3]; Bensoussan, Glowinski, and Rascanu [4]; Clark [8];

*Received by the editors May 14, 1993; accepted for publication (in revised form) December 21, 1995. This research was supported in part by Office of Naval Research grant N00014-95-1-0229 and Army Research Office grant DAAH 04-95-1-0164. A version of this paper was presented at the 35th IEEE Conference on Decision and Control, Kobe, Japan, December 11–13, 1996.

http://www.siam.org/journals/sicon/35-2/24891.html

†Center for Applied Mathematical Sciences, University of Southern California, Los Angeles, CA 90089-1113 (rozovski@cams.usc.edu).

‡Institute of Mathematics and Informatics, Akademijos 4, Vilnius 2600, Lithuania.

DiMasi and Runggaldier [9]; Elliott and Glowinski [10]; Florchinger and LeGland [12]; Krylov and Rozovskii [20]; Kunita [22]; Pardoux [33]; Rozovskii [34]; Zakai [37]; etc.).

Another comparatively recent approach is based on the Wiener chaos expansion (WCE) (see references below). In this paper we further develop a version of this approach based on the Cameron–Martin orthogonal decomposition of L_2-functionals of a Gaussian process (see Cameron and Martin [7]). We prove that the UFD can be written in the form

$$u(t,x) = \sum_{\alpha} \frac{1}{\sqrt{\alpha!}} \varphi_\alpha(t,x)\xi_\alpha(y), \tag{1.3}$$

where $\xi_\alpha(y)$ are Wick polynomials (certain products of Hermite polynomials; see, e.g., [14]) of Wiener integrals $\int_0^t m_i(s)dy(s)$, where $\{m_k\}$ is a complete orthonormal system in $L_2([0,t])$, and $\varphi_\alpha(t,x)$ are deterministic Hermite–Fourier coefficients in the Cameron–Martin orthogonal decomposition of $u(t,x)$ (see Mikulevicius and Rozovskii [30, 31]). The Wick series expansion (1.3) converges in L_2-sense on the reference probability space.

We prove that the set of functions $\{\varphi_\alpha(t,x)\}$ is a solution to a simple recursive system of Kolmogorov-like equations (see (2.6)). Below it will be referred to as the S-system.

Our interest in the WCE was motivated mainly by computational purposes. One important feature of the expansion (1.3) is that it separates observations and parameters in that the Wick polynomials are completely defined by the observation process $y(t)$ but the Hermite–Fourier coefficients $\varphi_\alpha(t,x)$ are determined only by the coefficients of the signal process $x(t)$, its initial distribution, and the observation function h.

Unfortunately, direct application of the above expansion for numerical computations is impractical, limited, at best, to short time intervals. The main reason is possible exponential growth of the errors inflicted by truncation of the infinite series (1.3) as the time interval $[0,t]$ increases (Theorem 2.2).

One important objective of the paper is to develop a numerical approximation scheme for the UFD which retains the separation of observations and parameters but is not subject to the aforementioned limitations (Theorem 2.5 and the accompanying algorithm).

This recursive scheme splits into two parts: deterministic and stochastic. The deterministic part (solving the S-system) might be time consuming but can be performed off-line since in many applications the coefficients of the processes $x(t), y(t)$ and also of the S-system are known a priori. The stochastic part (determining the Wick polynomials $\xi_\alpha(y)$) is computationally simple and can be performed in real time. In this paper this scheme is referred to as the spectral separating scheme (S^3).

We prove the strong convergence of S^3 both in L_2 and $\mathbf{C}$ spaces and demonstrate that the overall rate of convergence (on- and off-line) is of order $O(\Delta)$, where Δ is the time step (Theorems 2.2 and 2.4).

S^3 can be also viewed as a time-discretization scheme for a solution of the Zakai equation. In section 4 we demostrate that some well-known discretization algorithms for this equation (e.g., explicit Euler scheme, splitting-up method (see [4, 26])) can be derived from a multistep version of (1.3). In this section we also discuss the computational complexity of S^3, compare it with the complexity of the splitting-up method, and present some results of numerical simulations.

We conclude the introduction with some historical remarks. The idea of obtaining an "explicit" WCE solution of a stochastic (ordinary) differential equation can be

traced back to the paper [21] by Krylov and Veretennikov (see also Zvonkin and Krylov [38]). Kunita [22] applied this idea to prove uniqueness of the Zakai equation. Wong [35] obtained the solution of a special class of nonlinear filtering problems in the form of the WCE. Ocone [32] pioneered finite-order WCEs of normalized nonlinear filters (see also references therein).

In these works the multiple Wiener integral version of the WCE was used. The Cameron–Martin development is analytically equivalent to this version of the WCE (see, e.g., Ito [17]). However, it has some computational advantage since only ordinary Wiener integrals are required in this approach. Lo and Ng [28] were the first ones to utilize the above fact. They modified Ocone's approximation using the Cameron–Martin expansion. Unfortunately, the equations for the deterministic coefficients of the finite-order approximations in [28] are quite complex. To solve them one needs to know the Hermite–Fourier coefficients for the corresponding unnormalized filters. Computation of the latter was not discussed in [28].

The S-system (2.6) was introduced by Mikulevicius and Rozovskii [30, 31]. The upper bound $ce^{ct}t^{N+1}/(N+1)!$ on the error of the Nth-order approximation to (1.3) was obtained in [30]. Recently, Budhiraja and Kallianpur [5] developed a different WCE-type approximation of the unnormalized filtering density using the Haar-type basis. They also established an upper bound on the error of truncation with respect to the stochastic and deterministic bases.

2. Main results. Let $(\Omega, \mathcal{F}, \mathrm{P})$ be a probability space and $w(t)$ be an r-dimensional Brownian motion on the space. Let $x(t)$ be a d-dimensional (unobservable) signal process and $y(t)$ be the r-dimensional observation process given by

$$y(t) = \int_0^t h(x(s))ds + w(t), \quad 0 \le t \le T, \tag{2.1}$$

where $h = (h^l)_{1\le l\le r}$ is an r-dimensional vector function on $\mathbf{R}^d$. We assume in addition that the signal $x(t)$ is a diffusion Markov process of the form[1]

$$\begin{aligned} dx^i(t) &= b^i(x(t))dt + \sigma^{ij}(x(t))d\tilde{w}^j(t), \quad 0 < t \le T, \\ x(0) &= x_0, \end{aligned} \tag{2.2}$$

where $b = (b^i)_{1\le i\le d}$ is a d-dimensional vector function on $\mathbf{R}^d$, $\sigma = (\sigma^{ij})_{1\le i\le d,\ 1\le j\le d_1}$ is a $d\times d_1$ dimensional matrix function on $\mathbf{R}^d$, and $\tilde{w} = (\tilde{w}^i)_{1\le i\le d_1}$ is a d_1-dimensional Brownian motion on $(\Omega, \mathcal{F}, \mathrm{P})$.

The following is assumed about the model (2.1), (2.2):

(A1) the functions b, σ, and h are infinitely differentiable and bounded with all derivatives;

(A2) the processes w and $\tilde{w}$ are independent;

(A3) the random vector x_0 is independent of both w and $\tilde{w}$ and has density[2] $p(x) \in \mathbf{H}^n$ for $n = 0, 1, 2, \ldots$.

(Then by the Sobolev embedding theorem, $p(x)$ is also in $\mathbf{C}_b^n$ for any n.) Some of these assumptions can be weakened, and we will discuss them at the end of this section.

Let $\mathcal{F}_t^y$ be the σ-algebra generated by $y(s)$, $s \le t$. Denote

$$\rho(t) = \exp\Big\{ -\int_0^t h^l(x(s))dw^l(s) - \frac{1}{2}\sum_{l=1}^r \int_0^t |h^l(x(s))|^2 ds\Big\}.$$

[1] When the sum is finite, we assume summation over repeated indices and omit the $\sum$ sign.

[2] Here and below $\mathbf{H}^n$ is the Sobolev space $W_2^n(\mathbf{R}^d)$ (see, e.g., [25]), and $\mathbf{C}_b^n$ is the space of n times continuously differentiable on $\mathbf{R}^d$ functions bounded with all the derivatives.

It is well known (see, e.g., [27] or [19]) that the measure $\tilde{\mathrm{P}}$ defined by $d\tilde{\mathrm{P}} = \rho(T)d\mathrm{P}$ is a probability measure on $(\Omega, \mathcal{F})$ with the following properties:

(i) on the reference probability space $(\Omega, \mathcal{F}, \tilde{\mathrm{P}})$, $y(\cdot)$ is a Brownian motion independent of $x(\cdot)$;

(ii) the optimal filter $\hat{f}(x(t)) = \mathbf{E}[f(x(t))|\mathcal{F}_t^y]$ is given by

$$\hat{f}(x(t)) = \frac{\tilde{\mathbf{E}}[f(x(t))\rho(t)^{-1}|\mathcal{F}_t^y]}{\tilde{\mathbf{E}}[\rho(t)^{-1}|\mathcal{F}_t^y]}, \tag{2.3}$$

where $\tilde{\mathbf{E}}$ is the expectation with respect to measure $\tilde{\mathrm{P}}$. If assumptions (A1)–(A3) hold, the unnormalized filtering measure $\Phi_t(dx) = \tilde{\mathbf{E}}[1_{\{x(t)\in dx\}}\rho(t)^{-1}|\mathcal{F}_t^y]$ admits the density $u(t,x) = \Phi_t(dx)/dx$, called the UFD, which is a solution of the Zakai equation

$$du(t,x) = \mathcal{L}^* u(t,x)dt + h^l(x)u(t,x)dy^l(t), \tag{2.4}$$

where $\mathcal{L}^* u := \frac{1}{2}\frac{\partial^2}{\partial x_i \partial x_j}((\sigma\sigma^*)^{ij}u) - \frac{\partial}{\partial x_i}(b^i u)$ and such that for every $n \in N$

$$\tilde{\mathbf{E}}\sup_{t\le T}||u(t,\cdot)||^2_{\mathbf{C}_b^n} < \infty,$$
$$\tilde{\mathbf{E}}\sup_{t\le T}||u(t,\cdot)||^2_{\mathbf{H}^n} < \infty.$$

Using the UFD $u(t,x)$, one can rewrite (2.3) in the form (1.1).

DEFINITION. *A collection $\alpha = (\alpha_k^l)_{1\le l\le r,\ k\ge 1}$ of nonnegative integers is called an r-dimensional multiindex if only finitely many of α_k^l are different from zero.*

The set of all r-dimensional multiindices will be denoted by J. For $\alpha \in J$ we use the following definitions:

$|\alpha| := \sum_{l,k}\alpha_l^k$, length of α;

$d(\alpha) := \max\{k \ge 1 :\ \alpha_k^l > 0 \text{ for some } 1 \le l \le r\}$, order of α.

We also write $\alpha! = \prod_{k,l}(\alpha_k^l!)$.

Let us fix an arbitrary orthonormal system $\{m_k\} = \{m_k(s)\}_{k\ge 1}$ in the space $L_2([0,t])$ of square integrable functions on $[0,t]$ and set

$$\xi_{k,l} = \int_0^t m_k(s)dy^l(s).$$

Note that due to property (i) of the measure $\tilde{\mathrm{P}}$, $\xi_{k,l}$ are independent Gaussian random variables with zero mean and unit variance.

Let H_n be the nth Hermite polynomial defined by

$$H_n(x) = (-1)^n e^{x^2/2}\frac{d^n}{dx^n}e^{-x^2/2}. \tag{2.5}$$

It is well known (see, e.g., [7] or Theorem A.1) that the collection

$$\left\{\xi_\alpha := \prod_{k,l}\left(\frac{H_{\alpha_k^l}(\xi_{k,l})}{\sqrt{\alpha_k^l!}}\right),\quad \alpha \in J\right\}$$

is a complete orthonormal system (CONS) in $L_2(\Omega, \mathcal{F}_t^y, \tilde{\mathrm{P}})$.

To illustrate how the system is constructed, consider the case $r = 1$. Then α is a multiindex of the form $(\alpha_1, \alpha_2, \dots)$. If $|\alpha| = 0$ (i.e., $\alpha = (0, 0, \dots)$), then obviously $\xi_\alpha \equiv 1$.

If $|\alpha| = 1$, then the multiindex α is of the form $(0, \dots, 0, 1, 0, \dots)$ (i.e., $\alpha_i = 1$, $\alpha_k = 0$, $k \neq i$). In this case, $\xi_\alpha = \int_0^t m_i(s)dy(s)$.

Similarly, if $|\alpha| = 2$, then α is of either the form

$$(0, \dots, 0, 1, 0, \dots, 0, 1, 0, \dots)$$

(if $i < j$ and $\alpha_i = \alpha_j = 1$, $\alpha_k = 0, k \neq i, j$) or the form $(0, \dots, 0, 2, 0, \dots)$ (if $i = j$). For such α we have

$$\xi_\alpha = \Big(\int_0^t m_i(s)dy(s) \Big) \Big(\int_0^t m_j(s)dy(s) \Big)$$

in the first case,

$$\xi_\alpha = \frac{1}{\sqrt{2}} \Big[\Big(\int_0^t m_i(s)dy(s) \Big)^2 - 1 \Big]$$

in the second case, and so on. See also Remark A.2.

First, we will focus on the expansion of the UFD in the Wick polynomials ξ_α. To determine the coefficients of the expansion we consider the following system of deterministic PDEs:

$$\begin{aligned} \frac{\partial \varphi_\alpha(s,x)}{\partial s} &= \mathcal{L}^* \varphi_\alpha(s,x) + \textstyle\sum_{k,l} \alpha_k^l m_k(s) h^l(x) \varphi_{\alpha(k,l)}(s,x), \quad 0 < s \leq t, \\ \varphi_\alpha(0,x) &= p(x) 1_{\{|\alpha|=0\}}, \end{aligned} \tag{2.6}$$

where $\alpha = (\alpha_k^l)_{1 \leq l \leq r,\ k \geq 1} \in J$ and $\alpha(i,j)$ stands for the multiindex $\tilde{\alpha} = (\tilde{\alpha}_k^l)_{1 \leq l \leq r,\ k \geq 1}$ with

$$\tilde{\alpha}_k^l = \begin{cases} \alpha_k^l & \text{if } k \neq i \text{ or } l \neq j \text{ or both}, \\ \max(0, \alpha_i^j - 1) & \text{if } k = i \text{ and } l = j. \end{cases} \tag{2.7}$$

This system is recursive in $|\alpha|$: once we know the functions φ_α for all α of length $|\alpha| = k$, we can compute all φ_α for $|\alpha| = k + 1$. To illustrate the idea, again consider the case $r = 1$. Let us write φ_0 for the φ_α with $\alpha = (0, 0, \dots, 0, \dots)$ ($|\alpha| = 0$). Then $\varphi_0(s,x)$ satisfies the forward Kolmogorov equation corresponding to the state process:

$$\begin{aligned} \frac{\partial \varphi_0(s,x)}{\partial s} &= \mathcal{L}^* \varphi_0(s,x), \\ \varphi_0(0,x) &= p(x). \end{aligned}$$

If $|\alpha| = 1$ with $\alpha_i = 1$ and we write φ_i for φ_α with this α, then the corresponding equation in (2.6) becomes

$$\begin{aligned} \frac{\partial \varphi_i(s,x)}{\partial s} &= \mathcal{L}^* \varphi_i(s,x) + m_i(s) h(x) \varphi_0(s,x), \\ \varphi_i(0,x) &= 0. \end{aligned}$$

For $|\alpha| = 2$, the corresponding function $\varphi_{ij},\ i \leq j$, satisfies the equation

$$\begin{aligned} \frac{\partial \varphi_{ij}(s,x)}{\partial s} &= \mathcal{L}^* \varphi_{ij}(s,x) + m_i(s) h(x) \varphi_j(s,x) + m_j(s) h(x) \varphi_i(s,x), \\ \varphi_{ij}(0,x) &= 0, \end{aligned}$$

and so on.

Under assumptions (A1) and (A3), system (2.6) has a unique smooth solution (see Proposition A.1 for details).

Our approach is based on the following expansion of the UFD.

THEOREM 2.1 (Mikulevicius and Rozovskii [30, 31]). *Assume* (A1)–(A3). *Then for each* $x \in \mathbf{R}^d$ *the UFD is given by*

$$u(t,x) = \sum_{\alpha \in J} \frac{1}{\sqrt{\alpha!}} \varphi_\alpha(t,x)\xi_\alpha \quad (\mathrm{P}-\text{a.s.}). \tag{2.8}$$

This series converges in $L_2(\Omega, \tilde{\mathrm{P}})$, *and* $L_1(\Omega, \mathrm{P})$, *and the following Parseval's equality holds:*

$$\tilde{\mathbf{E}}|u(t,x)|^2 = \sum_{\alpha \in J} \frac{1}{\alpha!} |\varphi_\alpha(t,x)|^2. \tag{2.9}$$

Proof of this theorem is given in the appendix.

For the computational purposes one needs to truncate the sum in the expansion of u. This sum is "double infinite." Writing

$$u(t,x) = \sum_{k=0}^{\infty} \sum_{|\alpha|=k} \frac{1}{\sqrt{\alpha!}} \varphi_\alpha(t,x)\xi_\alpha, \tag{2.10}$$

one can see that for $k \geq 1$ there are infinitely many multiindices α with $|\alpha| = k$. To make it finite, we have to bound the length $|\alpha|$ of α and also the order $d(\alpha)$ of α: if $d(\alpha) \leq n$, then there are at most $(nr)^k$ multiindices α with $|\alpha| = k$.

Recall that if $\alpha = (\alpha_k^l)_{1 \leq l \leq r,\ k \geq 1}$, then α_k^l defines the degree of the Hermite polynomial of $\int_0^t m_k(s)dy^l(s)$ used in the construction of ξ_α. If $d(\alpha) \leq n$, then $\alpha_k^l = 0$ for all $k > n$, so the truncation of the order of α is equivalent to keeping only the first n elements of the (deterministic) basis $\{m_k(s)\}_{k \geq 1}$.

On the other hand, by restricting the length of α, we eliminate a number of elements of the stochastic basis $\{\xi_\alpha\}$, which are otherwise available with the retained collection of $\{m_k\}$.

Thus, restriction of the order of α makes the inner sum in (2.10) finite and is equivalent to the truncation of the deterministic basis $\{m_k\}$, while restriction of the length of α makes the outer sum in (2.10) finite and is equivalent to the truncation of the stochastic basis ξ_α.

The following theorem gives the upper bound on the error that one makes by doing both truncations for a particular choice of the basis $\{m_k\}$.

THEOREM 2.2. *Suppose that assumptions* (A1)–(A3) *hold and the deterministic basis* $\{m_k\}$ *is chosen as follows:*

$$m_1(s) = \frac{1}{\sqrt{t}}, \quad m_k(s) = \sqrt{\frac{2}{t}} \cos\Big(\frac{\pi(k-1)s}{t}\Big),\ k > 1,\ 0 \leq s \leq t.$$

Write $J_N^n = \{\alpha \in J : |\alpha| \leq N,\ d(\alpha) \leq n\}$ *and define*

$$u_N^n(t,x) := \sum_{\alpha \in J_N^n} \frac{1}{\sqrt{\alpha!}} \varphi_\alpha(t,x)\xi_\alpha. \tag{2.11}$$

Then

$$\tilde{\mathbf{E}}||u_N^n(t,\cdot) - u(t,\cdot)||_{L_2}^2 \leq Be^{Bt}\Big(\frac{(h_0 t)^{N+1}}{(N+1)!}||p||_{L_2}^2 + \frac{t^3}{n}||p||_{\mathbf{H}^2}^2\Big), \tag{2.12}$$

$$\sup_{x\in\mathbf{R}^d}\tilde{\mathbf{E}}|u_N^n(t,x)-u(t,x)|^2\le Ce^{Ct}\Big(\frac{(h_0t)^{N+1}}{(N+1)!}||p||^2_{\mathbf{C}_b^0}+\frac{t^3}{n}||p||^2_{\mathbf{C}_b^2}\Big). \tag{2.13}$$

Constants B and C depend only on the coefficients b, σ, and h of the model and $h_0:=\sum_{l=1}^r\sup_{x\in\mathbf{R}^d}|h^l(x)|^2$.

This and the following theorems will be proven in section 3.

REMARK 2.1. *For different k and l, random variables $\int_0^t m_k(s)dy^l(s)$, which make the stochastic basis ξ_α, are idependent and identically distributed $\mathcal{N}(0,1)$ under measure $\tilde{\mathrm{P}}$ for any CONS $\{m_k\}$. This suggests that the part of the error due to the truncation in the length of α should be independent of the choice of $\{m_k\}$, and the analysis of the proof shows that this is indeed the case. On the other hand, the error due to the truncation of the order of α crucially depends on the choice of $\{m_k\}$ (see also Remark* 3.1).

Truncations in the order and in the length can be done independently of each other. If $n=\infty$, we have truncation in length only; this case was studied by Mikulevicius and Rozovskii [30].

The Hermite–Fourier coefficients φ_α in (2.10) and (2.11) can be computed off-line, since system (2.6) does not involve the observation process y. In spite of this important property, approximation (2.11) does not yet provide an effective numerical algorithm for computing the UFD. The major reason for this is that the error of truncation may grow exponentially with t, so we can expect (2.11) to give a good approximation only for sufficiently small t. The above is a typical problem for approximations of solutions of parabolic equations (both deterministic and stochastic). One can try to offset this effect by choosing a higher-order approximation (in our case by taking larger N and n). However, higher-order numerical schemes are slower and often numerically unstable. A standard way to overcome the exponential growth of the truncation errors is to develop a recursive procedure by iterating the one-step approximation.

REMARK 2.2. *Of course, for the recursive approximation to converge, it is necessary that the error of the one-step approximation converges to zero fast enough as $t\downarrow 0$. By Theorem* 2.2*, the short time asymptotics of the error of approximation* (2.11) *are of order t if $N=1$ and of order $t^{3/2}$ if $N>1$, so it is possible to use* (2.11) *to costruct a multistep approximation* (*Theorem* 2.4).

In what follows, we present a recursive version of the expansion (2.8). It will allow us to modify the corresponding numerical scheme and eliminate the possible error growth.

Let $0=t_0<t_1<\cdots<t_M=T$ be a uniform partition of the interval $[0,T]$ with step Δ (so that $t_i=i\Delta$, $i=0,\dots,M$). Let $\{m_k^i\}=\{m_k^i(s)\}_{k\ge1}$ be a CONS in $L_2([t_{i-1},t_i])$. We also define random variables

$$\left\{\xi_\alpha^i:=\prod_{k,l}\left(\frac{H_{\alpha_k^l}(\xi_{k,l}^i)}{\sqrt{\alpha_k^l!}}\right),\quad \alpha\in J\right\}, \tag{2.14}$$

where $\xi_{k,l}^i=\int_{t_{i-1}}^{t_i}m_k^i(s)dy^l(s)$ and H_n is the nth Hermite polynomial (2.5).

Consider the following system of equations:

$$\begin{aligned}\frac{\partial\varphi_\alpha^i(s,x,g)}{\partial s}&=\mathcal{L}^*\varphi_\alpha^i(s,x,g)+\sum_{k,l}\alpha_k^lm_k^i(s)h^l(x)\varphi_{\alpha(k,l)}^i(s,x,g),\qquad t_{i-1}<s\le t_i,\\ \varphi_\alpha^i(t_{i-1},x,g)&=g(x)1_{\{|\alpha|=0\}},\end{aligned} \tag{2.15}$$

where $g(x)$ is a function to be determined. For each $i = 1, \dots, M$ this system is similar to (2.6). The main new feature is that the initial time moment is no longer zero and we now allow *that an arbitrary initial condition g may be different for different i*; this dependence on g is indicated explicitly in the arguments of φ.

The following is the recursive version of Theorem 2.1.

THEOREM 2.3. *Define $u(t_0, x) := p(x)$. Then for each $x \in \mathbf{R}^d$ and each t_i, $i = 1, \dots, M$, the UFD is given by*

$$u(t_i, x) = \sum_{\alpha \in J} \frac{1}{\sqrt{\alpha!}} \varphi_\alpha^i(t_i, x, u(t_{i-1}, \cdot)) \xi_\alpha^i, \quad i = 1, \dots, M \ (P - \ a.s.). \tag{2.16}$$

This series converges in $L_2(\Omega, \tilde{\mathrm{P}})$ and $L_1(\Omega, \mathrm{P})$, and the following Parseval's equality holds:

$$\tilde{\mathbf{E}} |u(t_i, x)|^2 = \sum_{\alpha \in J} \frac{1}{\alpha!} \tilde{\mathbf{E}} |\varphi_\alpha(t, x, u(t_{i-1}, \cdot))|^2, \qquad i = 1, \dots, M.$$

This result follows easily from Theorem 2.1, since (2.4) is linear with a unique solution, and random variables $u(t_{i-1}, x)$ and ξ_α^i are independent under measure $\tilde{\mathrm{P}}$.

Again, for computational purposes, we need to perform truncations in (2.16). For that purpose, as in Theorem 2.2, we will use the special basis $\{m_k^i\}$:

$$\begin{aligned} &m_k^i(s) = m_k(s - t_{i-1}), \qquad t_{i-1} \le s \le t_i, \\ &m_1(s) = \frac{1}{\sqrt{\Delta}}, \quad m_k(s) = \sqrt{\frac{2}{\Delta}} \cos\Big(\frac{\pi(k-1)s}{\Delta}\Big), \ k > 1, \quad 0 \le s \le \Delta, \\ &m_k(s) = 0, \qquad k \ge 1, \ s \notin [0, \Delta]. \end{aligned} \tag{2.17}$$

THEOREM 2.4. *Suppose that basis $\{m_k^i\}$ is given by* (2.17) *and assumptions* (A1)–(A3) *hold. Define $u_N^n(t_0, x) := p(x)$ and by induction*

$$u_N^n(t_i, x) := \sum_{\alpha \in J_N^n} \frac{1}{\sqrt{\alpha!}} \varphi_\alpha^i(\Delta, x) \xi_\alpha^i, \tag{2.18}$$

where $J_N^n = \{\alpha \in J : |\alpha| \le N, \ d(\alpha) \le n\}$ and φ_α^i are solutions of the system

$$\begin{aligned} &\frac{\partial \varphi_\alpha^i(s, x)}{\partial s} = \mathcal{L}^* \varphi_\alpha^i(s, x) + \textstyle\sum_{k,l} \alpha_k^l m_k(s) h^l(x) \varphi_{\alpha(k,l)}^i(s, x), \quad 0 < s \le \Delta, \\ &\varphi_\alpha^i(0, x) = u_N^n(t_{i-1}, x) 1_{\{|\alpha| = 0\}}. \end{aligned} \tag{2.19}$$

Then

$$\max_{1 \le i \le M} \tilde{\mathbf{E}} ||u_N^n(t_i, \cdot) - u(t_i, \cdot)||_{L_2}^2 \le B e^{BT} \Big(\frac{(h_0 \Delta)^N}{(N+1)!} ||p||_{L_2}^2 + \frac{\Delta^2}{n} ||p||_{\mathbf{H}^2}^2 \Big), \tag{2.20}$$

$$\max_{1 \le i \le M} \sup_{x \in \mathbf{R}^d} \tilde{\mathbf{E}} |u_N^n(t_i, x) - u(t_i, x)|^2 \le C e^{CT} \Big(\frac{(h_0 \Delta)^N}{(N+1)!} ||p||_{\mathbf{C}_b^0}^2 + \frac{\Delta^2}{n} ||p||_{\mathbf{C}_b^2}^2 \Big). \tag{2.21}$$

Constants B and C depend only on the coefficients b, σ, and h of the model and $h_0 := \sum_{l=1}^r \sup_{x \in \mathbf{R}^d} |h^l(x)|^2$.[3]

[3] Of course, B and C here are, in general, different from constants B and C in Theorem 2.2.

The sequence $\{u_N^n(t_i,x)\}_{1\le i\le M}$ gives an approximation to the UFD at all points of the time grid. This is a flexible and comparatively universal approximation. Many well-known numerical schemes for the Zakai equation can be obtained as particular cases of (2.18). In section 4 we will demonstrate this for two well-known algorithms: the explicit Euler scheme and the splitting-up method.

REMARK 2.3. *Analysis of the proofs of Theorems* 2.2–2.4 *shows that the wavelet type structure of our "global" basis* $\mathcal{M}=\cup_{i=1}^M\cup_{k=1}^\infty\{m_k^i\}$ *is of central importance.*

Specifically the following two properties of the basis are crucial:

(1) *The global basis* $\mathcal{M}$ *is a direct sum of "local" bases* $\mathcal{M}^i=\cup_{k=1}^\infty\{m_k^i\}$ *with nonoverlapping supports* (*Theorems* 2.3–2.5).

(2) *The functions* $m_k^i(s)$ *are smooth and* $\int_{t_{i-1}}^{t_i} m_k^i(s)ds=0$ *for* $k\ge 2$, $i=1,\dots,M$ (*see Theorem* 2.4).

The recursive version (2.18) of the spectral approximation of the unnormalized filtering density has one important disadvantage as compared to the one-step approximation (2.11). Indeed, to compute $u_N^n(t_i,x)$ we have to solve a certain number of equations from system (2.19). Although these equations are the same on every time interval and their coefficients do not involve the observation process y, the initial condition for the first equation of the system, $u_N^n(t_{i-1},x)$, does. This fact of course rules out off-line computation of the Fourier–Hermite coefficients $\varphi_\alpha(t,x)$, which is one of the important objectives of our study. For this reason, we present below a modification of the expansion (2.18) which admits off-line computations. Loosely speaking, the idea is to expand the initial condition for the first equation of (2.19) in a Fourier series as a function of spatial variable x, $u_N^n(t_{i-1},x)=\sum_l c_l e_l(x)$, and to exploit the obvious relation

$$\varphi_\alpha(t_i,x,u(t_{i-1},x))=\sum_l c_l\varphi_\alpha(t_i,x,e_l).$$

Note that the functions $\varphi_\alpha(t_i,x,e_l)$ can be computed off-line.

THEOREM 2.5. *Let* $\{e_l\}=\{e_l(x)\}_{l\ge1}$, $e_l\in\cap_n\mathbf{H}^n$, *be a CONS in* $L_2(\mathbf{R}^d)$ *and* $(\cdot,\cdot)$ *be the inner product in that space. Suppose that assumptions* (A1)–(A3) *hold and* $\{m_k^i\}$ *are given by* (2.17). *Consider the following system of equations:*

(2.22)
$$\frac{\partial\varphi_\alpha(s,x,g)}{\partial s}=\mathcal{L}^*\varphi_\alpha(s,x,g)+\sum_{k,l}\alpha_k^l m_k(s)h^l(x)\varphi_{\alpha(k,l)}(s,x,g),\qquad 0<s\le\Delta,$$
$$\varphi_\alpha(0,x,g)=g(x)1_{\{|\alpha|=0\}}.$$

Define $q_{\alpha k}^l:=(\varphi_\alpha(\Delta,\cdot,e_k),e_l)$ *and then by induction*

(2.23)
$$\psi_l(0,N,n):=(p,e_l),$$
$$\psi_l(i,N,n):=\sum_{\alpha\in J_N^n}\sum_k\frac{1}{\sqrt{\alpha!}}\psi_k(i-1,N,n)q_{\alpha k}^l\xi_\alpha^i.$$

Then

(2.24)
$$u_N^n(t_i,x)=\sum_l\psi_l(i,N,n)e_l(x),\ 0\le i\le M\quad(P-\ a.s.).$$

Now we can describe an approximation algorithm which stems from Theorem 2.5.

(1) Before the observations become available, (a) choose a finite collection $\{e_l\}_{1\le l\le\kappa}$;

(b) compute $\psi_l(0,N,n,\kappa) := (p,e_l),\ 1 \le l \le \kappa$, where p is the initial density;
(c) for all $\alpha \in J_N^n$ and $l = 1,\dots,\kappa$ compute $\varphi_\alpha(\Delta,x,e_l)$;
(d) compute $q_{\alpha k}^l = (\varphi_\alpha(\Delta,\cdot,e_k),e_l)$.
(2) On the ith step, when the observations become available,
(a) compute ξ_α^i;
(b) compute $\psi_l(i,N,n,\kappa) := \sum_{\alpha\in J_N^n}\sum_{k=1}^{\kappa}(1/\sqrt{\alpha!})\psi_k(i-1,N,n,\kappa)q_{\alpha k}^l\xi_\alpha^i$ for $l = 1,\dots,\kappa$;
(c) compute

$$u_N^{n,\kappa}(t_i,x) := \sum_{l=1}^{\kappa}\psi_l(i,N,n,\kappa)e_l(x). \tag{2.25}$$

.

We refer to this algorithm as the spectral separating scheme (S^3).

REMARK 2.4. *The amount of on-line operations and the amount of information that has to be stored in each step of* S^3 *do not depend on the number of steps to be performed. Also in contrast to the standard time-discretization schemes for the Zakai equation,* S^3 *does not require computing of the UFD at all the grid points* t_i, $i = 1,\dots,M$. *Specifically, step* 2(c) *of the algorithm can be omitted on any subset of time grid points (e.g., everywhere except the final point* t_M*). Note that computing of* (2.25) *is time consuming since it has to be done at all points of the space mesh.*

The truncation of the basis $\{e_l\}$ assumed in the above algorithm is necessary for computational reasons. Obviously it adds an extra error to (2.20). It is also clear that the error depends on the choice of the basis $\{e_l\}$ and is very much related to the particular numerical scheme used to solve (2.22).

It is beyond the scope of this work to study the above questions in detail, so we restrict ourselves to one particular case.

THEOREM 2.6. *Suppose that* $\{e_l\}$ *is the Hermite basis in* $L_2(\mathbf{R}^d)$ [16].

Let $0 = t_0 < \cdots < t_M = T$ *be a uniform partition of* $[0,T]$ *and* $u_N^n(t_i,x)$ *and* $u_N^{n,\kappa}(t_i,x)$ *be defined by* (2.18) *and* (2.25)*, respectively. Assume that* (A1)–(A3) *hold and in addition the initial density* p *and all its derivatives decay faster than any negative power of* $|x|$ *as* $|x| \to \infty$.

Then for any positive integer γ *there is a real number* $C_\gamma > 0$ *depending only on* γ *and the parameters* σ, b, p, *and* d *of the model such that*

$$\max_{1\le i\le M}\sqrt{\tilde{\mathbf{E}}||u_N^n(t_i,\cdot) - u_N^{n,\kappa}(t_i,\cdot)||_{L_2}^2} \le \frac{MC_\gamma(e^{C_\gamma T}-1)}{T\kappa^{\gamma-1/2}}. \tag{2.26}$$

This theorem shows that for sufficiently smooth initial condition p and with appropriate choice of the basis $\{e_l\}$, the error due to the truncation of the basis decays faster than any power of κ; i.e., our approximation is of a "spectral quality" (see, e.g., [15]).

REMARK 2.5. *The overall error of approximation for the spectral separating scheme follows from* (2.20) *and* (2.26) *and is given by*

$$\max_{1\le i\le M}\tilde{\mathbf{E}}||u(t_i,\cdot) - u_N^{n,\kappa}(t_i,\cdot)||_{L_2}^2 \le C\Big(\frac{(h_0\Delta)^N}{(N+1)!} + \frac{\Delta^2}{n} + \frac{C(\gamma)}{\Delta^2\kappa^{2\gamma-1}}\Big), \tag{2.27}$$

where C *is a constant depending on the parameters of the model (including the initial density* p *and the length of the time interval* T*) and it is assumed that the Wiener integrals* $\int_{t_{i-1}}^{t_i} m_k(t)dy^l(t)$ *are computed exactly. If* $n = 1$, *then only increments of the*

observation process are needed at each step and the computation of the integrals does not introduce any additional error. For $n > 1$, the integrals $\int_{t_{i-1}}^{t_i} m_k(t)dy^l(t),\ k > 1$, can be reduced to Riemann integrals and then approximated by subdividing the interval $[t_{i-1}, t_i]$ with some step $\delta << \Delta$. The error of the corresponding approximation will depend on the new asymptotic parameter δ. Still, formula (2.27) *implies that, in the limit $\lim_{\Delta\to 0}\lim_{\kappa\to\infty}$, the schemes with $n = 1$ and $n > 1$ have the same rate of convergence.*

REMARK 2.6. *Another approximation based on the Haar basis was proposed by Budhiraja and Kallianpur* [5]. *The approximation in* [5] *converges when $N \uparrow \infty$ and $\Delta \downarrow 0$. For computational purposes, though, it can be difficult to take arbitrarily large values of N because of the growing complexity and possible numerical instability. On the other hand, it follows from* (2.27) *that the spectral separating scheme converges in the limit $\lim_{\Delta\to 0}\lim_{\kappa\to\infty}$ for every $N \geq 1$ and the rate of convergence is the same for all $N \geq 2$.*

We remark that Theorems 2.1–2.6 can be extended to the case of time-dependent coefficients. Theorems 2.1–2.6 hold if the coefficients belong to the Hölder space $\mathbf{C}^{2+\alpha}(\mathbf{R}^d)$ for each t. The generalization is straightforward yet a bit cumbersome. Theorems 2.1–2.2 can be carried over to the case of correlated noises without many changes in the proofs [29, 31]. On the other hand the error estimates in the latter case are more delicate.

By no means is our approach a universal one. For example, it requires advanced knowledge of the parameters of the system, which are not always readily available. Also, it is not clear if it could be extended to the case of a non-Markov state process.

3. Proofs. In this section we will prove Theorems 2.2, 2.4, and 2.5. Everywhere C stands for a positive constant depending only on the parameters of the system; its actual value may be different in different places.

We introduce the following notation:

$\{T_s\}_{s\geq 0}$, the semigroup generated by the operator $\mathcal{L}^*$;

s^k, the the ordered set $(s_1, \dots, s_k)$; $ds^k := ds_1 \dots ds_k$;

$F(t; s^k; x) := T_{t-s_k} h T_{s_k - s_{k-1}} \dots h T_{s_1} p(x),\ k \geq 1$;

$\int^{(k)}(\cdots)ds^k := \int_0^t \int_0^{s_k} \dots \int_0^{s_2} (\cdots) ds_1 \dots ds_k$.

When $r = 1$, each multiindex $\alpha = (\alpha_1, \alpha_2, \dots)$ of length $|\alpha| = k$ can be identified with a vector $K_\alpha = (i_1^\alpha, \dots, i_k^\alpha)$, where $i_1^\alpha \leq i_2^\alpha \leq \dots \leq i_k^\alpha$. The first entry i_1^α of K_α is the number of the first nonzero element of α. The second entry i_2^α is equal to i_1^α if that first nonzero element $\alpha_{i_1^\alpha}$ is greater than 1; otherwise i_2^α is the number of the second nonzero element and so on. As a result, if $\alpha_j > 0$, then exactly α_j entries of the vector K_α are equal to j. We will call this vector the *characteristic set* of multiindex α. For example, if $\alpha = (0, 1, 0, 2, 3, 0, \dots)$, then nonzero elements are $\alpha_2 = 1,\ \alpha_4 = 2,\ \alpha_5 = 3$, and the characteristic set is $(2, 4, 4, 5, 5, 5)$. A similar construction is possible for general $r > 1$. In the future, when there is no danger of confusion, we will omit the upper index in i (i.e., write i_j rather than i_j^α).

Let $\mathcal{P}^k$ be the permutation group of the set $\{1, \dots, k\}$. For a given $\alpha \in J$ with $|\alpha| = k$ and the characteristic set $(i_1, \dots, i_k)$ $(r = 1)$ define

$$E_\alpha(s^k) := \sum_{\sigma\in\mathcal{P}^k} m_{i_1}(s_{\sigma(1)}) \cdots m_{i_k}(s_{\sigma(k)}).$$

Proof of Theorem 2.2. We will prove inequality (2.12); the other can be proven in a similar way.

Set

$$u_N(t,x) := \sum_{|\alpha|\le N} \frac{\varphi_\alpha(t,x)\xi_\alpha}{\sqrt{\alpha!}}.$$

Suppose that we know that

$$\tilde{\mathbf{E}}||u(t,\cdot) - u_N(t,\cdot)||^2_{L_2} \le \frac{(h_0 t)^{N+1}}{(N+1)!} e^{Ct} ||p||^2_{L_2} \tag{3.1}$$

and

$$\tilde{\mathbf{E}}||u_N(t,\cdot) - u^n_N(t,\cdot)||^2_{L_2} \le C\frac{t^3}{n} e^{Ct} ||p||^2_{\mathbf{H}^2}. \tag{3.2}$$

Then (2.12) will follow immediately from the inequality $(a+b)^2 \le 2(a^2+b^2)$.

The problem is thus to prove (3.1) and (3.2). To simplify the presentation, we assume that $r = 1$.

Proof of (3.1). We will use the following results:

$$\sum_{|\alpha|=k} \frac{\varphi^2_\alpha(t,x)}{\alpha!} = \int^{(k)} |F(t;s^k;x)|^2 ds^k, \tag{3.3}$$

where φ_α is the solution of (2.6) with any CONS $\{m_k\}$, and

$$||T_s f||^2_{L_2} \le e^{Cs} ||f||^2_{L_2}. \tag{3.4}$$

The first equality is established in the appendix, Proposition A.1 (see also [30]); inequality (3.4) is a standard fact (see [25]).

Since ξ_α are uncorrelated under $\tilde{\mathrm{P}}$, we have

$$\tilde{\mathbf{E}}|u(t,x) - u_N(t,x)|^2 = \sum_{k>N} \sum_{|\alpha|=k} \frac{\varphi^2_\alpha(t,x)}{\alpha!} = \sum_{k>N} \int^{(k)} |F(t;s^k;x)|^2 ds^k.$$

Then by the Fubini theorem

$$\begin{aligned}\tilde{\mathbf{E}}||u(t,\cdot) - u_N(t,\cdot)||^2_{L_2} &= \sum_{k>N} \int^{(k)} \Big(\int_{\mathbf{R}^d} |F(t;s^k;x)|^2 dx \Big) ds^k \\ &= \sum_{k>N} \int^{(k)} ||F(t,s^k,\cdot)||^2_{L_2} ds^k.\end{aligned}$$

Since h is bounded, it follows from the definition of F and (3.4) that

$$\begin{aligned}||F(t;s^k;\cdot)||^2_{L_2} &\le h_0 e^{C(t-s_k)} ||T_{s_k - s_{k-1}} h \dots h T_{s_1} p||^2_{L_2} \\ &\le \dots \le h_0^k e^{t - s_k + s_k - \dots + s_2 - s_1 + s_1} ||p||^2_{L_2} = h_0^k e^{Ct} ||p||^2_{L_2}.\end{aligned}$$

Finally, from $\int^{(k)} ds^k = t^k/k!$, we conclude that

$$\begin{aligned}\tilde{\mathbf{E}}||u(t,\cdot) - u_N(t,\cdot)||^2_{L_2} &\le e^{Ct} \sum_{k>N} \frac{(th_0)^k}{k!} \\ &\le \frac{(th_0)^{N+1}}{(N+1)!} e^{(C+h_0)t},\end{aligned}$$

and (3.1) follows. Note that it holds for any CONS $\{m_k\}$.

Proof of (3.2). If α is a multiindex with $|\alpha| = k$ and the characteristic set $(i_1^\alpha, \dots, i_k^\alpha)$, then $i_k^\alpha = d(\alpha)$, the order of α, and so the set J_N^n can be described as $\{\alpha \in J : |\alpha| \le N;\ i_{|\alpha|}^\alpha \le n\}$. Thus

$$\tilde{\mathbf{E}}|u_N^n(t,x) - u_N(t,x)|^2 = \sum_{l=n+1}^{\infty} \sum_{k=1}^{N} \sum_{|\alpha|=k; i_k^\alpha = l} \frac{\varphi_\alpha^2(t,x)}{\alpha!}.$$

The problem is thus to estimate $\sum_{l=n+1}^{\infty} \sum_{k=1}^{N} \sum_{|\alpha|=k; i_k^\alpha=l} \frac{\varphi_\alpha^2(t,x)}{\alpha!}$.

By Proposition A.1 (see also [30]) the corresponding solution φ_α of (2.6) can be written as

$$\varphi_\alpha(t,x) = \int^{(k)} F(t; s^k; x) E_\alpha(s^k) ds^k. \tag{3.5}$$

Note that we can also write

$$E_\alpha(s^k) = \sum_{j=1}^{k} m_{i_k}(s_j) E_{\alpha(i_k)}(s_j^k),$$

where s_j^k denotes the same set $(s_1, \dots, s_k)$ with omitted s_j (e.g., $s_1^k = (s_2, \dots, s_k)$) and $\alpha(i_k)$ is the multiindex with this characteristic set $(i_1, \dots, i_{k-1})$ (cf. (2.7); recall that $r = 1$).

This allows us to write (3.5) as

$$\varphi_\alpha(t,x) = \sum_{j=1}^{k} \int^{(k-1)} \Big(\int_{s_{j-1}}^{s_{j+1}} F(t; s^k; x) m_{i_k}(s_j) ds_j \Big) E_{\alpha(i_k)}(s_j^k) ds_j^k, \tag{3.6}$$

where $s_0 := 0;\ s_{k+1} := t$. (We just change the order of integration in the multiple integral.)

Denote

$$M_k(s) := \frac{\sqrt{2t}}{\pi(k-1)} \sin\Big(\frac{\pi(k-1)}{t} s\Big), \qquad k > 1,\ 0 \le s \le t,$$

and also $F_j := \frac{\partial F(t; s^k; x)}{\partial s_j}$. Then, as long as $i_k = l > 1$, we can integrate by parts the inner integral on the right of (3.6) to get

$$\begin{aligned} &\int_{s_{j-1}}^{s_{j+1}} F(t; s^k; x) m_l(s_j) ds_j \\ &\quad = F(t; s^k; x) M_l(s_j) \Big|_{s_j = s_{j-1}}^{s_j = s_{j+1}} - \int_{s_{j-1}}^{s_{j+1}} F_j(t, s^k, x) M_l(s_j) ds_j. \end{aligned}$$

For each j, let us rename the remaining variables s_j^k in (3.6) as follows: $t_i := s_i, i \le j-1;\ t_i := s_{i+1}, i > j-1$, or, symbolically, $t^{k-1} := s_j^k$. We will set $t_0 := 0$, $t_k := t$ and denote by $t^{k-1,j}, j = 1, \dots, k-1$, the set t^{k-1} in which t_j is repeated twice (e.g., $t^{k-1,1} = (t_1, t_1, \dots, t_{k-1})$, etc.); also $t^{k-1,0} := (t_0, t_1, t_2, \dots, t_{k-1})$, $t^{k-1,k} := (t_1, \dots, t_{k-1}, t_k)$.

Then

$$F(t;s^k;x)M_l(s_j)\Big|_{s_j=s_{j-1}}^{s_j=s_{j+1}}$$
$$= F(t;t^{k-1,j};x)M_l(t_j) - F(t;t^{k-1,j-1};x)M_l(t_{j-1}), \qquad j=1,\dots,k.$$

As a result, since $M_l(t_0)=M_l(t_k)=0$ (and this is the only place where the choice of $\{m_k\}$ really makes the difference), all these terms will cancel out as we sum over j. What remains can be written as

$$\int^{(k-1)} f_l(t;t^{k-1};x)E_{\alpha(l)}(t^{k-1})dt^{k-1},$$

where

$$f_l(t;t^{k-1};x) = -\int_0^{t_1} F_1(t;\tau,t^{k-1};x)M_l(\tau)d\tau$$
$$-\sum_{j=2}^{k-1}\int_{t_{j-1}}^{t_j} F_j(t;\dots,t_{j-1},\tau,t_j,\dots;x)M_l(\tau)d\tau$$
$$-\int_{t_{k-1}}^{t_k} F_k(t;t^{k-1},\tau;x)M_l(\tau)d\tau.$$

Then, since $|\alpha(i_{|\alpha|})| = |\alpha|-1$, $\alpha! \geq \alpha(i_{|\alpha|})!$, we get

$$\sum_{|\alpha|=k;i_k^\alpha=l} \frac{\varphi_\alpha^2(t,x)}{\alpha!}$$
$$= \sum_{|\alpha|=k;i_k^\alpha=l} \Big(\frac{1}{\sqrt{\alpha!}}\int^{(k-1)} f_l(t;t^{k-1};x)E_{\alpha(l)}(t^{k-1})dt^{k-1}\Big)^2$$
$$\leq \sum_{|\beta|=k-1} \Big(\frac{1}{\sqrt{\beta!}}\int^{(k-1)} f_l(t;t^{k-1};x)E_{\beta}(t^{k-1})dt^{k-1}\Big)^2,$$

and arguments similar to those used in the proof of Proposition A.1 show that the last expression is equal to

$$\int^{(k-1)} |f_l(t;t^{k-1};x)|^2 dt^{k-1}. \tag{3.7}$$

Direct computations yield

$$F_j(t,s^k,x) = T_{t-s_k}h\dots T_{s_{j+1}-s_j}h\mathcal{L}^*T_{s_j-s_{j-1}}\dots T_{s_1}p(x)$$
$$-T_{t-s_k}h\dots\mathcal{L}^*T_{s_{j+1}-s_j}hT_{s_j-s_{j-1}}\dots T_{s_1}p(x).$$

Since $\mathcal{L}^*$ is a continuous operator from $\mathbf{H}^2$ to L_2, it follows from (3.4) and a similar inequality for $\mathbf{H}^2$-norms that

$$||F_j(t;s^k;\cdot)||_{L_2}^2 \leq e^{Ct}C^k||p||_{\mathbf{H}^2}^2.$$

Then the definition of f_l and obvious inequalities

$$(a_1+\dots+a_k)^2 \leq k(a_1^2+\dots+a_k^2)$$

and

$$\left(\int_0^x f(y)dy\right)^2 \le x\int_0^x (f(y))^2 dy$$

imply

$$\begin{aligned}||f_l(t;t^{k-1};\cdot)||_{L_2}^2 &\le kC^k e^{Ct}||p||_{\mathbf{H}^2}^2 t\int_0^t (M_l(s))^2 ds\\ &\le \frac{kC^k t^3 e^{Ct}}{(l-1)^2}||p||_{\mathbf{H}^2}^2;\end{aligned}$$

so, since $\int^{(k-1)} dt^{k-1} = t^{k-1}/(k-1)!$, (3.7) and the last inequality yield

$$\sum_{|\alpha|=k, i_k^\alpha = l} \frac{||\varphi_\alpha^2(t,\cdot)||_{L_2}^2}{\alpha!} \le e^{Ct}||p||_{\mathbf{H}^2}^2 t^3 \frac{kC^k}{(l-1)^2(k-1)!}.$$

Now we collect everything to get

$$\begin{aligned}\tilde{\mathbf{E}}||u_N(t,\cdot) - u_N^n(t,\cdot)||_{L_2}^2 &= \sum_{l\ge n+1}\sum_{k=1}^N \sum_{|\alpha|=k; i_k^\alpha = l} \frac{\varphi_\alpha^2(t,x)}{\alpha!}\\ &\le Ct^3 e^{Ct}\Big(\sum_{k\ge 1}\frac{k(Ct)^{k-1}}{(k-1)!}\Big)\sum_{l\ge n}\frac{1}{l^2} \le C\frac{t^3}{n}e^{Ct}||p||_{\mathbf{H}^2}^2.\end{aligned}$$

This completes the proof of (3.2) and the theorem as a whole. □

Proof of Theorem 2.4. We again prove only the first inequality.

First of all notice that time homogeneity of (2.15) and the special choice of $\{m_k^i\}$ as $m_k^i(s) = m_k(s - t_{i-1})$ imply

$$\varphi_\alpha^i(\Delta, x) = \varphi_\alpha^i(t_i, x, u(t_{i-1},\cdot))$$

(see (2.15) and (2.19)). Then by Fubini's theorem and Theorem 2.3 and due to linearity of system (2.15),

$$\begin{aligned}\tilde{\mathbf{E}}||u_N^n(t_i,\cdot) - u(t_i,\cdot)||_{L_2}^2 &= \sum_{\alpha\in J_N^n}\frac{1}{\alpha!}\tilde{\mathbf{E}}||\varphi_\alpha^i(t_i,\cdot,u_N^n(t_{i-1},\cdot) - u(t_{i-1},\cdot))||_{L_2}^2\\ &\quad + \sum_{\alpha\notin J_N^n}\frac{1}{\alpha!}\tilde{\mathbf{E}}||\varphi_\alpha^i(t_i,\cdot,u(t_{i-1},\cdot))||_{L_2}^2\\ &\le \sum_{\alpha\in J}\frac{1}{\alpha!}\tilde{\mathbf{E}}||\varphi_\alpha^i(t_i,\cdot,u_N^n(t_{i-1},\cdot) - u(t_{i-1},\cdot))||_{L_2}^2\\ &\quad + \sum_{\alpha\notin J_N^n}\frac{1}{\alpha!}\tilde{\mathbf{E}}||\varphi_\alpha^i(t_i,\cdot,u(t_{i-1},\cdot))||_{L_2}^2.\end{aligned} \tag{3.8}$$

By Theorem 2.3 and linearity of equation (2.4), we have

$$\begin{aligned}&\sum_{\alpha\in J}\frac{1}{\alpha!}\tilde{\mathbf{E}}||\varphi_\alpha^i(t_i,\cdot,u_N^n(t_{i-1},\cdot) - u(t_{i-1},\cdot))||_{L_2}^2\\ &\qquad = \tilde{\mathbf{E}}||U(t_i,x;u_N^n(t_{i-1},\cdot) - u(t_{i-1},\cdot))||_{L_2}^2,\end{aligned} \tag{3.9}$$

where $U(t,x;u_N^n(t_{i-1},\cdot)-u(t_{i-1},\cdot))$ is the solution of

$$\begin{aligned} dv(t,x) &= \mathcal{L}^* v(t,x)dt + h^l(x)v(t,x)dy^l(t), \quad t\in(t_{i-1},t_i],\\ v(t_{i-1},x) &= u_N^n(t_{i-1},x)-u(t_{i-1},x). \end{aligned}$$

It is a standard fact that under assumptions (A1) and (A3),

$$\tilde{\mathbf{E}}||U(t_i,\cdot;u_N^n(t_{i-1},\cdot)-u(t_{i-1},\cdot))||_{L_2}^2 \le e^{C\Delta}\tilde{\mathbf{E}}||u_N^n(t_{i-1},\cdot)-u(t_{i-1},\cdot)||_{L_2}^2 \tag{3.10}$$

(see, e.g., [34]).

Repeating the same arguments as in the proof of Theorem 2.2, one can check that

$$\begin{aligned} &\sum_{\alpha\notin J_N^n}\frac{1}{\alpha!}\tilde{\mathbf{E}}||\varphi_\alpha^i(t_i,\cdot,u(t_{i-1},\cdot))||_{L_2}^2\\ &\qquad \le Ce^{C\Delta}\Big(\frac{(h_0\Delta)^{N+1}}{(N+1)!}\tilde{\mathbf{E}}||u(t_{i-1},\cdot)||_{L_2}^2+\frac{\Delta^3}{n}\tilde{\mathbf{E}}||u(t_{i-1},\cdot)||_{\mathbf{H}^2}^2\Big). \end{aligned} \tag{3.11}$$

Finally, we use the inequalities

$$\tilde{\mathbf{E}}||u(t_{i-1},\cdot)||_{L_2}^2 \le e^{CT}||p||_{L_2}^2 \tag{3.12}$$

and

$$\tilde{\mathbf{E}}||u(t_{i-1},\cdot)||_{\mathbf{H}^2}^2 \le e^{CT}||p||_{\mathbf{H}^2}^2. \tag{3.13}$$

These inequalities are similar to (3.10) and can also be found in [34].

If we now denote $\tilde{\mathbf{E}}||u_N^n(t_i,\cdot)-u(t_i,\cdot)||_{L_2}^2$ by ε_i, then, combining (3.8)–(3.13), we arrive at

$$\varepsilon_i \le \left(\varepsilon_{i-1}+Ce^{CT}\Big(\frac{(h_0\Delta)^{N+1}}{(N+1)!}||p||_{L_2}^2+\frac{\Delta^3}{n}\tilde{\mathbf{E}}||p||_{\mathbf{H}^2}^2\Big)\right)e^{C\Delta},$$

and since $\varepsilon_0=0$, the statement of the theorem follows from the discrete Gronwall lemma. □

Proof of Theorem 2.5. By construction, $u_N^n(t_i,\cdot)\in L_2(\mathbf{R}^d)$ (P − a.s.), so

$$u_N^n(t_i,x)=\sum_{l\ge1}\psi_l(i,N,n)e_l(x) \quad (\mathrm{P}-\text{a.s.})$$

with some $\psi_l(i,N,n)$. Then all we have to do is to establish (2.23), which means

$$\sum_{\alpha\in J_N^n}\sum_k\frac{1}{\sqrt{\alpha!}}\psi_k(i-1,N,n)q_{\alpha k}^l\xi_\alpha^i=(u_N^n(t_i,\cdot),e_l). \tag{3.14}$$

We will prove this by induction. For $i=0$, $\psi_l(0,N,n)=(u_N^n(t_0,\cdot),e_l)$ by definition. Assume that $u_N^n(t_{i-1},x)=\sum_l\psi_l(i-1,N,n)e_l(x)$ for some $i\ge1$.

The proof of Theorem 2.2 shows that operator $g\mapsto\varphi_\alpha(t_i,\cdot,g)$ is continuous and linear from $L_2(\mathbf{R}^d)$ to $L_2(\mathbf{R}^d)$ for all $\alpha\in J$, where $\varphi_\alpha(t_i,\cdot,g)$ is the solution of (2.22). Then

$$\begin{aligned} \sum_k\psi_k(i-1,N,n)q_{\alpha k}^l &= \sum_k\psi_k(i-1,N,n)(\varphi_\alpha(\Delta,\cdot,e_k),e_l)\\ &= \Big(\varphi_\alpha\Big(\Delta,\cdot,\sum_k\psi_k(i-1,N,n)e_k\Big),e_l\Big), \end{aligned}$$

and by an induction assumption the right-hand side of the above formula is equal to

$$(\varphi_\alpha(\Delta, \cdot, u_N^n(t_{i-1}, \cdot)), e_l).$$

On the other hand, comparing (2.19) and (2.22) we conclude that

$$\varphi_\alpha(\Delta, x, u_N^n(t_{i-1}, \cdot)) = \varphi_\alpha^i(\Delta, x).$$

As a result,

$$\sum_{\alpha \in J_N^n} \sum_k \frac{1}{\sqrt{\alpha!}} \psi_k(i-1, N, n) q_{\alpha k}^l \xi_\alpha^i = \Big(\sum_{\alpha \in J_N^n} \frac{1}{\sqrt{\alpha!}} \varphi_\alpha^i(t_i, \cdot, u_N^n(t_{i-1}, \cdot)) \xi_\alpha^i, e_l \Big),$$

and by (2.18) this is equal to $(u_N^n(t_i, \cdot), e_l)$. This completes the proof of (3.14) and the theorem as a whole. □

REMARK 3.1. *Analysis of the proof shows that the result (with obvious modifications) is also true for the exact solution $u(t_i, x)$.*

Proof of Theorem 2.6. In what follows, C_γ denotes a constant depending on γ and (maybe) the parameters of the model. As before, C is a constant depending only on the parametes of the model. Values of C and C_γ may be different in different places.

If $d = 1$, then

$$e_l(x) = \frac{1}{\sqrt{(2\pi)^{1/2} l!}} e^{-x^2/4} H_l(x), \tag{3.15}$$

where H_l is the lth Hermite polynomial (2.5) [15, 16].

For $d > 1$, the elements of the basis are

$$e_l(x_1, \dots, x_d) = e_{l_1}(x_1) \dots e_{l_d}(x_d),$$

where $l_i \geq 0$ and e_{l_i} are given by (3.15), $i = 1, \dots, d$ [16]. The system $\{e_l\}$ is thus indexed by the set of d-dimensional multiindices $l = (l_1, \dots, l_d)$ ordered in some natural way. We will say that $l \leq \kappa$ if $\max_{1 \leq i \leq d} l_i \leq \kappa$.

To simplify the presentation, we assume from now on that $d = 2$. Then $l = (l_1, l_2)$.

Direct computations show that e_{l_i} satisfies

$$\mathcal{A}_i e_{l_i} = (l_i + 1) e_{l_i}, \quad i = 1, 2,$$

where operator $\mathcal{A}_i$ is defined by

$$\mathcal{A}_i f(x) = -\frac{\partial^2 f(x)}{\partial x_i^2} + \frac{2 + x_i^2}{4} f(x).$$

As a result, $\mathcal{A}_2 \mathcal{A}_1 e_l(x_1, x_2) = (l_1 + 1) \mathcal{A}_2 (e_{l_1}(x_1) e_{l_2}(x_2)) = (l_1 + 1)(l_2 + 1) e_l(x_1, x_2)$. This means that if f and all its derivative decay fast enough, then

$$|(f, e_l)| \leq \frac{||\mathcal{A}_1 \mathcal{A}_2 f||_{L_2}}{(l_1 + 1)(l_2 + 1)} \leq \dots \leq \frac{||(\mathcal{A}_1 \mathcal{A}_2)^\gamma f||_{L_2}}{(l_1 + 1)^\gamma (l_2 + 1)^\gamma}. \tag{3.16}$$

If $\mathbf{H}^s(r)$, $s, r \in \mathbf{R}$, is the weighted Sobolev space $W_2^s(r, \mathbf{R}^2)$ ([34]; see also [16]), then definition of $\mathcal{A}_i$ implies that

$$||(\mathcal{A}_1 \mathcal{A}_2)^\gamma f||_{L_2} \leq C_\gamma ||f||_{\mathbf{H}^{4\gamma}(4\gamma)},$$

and (3.16) becomes

$$|(f,e_k)| \le \frac{C_\gamma ||f||_{\mathbf{H}^{4\gamma}(4\gamma)}}{(l_1+1)^\gamma (l_2+1)^\gamma}. \tag{3.17}$$

Introduce the following notation:
$|||f||| := \sqrt{\tilde{\mathbf{E}}||f||^2_{L_2}}$,
$\varepsilon_i := |||u^n_N(t_i,\cdot) - u^{n,\kappa}_N(t_i,\cdot)|||$,
Π^κ, the L_2-orthogonal projection on the subspace generated by $e_l,\ l \le \kappa$,
$V^{n,i}_N(f)$, the operator $f \mapsto \sum_{\alpha\in J^n_N} \varphi_\alpha(\Delta,\cdot,f)\xi^i_\alpha;\ f \in L_2(\Omega,\mathcal{F}^y_{t_{i-1}},\tilde{\mathrm{P}})$ and $U^i(f) := V^{\infty,i}_\infty$. Since $V^{n,i}_N(f)$ is the $L_2(\Omega,\tilde{\mathrm{P}})$-orthogonal projection of $U^i(f)$ on the subspace generated by $\{\xi_\alpha,\ \alpha \in J^n_N\}$,

$$|||V^{n,i}_N(f)||| \le |||U^i(f)|||. \tag{3.18}$$

Below we will be dealing with a fixed set (n,i,N) and to simplify notation will write V instead of $V^{n,i}_N$. We also omit the dot in $u^n_N(t_i,\cdot)$, etc.

Since the coefficients of the model are time independent,

$$u^n_N(t_i) = V(u^n_N(t_{i-1})), \qquad u^{n,\kappa}_N(t_i) = \Pi^\kappa V(u^{n,\kappa}_N(t_{i-1})).$$

The second equality follows from (2.25), the definition of $\psi_l(i,N,n,\kappa)$, and the linearity of the map $f \mapsto \varphi_\alpha(\Delta,f)$. Then by the triangle inequality

$$\begin{aligned}\varepsilon_i \le {} & |||\Pi^\kappa V(u^n_N(t_{i-1})) - \Pi^\kappa V(u^{n,\kappa}_N(t_{i-1}))||| \\ & + |||u^n_N(t_i) - \Pi^\kappa u^n_N(t_i)|||.\end{aligned} \tag{3.19}$$

By the definition of Π^κ,

$$\begin{aligned}&|||\Pi^\kappa V(u^n_N(t_{i-1})) - \Pi^\kappa V(u^{n,\kappa}_N(t_{i-1}))||| \\ &\quad \le |||V(u^n_N(t_{i-1})) - V(u^{n,\kappa}_N(t_{i-1}))||| \le e^{C\Delta}\varepsilon_{i-1},\end{aligned} \tag{3.20}$$

$\Delta = t_i - t_{i-1}$, where the last inequality follows from (3.18) and (3.10).

Under the assumptions of the theorem it is easy to show, using the standard estimates from [25] or [34], that for any $i = 1,\dots,M$, $u^n_N(t_i) \in \cap_s \mathbf{H}^s(r)$ $(\mathrm{P}-\text{a.s.})$ for any $r \in \mathbf{R}$. In addition,

$$\sum_{\alpha\in J^n_N} \frac{\tilde{\mathbf{E}}||\varphi^i_\alpha(\Delta)||^2_{\mathbf{H}^\gamma(\gamma)}}{\alpha!} \le e^{C_\gamma T}||p||^2_{\mathbf{H}^\gamma(\gamma)} \tag{3.21}$$

for any positive integer γ, where $\varphi^i_\alpha(\Delta) = \varphi_\alpha(\Delta,u(t_{i-1}))$. As a result, from (2.18), (3.17), (3.21), and the obvious estimates $\sum_{j>\kappa} 1/(j+1)^\gamma \le C_\gamma/(\kappa+1)^{\gamma-1} \le C_\gamma/\kappa^{\gamma-1}$ (valid for $\gamma > 1$), we conclude that

$$\begin{aligned}|||u^n_N(t_{i-1}) - \Pi^\kappa u^n_N(t_{i-1})|||^2 &= \sum_{\alpha\in J^n_N}\sum_{l>\kappa} \frac{\tilde{\mathbf{E}}(\varphi^i_\alpha(\Delta),e_l)^2}{\alpha!} \\ &\le 2\sum_{\alpha\in J^n_N}\Big(\sum_{l_1>\kappa,l_2\ge 0}\frac{1}{(l_1+1)^{2\gamma}(l_2+1)^{2\gamma}}\Big)\frac{\tilde{\mathbf{E}}||\varphi^i_\alpha(\Delta)||^2_{\mathbf{H}^{4\gamma}(4\gamma)}}{\alpha!} \\ &\le \frac{C_\gamma e^{C_\gamma T}}{\kappa^{2\gamma-1}}||p||^2_{\mathbf{H}^{4\gamma}(4\gamma)}.\end{aligned} \tag{3.22}$$

Combining (3.19), (3.20), and (3.22), we arrive at

$$\varepsilon_i \le e^{C\Delta}\varepsilon_{i-1} + \frac{C_\gamma e^{C_\gamma T}}{\kappa^{\gamma-1/2}}||p||_{\mathbf{H}^{4\gamma}(4\gamma)},$$

which by the discrete Gronwall lemma implies

$$\varepsilon_i \le \frac{C_\gamma(e^{C_\gamma T}-1)}{\Delta\ \kappa^{\gamma-1/2}}||p||_{\mathbf{H}^{4\gamma}(4\gamma)}.$$

Since $\Delta = T/M$ and by assumption $||p||_{\mathbf{H}^{4\gamma}(4\gamma)} \le C_\gamma$, (2.26) follows. □

4. Comparison with other algorithms and numerical simulations. The Wiener chaos approximations (2.8), (2.16) can be viewed as higher-order time-discretization schemes for the Zakai equation.

For $N \ge 2$, the rate of convergence of the Wiener chaos approximation u_N^n is $O(\Delta)$, where Δ is the time step (Theorem 2.4). This is similar to the rates of convergence of the splitting-up algorithm (see [26]) and the implicit Euler–Milstein scheme (see [18]) for the Zakai equation.

In fact, many well-known time-discretization schemes can be obtained as particular cases of the Wiener chaos approximation.

One of the simplest is the explicit Euler scheme. Take a uniform partition of the interval $[0,T]$ with step Δ. Then the explicit Euler approximation $u_i(x)$ to the Zakai equation is obtained from

(4.1)
$$u_0(x) = p(x), \quad u_i(x) = (1+\Delta\cdot\mathcal{L}^*)u_{i-1}(x) + \sum_{l=1}^{r} h^l(x)u_{i-1}(x)(y^l(t_i)-y^l(t_{i-1})).$$

Now we will derive the same result from Theorem 2.4. Take $n = N = 1$. Then set J_1^1 contains $r+1$ elements, and on each step we need to solve $r+1$ equations from (2.19):

$$\frac{\partial\varphi_0^i(s,x)}{\partial s} = \mathcal{L}^*\varphi_0^i(s,x), \quad 0 < s \le \Delta,$$
$$\varphi_0^i(0,x) = u_1^1(t_{i-1},x)$$

(for $|\alpha| = 0$);

$$\frac{\partial\varphi_l^i(s,x)}{\partial s} = \mathcal{L}^*\varphi_l^i(s,x) + \textstyle\sum_l \frac{h^l(x)}{\sqrt{\Delta}}\varphi_0^i(s,x), \quad 0 < s \le \Delta,$$
$$\varphi_l^i(0,x) = 0, \qquad l = 1,\dots,r$$

(for $|\alpha| = 1$ with $\alpha_1^l = 1$) and $u_1^1(t_0,x) = p(x)$.

We solve these equations using the explicit Euler scheme; the (approximate) solutions are then given by

$$\varphi_0^i(\Delta,x) = (1+\Delta\cdot\mathcal{L}^*)u_1^1(t_{i-1},x),$$

$$\varphi_l^i(\Delta,x) = h^l(x)\sqrt{\Delta}u_1^1(t_{i-1},x).$$

By definition,

$$\xi_l^i = \int_{t_{i-1}}^{t_i} m_1(s)dy^l(s) = \frac{y^l(t_i)-y^l(t_{i-1})}{\sqrt{\Delta}},$$

and by Theorem 2.4,

$$u_1^1(t_i,x)=\varphi_0^i(\Delta,x)+\sum_{l=1}^{r}\varphi_l^i(\Delta,x)\xi_l^i,$$

and this, due to the above relations, coincides with (4.1).

Another well-known algorithm for solving the Zakai equation (2.4) is the splitting-up approximation (see Bensoussan, Glowinski, and Rascanu [4]; Florchinger and LeGland [12]; etc.). For simplicity, we consider the case $r=1$. Take a uniform partition of $[0,T]$ with step Δ and let $\{T_t\}$ be the semigroup generated by operator $\mathcal{L}^*$ (or some approximation of that semigroup). Then the splitting-up approximation $u_i(x)$ to $u(t_i,x)$ is computed from the recursion

$$u_0(x)=p(x),\qquad u_i(x)=T_\Delta\exp([y(t_i)-y(t_{i-1})]h-0.5h^2\Delta)u_{i-1}(x). \tag{4.2}$$

Let us see how the same result can be obtained from Theorem 2.4. Set $n=1$, $N=\infty$. Then the set J_N^n consists of multiindices $\alpha=(k,0,0,\dots)$; the corresponding φ_α will be denoted by φ_k. We need to solve the following system:

$$\frac{\partial\varphi_0^i(s,x)}{\partial s}=\mathcal{L}^*\varphi_0^i(s,x),\quad 0<s\le\Delta,$$
$$\varphi_0^i(0,x)=u_\infty^1(t_{i-1},x)$$

(for $|\alpha|=0$);

$$\frac{\partial\varphi_k^i(s,x)}{\partial s}=\mathcal{L}^*\varphi_k^i(s,x)+k\tfrac{h(x)}{\sqrt{\Delta}}\varphi_{k-1}^i(s,x),\quad 0<s\le\Delta,$$
$$\varphi_k^i(0,x)=0,\quad k\ge1$$

(for $|\alpha|=k$) and $u_\infty^1(t_0,x)=p(x)$. An approximate solution to this system is given by

$$\varphi_k^i(t,x)=T_t\Big(\frac{th}{\sqrt{\Delta}}\Big)^k u_\infty^1(t_{i-1},\cdot)\ (x),\qquad k\ge0. \tag{4.3}$$

Indeed, for $k=0$, this is the exact solution (if T_t is exact); assuming (4.3) for some $k=n-1\ge0$, we get for $k=n$

$$\begin{aligned}\varphi_n^i(t,x)&=n\int_0^t T_{t-s}\frac{h}{\sqrt{\Delta}}\varphi_{n-1}^i(s,\cdot)(x)ds\\&=\frac{n}{\Delta^{n/2}}\int_0^t T_{t-s}h^nT_su_\infty^1(t_{i-1},\cdot)(x)s^{n-1}ds\\&\approx\frac{n}{\Delta^{n/2}}T_th^nu_\infty^1(t_{i-1},\cdot)(x)\int_0^t s^{n-1}ds\\&=T_t\Big(\frac{th}{\sqrt{\Delta}}\Big)^n u_\infty^1(t_{i-1},\cdot)\ (x),\end{aligned}$$

so (4.3) follows by induction. Note that, if $T_t(hf)(x)=hT_t(f)$ for all $f(x)$, it would be an exact solution.

Clearly, (4.3) implies that

$$\varphi_k^i(\Delta,x)=T_\Delta(h\sqrt{\Delta})^ku_\infty^1(t_{i-1},\cdot)\ (x),\qquad k\ge0.$$

It is also clear that

$$\xi_k^i = \frac{1}{\sqrt{k!}} H_k\Big(\frac{y(t_i) - y(t_{i-1})}{\sqrt{\Delta}}\Big), \tag{4.4}$$

and then by Theorem 2.4

$$\begin{aligned} u_\infty^1(t_i, x) &= T_\Delta u_\infty^1(t_{i-1}, \cdot)(x) + T_\Delta \sum_{k\geq 1} \frac{1}{k!} (h\sqrt{\Delta})^k H_k\Big(\frac{y(t_i) - y(t_{i-1})}{\sqrt{\Delta}}\Big) \\ &= T_\Delta \exp([y(t_i) - y(t_{i-1})]h - 0.5h^2\Delta) u_\infty^1(t_i, \cdot)(x). \end{aligned}$$

(The last equality follows from the well-known expansion

$$\exp(ax - 0.5x^2) = \sum_{k\geq 0} \frac{1}{k!} H_k(a) x^k$$

if we set $a = (y(t_i) - y(t_{i-1}))/\sqrt{\Delta},\ x = h\sqrt{\Delta}$.)

An alternative form of the splitting-up approximation, namely,

$$u_o(x) = p(x), \qquad u_i(x) = \exp((y(t_i) - y(t_{i-1}))h(x) - 0.5|h(x)|^2\Delta) T_\Delta u_{i-1}(\cdot)(x)\ , \tag{4.5}$$

can be obtained by Theorem 2.4 in the same way.

Next, we present an estimate on the number of on-line operations required by S^3 and compare it with a corresponding estimate for the splitting-up method.

We introduce the following parameters: N_s, the number of grid points in the spatial domain; N_J, the number of elements in J_N^n; κ, the number of basis functions e_l.

Assume that one needs to compute an approximation to the solution of (2.4) at moment $t = N_\tau \Delta$.

To do this using S^3, one has to find $\psi_l(i, N, n, \kappa)$, $i = 1, \dots, N_\tau$, for every $l = 1, \dots, \kappa$, which requires about $2\kappa^2 N_J N_\tau$ flops, and then compute the sum in (2.25)—κN_s more flops. The Wiener integral $\xi_{k,l} = \int_0^\Delta m_k(s) dy^l(s)$ reduces to a one-dimensional Riemann integral by integrating by parts. In addition, computations of the integrals $\xi_{k,l}$ for different k and l can be performed in parallel. As a result, computational complexity of the Wick polynomials ξ_α is negligible as compared to other procedures of S^3.

The total number of flops N_{S^3} is then $N_{S^3} = 2\kappa^2 N_J N_\tau + N_s \kappa$. Given the precision of the approximation, the number κ^2 will grow with d as C^d, where C is some constant depending on the type of the basis (but not on d), so $N_{S^3} \leq C^d(2N_J N_\tau + N_s)$.

If the splitting-up algorithm is used, one has to perform N_τ steps of the type (4.2). Each step requires solving a parabolic equation. To estimate the corresponding number of operations, assume that a finite element method is used and the resulting linear system is solved using an iterative procedure without preconditioning. The matrix of the system is of dimension $N_s \times N_s$, sparse and nonsymmetric (since operator $\mathcal{L}^*$ is not self-adjoint). Then one iteration requires about $C_d N_s$ flops, where C_d is a constant depending on d and on the particular numerical algorithm (see [1]), and the total number of iterations is proportional to the condition number of the matrix [36]. For nonsymmetric matrices, the condition number is proportional to at least $(\ln N_s)^{d-1}$ [1, 6]. Thus the total number of operations required to solve the equation on one step is $C_d N_s (\ln N_s)^{d-1}$. One also has to compute a certain number of exponential

TABLE 4.1
Comparison of the splitting-up approximation and the S^3.

	$t = 1$ (Step 100)		$t = 2$ (Step 200)	
	Splitting-up	S^3	Splitting-up	S^3
Flops	8161001	397431	16321601	788431
N_{50}	32	27	20	16
N_{75}	61	55	47	36
N_{95}	93	90	85	81

functions, but this can be done much faster and we disregard it. The total number of on-line operations is then $N_{sp-up} = N_\tau C_d N_s (\ln N_s)^{d-1}$.

As a result,

$$\frac{N_{\mathrm{S}^3}}{N_{sp-up}} \le \frac{C}{C_d}\Big(\frac{C}{\ln N_s}\Big)^{d-1}\Big[\frac{2N_J}{N_s} + \frac{1}{N_\tau}\Big].$$

Theorem 2.4 shows that for any d the splitting-up algorithm and S^3 have errors of the same order in Δ already for $N = 2,\ n = 1$, so we can take $1 + 2r + r(r-1)/2$ as the lower bound on N_J, where r is the dimension of the observation process. Since N_s usually grows with d, we can expect S^3 to have an advantage over the splitting-up algorithm in the following situations:

(1) when the estimation of u is required at one time moment after a long observation period ($N_\tau \gg 1$). This is characteristic for some tracking problems.

(2) when the dimension d of the state process is large.

To conclude this section we compare (numerically) the on-line performance of S^3 and the splitting-up method for one simple example.

For the test model, both signal and observation processes were chosen one-dimensional with the signal

$$dx(t) = 0.1\cos(2x(t))dt + 0.14 d\tilde{w}(t), \quad x(0) \sim \mathcal{N}(0, 0.1),$$

and the observations

$$y(t) = \int_0^t \arctan(x(s))ds + 0.04 w(t);$$

obvious modifications were made to reduce the last equation to the standard form (2.1). We took $T = 2$ and $\Delta = 0.01$.

The interval $[-1, 1]$ was taken as the spatial domain; it was discretized uniformly with step 0.01. Functions $\sin(\pi l(x-1)/2),\ 1 \le l \le 15$, sampled at the points of the spatial grid served as the basis $\{e_l\}$.

For the S^3, multiindices α with $|\alpha| \le 8,\ d(\alpha) \le 1$ were used. (This corresponds to the set J_8^1 in Theorem 2.4.)

Given the trajectory of the signal process, 100 independent observation trajectories were simulated; for each trajectory, the filtering density was computed at moments $25\Delta,\ 50\Delta, \dots, 200\Delta$, using both the S^3 and the splitting-up method.

The results are presented in Table 4.1. They are borrowed from [13]. In the table, "flops" stands for the total number of the *on-line* floating point operations (additions and multiplications) that it took to compute the filtering density at the given time moment; N_{50} (resp., N_{75}, N_{95}) is the number of times the value of the signal process was in the 50% (resp., 75%, 95%) confidence interval defined by the *computed* density.

We see that S^3 results in substantial reduction (up to 20 times) in the number of on-line computations without significant loss of accuracy. The decrease in the number of on-line computations should be even more conspicuous as the dimension of the observation process grows.

Appendix. To make the exposition as self-contained as possible, we will prove Theorem 2.1 and give some other results used in the proof of Theorem 2.2. Most of the results come from [30].

The summation over repeated indices convention is still in force. We also use the notations introduced at the beginning of section 3.

To begin with we recall the celebrated Cameron–Martin development (see, e.g., [7] and also [16, 19]).

THEOREM A.1 (Cameron–Martin development). *Let $B_s = (B_s^1, \dots, B_s^r)$, $0 \le s \le T$, be an r-dimensional Brownian motion and η be a measurable functional of the path $\{B_s, s \le T\}$ such that $\mathbf{E}\eta^2 < \infty$. Let $\{c_i(t)\}_{i\ge 1}$ be an arbitrary complete orthonormal system in $L_2([0,T])$. For $\alpha = \{\alpha_k^l\} \in J$ set*

$$\xi_\alpha(B) = \prod_{k,l} \frac{H_{\alpha_k^l}\Big(\int_0^T c_k(s)dB^l(s)\Big)}{\sqrt{\alpha_k^l!}}.$$

Then $(\xi_\alpha)_{\alpha\in J}$ is a CONS in $L_2(\Omega, \mathcal{F}_T^B, \mathrm{P})$, where $\mathcal{F}_T^B = \sigma(B_s,\ s \le T)$, and

$$\eta = \sum_{\alpha\in J} \mathbf{E}[\eta\xi_\alpha(B)]\xi_\alpha(B), \tag{A.1}$$

$$\mathbf{E}\eta^2 = \sum_{\alpha\in J} (\mathbf{E}[\eta\xi_\alpha(B)])^2. \tag{A.2}$$

The series (A.1) *converges in $L_2(\Omega, \mathrm{P})$.*

Let $\{z_k^l\},\ l = 1, \dots, r,\ k = 1, 2, \dots$, be a sequence of real numbers such that $\sum_{k,l} |z_k^l|^2 < \infty$. Set $m_z^l = m_k(s)z_k^l$, where $\{m_k\}$ is a CONS in $L_2([0,t])$. We also define

$$P_s(z) = \exp\Big\{ \int_o^s m_z^l(\tau)dy^l(\tau) - 0.5 \int_0^s \sum_{l=1}^r |m_z^l(\tau)|^2 d\tau \Big\}$$

and denote

$$\frac{\partial^\alpha}{\partial z^\alpha} := \prod_{k,l} \frac{\partial^{\alpha_k^l}}{(\partial z_k^l)^{\alpha_k^l}}.$$

Proof of Theorem 2.1. It is known (see, e.g., [34]) that for every t, x the UFD $u(t,x)$ is a measurable functional of the observation process $y(s),\ s \le t$. By Girsanov's theorem $y(s)$ is a Brownian motion on the new probability space $(\Omega, \mathcal{F}, \tilde{\mathrm{P}})$ (recall that $d\tilde{\mathrm{P}} = \rho(T)d\mathrm{P}$). Then by Theorem A.1 we have

$$u(t,x) = \sum_{\alpha\in J} \tilde{\mathbf{E}}[u(t,x)\xi_\alpha(y)]\xi_\alpha(y), \tag{A.3}$$

$$\tilde{\mathbf{E}}|u(t,x)|^2 = \sum_{\alpha\in J} (\tilde{\mathbf{E}}[u(t,x)\xi_\alpha(y)])^2, \tag{A.4}$$

where $\tilde{\mathbf{E}}$ stands for the expectation symbol with respect to measure $\tilde{\mathrm{P}}$, and the right-hand side of (A.3) converges in $L_2(\Omega, \tilde{\mathrm{P}})$.

Let us denote

$$\varphi_\alpha(s,x) := \sqrt{\alpha!}\tilde{\mathbf{E}}[u(s,x)\xi_\alpha(y)].$$

It is a standard fact (see, e.g., [16]) that

$$\xi_\alpha(y) = \frac{1}{\sqrt{\alpha!}}\frac{\partial^\alpha}{\partial z^\alpha}P_t(z)|_{z=0},$$

and so for every $s \le t$

$$\varphi_\alpha(s,x) = \frac{\partial^\alpha}{\partial z^\alpha}\tilde{\mathbf{E}}[u(s,x)P_t(z)]|_{z=0} = \frac{\partial^\alpha}{\partial z^\alpha}\tilde{\mathbf{E}}[u(s,x)P_s(z)]|_{z=0},$$

where the second equality follows from the martingale property of $P_s(z)$ on $(\Omega, \tilde{\mathrm{P}})$. Now to prove (2.8) and (2.9) it remains to show that the system of functions $\{\varphi_\alpha\}$, $\alpha \in J$, is a solution to the S-system (2.6). For this purpose it will be convenient to treat the UFD $u(t,x)$ as the solution of the Zakai equation (2.4). Since $P_s(z)$ satisfies the Ito stochastic differential equation

$$dP_s(z) = m_z^l(s)P_s(z)dy^l(s), \quad s \le t;\ P_0(z) = 1, \tag{A.5}$$

by the Ito chain rule

$$\begin{aligned} u(t,x)P_t(z) = p(x) &+ \int_0^t (\mathcal{L}^* u(s,x)P_s(z) + h^l(x)m_z^l(z)u(s,x)P_s(z))ds \\ &+ \int_0^t (h^l(x)u(s,x)P_s(z) + u(s,x)m_z^l(s)P_s(z))dy^l(s). \end{aligned}$$

Taking expectation $\tilde{\mathbf{E}}$ on both sides of the last equality and setting $\varphi(s,x,z) := \tilde{\mathbf{E}}u(s.x)P_s(z)$ we obtain

$$\begin{aligned} &\frac{\partial\varphi(s,x,z)}{\partial s} = \mathcal{L}^*\varphi(s,x,z) + m_z^l(s)h^l(x)\varphi(s,x,z), \quad 0 < s \le t, \\ &\varphi(0,x,z) = p(x)1_{\{|\alpha|=0\}}. \end{aligned} \tag{A.6}$$

Applying the operator $\frac{1}{\sqrt{\alpha!}}\frac{\partial^\alpha}{\partial z^\alpha}$ on both sides of (A.6) and setting $z = 0$ we get (2.6).

To complete the proof of Theorem 2.1 one needs to prove that the right-hand side of (2.8) converges also in $L_1(\Omega, \mathrm{P})$. This follows in a simple way from the convergence in $L_2(\Omega, \tilde{\mathrm{P}})$ and Cauchy–Schwartz inequality [30]. □

In what follows we give some additional properties of the solution of (2.6) in the case $r = 1$; these properties are used in the proof of Theorem 2.2. Generalizations to the general case $r > 1$ are straightforward.

PROPOSITION A.1 (see [30]). *Let $\{\varphi_\alpha(t,x)\}_{\alpha\in J}$ be a solution of* (2.6). *Then for each α with $|\alpha| = k$*

$$\begin{aligned} &\varphi_\alpha(t,x) = \sum_{\sigma\in\mathcal{P}^k}\int^{(k)} F(t;s^k;x)m_{i_{\sigma(k)}}(s_k)\dots m_{i_{\sigma(1)}}(s_1)ds^k, \qquad k > 1, \\ &\varphi_\alpha(t,x) = \int_0^t T_{t-s_1}hT_{s_1}p(x)m_i(s_1)ds_1, \qquad k = 1, \\ &\varphi_\alpha(t,x) = T_t p(x), \qquad k = 0, \end{aligned} \tag{A.7}$$

where $(i_1, \dots, i_k)$ *is the characteristic set of* α *(see the beginning of section* 3 *for notation).*

In addition,

$$\sum_{|\alpha|=k} \frac{\varphi_\alpha^2(t,x)}{\alpha!} = \int^{(k)} |F(t;s^k;x)|^2 ds^k. \tag{A.8}$$

Proof. Representation (A.7) is obviously true for $|\alpha| = 0$. Then the general case $|\alpha| \geq 1$ follows by induction from the variation of parameters formula.

To prove (A.8), first of all note that

$$\sum_{\sigma\in\mathcal{P}^k} m_{i_{\sigma(k)}}(s_k) \dots m_{i_{\sigma(1)}}(s_1) = \sum_{\sigma\in\mathcal{P}^k} m_{i_k}(s_{\sigma(k)}) \dots m_{i_1}(s_{\sigma(1)}).$$

Indeed, any term on the left corresponding to a given $\sigma_0 \in \mathcal{P}^k$ is equal to the term on the right corresponding to $\sigma_0^{-1} \in \mathcal{P}^k$.

Then we can write (A.7) as

$$\varphi_\alpha(t,x) = \int^{(k)} F(t;s^k;x) E_\alpha(s^k) ds^k.$$

Introducing

$$G(s^k;x) := \sum_{\sigma\in\mathcal{P}^k} T_{t-s_{\sigma(k)}} h \dots T_{s_{\sigma(2)}-s_{\sigma(1)}} h T_{s_{\sigma(1)}} p(x) 1_{s_{\sigma(1)}<\dots<s_{\sigma(k)}},$$

we can rewrite it further as

$$\varphi_\alpha(t,x) = \frac{1}{k!} \int\limits_{[0,t]^k} G(s^k) E_\alpha(s^k) ds^k. \tag{A.9}$$

Since for each x G is a symmetric function from $L_2([0,t]^k)$ and $\{E_\alpha/\sqrt{\alpha!k!}, |\alpha| = k\}$ form a CONS for the symmetric part of the space, we have

$$G = \sum_{|\alpha|=k} \frac{c_\alpha E_\alpha}{\sqrt{\alpha!k!}}$$

with some $c_\alpha \in \mathbf{R}$. Then from (A.9) $\varphi_\alpha^2/\alpha! = c_\alpha^2/k!$ and so

$$\sum_{|\alpha|=k} \frac{\varphi_\alpha^2(t,x)}{\alpha!} = \frac{1}{k!} \sum_{|\alpha|=k} c_\alpha^2 = \frac{1}{k!} \int\limits_{[0,t]^k} |G(s^k;x)|^2 ds^k$$

$$= \frac{1}{k!} \int\limits_{[0,t]^k} \Big| \sum_{\sigma\in\mathcal{P}^k} T_{t-s_{\sigma(k)}} h \dots T_{s_{\sigma(2)}-s_{\sigma(1)}} h T_{s_{\sigma(1)}} p(x) 1_{s_{\sigma(1)}<\dots<s_{\sigma(k)}} \Big|^2 ds^k$$

$$= \int^{(k)} |F(t;s^k;x)|^2 ds^k,$$

which proves (A.6). □

REMARK A.1. *In this article we needed WCE* (2.8) *only at the final point of the time interval. However, it is readily checked that, due to* $\mathcal{F}_t^y$*-measurability of UFD*

$u(s,x)$ for all $s \le t$, the statement and the proof of Theorem 2.1 *remain virtually unchanged if, in* (2.8) *and* (2.9)*, we replace t by any $s \le t$. This implies in particular that equality* (2.16) *holds not only for grid points t_i but for every $s \in [t_{i-1}, t_i]$.*

REMARK A.2. *If $r = 1$ and $|\alpha| = k$, then, by* [17, *Theorem* 3.1],

$$\xi_\alpha = \frac{1}{\sqrt{\alpha!}} \int_0^t \int_0^{s_k} \dots \int_0^{s_2} E_\alpha(s^k) dy(s_1) \dots dy(s_k).$$

This gives an alternative (but equivalent) form of WCE (2.8) *in terms of multiple Wiener integrals. A similar expansion holds for an arbitrary r.*

Acknowledgments. Very helpful comments and suggestions from Drs. K. Ito and J. Dym and from the reviewers are acknowledged.

REFERENCES

[1] O. AXELSSON AND V. A. BARKER, *Finite Element Solution of Boundary Value Problems*, Academic Press, New York, 1984.
[2] J. BARAS, *Real-time architectures for the Zakai equations and applications*, in Stochastic Analysis, E. Mayer-Wolf et al., eds., Academic Press, Boston, 1991.
[3] V. E. BENESH, *Exact finite-dimensional filters for certain diffusions with nonlinear drift*, Stochastics, 5 (1981), pp. 65–92.
[4] A. BENSOUSSAN, R. GLOWINSKI, AND R. RASCANU, *Approximations of the Zakai equation by splitting up method*, SIAM J. Control Optim., 28 (1990), pp. 1420–1431.
[5] A. BUDHIRAJA AND G. KALLIANPUR, *Approximations to the Solution of the Zakai Equations Using Multiple Wiener and Stratonovich Integral Expansions*, Tech. report 447, Center for Stochastic Processes, University of North Carolina, Chapel Hill, NC, Jan. 1995.
[6] Z. CAI AND WEINAN E, *Hierarchical method for elliptic problems using wavelet*, Comm. Appl. Numer. Methods, 8 (1992), pp. 819–825.
[7] R. H. CAMERON AND W. T. MARTIN, *The orthogonal development of non-linear functionals in a series of Fourier-Hermite functions*, Ann. Math., 48 (1947), pp. 385–392.
[8] J. M. C. CLARK, *The design of robust approximations to the stochastic differential equations of nonlinear filtering*, in Communication Systems and Random Processes Theory, NATO Adv. Sci. Inst. Ser. E Appl. Sci. 25, J. K. Skwirzynski, ed., Springer-Verlag, Berlin, 1977, pp. 721–734.
[9] G. B. DIMASI AND W. J. RUNGGALDIER, *On measure transformations for combined filtering and parameter estimation in discrete time*, Systems Control Lett., 2 (1982), pp. 57–62.
[10] R. J. ELLIOTT AND R. GLOWINSKI, *Approximations to solutions of the Zakai filtering equation*, Stochastic Anal. Appl., 7 (1989), pp. 145–168.
[11] W. H. FLEMING AND E. PARDOUX, *Optimal control for partially observed diffusions*, SIAM J. Control Optim., 20 (1982), pp. 261–285.
[12] P. FLORCHINGER AND F. LEGLAND, *Time discretization of the Zakai equation for diffusion processes observed in correlated noise*, Stochastics Stochastics Rep., 35 (1991), pp. 233–256.
[13] C. P. FUNG, *Solving Hidden Markov Problems by Spectral Methods*, Ph.D. thesis, University of Southern California, Los Angeles, CA, Nov. 1995.
[14] J. GLIMM AND A. JAFFE, *Quantum Physics*, Springer, New York, 1987.
[15] D. GOTTLIEB AND S. A. ORSZAG, *Numerical Analysis of Spectral Methods: Theory and Applications*, CBMS-NSF Regional Conf., Ser. in Appl. Math., 26, SIAM, Philadelphia, 1977.
[16] T. HIDA, H.-H. KUO, J. POTTHOFF, AND L. SREIT, *White Noise*, Kluwer, Boston, 1993.
[17] K. ITO, *Multiple Wiener integral*, J. Math. Soc. Japan, 3 (1951), pp. 157–169.
[18] K. ITO, *Approximation of the Zakai equation for nonlinear filtering*, SIAM J. Control Optim., 34 (1996), pp. 620–634.
[19] G. KALLIANPUR, *Stochastic Filtering Theory*, Springer-Verlag, New York, 1980.
[20] N. V. KRYLOV AND B. L. ROZOVSKII, *On conditional distributions of diffusion processes*, Math. USSR-Izv., 42 (1978), pp. 336–356.
[21] N. V. KRYLOV AND A. VERETENNIKOV, *On explicit formulae for solutions of stochastic equations*, Math. USSR-Sb., 29 (1976), pp. 239–256.
[22] H. KUNITA, *Cauchy problem for stochastic partial differential equations arising in non-linear filtering theory*, Systems Control Lett., 1 (1981), pp. 37–41.

[23] H. Kunita, *Stochastic Flows and Stochastic Differential Equations*, Cambridge University Press, Cambridge, UK, 1982.
[24] H. J. Kushner, *Probability Methods for Approximations in Stochastic Control and for Elliptic Equations*, Academic Press, New York, 1977.
[25] O. A. Ladyzhenskaia, V. A. Solonikov, and N. N. Ural'tseva, *Linear and Quasi-linear Equations of Parabolic Type*, American Mathematical Society, Providence, RI, 1968.
[26] F. LeGland, *Time discretizations of nonlinear filtering equations*, in Proc. 28th IEEE Conf. on Decision and Control, Tampa, 1989, IEEE Control Systems Society, Piscataway, NJ, 1989, pp. 2601–2606.
[27] R. S. Liptser and A. N. Shiryayev, *Statistics of Random Processes*, Springer-Verlag, New York, 1992.
[28] J. T.-H. Lo and S.-K. Ng, *Optimal orthogonal expansion for estimation* I*: Signal in white Gaussian noise*, in Nonlinear Stochastic Problems, R. Bucy and J. Moura, eds., D. Reidel, Dordrecht, 1983, pp. 291–309.
[29] R. Mikulevicius and B. L. Rozovskii, *Linear parabolic stochastic PDEs and Wiener chaos*, unpublished.
[30] R. Mikulevicius and B. L. Rozovskii, *Separation of observations and parameters in nonlinear filtering*, in Proc. 32nd IEEE Conf. on Decision and Control, Part 2, San Antonio, 1993, IEEE Control Systems Society, Piscataway, NJ, 1993, pp. 1564–1569.
[31] R. Mikulevicius and B. L. Rozovskii, *Soft solutions of linear parabolic SPDE's and the Wiener chaos expansion*, in Stochastic Analysis on Infinite Dimensional Spaces, Pitman Res. Notes Math. Sec. 310, H. Kunita and H.-H. Kuo, eds., Longman, Harlow, UK, 1994.
[32] D. Ocone, *Multiple integral expansions for nonlinear filtering*, Stochastics, 10 (1983), pp. 1–30.
[33] E. Pardoux, *Filtrage non lineaire et equations aux derivees partielles stochastiques associees*, in Ecole d'été de Probabilités de Saint-Flour, Springer-Verlag, New York, 1989.
[34] B. L. Rozovskii, *Stochastic Evolution Systems*, Kluwer, Amsterdam, 1990.
[35] E. Wong, *Explicit solutions to a class of nonlinear filtering problems*, Stochastics, 5 (1981), pp. 311–321.
[36] D. M. Young, *Iterative Solution of Large Linear Systems*, Academic Press, New York, 1971.
[37] M. Zakai, *On the optimal filtering of diffusion processes*, Z. Wahrscheinlichkeitstheorie und verw. Gebiete, 4 (1969), pp. 230–233.
[38] A. K. Zvonkin and N. V. Krylov, *Strong solutions of stochastic differential equations*, in Proc. of the School Seminar on the Theory of Random Processes, Druskinikai, November 25–30, 1974, Vilnius, 1975 (in Russian).

SIAM J. CONTROL OPTIM.
Vol. 35, No. 2, pp. 462–479, March 1997

005

DESCRIPTOR SYSTEMS WITHOUT CONTROLLABILITY AT INFINITY*

RALPH BYERS†, TON GEERTS‡, AND VOLKER MEHRMANN§

Abstract. This paper concerns the structure that can be achieved by feedback in descriptor systems that lack controllability at infinity. Staircase and double staircase condensed forms obtained through a sequence of orthogonal state transformations display when and how feedback can be used to achieve minimal index. Furthermore, they reveal that the modes that are uncontrollable at infinity have a fixed minimal index that cannot be reduced by feedback. However, this fixed higher index part of the control system is constrained to be zero in an appropriate coordinate system, provided the initial conditions are consistent. The remainder is a reduced order system that is controllable at infinity that can be made to have index one by feedback.

Key words. descriptor system, controllability, numerical methods, impulse

AMS subject classifications. 93B11, 93B52, 93B05, 93B27, 93B40

PII. S0363012994269818

1. Introduction.

Consider the linear, time-invariant descriptor system

$$\begin{aligned} E\dot{x} &= Ax + Bu, \qquad Ex(0) = Ex^0, \\ y &= Cx, \end{aligned} \tag{1}$$

with system matrices $E \in \mathbf{C}^{n\times n}$, $A \in \mathbf{C}^{n\times n}$, $B \in \mathbf{C}^{n\times m}$, $C \in \mathbf{C}^{p\times n}$, state $x = x(t) \in \mathbf{C}^n$, input $u = u(t) \in \mathbf{C}^m$, and output $y = y(t) \in \mathbf{C}^p$. Descriptor systems arise naturally in circuit design, mechanical multibody systems, and a variety of other applications [25, 32, 33]. They have recently attracted the attention of many authors to all aspects of control, including pole placement, filtering, stabilization, controllability, observability, optimal control problems, invertibility, duality, realization, etc. See, for example, [6, 14, 13, 27] and the references therein.

In contrast to standard systems in which $E = I$, continuous inputs to a descriptor system can give rise to discontinuities or impulsive modes in the state trajectories. A detailed analysis of solvability aspects of the distributional version of (1) is given in [14]. A simple example is the single input system determined by

$$E = \begin{bmatrix} 0 & 1 \\ 0 & 0 \end{bmatrix}, \quad A = \begin{bmatrix} 1 & 0 \\ 0 & 1 \end{bmatrix}, \quad B = \begin{bmatrix} 0 \\ -1 \end{bmatrix}, \quad C = \begin{bmatrix} 1 & 0 \\ 0 & 1 \end{bmatrix}. \tag{2}$$

*Received by the editors June 20, 1994; accepted for publication (in revised form) January 10, 1996.

http://www.siam.org/journals/sicon/35-2/26981.html

†Department of Mathematics, University of Kansas, Lawrence, KS 66045 (byers@ariel.math.ukans.edu). The research of this author was supported in part by National Science Foundation grant INT-8922444, NSF award CCR-9404425, and University of Kansas General Research Allocation 3514-20-0038.

‡Marensedijk 18, NL-5398 KM Maren-Kessel, the Netherlands. The research of this author was supported by the Dutch Organization for Scientific Research (N.W.O.).

§Fakultät für Mathematik, Technische Universität Chemnitz-Zwickau, D-09107 Chemnitz, Germany (mehrmann@mathematik.tu-chemnitz.de). The research of this author was supported in part by Sonderforschungsbereich Diskrete Strukturen in der Mathematik and Forschungsschwerpunkt Mathematisierung, Universität Bielefeld, and Deutsche Forschungsgemeinschaft Projekt Me 790/5-1.

The input $u(t)$ induces the state $x(t)$:

$$x(t) = \begin{bmatrix} x_1(t) \\ x_2(t) \end{bmatrix} = \begin{bmatrix} \dot{u}(t) \\ u(t) \end{bmatrix}.$$

If $x^0 = x(0^-) = \left[\begin{smallmatrix} x_1^0 \\ x_2^0 \end{smallmatrix}\right]$ and u is smooth but $x_2^0 \neq u(0^+)$, then x_1 will still exhibit a pulse $[u(0^+) - x_2^0]\delta$, where δ is the Dirac delta distribution [14]. (If $u(0^+) = x_2^0$ but $\dot{u}(0^+) \neq x_1^0$, i.e., $x_1(0^+) \neq x_1(0^-)$, then x_1 will still be impulse free [14, 40].)

If (1) is controllable and observable at infinity, i.e., $\operatorname{rank}[E,\, AS_\infty,\, B] = n$, where

$$\operatorname{Im}(S_\infty) = \ker(E) \quad \text{and} \quad \operatorname{rank} \begin{bmatrix} E \\ T_\infty^H A \\ C \end{bmatrix} = n,$$

where $\operatorname{Im}(T_\infty) = \ker(E^H)$, then the problem of impulses can be avoided (or at least disguised) by using an appropriate feedback [4, 3]. Here, $\operatorname{Im}(\cdot)$ denotes the image (or range) and $\ker(\cdot)$ is the kernel (or null space).

For example, using the feedback control $u = FCx + v$, where $F = \begin{bmatrix} -1 & 0 \end{bmatrix}$ in (2) gives the closed loop system matrices

$$E = \begin{bmatrix} 0 & 1 \\ 0 & 0 \end{bmatrix}, \quad A = \begin{bmatrix} 1 & 0 \\ 1 & 1 \end{bmatrix}, \quad B = \begin{bmatrix} 0 \\ -1 \end{bmatrix}, \quad C = \begin{bmatrix} 1 & 0 \\ 0 & 1 \end{bmatrix}. \tag{3}$$

Here the inputs $v(t)$ and the resulting state trajectories exhibit the same impulsive behavior, and thus x will be impulse free if v is. In addition if v is q-times continuously differentiable for $t > 0$, then x is as well.

Notice that the closed loop system matrices (3) have a stronger robustness property than the ones in (2). If the closed loop system is perturbed by some unmodeled dynamic forcing function $f(t)$ giving

$$E\dot{x} = (A + BFC)x + Bv + f(t),$$

then the resulting state $x(t)$ still has as many derivatives as $\left[\begin{smallmatrix} f(t) \\ v(t) \end{smallmatrix}\right]$, even if $x(0^-) = x^0$ is not consistent. Using distributions, we get

$$\begin{aligned} x_1 &= -x_2 + v - f_2, \\ x_2 &= (\delta^{(1)} + \delta)^{*^{-1}}[x_2(0^-)\delta + v + f_1 - f_2]. \end{aligned}$$

In time domain, if v and f are functions, we have

$$\begin{aligned} x_1(t) &= -x_2(t) + v(t) - f_2(t), \\ x_2(t) &= e^{-t}x_2(0^-) + \int_0^t e^{-(t-s)}[v(s) + f_1(s) - f_2(s)]ds. \end{aligned}$$

Hence, $x = \left[\begin{smallmatrix} x_1(t) \\ x_2(t) \end{smallmatrix}\right]$ has indeed as many derivatives for $t > 0$ as $\left[\begin{smallmatrix} f \\ v \end{smallmatrix}\right]$, even if $x(0^-) \neq x(0^+)$.

We call this property *index-one robustness* because it is shared by regular descriptor systems of index at most one. (Regularity and index are defined in the next

section.) Even smooth perturbations f in (2) will in general give rise to extra pulses in the solution, whereas this cannot happen in systems that are index-one robust. It is implicit in the results of [4, 3] that systems that are controllable and observable at infinity can be made to be index-one robust by feedback.

In several applications including mechanical multibody systems [19, 29, 28, 33, 34] the assumptions of controllability and/or observability at infinity do not hold. Consider for example the planar model of a three-link manipulator introduced in [19],

$$E = \begin{bmatrix} I & 0 & 0 \\ 0 & M_0 & 0 \\ 0 & 0 & 0 \end{bmatrix}, \; A = \begin{bmatrix} 0 & I & 0 \\ -K_0 & -D_0 & F_0^T \\ F_0 & 0 & 0 \end{bmatrix}, \; B = \begin{bmatrix} 0 \\ S_0 \\ 0 \end{bmatrix}.$$

This system is not controllable at infinity. With output $y = \begin{bmatrix} C_1 & C_2 & 0 \end{bmatrix}$ it is not observable at infinity either. Due to the special structure of this mechanical multibody system, however, the part of the system that is characterized by the uncontrollable modes at infinity can be neglected. The remaining system can be made index-one robust [28]. If this simplification is not carried out, then undesired phenomena as in the following example may occur. Let

$$E = \begin{bmatrix} 0 & 1 & 0 & 0 \\ 0 & 0 & 0 & 0 \\ 0 & 0 & 0 & 1 \\ 0 & 0 & 0 & 0 \end{bmatrix}, \quad A = I, \quad B = \begin{bmatrix} 0 \\ -1 \\ 0 \\ 0 \end{bmatrix}, \quad C = I. \tag{4}$$

In this case, the first two components of the state obey (2), while the second two components of the state are confined to be zero (assuming consistent initial conditions). Choosing an appropriate feedback causes the first two components of the state to obey (3) while the second two remain zero. However, regardless of what feedback is chosen, the closed loop system does not have index-one robustness. If the forcing function $f(t) = \begin{bmatrix} f_1(t), & f_2(t), & f_3(t), & f_4(t) \end{bmatrix}^T$ were added to the closed loop system, then the third state component is $x_3(t) = -f_3(t) - \dot{f}_4$. In the third component of the state, we will obtain an impulse of the form $(-f_4(0^+) - x_{04})\delta$, where $x_{04} = x_4(0^-)$. Lack of differentiability in $f_4(t)$ may translate into lack of continuity in x_3. A jump discontinuity in $f(t)$ may cause an impulse.

This paper concerns the properties that can be achieved without controllability and observability at infinity, including when and how feedback can be used to achieve minimal index by numerically stable methods. All these properties are displayed by the Kronecker-like feedback canonical form introduced in [24]. Extracting this canonical form may, however, require ill-conditioned transformations which are sensitive to rounding errors. For this reason, following the approaches of [4, 3, 37, 36], we derive condensed staircase and double staircase forms through a sequence of unitary state space transformations. They display when and how feedback can be used to achieve minimal index. They also reveal that (4) is typical of systems which lack controllability at infinity. The parts of the state which are uncontrollable at infinity are constrained to be zero in an appropriate coordinate system, provided the initial conditions are consistent and may be decoupled from the rest of the system. This leaves a reduced order system that is controllable at infinity to which the work of [4, 3] applies. A similar argument applies to parts of the state which are not observable at infinity. By choosing an appropriate basis, these parts can be decoupled from the rest, and since they cannot be observed, they can be removed without changing the dynamics of the system.

2. Definitions and lemmas. The control system (1) and the associated matrix pencil $\lambda E - A$ are said to be *regular* if the characteristic polynomial $\det(\lambda E - A)$ is not identically zero. If the pencil $\lambda E - A$ is not regular, then the system of differential algebraic equations

$$E\dot{x} = Ax + f(t) \tag{5}$$

is underdetermined in the sense that consistent initial conditions do not uniquely determine solutions [12]. If the pencil $\lambda E - A$ is regular, then the roots of the characteristic polynomial are the finite eigenvalues of the pencil and include the poles of the transfer function of (1). In addition, if E is singular, the pencil is said to have infinite eigenvalues which may be identified as the zero eigenvalues of the inverse pencil $E - \lambda A$.

The eigenstructure of regular pencils is displayed by the Weierstraß canonical form (WCF).

THEOREM 2.1 (Weierstraß canonical form [12]). *If $\lambda E - A$ is regular, then there exist nonsingular matrices $X = [X_r, X_\infty] \in \mathbf{C}^{n\times n}$ and $Y = [Y_r, Y_\infty] \in \mathbf{C}^{n\times n}$ for which*

$$Y^H E X = \begin{bmatrix} Y_r^H \\ Y_\infty^H \end{bmatrix} E \begin{bmatrix} X_r & X_\infty \end{bmatrix} = \begin{bmatrix} I & 0 \\ 0 & N \end{bmatrix} \tag{6}$$

and

$$Y^H A X = \begin{bmatrix} Y_r^H \\ Y_\infty^H \end{bmatrix} A \begin{bmatrix} X_r & X_\infty \end{bmatrix} = \begin{bmatrix} J & 0 \\ 0 & I \end{bmatrix}, \tag{7}$$

where J is a matrix in Jordan form whose diagonal elements are the finite eigenvalues and N is a nilpotent matrix also in Jordan form. J and N are unique up to permutation of Jordan blocks. □

The *index* of the pencil $\lambda E - A$ and of the descriptor system (1) is the index of nilpotency of the nilpotent block N in the WCF; i.e., the index of the pencil is μ if and only if $N^{\mu-1} \neq 0$ and $N^\mu = 0$. By convention, if E is nonsingular, then the pencil is said to have index zero. We denote the index of the pencil $\lambda E - A$ by $\operatorname{index}(\lambda E - A)$. If E is a nilpotent matrix and A nonsingular, then we write $\operatorname{index}(E)$ instead of $\operatorname{index}(\lambda E - A)$.

Most of the information displayed by the WCF is also easily obtained from triangular pencils or block triangular pencils. It often simplifies derivations to use triangular or block triangular pencils. Furthermore, numerical algorithms that transform pencils to triangular form are usually more reliable than those that reduce to the WCF (6)–(7) [7, 8, 20].

LEMMA 2.2. *The eigenvalues of the block triangular pencil*

$$\lambda \begin{bmatrix} E_{11} & E_{12} \\ 0 & E_{22} \end{bmatrix} - \begin{bmatrix} A_{11} & A_{12} \\ 0 & A_{22} \end{bmatrix} \tag{8}$$

are the union of the eigenvalues of the diagonal blocks

$$\lambda E_{11} - A_{11}, \tag{9}$$

$$\lambda E_{22} - A_{22}. \tag{10}$$

In particular, (8) *is regular if and only if* (9) *and* (10) *are regular.*

Moreover, if (9) *and* (10) *have disjoint eigenvalues, then the Jordan and nilpotent parts of the WCF of* (8) *are the union of the Jordan and nilpotent parts of the WCFs of* (9) *and* (10). □

The next lemma gives a useful characterization of regular, index-one pencils.

LEMMA 2.3 (see [21]). *The pencil $\lambda E - A$ is regular and has index at most one if and only if*

$$\operatorname{rank}\left(\begin{bmatrix} E \\ T_\infty^H A \end{bmatrix}\right) = \operatorname{rank}\left(E + T_\infty T_\infty^H A\right) = n,$$

where the columns of T_∞ span the null space of E^H. Equivalently, the pencil $\lambda E - A$ is regular and of index less than or equal to one if and only if

$$\operatorname{rank}\left([E,\ AS_\infty]\right) = \operatorname{rank}\left(E + AS_\infty S_\infty^H\right) = n,$$

where the columns of S_∞ span the null space of E. □

Similar statements for arbitrary linear systems are given in [14].

If $\lambda E - A$ is regular, then in terms of the WCF (6)–(7), the solutions of (5) take the form

$$x(t) = X_r z_1(t) + X_\infty z_2(t),$$

where

$$z_1(t) = e^{tJ} z_1(0) + \int_0^t e^{(t-s)J} Y_r^H f(s)\, ds,$$

$$z_2(t) = -\sum_{i=0}^{\mu-1} \frac{d^i}{dt^i}(N^i Y_\infty^H f(t)). \tag{11}$$

From this we see that in order to have a smooth solution $x(t)$, the initial condition $x(0^-)$ must be a member of the set of *admissible* initial conditions

$$\left\{ X_r z_1 + X_\infty z_2 \;\middle|\; z_1 \in \mathbf{C}^r,\ z_2 = -\sum_{i=0}^{\mu-1} (N^i Y_\infty^H f^{(i)}(0)) \right\}.$$

It may be worthwhile to use feedback to minimize the index of a control system even when it cannot be reduced to index one in order to minimize the effect of discontinuities in the derivatives of unmodeled or perturbing forcing functions. One of the goals of this study is to determine what is the minimal index that can be achieved and to determine a feedback that achieves it. It turns out that according to the linear model (1), the modes that cannot be made to be index one by feedback are constrained to be zero in an appropriate coordinate system, provided the initial conditions are consistent. The remaining active modes may be made to have index one.

We now introduce some further definitions and notation. A system of the form (1) is *regularizable by state feedback*, if there exists a feedback $F \in \mathbf{C}^{m\times n}$ such that the pencil $\lambda E - (A + BF)$ is regular [3, 31]. Similarly, it is *regularizable by output feedback* if there exists $G \in \mathbf{C}^{m\times p}$ such that the pencil $\lambda E - (A + BGC)$ is regular. A system (1) is *controllable at infinity or impulse controllable* if $\operatorname{rank}[E,\ AS_\infty,\ B] = n$, where $\operatorname{Im}(S_\infty) = \ker(E)$. It is called *observable at infinity or impulse observable* if

$$\operatorname{rank}\begin{bmatrix} E \\ T_\infty^H A \\ C \end{bmatrix} = n,$$

where $\mathrm{Im}(T_\infty) = \ker(E^H)$ [5, 40]. In geometric terms, controllability at infinity is equivalent to

$$\mathrm{Im}(E) + A\ker(E) + \mathrm{Im}(B) = \mathbf{C}^n.$$

(See [5, 14, 17, 40].)

A regular descriptor system with index of at most one is a fortiori controllable at infinity. Systems that are controllable at infinity admit a state feedback control which makes the closed loop system be regular and have index of at most one [4, 3]. Moreover, the system transformations may be chosen to minimize the effects of rounding error [3, 10].

Let $P \in \mathbf{C}^{n\times n}$, $Q \in \mathbf{C}^{n\times n}$, $R \in \mathbf{C}^{m\times m}$, and $S \in \mathbf{C}^{p\times p}$ be nonsingular. If

$$\begin{aligned} \tilde{E} = PEQ\tilde{A} = PAQ\tilde{B} = PBR, \\ \tilde{C} = SCQ, \end{aligned} \tag{12}$$

then the descriptor system

$$\begin{aligned} \tilde{E}\dot{\tilde{x}} &= \tilde{A}\tilde{x} + \tilde{B}\tilde{u}, \\ \tilde{y} &= \tilde{C}\tilde{x} \end{aligned}$$

is equivalent to (1) in the sense that

$$\begin{aligned} x &= Q\tilde{x}, \\ u &= R\tilde{u}, \\ y &= S\tilde{y}. \end{aligned}$$

The transformation (12) is a *generalized state transformation*. Such transformations establish an equivalence relation among descriptor systems. Controllability at infinity, observability at infinity, regularity, eigenvalues, and index are preserved by generalized state transformations. Canonical forms under these and other state transformations are discussed in [24, 31]. However, these canonical forms are not easily computed, because modeling errors, measurement errors, or rounding errors may sometimes change them completely. In the next section we use a sequence of state transformations via unitary matrices to bring $\tilde{E}$, $\tilde{A}$, $\tilde{B}$, and $\tilde{C}$ into a staircase-like form in the style of [4, 3, 37, 36]. Although our canonical forms display less information than those of [24, 31], they are less sensitive to data perturbations and rounding errors.

The proofs of the staircase-like form in this paper are constructive and form the basis of a numerically stable algorithm for computing the factorization. The basic operations are *row compressions* and *column compressions*. A row compression of a matrix $M \in \mathbf{C}^{h\times k}$ of rank r is the factorization

$$UM = \begin{matrix} \\ r \\ h-r \end{matrix} \begin{matrix} \begin{matrix} r & k-r \end{matrix} \\ \begin{bmatrix} M_1 & M_2 \\ 0 & 0 \end{bmatrix} \end{matrix},$$

where $U \in \mathbf{C}^{h\times h}$ is unitary and $\begin{bmatrix} M_1, & M_2 \end{bmatrix}$ has full row rank r. The unitary matrix U may be obtained from a QR factorization or the singular value decomposition (SVD) of M or a combination of both [18]. If necessary, U may be chosen so that M_1 is upper triangular. Excellent software for computing the SVD and QR factorizations is widely available [1, 9, 35]. A column compression is a row compression of M^H.

3. Reduction to condensed form. In this section we construct a unitary state transformation that reduces the system matrices of (1) to a staircase and double staircase form similar to those constructed in [2, 4, 3, 37, 36, 38, 39]. A feedback that minimizes the index can be constructed from this staircase form. In addition, the staircase form reveals which modes are uncontrollable at infinity and cannot be reduced to index one by feedback. According to the linear model (1) these modes are not excited and play no role in the system dynamics.

In what follows, it is convenient to allow partitioned matrices which in some special cases may have submatrices with no rows or no columns. In this case, of course, those submatrices are vacuous and simply do not appear. By convention, "0−by−0 matrices" are nonsingular.

LEMMA 3.1. *There exists a state transformation of* (1) *by unitary matrices* $P \in \mathbf{C}^{n\times n}$ *and* $Q \in \mathbf{C}^{n\times n}$ *such that*

$$PEQ = \begin{array}{c} \\ r \\ s \\ q \end{array}\!\!\begin{array}{c} \begin{array}{ccc} r & s & q \end{array} \\ \begin{bmatrix} E_{11} & 0 & E_{13} \\ 0 & 0 & 0 \\ 0 & 0 & 0 \end{bmatrix} \end{array}, \tag{13}$$

$$PAQ = \begin{array}{c} \\ r \\ s \\ q \end{array}\!\!\begin{array}{c} \begin{array}{ccc} r & s & q \end{array} \\ \begin{bmatrix} A_{11} & A_{12} & A_{13} \\ A_{21} & A_{22} & A_{23} \\ 0 & 0 & A_{33} \end{bmatrix} \end{array}, \tag{14}$$

$$PB = \begin{array}{c} \\ r \\ s \\ q \end{array}\!\!\begin{array}{c} \begin{array}{c} m \end{array} \\ \begin{bmatrix} B_1 \\ B_2 \\ 0 \end{bmatrix} \end{array}, \tag{15}$$

$$CQ = \begin{array}{c} \\ p \end{array}\!\!\begin{array}{c} \begin{array}{ccc} r & s & q \end{array} \\ \begin{bmatrix} C_1 & C_2 & C_3 \end{bmatrix} \end{array}, \tag{16}$$

where $r = \operatorname{rank}(E)$, $s = \operatorname{rank}(B_2)$, *and* $q = n - r - s$.

Proof. The proof is by construction. First choose a row compression of the augmented matrix $\begin{bmatrix} E, B, A \end{bmatrix}$,

$$P\begin{bmatrix} E, B, A \end{bmatrix} = \begin{array}{c} \\ r \\ s \\ q \end{array}\!\!\begin{array}{c} \begin{array}{ccccccc} r & s & q & m & r & s & q \end{array} \\ \begin{bmatrix} \tilde{E}_{11} & \tilde{E}_{12} & \tilde{E}_{13} & B_1 & \tilde{A}_{11} & \tilde{A}_{12} & \tilde{A}_{13} \\ 0 & 0 & 0 & B_2 & \tilde{A}_{21} & \tilde{A}_{22} & \tilde{A}_{23} \\ 0 & 0 & 0 & 0 & \tilde{A}_{31} & \tilde{A}_{32} & \tilde{A}_{33} \end{bmatrix} \end{array}, \tag{17}$$

where $r = \operatorname{rank}(E)$, $s = \operatorname{rank}(B_2)$, and $q = n - r - s$. If the column space of B is contained in the column space of E, then B_2 is vacuous and $s = 0$. Now, choose a column compression of the permuted submatrix

$$\begin{array}{c} \\ q \\ r \end{array}\!\!\begin{array}{c} \begin{array}{ccc} q & r & s \end{array} \\ \begin{bmatrix} \tilde{A}_{33} & \tilde{A}_{31} & \tilde{A}_{32} \\ \tilde{E}_{13} & \tilde{E}_{11} & \tilde{E}_{12} \end{bmatrix} \end{array}$$

to get

$$\begin{array}{c} \\ q \\ r \end{array}\!\!\begin{array}{c} \begin{array}{ccc} q & r & s \end{array} \\ \begin{bmatrix} \tilde{A}_{33} & \tilde{A}_{31} & \tilde{A}_{32} \\ \tilde{E}_{13} & \tilde{E}_{11} & \tilde{E}_{12} \end{bmatrix} \end{array}\tilde{Q} = \begin{array}{c} \\ q \\ r \end{array}\!\!\begin{array}{c} \begin{array}{ccc} q & r & s \end{array} \\ \begin{bmatrix} A_{33} & 0 & 0 \\ E_{13} & E_{11} & 0 \end{bmatrix} \end{array}.$$

If $K \in \mathbf{C}^{n\times n}$ is the permutation matrix

$$K = \begin{array}{c} \\ r \\ s \\ q \end{array} \begin{array}{c} \begin{array}{ccc} q & r & s \end{array} \\ \begin{bmatrix} 0 & I & 0 \\ 0 & 0 & I \\ I & 0 & 0 \end{bmatrix} \end{array},$$

then P as in (17) and $Q = K\tilde{Q}K^T$ satisfy the statement of the lemma. □

Note that E_{11} in (13) is not necessarily of full row rank. To achieve this we apply the lemma recursively to construct the following staircase-like condensed form, which generalizes the staircase form in [37] to three and four matrices.

THEOREM 3.2. *There exists a state transformation of* (1) *by unitary matrices* $P \in \mathbf{C}^{n\times n}$ *and* $Q \in \mathbf{C}^{n\times n}$ *which puts the system pencil in the form*

$$PEQ = \begin{array}{c} \\ t_1 \\ t_2 \\ t_3 \end{array} \begin{array}{c} \begin{array}{ccc} t_1 & t_2 & t_3 \end{array} \\ \begin{bmatrix} E_{11} & 0 & E_{13} \\ 0 & 0 & E_{23} \\ 0 & 0 & E_{33} \end{bmatrix} \end{array}, \tag{18}$$

$$PAQ = \begin{array}{c} \\ t_1 \\ t_2 \\ t_3 \end{array} \begin{array}{c} \begin{array}{ccc} t_1 & t_2 & t_3 \end{array} \\ \begin{bmatrix} A_{11} & A_{12} & A_{13} \\ A_{21} & A_{22} & A_{23} \\ 0 & 0 & A_{33} \end{bmatrix} \end{array}, \tag{19}$$

$$PB = \begin{array}{c} \\ t_1 \\ t_2 \\ t_3 \end{array} \begin{array}{c} \begin{array}{c} m \end{array} \\ \begin{bmatrix} B_1 \\ B_2 \\ 0 \end{bmatrix} \end{array}, \tag{20}$$

$$CQ = \begin{array}{c} \\ p \end{array} \begin{array}{c} \begin{array}{ccc} t_1 & t_2 & t_3 \end{array} \\ \begin{bmatrix} C_1 & C_2 & C_3 \end{bmatrix} \end{array}, \tag{21}$$

where

(1) $\operatorname{rank}(E_{11}) = t_1$,

(2) $\operatorname{rank}(B_2) = t_2$,

(3) A_{33} *is block upper triangular, and*

(4) E_{33} *is block upper triangular, has zero diagonal blocks, and is partitioned conformally with* A_{33}.

Proof. The proof uses Lemma 3.1 inductively. Initially, apply Lemma 3.1 to get unitary matrices $P^{(1)}$ and $Q^{(1)}$ such that

$$P^{(1)}EQ^{(1)} = \begin{bmatrix} E_{11}^{(1)} & 0 & E_{13}^{(1)} \\ 0 & 0 & 0 \\ 0 & 0 & 0 \end{bmatrix},$$

$$P^{(1)}AQ^{(1)} = \begin{bmatrix} A_{11}^{(1)} & A_{12}^{(1)} & A_{13}^{(1)} \\ A_{21}^{(1)} & A_{22}^{(1)} & A_{23}^{(1)} \\ 0 & 0 & A_{33}^{(1)} \end{bmatrix},$$

$$P^{(1)}B = \begin{bmatrix} B_1^{(1)} \\ B_2^{(1)} \\ 0 \end{bmatrix},$$

$$CQ^{(1)} = \begin{bmatrix} C_1^{(1)} & C_2^{(1)} & C_3^{(1)} \end{bmatrix}.$$

For the inductive step, assume that we have constructed a unitary state transformation $P^{(k)}$ and $Q^{(k)}$ such that the transformed system is in the form of (18)–(21) with the exception that $\operatorname{rank}(E_{11}^{(k)}) < t_1$. Apply Lemma 3.1 to the subsystem

$$\tilde{E} = \begin{bmatrix} E_{11}^{(k)} & 0 \\ 0 & 0 \end{bmatrix}, \quad \tilde{A} = \begin{bmatrix} A_{11}^{(k)} & A_{12}^{(k)} \\ A_{21}^{(k)} & A_{22}^{(k)} \end{bmatrix}, \quad \tilde{B} = \begin{bmatrix} B_1^{(k)} \\ B_2^{(k)} \end{bmatrix},$$
$$\tilde{C} = \begin{bmatrix} C_1^{(k)} & C_2^{(k)} \end{bmatrix}$$

to obtain a state transformation of the subsystem $\tilde{P}$ and $\tilde{Q}$ which brings the subsystem into the form of (13)–(16). Embed this transformation by defining

$$P^{(k+1)} = \begin{bmatrix} \tilde{P} & 0 \\ 0 & I \end{bmatrix} P^{(k)},$$
$$Q^{(k+1)} = Q^{(k)} \begin{bmatrix} \tilde{Q} & 0 \\ 0 & I \end{bmatrix}.$$

If the $(1,1)$ block of $P^{(k+1)}EQ^{(k+1)}$ is nonsingular, then the pencil is in the required form. Otherwise, Lemma 3.1 may be applied again to further refine the block structure. Each application of Lemma 3.1 reduces t_1 by at least one. After at most n steps either the $(1,1)$ block of the transformed E is nonsingular or $t_1 = 0$. In either case, the pencil reaches the required form in at most n steps. ∎

Theorem 3.2 essentially separates the uncontrollable infinite modes from the others. We have the following corollary.

COROLLARY 3.3. *In Theorem* 3.2*, the subsystem obtained from the first two block rows and columns of* (18)–(21)*,*

$$\begin{bmatrix} E_{11} & 0 \\ 0 & 0 \end{bmatrix} \begin{bmatrix} \dot{x}_1 \\ \dot{x}_2 \end{bmatrix} = \begin{bmatrix} A_{11} & A_{12} \\ A_{21} & A_{22} \end{bmatrix} \begin{bmatrix} x_1 \\ x_2 \end{bmatrix} + \begin{bmatrix} B_1 \\ B_2 \end{bmatrix} u,$$

is controllable at infinity.

Proof. The proof follows directly from the definition. ∎

To understand the output feedback case, it is helpful to condense (18)–(21) somewhat further to a double staircase form that decouples both the uncontrollable at infinity modes and the unobservable at infinity modes. (See also [6].)

THEOREM 3.4. *There exists a state transformation of* (1) *by unitary matrices* $\tilde{P} \in \mathbf{C}^{n\times n}$ *and* $\tilde{Q} \in \mathbf{C}^{n\times n}$ *which puts the system pencil in the form*

$$\tilde{P}E\tilde{Q} = \begin{array}{c} \\ \tilde{t}_1 \\ \tilde{t}_2 \\ \tilde{t}_3 \\ \tilde{t}_4 \end{array} \begin{array}{c} \begin{array}{cccc} \tilde{t}_1 & \tilde{t}_2 & \tilde{t}_3 & \tilde{t}_4 \end{array} \\ \begin{bmatrix} \tilde{E}_{11} & 0 & 0 & \tilde{E}_{14} \\ 0 & 0 & 0 & \tilde{E}_{24} \\ \tilde{E}_{31} & \tilde{E}_{32} & \tilde{E}_{33} & \tilde{E}_{34} \\ 0 & 0 & 0 & \tilde{E}_{44} \end{bmatrix} \end{array}, \tag{22}$$

$$\tilde{P}A\tilde{Q} = \begin{array}{c} \\ \tilde{t}_1 \\ \tilde{t}_2 \\ \tilde{t}_3 \\ \tilde{t}_4 \end{array} \begin{array}{c} \begin{array}{cccc} \tilde{t}_1 & \tilde{t}_2 & \tilde{t}_3 & \tilde{t}_4 \end{array} \\ \begin{bmatrix} \tilde{A}_{11} & \tilde{A}_{12} & 0 & \tilde{A}_{14} \\ \tilde{A}_{21} & \tilde{A}_{22} & 0 & \tilde{A}_{24} \\ \tilde{A}_{31} & \tilde{A}_{32} & \tilde{A}_{33} & \tilde{A}_{34} \\ 0 & 0 & 0 & \tilde{A}_{44} \end{bmatrix} \end{array}, \tag{23}$$

$$\tilde{P}B = \begin{matrix} \\ \tilde{t}_1 \\ \tilde{t}_2 \\ \tilde{t}_3 \\ \tilde{t}_4 \end{matrix} \begin{matrix} m \\ \left[\begin{matrix} \tilde{B}_1 \\ \tilde{B}_2 \\ \tilde{B}_3 \\ 0 \end{matrix}\right] \end{matrix}, \tag{24}$$

$$C\tilde{Q} = \begin{matrix} \\ p \end{matrix} \begin{matrix} \tilde{t}_1 & \tilde{t}_2 & \tilde{t}_3 & \tilde{t}_4 \\ \left[\begin{matrix} \tilde{C}_1 & \tilde{C}_2 & 0 & \tilde{C}_4 \end{matrix}\right] \end{matrix}, \tag{25}$$

with the following properties.

(1) $\operatorname{rank}(\tilde{E}_{11}) = \tilde{t}_1$*;*

(2) $\operatorname{rank}(\tilde{C}_2) = \tilde{t}_2$*;*

(3) $\tilde{A}_{33}$ *is block lower triangular;*

(4) $\tilde{E}_{33}$ *is block lower triangular with zero diagonal blocks, partitioned conformally with* $\tilde{A}_{33}$*;*

(5) $\tilde{A}_{44}$ *is block upper triangular;*

(6) $\tilde{E}_{44}$ *is block upper triangular with zero diagonal blocks, partitioned conformally with* $\tilde{A}_{44}$*;*

(7) *the subsystem obtained by deleting the last block row and column in* (22)–(25) *is controllable at infinity.*

Proof. Apply Theorem 3.2 to system (1) to get the state transformations P_1 and Q_1 and partitioning of (18)–(21). Apply Theorem 3.2 again to the transposed subsystem given by

$$\hat{E} = \begin{pmatrix} & t_1 & t_2 \\ t_1 & \left[\begin{matrix} E_{11} & 0 \end{matrix}\right] \\ t_2 & \left[\begin{matrix} 0 & 0 \end{matrix}\right] \end{pmatrix}^H,$$

$$\hat{A} = \begin{pmatrix} & t_1 & t_2 \\ t_1 & \left[\begin{matrix} A_{11} & A_{12} \end{matrix}\right] \\ t_2 & \left[\begin{matrix} A_{21} & A_{22} \end{matrix}\right] \end{pmatrix}^H,$$

$$\hat{B} = \begin{pmatrix} & t_1 & t_2 \\ p & \left[\begin{matrix} C_1 & C_2 \end{matrix}\right] \end{pmatrix}^H,$$

$$\hat{C} = \begin{pmatrix} & m \\ t_1 & \left[\begin{matrix} B_1 \end{matrix}\right] \\ t_2 & \left[\begin{matrix} B_2 \end{matrix}\right] \end{pmatrix}^H$$

to get orthogonal matrices $\hat{P}$, $\hat{Q} \in \mathbf{C}^{(t_1+t_2)\times(t_1+t_2)}$ that reduce the subsystem to the form of (18)–(21). Define P_2 and Q_2 by

$$P_2 = \begin{bmatrix} \hat{Q}^H & 0 \\ 0 & I_{t_3} \end{bmatrix},$$

$$Q_2 = \begin{bmatrix} \hat{P}^H & 0 \\ 0 & I_{t_3} \end{bmatrix}.$$

The state transformation given by $\tilde{P} = P_2 P_1$ and $\tilde{Q} = Q_1 Q_2$ achieves the condensed form of (22)–(25).

Properties (1)–(4) come directly from Theorem 3.2 applied to the subsystem. Properties (5) and (6) also follow from Theorem 3.2 because when the second state transformation P_2, Q_2 is applied to (18)–(21), E_{33} and A_{33} are unchanged; i.e., $\tilde{A}_{44}$ in (23) is just A_{33} in (19) and $\tilde{E}_{44}$ in (22) is just E_{33} in (19). Property (7) follows because the first three block rows and columns in (22)–(25) are a state transformation by $\hat{P}$ and $\hat{Q}$ of the first two block rows and columns of (18)–(21). ∎

We have the following corollary.

COROLLARY 3.5. *The subsystem obtained by deleting the last two block rows and columns from* (22)–(25) *is controllable and observable at infinity.*

Proof. It is clear that the subsystem is observable at infinity by construction. By Theorem 3.4, we have that the subsystem given by

$$\begin{array}{c} \\ \tilde{t}_1 \\ \tilde{t}_2 \\ \tilde{t}_3 \end{array} \begin{array}{c} \begin{array}{ccc} \tilde{t}_1 & \tilde{t}_2 & \tilde{t}_3 \end{array} \\ \begin{bmatrix} \tilde{E}_{11} & 0 & 0 \\ 0 & 0 & 0 \\ \tilde{E}_{31} & \tilde{E}_{32} & \tilde{E}_{33} \end{bmatrix} \end{array}, \quad \begin{array}{c} \\ \tilde{t}_1 \\ \tilde{t}_2 \\ \tilde{t}_3 \end{array} \begin{array}{c} \begin{array}{ccc} \tilde{t}_1 & \tilde{t}_2 & \tilde{t}_3 \end{array} \\ \begin{bmatrix} \tilde{A}_{11} & \tilde{A}_{12} & 0 \\ \tilde{A}_{21} & \tilde{A}_{22} & 0 \\ \tilde{A}_{31} & \tilde{A}_{32} & \tilde{A}_{33} \end{bmatrix} \end{array}, \quad \begin{array}{c} \\ \tilde{t}_1 \\ \tilde{t}_2 \\ \tilde{t}_3 \end{array} \begin{array}{c} \begin{array}{c} m \end{array} \\ \begin{bmatrix} \tilde{B}_1 \\ \tilde{B}_2 \\ \tilde{B}_3 \end{bmatrix} \end{array},$$

$$\begin{array}{c} \\ p \end{array} \begin{array}{c} \begin{array}{ccc} \tilde{t}_1 & \tilde{t}_2 & \tilde{t}_3 \end{array} \\ \begin{bmatrix} \tilde{C}_1 & \tilde{C}_2 & 0 \end{bmatrix} \end{array}$$

is controllable at infinity. This directly implies that the subsystem given by

$$\begin{array}{c} \\ \tilde{t}_1 \\ \tilde{t}_2 \end{array} \begin{array}{c} \begin{array}{cc} \tilde{t}_1 & \tilde{t}_2 \end{array} \\ \begin{bmatrix} \tilde{E}_{11} & 0 \\ 0 & 0 \end{bmatrix} \end{array}, \quad \begin{array}{c} \\ \tilde{t}_1 \\ \tilde{t}_2 \end{array} \begin{array}{c} \begin{array}{cc} \tilde{t}_1 & \tilde{t}_2 \end{array} \\ \begin{bmatrix} \tilde{A}_{11} & \tilde{A}_{12} \\ \tilde{A}_{21} & \tilde{A}_{22} \end{bmatrix} \end{array}, \quad \begin{array}{c} \\ \tilde{t}_1 \\ \tilde{t}_2 \end{array} \begin{array}{c} \begin{array}{c} m \end{array} \\ \begin{bmatrix} \tilde{B}_1 \\ \tilde{B}_2 \end{bmatrix} \end{array},$$

$$\begin{array}{c} \\ p \end{array} \begin{array}{c} \begin{array}{cc} \tilde{t}_1 & \tilde{t}_2 \end{array} \\ \begin{bmatrix} \tilde{C}_1 & \tilde{C}_2 \end{bmatrix} \end{array}$$

is also controllable at infinity. ∎

4. Regularization and index minimization by state feedback. The following theorem answers the question of when state feedback may be used to make a descriptor system regular.

THEOREM 4.1. *If system* (1) *is in the form of Theorem* 3.2*, then the system is regularizable by state feedback; i.e., there exists a state feedback gain matrix* $F \in \mathbf{C}^{m\times n}$ *such that the pencil* $\lambda E - (A + BF)$ *is regular if and only if* A_{33} *is nonsingular.*

Proof. Let $F \in \mathbf{C}^{m\times n}$ be partitioned as

$$\begin{array}{c} \\ F = m \end{array} \begin{array}{c} \begin{array}{ccc} t_1 & t_2 & t_3 \end{array} \\ \begin{bmatrix} F_1 & F_2 & F_3 \end{bmatrix} \end{array}.$$

The pencil $\lambda E - (A + BF)$ is block upper triangular, so its characteristic polynomial is

$$\begin{aligned} &\det(\lambda E - (A + BF)) \\ &\quad = \det\left(\lambda \begin{bmatrix} E_{11} & 0 \\ 0 & 0 \end{bmatrix} - \left(\begin{bmatrix} A_{11} + B_1F_1 & A_{12} + B_1F_2 \\ A_{21} + B_2F_1 & A_{22} + B_2F_2 \end{bmatrix}\right)\right) \det(\lambda E_{33} - A_{33}) \\ &\quad = \det\left(\lambda \begin{bmatrix} E_{11} & 0 \\ 0 & 0 \end{bmatrix} - \left(\begin{bmatrix} A_{11} + B_1F_1 & A_{12} + B_1F_2 \\ A_{21} + B_2F_1 & A_{22} + B_2F_2 \end{bmatrix}\right)\right) \det(-A_{33}). \end{aligned} \tag{26}$$

The last equality follows because $\lambda E_{33} - A_{33}$ is block triangular with diagonal blocks that are independent of λE_{33}.

If A_{33} is singular, then (26) is zero independent of λ, and the pencil is not regular.

Suppose that A_{33} is nonsingular. Because B_2 has full row rank, there exists a matrix $F_2 \in \mathbf{C}^{m\times t_2}$ such that $A_{22} + B_2F_2$ is nonsingular. Let

$$F = \begin{matrix} \\ m \end{matrix} \begin{matrix} \begin{matrix} t_1 & t_2 & t_3 \end{matrix} \\ \begin{bmatrix} 0 & F_2 & 0 \end{bmatrix} \end{matrix}.$$

The first factor on the right-hand side of (26) is the characteristic polynomial of the subpencil

$$\lambda \begin{bmatrix} E_{11} & 0 \\ 0 & 0 \end{bmatrix} - \left(\begin{bmatrix} A_{11} & A_{12} \\ A_{21} & A_{22} \end{bmatrix} + \begin{bmatrix} B_1 \\ B_2 \end{bmatrix} \begin{bmatrix} 0 & F_2 \end{bmatrix} \right). \tag{27}$$

Since $A_{22} + B_2F_2$ is nonsingular, the pencil in (27) is equivalent to a pencil of the form

$$\lambda \begin{bmatrix} E_{11} & 0 \\ 0 & 0 \end{bmatrix} - \begin{bmatrix} \tilde{A}_{11} & \tilde{A}_{12} \\ 0 & A_{22} + B_2F_2 \end{bmatrix}. \tag{28}$$

It then follows from the nonsingularity of E_{11} and Lemma 2.3 that (28) is regular and has index one. Hence, neither factor in (26) is identically zero and the pencil is regular. ∎

The next theorem shows what index can be achieved by state feedback.

THEOREM 4.2. *If system* (1) *is in the form of Theorem* 3.2 *and* A_{33} *is nonsingular, then there exists a state feedback gain matrix* $F \in \mathbf{C}^{m\times n}$ *such that* $\lambda E - (A + BF)$ *is regular and*

$$\operatorname{index}(\lambda E - (A + BF)) = \operatorname{index} \begin{pmatrix} & \begin{matrix} t_2 & t_3 \end{matrix} \\ \begin{matrix} t_2 \\ t_3 \end{matrix} & \begin{bmatrix} 0 & E_{23} \\ 0 & E_{33} \end{bmatrix} \end{pmatrix}.$$

Proof. Choose $F_1 \in \mathbf{C}^{m\times t_1}$ so that $A_{21} + B_2F_1 = 0$ and choose $F_2 \in \mathbf{C}^{m\times t_2}$ so that $A_{22} + B_2F_2$ is nonsingular. Both F_1 and F_2 exist, because B_2 has full row rank. Define $F \in \mathbf{C}^{m\times n}$ by

$$F = \begin{matrix} \\ m \end{matrix} \begin{matrix} \begin{matrix} t_1 & t_2 & t_3 \end{matrix} \\ \begin{bmatrix} F_1 & F_2 & 0 \end{bmatrix} \end{matrix}.$$

The pencil $\lambda E - (A + BF)$ is block upper triangular with diagonal blocks

$$\lambda E_{11} - (A_{11} + B_1F_1), \tag{29}$$

$$\lambda \begin{bmatrix} 0 & E_{23} \\ 0 & E_{33} \end{bmatrix} - \begin{bmatrix} A_{22} + B_2F_2 & A_{32} \\ 0 & A_{33} \end{bmatrix}. \tag{30}$$

Pencil (29) has only finite eigenvalues because E_{11} is nonsingular. Pencil (30) has only infinite eigenvalues, because the left-hand side is nilpotent and the right-hand side is nonsingular. Lemma 2.2 implies that

$$\begin{aligned} \operatorname{index}(\lambda E - (A + BF)) &= \operatorname{index} \left(\lambda \begin{bmatrix} 0 & E_{23} \\ 0 & E_{33} \end{bmatrix} - \begin{bmatrix} A_{22} + B_2F_2 & A_{32} \\ 0 & A_{33} \end{bmatrix} \right) \\ &= \operatorname{index} \left(\begin{bmatrix} 0 & E_{23} \\ 0 & E_{33} \end{bmatrix} \right). \end{aligned}$$

Here we have used properties 3 and 4 of Theorem 3.2. ∎

The index of nilpotency of $\left(\begin{bmatrix} 0 & E_{23} \\ 0 & E_{33} \end{bmatrix}\right)$ can be displayed by applying another staircase algorithm.

The Kronecker-like feedback canonical form of [24] also displays this minimal index, but this canonical form is not suitable for numerical computation.

An obvious consequence of Theorem 4.2 is that the index of nilpotency of E_{33} differs from the minimal obtainable index by at most one.

THEOREM 4.3. *Suppose that system* (1) *is in the form of Theorem* 3.2. *If* $F \in \mathbf{C}^{m\times n}$ *is a state feedback gain matrix for which* $\lambda E - (A + BF)$ *is regular, then*

$$\operatorname{index}(E_{33}) + 1 \geq \operatorname{index}(\lambda E - (A + BF)) \geq \operatorname{index}(E_{33}).$$

Proof. The proof is an immediate consequence of Theorem 4.2. ∎

If (1) is in the form of Theorem 3.2, then the subsystem $E_{33}\dot{z} = A_{33}z$ represents the uncontrollable infinite eigenvalue modes. Being uncontrollable, as Theorem 4.3 shows, there is nothing that feedback can do to lower $\operatorname{index}(\lambda E_{33} - A_{33})$. However, the next theorem shows that, regardless of initial condition, $z(t) = 0$ for $t > 0$ and $z(0^+) = 0$.

If $z(0^-) = 0$, then z is impulse free and constrained to be zero. Hence, the uncontrollable, infinite modes are not involved in the dynamics! To prove this result we use the trivial fact that if $N \in \mathbf{C}^{n\times n}$ is nilpotent, then the only smooth function $x = x(t) \in \mathbf{C}^n$ satisfying $N\dot{x} = x$ is $x = 0$. It follows that any distributional solution of $N\dot{x} = x$ is purely impulsive and for $t > 0$, $x(t) = 0$. The distributional solution is impulse free if and only if $Nx(0^-) = 0$. Moreover, $x(0^-) = x(0^+)$ if and only if $x(0^-) = x(0^+) = 0$.

THEOREM 4.4. *If the descriptor system in the form of Theorem* 3.2,

$$\begin{bmatrix} E_{11} & 0 & E_{13} \\ 0 & 0 & E_{23} \\ 0 & 0 & E_{33} \end{bmatrix} \begin{bmatrix} \dot{x}_1 \\ \dot{x}_2 \\ \dot{x}_3 \end{bmatrix} = \begin{bmatrix} A_{11} & A_{12} & A_{13} \\ A_{21} & A_{22} & A_{23} \\ 0 & 0 & A_{33} \end{bmatrix} \begin{bmatrix} x_1 \\ x_2 \\ x_3 \end{bmatrix} + \begin{bmatrix} B_1 \\ B_2 \\ 0 \end{bmatrix} u, \tag{31}$$

is regularizable, then for $t > 0$, $x_3(t) = 0$ *independent of the control* u. *If* $E_{33}x_3(0^-) = 0$, *then* x_3 *is impulse free.*

Proof. The third equation of (31) is $E_{33}\dot{x}_3 = A_{33}x_3$. Theorem 4.1 implies that A_{33} is nonsingular, so this is equivalent to $A_{33}^{-1}E_{33}\dot{x}_3 = x_3$ and $A_{33}^{-1}E_{33}x_3(0^-) = 0$. By hypothesis, properties 3 and 4 of Theorem 3.2 hold, so A_{33} and E_{33} are block upper triangular and $A_{33}^{-1}E_{33}$ is nilpotent. The theorem follows. ∎

It follows from Theorems 4.1, 4.2, and 4.4 that a regularizable system decouples into the uncontrollable infinite modes and the subsystem

$$\begin{bmatrix} E_{11} & 0 \\ 0 & 0 \end{bmatrix} \begin{bmatrix} \dot{x}_1 \\ \dot{x}_2 \end{bmatrix} = \begin{bmatrix} A_{11} & A_{12} \\ A_{21} & A_{22} \end{bmatrix} \begin{bmatrix} x_1 \\ x_2 \end{bmatrix} + \begin{bmatrix} B_1 \\ B_2 \end{bmatrix} u. \tag{32}$$

The infinite uncontrollable modes play no role in the system dynamics. With consistent initial conditions, they are constrained to be zero. Only the modes involved in (32) are active. The subsystem (32) is controllable at infinity, so the results of [4, 3] apply. It follows that those modes constrained to be zero may be eliminated from the system. In this way, all regularizable descriptor systems may be made to be controllable at infinity. Hence, the methods designed for linear quadratic control, pole assignment, stabilization, etc., under the assumption of controllability at infinity, may be used [27].

A similar result in the context of linear quadratic control of a particular mechanical multibody system was obtained by explicit transformation in [28].

5. Geometric proofs. In this section we will provide *geometric* proofs for the results in the previous sections.

Let $\mathcal{I}_s$ denote the *largest* subspace $\mathcal{L}$ that satisfies

$$\mathcal{L} \subset A^{(-1)}(E\mathcal{L} + \mathrm{Im}(B)).$$

For a discussion of this space see [13, 23, 26]. The subspace $\mathcal{I}_s = A^{(-1)}(E\mathcal{L}+\mathrm{Im}(B))$ is called the *consistent* subspace, since every point in $\mathcal{I}_s$ is consistent; i.e., for every point $x_0 \in \mathcal{I}_s$ there exists a smooth input $u(t)$ and an associated smooth state trajectory $x(t)$ of system (1) satisfying $x(0) = x_0$ [13].

Let $\mathcal{X}_1$ be such that $\mathcal{X}_1 \oplus (\mathcal{I}_s \cap \ker(E)) = \mathcal{I}_s$; let $\mathcal{X}_3$ be such that $\mathcal{I}_s \oplus \mathcal{X}_3 = \mathbf{C}^n$; and let $\mathcal{Y}_2, \mathcal{Y}_3, \mathcal{Y}_4$ be spaces chosen such that $E\mathcal{I}_s \oplus \mathcal{Y}_2 = E\mathcal{I}_s + \mathrm{Im}(B)$, $(E\mathcal{I}_s + \mathrm{Im}(B)) \oplus \mathcal{Y}_3 = E\mathcal{I}_s + \mathrm{Im}(B) + \mathrm{Im}(A)$, and $(E\mathcal{I}_s + \mathrm{Im}(B) + \mathrm{Im}(A)) \oplus \mathcal{Y}_4 = \mathbf{C}^n$. Choose $\mathcal{U}_2$ so that $B^{(-1)}(E\mathcal{I}_s) \oplus \mathcal{U}_2 = \mathbf{C}^m$. With respect to suitably chosen bases, (1) transforms to

$$\begin{bmatrix} E_{11} & 0 & E_{13} \\ 0 & 0 & E_{23} \\ 0 & 0 & E_{33} \\ 0 & 0 & E_{43} \end{bmatrix} \dot{x}(t) = \begin{bmatrix} A_{11} & A_{12} & A_{13} \\ A_{21} & A_{22} & A_{23} \\ 0 & 0 & A_{33} \\ 0 & 0 & 0 \end{bmatrix} x(t) + \begin{bmatrix} B_{11} & B_{12} \\ 0 & B_{22} \\ 0 & 0 \\ 0 & 0 \end{bmatrix} u(t). \tag{33}$$

By construction, $\ker(B_{22}) = 0$, B_{22} is right invertible, E_{11} is invertible, and

$$M(\lambda) = \left[\lambda \begin{bmatrix} E_{23} \\ E_{33} \\ E_{43} \end{bmatrix} - \begin{bmatrix} A_{23} \\ A_{33} \\ 0 \end{bmatrix}, \quad \begin{bmatrix} B_{22} \\ 0 \\ 0 \end{bmatrix} \right]$$

has full column rank for all $\lambda \in \mathbf{C}$ [15, Appendix, Lemma 1]. In addition, we can show that A_{33} is square and therefore nonsingular.

LEMMA 5.1. *If the system* (1) *is in the form* (33)*, then* A_{33} *is invertible.*

Proof. We have that $A\mathcal{I}_s = (E\mathcal{I}_s + \mathrm{Im}(B)) \cap \mathrm{Im}(A)$. (See, for example, [16].) In addition [30]

$$\begin{aligned} \dim(A\mathcal{I}_s) &= \dim(\mathcal{I}_s) - \dim(\mathcal{I}_s \cap \ker(A)) \\ &= \dim(\mathcal{I}_s) - \dim(\ker(A)) \\ &= \dim(\mathcal{I}_s) - n + \mathrm{rank}(A). \end{aligned}$$

Thus,

$$\begin{aligned} &\dim(E\mathcal{I}_s + \mathrm{Im}(B)) + \dim(\mathrm{Im}(A)) \\ &= \dim(E\mathcal{I}_s + \mathrm{Im}(B) + \mathrm{Im}(A)) + \dim(E\mathcal{I}_s + \mathrm{Im}(B)) \cap \mathrm{Im}(A)) \\ &= \dim(E\mathcal{I}_s + \mathrm{Im}(B) + \mathrm{Im}(A)) + \dim(\mathcal{I}_s) - n + \dim(\mathrm{Im}(A)) \end{aligned}$$

and $\dim(E\mathcal{I}_s + \mathrm{Im}(B) + \mathrm{Im}(A)) = n - \dim(\mathcal{I}_s)$. Therefore, A_{33} is square and hence invertible. ∎

The next step is to relate the condensed form (33) to the regularizability of the system. We have the following well-known result [11, 31].

THEOREM 5.2. *The following are equivalent.*

(i) $E\mathcal{I}_s + \mathrm{Im}(B) + \mathrm{Im}(A) = \mathbf{C}^n$.

(ii) $\begin{bmatrix} \lambda E - A, & B \end{bmatrix}$ *is right invertible as a rational matrix.*

(iii) *The system* (1) *is regularizable by proportional state feedback.*

Proof. For completeness we give a short proof of this result. Statement (i) implies that $\mathcal{Y}_4 = \{0\}$. Hence, the last block row in (33) does not occur. From Lemma 5.1, it follows that A_{33} is invertible. By choosing feedback $u = Fx + v$ with

$$F = \begin{bmatrix} 0 & 0 & 0 \\ -B_{22}^{-1}A_{21} & B_{22}^{-1}(I - A_{22}) & 0 \end{bmatrix}, \tag{34}$$

we obtain the closed loop system

$$E\dot{x} = (A + BF)x + Bv \tag{35}$$

for which $\begin{bmatrix} \lambda E - (A + BF), & B \end{bmatrix}$ is right invertible. This implies (ii).

Conversely (ii) implies that $M(\lambda)$ is right invertible. Hence, $\mathcal{Y}_4 = \{0\}$ and we have (i).

Statement (i) implies statement (iii) because the feedback (34) makes the pencil (35) regular. The converse is clear. If (i) did not hold, then for every feedback F, the pencil of the closed loop system would be singular, which contradicts (iii). ∎

From this we see that for regularizable systems the condensed form (18)–(21), which is constructible in a numerically stable way, coincides with the form (33).

6. Derivative and output feedback. In this section we give a few results about derivative and output feedback.

If we use state derivative feedback, the minimal attainable index is index(E_{33}) in (18).

THEOREM 6.1. *If* (1) *is in the form of Theorem* 3.2 *and* A_{33} *is nonsingular, then then there exists a derivative feedback gain matrix* $G \in \mathbf{C}^{m\times n}$ *such that the pencil* $\lambda(E + BG) - A$ *is regular with* $t_1 + t_2$ *finite eigenvalues and* index$(\lambda(E + BG) - A =$ index(E_{33}).

Proof. Let $G_2 \in \mathbf{C}^{m\times t_2}$ be chosen so that B_2G_2 is nonsingular. Define G by

$$G = \begin{matrix} \\ m \end{matrix} \begin{matrix} \begin{matrix} t_1 & t_2 & t_3 \end{matrix} \\ \begin{bmatrix} 0 & G_2 & 0 \end{bmatrix} \end{matrix}.$$

Then

$$\lambda(E + BG) - A = \lambda \begin{bmatrix} E_{11} & B_1G_2 & E_{13} \\ 0 & B_2G_2 & E_{23} \\ 0 & 0 & E_{33} \end{bmatrix} - \begin{bmatrix} A_{11} & A_{12} & A_{13} \\ A_{21} & A_{22} & A_{23} \\ 0 & 0 & A_{33} \end{bmatrix}.$$

This is a block triangular pencil with diagonal blocks

$$\lambda \begin{bmatrix} E_{11} & B_1G_2 \\ 0 & B_2G_2 \end{bmatrix} - \begin{bmatrix} A_{11} & A_{12} \\ A_{21} & A_{22} \end{bmatrix}, \qquad \lambda E_{33} - A_{33}.$$

By Lemma 2.2, the infinite eigenvalues are the eigenvalues of $\lambda E_{33} - A_{33}$ and

$$\begin{aligned} \text{index}(\lambda(E + BG) - A) &= \text{index}(\lambda E_{33} - A_{33}) \\ &= \text{index}(E_{33}). \end{aligned}$$ ∎

The conclusion of Theorem 6.1 also holds if both state and derivative feedback are used.

An extra hypothesis is needed to obtain the same results in the output feedback case.

THEOREM 6.2. *If equation* (1) *is in the form of Theorem* 3.2*, then the system is regularizable by output feedback; i.e., there exists an output feedback gain matrix* F *such that* $\lambda E - (A + BFC)$ *is regular and*

$$\operatorname{index}(\lambda E - (A + BFC)) = \operatorname{index} \begin{pmatrix} & t_2 & t_3 \\ t_2 & \begin{bmatrix} 0 & E_{23} \end{bmatrix} \\ t_3 & \begin{bmatrix} 0 & E_{33} \end{bmatrix} \end{pmatrix}$$

if and only if A_{33} *is nonsingular and* $\begin{bmatrix} A_{22} \\ C_2 \end{bmatrix}$ *has full column rank.*

Proof. Since B_2 has full row rank there is a matrix F such that $A_{22} + B_2FC_2$ is nonsingular if and only if $\begin{bmatrix} A_{22} \\ C_2 \end{bmatrix}$ has full column rank. Applying the argument in the proof of Theorem 4.1, the result follows. ∎

It follows immediately from Theorem 3.4 that the part of the system which is unobservable at infinity can be completely decoupled from the rest of the system. This part of the system can be removed because, according to the linear model, it does not influence the dynamics of the system and the possible impulsive behavior cannot be observed.

7. Discrete time systems and linear quadratic control. So far, we have discussed continuous time systems only. It should be noted that Lemma 3.1, Corollary 3.5, and Theorems 3.2, 3.4, 4.1, 4.2, 4.3, and 6.1 are independent of the origin of the matrices and thus also hold for discrete time systems. There exists an easily formulated, analogous discrete time version of Theorem 4.4.

The results apply to linear quadratic optimal control problems of the following form:

minimize the cost functional

$$J(x, u) := \int_{t_0}^{t_1} (x^T Q x + u^T R u)\, dt$$

subject to the descriptor system (1).

After transforming to the reduced form of Theorem 3.2, we may just omit the components which are uncontrollable at infinity from the cost functional and the constraint to obtain a reduced order problem which is controllable at infinity. For such systems, the methods described in [27] apply. For a detailed analysis of general linear quadratic optimal control problems for descriptor systems see [15].

8. Conclusions. According to the linear model (1), problems associated with uncontrollable infinite modes in a regularizable system do not occur. With consistent initial conditions, they are constrained to be zero. The only active dynamics in (1) are controllable at infinity. The active dynamics may be made to be index one by state feedback and the entire system may be treated as if it were controllable at infinity as in [4, 3]. However, the resulting system is not index-one robust. If there is an unmodeled forcing function that excites modes that are uncontrollable at infinity, then it may generate impulses.[1]

[1] The situation is similar to one described in a nonsense verse [22]:

Yesterday, upon the stair
I saw a man who wasn't there.
He wasn't there again today.
Gee, I wish he'd go away!

The uncontrollable high-index infinite modes are the man upon the stair. He is not there, but he is disturbing nevertheless.

Acknowledgment. We thank P.C. Müller for bringing the systems that are not controllable at infinity to our attention and A. Bunse-Gerstner, J. Demmel, and P. Van Dooren for helpful discussions in an early stage of this paper. We also thank an anonymous referee for many helpful comments in preparing a revised version of the paper.

REFERENCES

[1] E. Anderson, Z. Bai, C. Bischof, J. Demmel, J. Dongarra, J. D. Croz, A. Greenbaum, S. Hammarling, A. McKenney, S. Ostrouchov, and D. Sorensen, *LAPACK Users' Guide*, SIAM, Philadelphia, PA, 1992.

[2] T. Beelen and P. M. Van Dooren, *An improved algorithm for the computation of Kronecker's canonical form of a singular pencil*, Linear Algebra Appl., 105 (1988), pp. 9–65.

[3] A. Bunse-Gerstner, V. Mehrmann, and N. Nichols, *Regularization of descriptor systems by derivative and proportional state feedback*, SIAM J. Matrix Anal. Appl., 13 (1992), pp. 46–67.

[4] A. Bunse-Gerstner, V. Mehrmann, and N. Nichols, *Regularization of descriptor system by output feedback*, IEEE Trans. Automat. Control, AC–39 (1994), pp. 1742–1748.

[5] D. Cobb, *Controllability, observability and duality in singular systems*, IEEE Trans. Automat. Control, 29 (1984), pp. 1076–1082.

[6] L. Dai, *Singular Control Systems*, Lecture Notes in Control and Information Sciences 118, Springer-Verlag, Berlin, New York, 1989.

[7] J. W. Demmel and B. Kågström, *Stable eigendecompositions of matrix pencils*, Linear Algebra Appl., 88/89 (1987), pp. 137–186.

[8] J. W. Demmel and B. Kågström, *The generalized Schur decomposition of an arbitrary pencil $A - \lambda B$, Parts* I *and* II, ACM Trans. Math. Software, 19 (1993), pp. 160–174, 175–201.

[9] J. J. Dongarra, J. R. Bunch, C. B. Moler, and G. W. Stewart, *LINPACK Users' Guide*, SIAM, Philadelphia, PA, 1979.

[10] L. Elsner, C. He, and V. Mehrmann, *Completion of a matrix so that the inverse has minimal norm. Application to the regularization of descriptor control problems*, in Linear Algebra for Control Theory, P. M. Van Dooren, ed., Springer-Verlag, Berlin, New York, 1993.

[11] L. Fletcher, *Regularizability of descriptor systems*, Internat. J. Systems Sci., 17 (1986), pp. 843–847.

[12] F. Gantmacher, *Theory of Matrices*, Vols. I, II, Chelsea, New York, 1959.

[13] T. Geerts, *Invariant subspaces and invertibility properties for singular systems: The general case*, Linear Algebra Appl., 183 (1993), pp. 61–88.

[14] T. Geerts, *Solvability conditions, consistency, and weak consistency for linear differential-algebraic equations and time-invariant linear systems: The general case*, Linear Algebra Appl., 181 (1993), pp. 111–130.

[15] T. Geerts, *Linear-quadratic control with and without stability subject to general implicit continuous-time systems: Coordinate-free interpretations of the optimal costs in terms dissipation inequality and linear matrix inequality; existence and uniqueness of optimal controls and state trajectories*, Linear Algebra Appl., 203/204 (1994), pp. 607–658.

[16] T. Geerts, *Output consistency and weak output consistency for continuous-time implicit systems*, in Systems and Networks: Mathematical Theory and Applications, Proceedings of the International Symposium MTNS '93 held in Regensburg, Germany, August 2–6, 1993, Vol. II, U. Helmke, R. Mennicken, and J. Saurer, eds., Akademie Verlag, Berlin, 1994, pp. 123–128.

[17] T. Geerts and V. Mehrmann, *Linear Differential Equations with Constant Coefficients: A Distributional Approach*, Tech. Report 90-073, Sonderforschungsbereich 343, Diskrete Strukturen in der Mathematik, Universität Bielefeld, Bielefeld, Germany, 1990.

[18] G. Golub and C. Van Loan, *Matrix Computations*, 2nd ed., The Johns Hopkins University Press, Baltimore, MD, 1989.

[19] M. Hou and P. Müller, *LQ and Tracking Control of Descriptor Systems with Application to Constrained Manipulator*, Tech. report, Sicherheitstechnische Regelungs- und Meßtechnik, Universität Wuppertal, Gauß-Straße 20, D-5600 Wuppertal 1, Germany, 1994.

[20] B. Kågström, *RGSVD—an algorithm for computing the Kronecker structure and reducing subspaces of singular $A-\lambda B$ pencils*, SIAM J. Sci. Statist. Comput., 7 (1986), pp. 185–211.

[21] J. Kautsky, N. K. Nichols, and E. K.-W. Chu, *Robust pole assignment in singular control systems*, Linear Algebra Appl., 121 (1989), pp. 9–37.

[22] B. Lee, *The Man Who Wasn't There. A First Poetry Book*, Oxford University Press, Oxford, UK, 1979.
[23] F. Lewis and K. Özçaldiran, *Geometric structure and feedback in singular systems*, IEEE Trans. Automat. Control, AC-34 (1989), pp. 450–455.
[24] J. J. Loiseau, K. Özçaldiran, M. Malabre, and N. Karcanias, *Feedback canonical forms of singular systems*, Kybernetika, 27 (1991), pp. 289–305.
[25] D. G. Luenberger, *Dynamic equations in descriptor form*, IEEE Trans. Automat. Control, AC-22 (1977), pp. 312–321.
[26] M. Malabre, *Generalized linear systems: Geometric and structural approaches*, Linear Algebra Appl., 1222/123/124 (1989), pp. 591–621.
[27] V. Mehrmann, *The Autonomous Linear Quadratic Control Problem: Theory and Numerical Algorithms*, Lecture Notes in Control and Information Sciences 163, Springer-Verlag, Berlin, New York, Heidelberg, 1991.
[28] P. Müller, *Linear quadratic optimal control of mechanical descriptor systems*, in Systems and Networks: Mathematical Theory and Applications, Proceedings of the International Symposium MTNS '93 held in Regensburg, Germany, August 2–6, 1993, Vol. II, U. Helmke, R. Mennicken, and J. Saurer, eds., Akademie Verlag, Berlin, 1994, pp. 361–366.
[29] P. Müller and T. Schmidt, *Parameterschätzung komplexer mechanischer Regelungssysteme mit Zwangsbedingungen*, Tech. Report 91, Sicherheitstechnische Regelungs- und Meßtechnik, Universität Wuppertal, Gauß-Straße 20, D-5600 Wuppertal 1, Germany, 1991.
[30] D. Owens and D. Debeljkovic, *Consistency and Lyapunov stability of linear descriptor systems: A geometric analysis*, IMA J. Math. Control Inform., 2 (1985), pp. 139–151.
[31] K. Özçaldiran and F. Lewis, *On regularizability of singular systems*, IEEE Trans. Automat. Control, AC-35 (1990), pp. 1156–1160.
[32] H. Rosenbrock, *Structural properties of linear dynamic systems*, Internat. J. Control, 20 (1974), pp. 191–202.
[33] T. Schmidt and M. Hou, *Rollringgetriebe*, Tech. report, Sicherheitstechnische Regelungs- und Meßtechnik, Universität Wuppertal, Gauß-Straße 20, D-5600 Wuppertal 1, Germany, 1992.
[34] B. Simeon, F. Grupp, C. Führer, and P. Rentrop, *A Nonlinear Truck Model and Its Treatment as a Multibody System*, Tech. report, Mathematisches Institut, Technische Universität München, 1992.
[35] B. T. Smith, J. M. Boyle, J. J. Dongarra, B. S. Garbow, Y. Ikebe, V. C. Klema, and C. B. Moler, *Matrix Eigensystem Routines—EISPACK Guide*, Lecture Notes in Computer Science, Springer-Verlag, Berlin, New York, 1976.
[36] P. M. Van Dooren, *The computation of Kronecker's canonical form of a singular pencil*, Linear Algebra Appl., 27 (1979), pp. 103–140.
[37] P. M. Van Dooren, *The generalized eigenstructure problem in linear system theory*, IEEE Trans. Automat. Control, 6 (1981), pp. 111–129.
[38] P. M. Van Dooren, A. Emami-Naeini, and L. Silverman, *Stable extraction of the Kronecker structure of pencils*, in Proc. 17th IEEE Conf. on Decision and Control, San Diego, IEEE, New York, 1979, pp. 521–524.
[39] P. M. Van Dooren and M. Verhaegen, *On the use of unitary state-space transformations*, Contemp. Math., 47 (1985), pp. 447–463.
[40] G. C. Verghese, B. C. Lévy, and T. Kailath, *A general state space for singular systems*, IEEE Trans. Automat. Control, 26 (1981), pp. 118–831.

SIAM J. CONTROL OPTIM.
Vol. 35, No. 2, pp. 480–499, March 1997

006

A BEHAVIORAL APPROACH TO DELAY-DIFFERENTIAL SYSTEMS*

HEIDE GLÜSING-LÜERSSEN†

Abstract. We will study linear time-invariant delay-differential systems from the behavioral point of view as it was introduced for dynamical systems by Willems [*Dynam. Report.*, 2 (1989), pp. 171–269]. A ring $\mathcal{H}$ which lies between $\mathbb{R}[s, z, z^{-1}]$ and $\mathbb{R}(s)[z, z^{-1}]$ will be presented, whose elements can be interpreted as a generalized version of delay-differential operators on $\mathcal{C}^\infty(\mathbb{R}, \mathbb{R})$. In this framework, a behavior is the kernel of such an operator. Using the ring $\mathcal{H}$, an algebraic characterization of inclusion, respectively, equality of the behaviors under consideration, is given. Finally, controllability of the behaviors is characterized in terms of the rank of the associated matrices. In the case of time-delay state-space systems this criterion becomes the known Hautus criterion for spectral controllability.

Key words. time-delay systems, behaviors, polynomial matrices

AMS subject classifications. 93B25, 93C35, 93B05

PII. S0363012995281869

1. Introduction.

The purpose of this paper is an approach to linear time-invariant delay-differential systems with algebraic methods. In contrast to the work of, e.g., Morse [16], Sontag [21], and more recently Habets [8], we will not consider these systems as systems over (polynomial) rings. Instead we will use the behavioral viewpoint for dynamical systems as it was introduced by Willems [22]: our objects will be behaviors, which are defined by linear time-invariant delay-differential equations over the time axis $\mathbb{R}$ (for the definition of a behavior, see [22]). In the scalar case such equations are given by

$$\sum_{j=0}^{L}\sum_{i=0}^{N} p_{ij} w^{(i)}(t-j) = 0, \qquad t \in \mathbb{R}, \tag{1.1}$$

where $p_{ij} \in \mathbb{R}$ and $w^{(i)}$ denotes the ith derivative of the function w. In our approach only functions w in $\mathcal{C}^\infty(\mathbb{R}, \mathbb{R})$ will be considered. In the multivariable case, linear subspaces $\mathcal{B}$ of $\mathcal{C}^\infty(\mathbb{R}, \mathbb{R}^m)$ are investigated that are the solution space of a system of delay-differential equations, i.e., for which there exist n, L, $N \in \mathbb{N}$, and matrices $P_{ij} \in \mathbb{R}^{n\times m}$ so that

$$\mathcal{B} = \left\{ w \in \mathcal{C}^\infty(\mathbb{R}, \mathbb{R}^m) \middle| \sum_{j=0}^{L}\sum_{i=0}^{N} P_{ij} w^{(i)}(t-j) = 0,\ t \in \mathbb{R} \right\}. \tag{1.2}$$

The behavior in (1.2) can be written as $\mathcal{B} = \ker \widetilde{P}$, where $P = \sum_{j=0}^{L}\sum_{i=0}^{N} P_{ij} s^i z^j \in \mathbb{R}[s, z]^{n\times m}$ and $\widetilde{P}$ denotes the associated delay-differential operator from $\mathcal{C}^\infty(\mathbb{R}, \mathbb{R}^m)$ to $\mathcal{C}^\infty(\mathbb{R}, \mathbb{R}^n)$; i.e., $\widetilde{P}w(t) = \sum_{j=0}^{L}\sum_{i=0}^{N} P_{ij} w^{(i)}(t-j)$. Note that (1.2) includes ordinary differential equations ($P \in \mathbb{R}[s]$) as well as the case of a pure delay equation

*Received by the editors February 21, 1995; accepted for publication (in revised form) January 16, 1996.

http://www.siam.org/journals/sicon/35-2/28186.html

†Universität Oldenburg, Fachbereich 6 – Mathematik, Postfach 2503, 26111 Oldenburg, Germany (gluesing@mathematik.uni-oldenburg.de).

($P \in \mathbb{R}[z]$). Since the shift yields an isomorphism on $\mathcal{C}^\infty(\mathbb{R},\mathbb{R})$, it is algebraically more adequate to consider the polynomial ring $\mathbb{R}[s,z,z^{-1}]$ instead of $\mathbb{R}[s,z]$.

Although the space $\mathcal{B}$ is in general infinite dimensional, via polynomial matrices it is given a description with finitely many parameters. This leads to the possibility of studying special aspects of this type of equations with mainly algebraic methods.

The polynomial approach to time-delay systems was already introduced by Kamen [10]. He considered delay-differential operators as special convolution operators in the distributional sense and presented, within this set-up, procedures for the solution of input/output equations and for the internal description (state-space realizations) of such equations.

In the present paper our starting point will be the solution spaces (or behaviors) $\ker \widetilde{P}$ as given in (1.2). We will not investigate the question as to which subspaces of $\mathcal{C}^\infty(\mathbb{R},\mathbb{R}^m)$ occur as such behaviors. The main ideas for an answer to this question are contained in the thesis of Soethoudt [20]. He characterizes behaviors which have an AR-representation (that is, a representation via autoregressive equations) in the purely differential sense. Instead of attacking this (nevertheless interesting) problem of the existence of polynomial representations, we will consider the question of uniqueness: for what pairs of matrices P, Q over $\mathbb{R}[s,z,z^{-1}]$ does $\ker \widetilde{P} = \ker \widetilde{Q}$ hold? It should be obvious that an answer to this question is necessary for the development of a "behavioral theory" using polynomial (AR-) representations for time-delay systems. Simple examples show that the above problem cannot be satisfactorily solved with the help of the ring $\mathbb{R}[s,z,z^{-1}]$ or even $\mathbb{R}(s)[z,z^{-1}]$. The appropriate domain in order to translate relations between behaviors into relations between the associated polynomial matrices lies between these two rings and turns out to be

$$\mathcal{H} = \{p \in \mathbb{R}(s)[z,z^{-1}] \,|\, p(s,e^{-s}) \text{ is an entire function}\}.$$

In the preliminaries an interpretation of the elements of $\mathcal{H}$ as operators on $\mathcal{C}^\infty(\mathbb{R},\mathbb{R})$ is given. It generalizes the interpretation of polynomials in $\mathbb{R}[s,z,z^{-1}]$ as delay-differential operators. Therefore we will refer to these associated operators as delay-differential operators as well.

A similar construction occurred already in the work of Kamen, Khargonekar, and Tannenbaum [11], where the ring Θ generated by the entire functions $(1-e^{-s}e^{\sigma})(s-\sigma)^{-1}$, $\sigma \in \mathbb{C}$, and their derivatives is considered. One can easily see that the ring $\Theta[s,z]$ in [11, p. 841] is contained in $\mathcal{H}$. Kamen, Khargonekar, and Tannenbaum also gave an interpretation of the functions $(1-e^{-s}e^{\sigma})(s-\sigma)^{-1}$ as transfer functions of distributed-delay systems.

One main tool in the present approach is the fact that the division properties in the ring $\mathcal{H}$ correspond to the division properties in the ring of entire functions, i.e., for p, $q \in \mathcal{H}$ it holds: p divides q in $\mathcal{H}$ iff $q(s,e^{-s})p(s,e^{-s})^{-1}$ is an entire function. For the associated delay-differential equations this has as a consequence that it suffices to consider *fundamental solutions*, i.e., functions of the type $w(t) = t^k e^{\lambda t}$ instead of the full solution space. This fits with a result of Malgrange [14, p. 318], who proved that the space of all linear combinations of fundamental solutions of a delay-differential equation lies dense in the full space of smooth solutions (with respect to the topology of uniform convergence of all derivatives on all compact subsets in $\mathbb{R}$).

Another important result in our framework is the fact that $\mathcal{H}$ is a so-called elementary divisor ring. This means first that $\mathcal{H}$ is a Bézout domain, i.e., every finitely generated ideal in $\mathcal{H}$ is principal. Second, every matrix over $\mathcal{H}$ can be brought into diagonal form via multiplication with unimodular matrices from the left and from the right. With this type of normal form (which cannot be achieved, e.g., over the

ring $\mathbb{R}[s, z, z^{-1}]$), the results for multivariable delay-differential equations can easily be derived from the scalar case.

With this information about the ring $\mathcal{H}$, which is derived in section 3, we will show in the fourth section how the relations between behaviors as given in (1.2) can be put into correspondence with the division relations of the associated matrices over $\mathcal{H}$. In particular, we prove for $P \in \mathcal{H}^{n\times m}$, $Q \in \mathcal{H}^{r\times m}$: $\ker \widetilde{P} \subseteq \ker \widetilde{Q}$ iff $Q = AP$ for some $A \in \mathcal{H}^{n\times r}$, which yields $\ker \widetilde{P} = \ker \widetilde{Q}$ iff A is unimodular over $\mathcal{H}$.

Finally, in section 5, controllability of delay-differential systems is considered. In this set-up it is natural to use the notion of controllability for behaviors as introduced by Willems [22]. Using a diagonal form for matrices $P \in \mathcal{H}^{n\times m}$, it will be proven that $\ker \widetilde{P}$ is controllable iff $\mathrm{rk}_{\mathbb{C}} P(s, e^{-s}) = \mathrm{rk}_{\mathcal{H}} P$ for all $s \in \mathbb{C}$. Recently, this characterization has been obtained independently for the same situation of delay-differential equations by Rocha and Willems [19]. The given criterion is a generalization of the Hautus test for time-delay state-space systems which characterizes the so-called *spectral controllability*; see, e.g., Pandolfi [18], Bhat and Koivo [2], Manitius and Triggiani [15], and Kamen, Kargonekar, and Tannenbaum [11].

2. Preliminaries. In this section we present the framework for our study of delay-differential equations and introduce the notations. Starting with the interpretation of polynomials in $\mathbb{R}[s, z, z^{-1}]$ as delay-differential operators on $\mathcal{C}^\infty(\mathbb{R}, \mathbb{R})$, we first have a glance at the fundamental solutions of the associated equations. This leads us to the corresponding characteristic function and its zeros. Simple examples suggest the introduction of a larger space $\mathcal{H}$ of operators which are closely related to the delay-differential operators. Finally we state the surjectivity of the operators under consideration.

DEFINITION 2.1.

(1) *Put* $\mathcal{R} := \mathbb{R}[s, z, z^{-1}]$ *and let* $\mathcal{C}^\infty(\mathbb{R}^m) := \mathcal{C}^\infty(\mathbb{R}, \mathbb{R}^m)$ *for* $m \geq 1$.

(2) *For* $m \geq 1$ *and* $t_0 \in \mathbb{R}$ *define the shift* $\sigma^{t_0} : \mathcal{C}^\infty(\mathbb{R}^m) \to \mathcal{C}^\infty(\mathbb{R}^m)$ *by* $(\sigma^{t_0} w)(t) = w(t - t_0)$ *for* $w \in \mathcal{C}^\infty(\mathbb{R}^m)$. *In particular, let* $\sigma := \sigma^1$.

(3) *With* $P = \sum_{j=l}^{L} \sum_{i=0}^{N} P_{ij} s^i z^j \in \mathcal{R}^{n\times m}$ *associate the following delay-differential operator:*

$$\begin{aligned} \widetilde{P} : \mathcal{C}^\infty(\mathbb{R}^m) &\longrightarrow \mathcal{C}^\infty(\mathbb{R}^n), \\ w &\longmapsto \sum_{j=l}^{L} \sum_{i=0}^{N} P_{ij} \sigma^j w^{(i)}, \end{aligned} \tag{2.1}$$

where $w^{(i)} = \frac{d^i}{dt^i} w$.

(4) *For* $p = \sum_{i=0}^{N} p_i s^i \in \mathbb{R}[s]$ *and* $w \in \mathcal{C}^\infty([a, b], \mathbb{R}) =: \mathcal{C}^\infty[a, b]$ *we use analogously the notion* $\widetilde{p}(w)(t) = \sum_{i=0}^{N} p_i w^{(i)}(t)$, *hence* $\widetilde{p}(w) \in \mathcal{C}^\infty[a, b]$.

Note that part (3) indeed makes sense, since on $\mathcal{C}^\infty(\mathbb{R})$ the operators σ and $\frac{d}{dt}$ commute.

In this context, the solution space in $\mathcal{C}^\infty(\mathbb{R})$ of the scalar equation (1.1) is just $\ker \widetilde{p}$, a linear *shift-invariant* subspace of $\mathcal{C}^\infty(\mathbb{R})$; i.e., $\sigma^t(\ker \widetilde{p}) = \ker \widetilde{p}$ for all $t \in \mathbb{R}$. In this section we will only study the scalar equation (1.1). We will come to the multivariable situation in section 4.

Remark 2.2. The map

$$\begin{aligned} T : \mathcal{R} &\longrightarrow \mathrm{End}_{\mathbb{R}}(\mathcal{C}^\infty(\mathbb{R})), \\ p &\longmapsto \widetilde{p} \end{aligned}$$

is an injective algebra homomorphism. The homomorphism properties $\widetilde{p+q} = \widetilde{p} + \widetilde{q}$, $\widetilde{pq} = \widetilde{p} \circ \widetilde{q}$ can easily be verified. To prove injectivity of T, let $p = \sum_{i,j} p_{ij} s^i z^j \in \mathcal{R}$ and assume that $\widetilde{p} = 0$. Then for arbitrary $\lambda \in \mathbb{C}$ and $w \in \mathcal{C}^\infty(\mathbb{R})$ with $w(t) = e^{\lambda t}$ we obtain $0 = \widetilde{p}(w)(t) = \sum_{i,j} p_{ij} \lambda^i e^{\lambda(t-j)} = \sum_{i,j} p_{ij} \lambda^i e^{-\lambda j} e^{\lambda t}$ for all $t \in \mathbb{R}$, hence $\sum_{i,j} p_{ij} \lambda^i e^{-\lambda j} = 0$. Since this holds true for all $\lambda \in \mathbb{C}$, the linear independence of the functions $\lambda \mapsto \lambda^i e^{\lambda j}$ yields in fact $p_{ij} = 0$ for all i, j.

One question we want to attack in this paper is how to characterize the inclusion $\ker \widetilde{p} \subseteq \ker \widetilde{q}$ in terms of the elements p, $q \in \mathcal{R}$. Let us first have a look at a simple example.

Example 2.3.

(1) For p, $q \in \mathbb{R}[s] \subset \mathcal{R}$ the theory of ordinary differential equations leads to $\ker \widetilde{p} \subseteq \ker \widetilde{q}$ iff p divides q in $\mathbb{R}[s]$, hence iff p divides q in $\mathcal{R}$.

(2) It is easily seen that

$$\ker \widetilde{s} = \{\text{constants}\} \subset \ker \widetilde{z-1} = \{w \in \mathcal{C}^\infty(\mathbb{R}) | w \text{ is of period } 1\}.$$

But s does not divide $z - 1$ in $\mathcal{R}$. Of course, s divides $z - 1$ in $\mathbb{R}(s)[z, z^{-1}]$.

The above shows that the division properties of the two rings $\mathcal{R}$ and $\mathbb{R}(s)[z, z^{-1}]$ are not useful in the algebraic description of $\ker \widetilde{p} \subseteq \ker \widetilde{q}$.

As with ordinary differential equations, more information about the solution space of (1.1) is obtainable by studying fundamental solutions $w(t) = t^k e^{\lambda t}$, where $k \in \mathbb{N}_0$ and $\lambda \in \mathbb{C}$. In the present case this leads to the characteristic function of (1.1), which will be an entire function. We will need the concept of a characteristic function in a slightly more general situation, which is handled in the next definition. In the special case of part (2) of the definition, these functions are often called quasi polynomials (see, e.g., [7, p. 63]) or exponential polynomials (see [1, Chap. 12]). In parts (3) and (4) we introduce some notations useful for what follows.

DEFINITION 2.4.

(1) *For* $p = \sum_{j=l}^{L} p_j z^j \in \mathbb{R}(s)[z, z^{-1}]$ *with* $p_j \in \mathbb{R}(s)$ *and* $p_l \neq 0 \neq p_L$ *define the degree of* p *to be* $\deg_z p := L - l$. *Further, let*

$$p^*(s) := \sum_{j=l}^{L} p_j(s) e^{-js} \text{ for all } s \in \mathbb{C} \text{ not being a pole of } p_j,\ j = l, \ldots, L.$$

Then $p^* \in M(\mathbb{C})$, *the set of all meromorphic functions on* $\mathbb{C}$.

(2) *If* $p = \sum_{j=l}^{L} \sum_{i=0}^{N} p_{ij} s^i z^j \in \mathcal{R}$, *then* $p^* \in H(\mathbb{C})$, *the set of entire functions.* p^* *is called the* characteristic function of the delay-differential equation

$$\sum_{j=l}^{L} \sum_{i=0}^{N} p_{ij} w^{(i)}(t-j) = 0, \qquad t \in \mathbb{R}.$$

(3) *For* $f \in M(\mathbb{C})$ *and* $\alpha \in \mathbb{C}$ *denote the* order of the zero *(resp.,* pole*)* α *of* f *by*

$$\mu_\alpha(f) := \min\{k \in \mathbb{Z} \,|\, (s-\alpha)^{-k} f \text{ holomorphic and not zero around } \alpha\}.$$

(4) *For* $f_1, \ldots, f_r \in M(\mathbb{C})$ *let*

$$\mathcal{V}(f_1, \ldots, f_r) = \{\alpha \in \mathbb{C} \,|\, \mu_\alpha(f_i) \geq 1,\ i = 1, \ldots, r\}$$

be the set of common zeros of $f_1, \ldots, f_r$.

Note that we interpret here s as an algebraic indeterminate over $\mathbb{R}$ as well as a complex variable.

Remark 2.5. The map $\mathbb{R}(s)[z, z^{-1}] \to M(\mathbb{C})$, $p \mapsto p^*$ is an injective ring homomorphism. The injectivity follows, as in Remark 2.2, from the linear independence of the functions $s \mapsto s^k e^{js}$.

With this notation, we get from the theory of delay-differential equations for $p \in \mathcal{R}$ and for the function $w \in \mathcal{C}^\infty(\mathbb{R}, \mathbb{C})$, $w(t) = t^k e^{\lambda t}$,

$$w \in \ker \widetilde{p} \iff \mu_\lambda(p^*) > k \tag{2.2}$$

(see [1, pp. 54–55] for a special case). This can also be proven directly by showing that $\widetilde{p}w(t) = \frac{d^k}{ds^k}(p^*(s)e^{st})|_{s=\lambda}$. As with ordinary differential equations it is true that with $w \in \mathcal{C}^\infty(\mathbb{R}, \mathbb{C})$ also $\operatorname{Re} w$, $\operatorname{Im} w \in \mathcal{C}^\infty(\mathbb{R})$ are in $\ker \widetilde{p}$.

The foregoing consideration indicates that a first knowledge about the dimension of $\ker \widetilde{p}$ can be obtained by calculating the number of zeros of the associated characteristic function p^*. Using the theory of entire functions this can be done in the following sense.

PROPOSITION 2.6. *Let $p \in \mathcal{R}$. Then*

$$\#\mathcal{V}(p^*) < \infty \iff p = z^k \phi \text{ for some } k \in \mathbb{Z} \text{ and } \phi \in \mathbb{R}[s]\backslash\{0\}.$$

This result can be proven by use of some facts about the order of entire functions, as they can be found, e.g., in [9]. Since we are not aware of an explicit proof in the literature, we present here a short sketch of how to establish the result with the help of [9].

Proof. "$\Leftarrow$" is obvious.

"$\Rightarrow$" Let $p = \sum_{j=l}^{L} p_j z^j \in \mathcal{R}$ with $p_j \in \mathbb{R}[s]$. If $\#\mathcal{V}(p^*) < \infty$, then $p^* = ae^g$ with $a \in \mathbb{C}[s]$ and $g \in H(\mathbb{C})$. Suppose that g is not a constant. From [9, Lemmas 2.7.3 and 2.7.4 and Theorem 4.2.1] it follows that $\operatorname{ord}(p^*) = \operatorname{ord}(\sum_{j=l}^{L} p_j e^{-j\cdot}) \leq 1$, where the order $\operatorname{ord}(f)$ of an entire function f is defined as in [9, Definition 1.11.1]. But then [9, Lemmas 2.7.3 and 2.7.5] implies $g \in \mathbb{C}[s]$, and moreover $g(s) = \alpha s + \beta$ with some $\alpha, \beta \in \mathbb{C}$. Hence $p^*(s) = \sum_{j=l}^{L} p_j(s) e^{-js} = a(s)e^{\beta} e^{\alpha s}$. Now, from the independence of the functions $s^k e^{\alpha s}$, we get $\alpha \in \{-L, \ldots, -l\}$ and $p_j = 0$ for $j \neq -\alpha$. Thus $p = p_{-\alpha} z^\alpha$. ∎

Note the simple fact that for $p = z^k \phi \in \mathcal{R}$ with $\phi \in \mathbb{R}[s]$ and $k \in \mathbb{Z}$ one has $\ker \widetilde{p} = \ker \widetilde{\phi}$, which is just the solution space of an ordinary linear homogeneous differential equation with constant coefficients over $\mathbb{R}$. Hence, as an immediate consequence of (2.2) and Proposition 2.6 we get

$$\#\mathcal{V}(p^*) = \infty \iff \dim \ker \widetilde{p} = \infty$$

for arbitrary $p \in \mathcal{R}$. In other words, $\ker \widetilde{p}$ is finite dimensional iff $\widetilde{p}$ is a (shifted) ordinary differential operator. Moreover, for $q \in \mathcal{R}$ and $\phi \in \mathbb{R}[s]\backslash\{0\}$ the finite dimensionality of $\ker \widetilde{\phi}$ together with (2.2) implies the crucial fact that

$$\frac{q^*}{\phi} \in H(\mathbb{C}) \iff \ker \widetilde{\phi} \subseteq \ker \widetilde{q}. \tag{2.3}$$

This easy equivalence is central for our framework, as it allows us to introduce a bigger class $\mathcal{H}$ of linear operators on $\mathcal{C}^\infty(\mathbb{R})$ which are closely related to delay-differential operators. More precisely, for $p = q\phi^{-1} \in \mathbb{R}(s)[z, z^{-1}]$, where $p^* = q^* \phi^{-1} \in H(\mathbb{C})$, it is possible to define $\widetilde{p} = \widetilde{q} \circ \widetilde{\phi}^{-1}$.

We introduce precisely these objects in the following definition and show their well-definedness as well as some elementary properties in Remark 2.8.

DEFINITION 2.7.

(1) *Put* $\mathcal{H} := \{p \in \mathbb{R}(s)[z, z^{-1}] \mid p^* \in H(\mathbb{C})\}$.

(2) *For* $p = q\phi^{-1} \in \mathcal{H}$ *with* $q \in \mathcal{R}$ *and* $\phi \in \mathbb{R}[s]\backslash\{0\}$ *define the operator*

$$\widetilde{p}: \mathcal{C}^\infty(\mathbb{R}) \longrightarrow \mathcal{C}^\infty(\mathbb{R}),$$
$$w \longmapsto \widetilde{p}(w) := \widetilde{q}(v), \quad \text{where } v \in \mathcal{C}^\infty(\mathbb{R}) \text{ with } \widetilde{\phi}(v) = w.$$

We call $\widetilde{p}$ *a* delay-differential operator *also if* $p \in \mathcal{H}$.

Remark 2.8.

(1) From Remark 2.5 it follows that $\mathcal{H}$ is a commutative domain.

(2) One has to establish the well-definedness of the map $\widetilde{p}$. First, for fixed $q \in \mathcal{R}$ and $\phi \in \mathbb{R}[s]$ with $q\phi^{-1} \in \mathcal{H}$ the well-definedness of the map $w \mapsto \widetilde{q}(v)$, where $v \in \mathcal{C}^\infty(\mathbb{R})$ satisfies $\widetilde{\phi}(v) = w$, is a consequence of (2.3). Next, to see that the map $\widetilde{p}$ does not depend on the special representation of p, let $p = q\phi^{-1} = q'\psi^{-1} \in \mathcal{H}$. For $w \in \mathcal{C}^\infty(\mathbb{R})$ put $\widetilde{\phi}(v) = w = \widetilde{\psi}(v')$ and $\widetilde{\phi}(h) = v'$ with suitable v, v', $h \in \mathcal{C}^\infty(\mathbb{R})$. Then $\widetilde{\psi}(h) - v \in \ker\widetilde{\phi} \subseteq \ker\widetilde{q}$ and therefore $\widetilde{q}(v) = \widetilde{q}(\widetilde{\psi}(h)) = \widetilde{q'}(\widetilde{\phi}(h)) = \widetilde{q'}(v')$.

(3) It can easily be verified that $\widetilde{p}$ is an endomorphism on $\mathcal{C}^\infty(\mathbb{R})$. Moreover, the ring $\mathcal{H}$ can be viewed as a subring of $\mathrm{End}_{\mathbb{R}}(\mathcal{C}^\infty(\mathbb{R}))$. To see this, we need to prove that the map $p \mapsto \widetilde{p}$ is an injective ring homomorphism. For this, let $p = a\phi^{-1}$, $q = b\psi^{-1} \in \mathcal{H}$ with a, $b \in \mathcal{R}$, and ϕ, $\psi \in \mathbb{R}[s]$. For $w \in \mathcal{C}^\infty(\mathbb{R})$ define $v \in \mathcal{C}^\infty(\mathbb{R})$ such that $\widetilde{\phi\psi}(v) = w$. Then $\widetilde{p+q}(w) = (\widetilde{a\psi + b\phi})(v) = \widetilde{a\psi}(v) + \widetilde{b\phi}(v) = \widetilde{p}(w) + \widetilde{q}(w)$ and from $\widetilde{\psi}(\widetilde{\phi}(v)) = w$ it follows $\widetilde{p} \circ \widetilde{q}(w) = \widetilde{p}(\widetilde{b}(\widetilde{\phi}(v))) = \widetilde{p} \circ \widetilde{\phi}(\widetilde{b}(v)) = \widetilde{a} \circ \widetilde{b}(v) = \widetilde{ab}(v) = \widetilde{pq}(w)$, where we used the homomorphism properties of T as defined in Remark 2.2. The injectivity of $p \mapsto \widetilde{p}$ follows from the same remark.

(4) A special case of the homomorphism property of $p \mapsto \widetilde{p}$ is the following: from $p = q\phi^{-1} \in \mathcal{H}$ one has obviously $p\phi = q = \phi p$ in the ring $\mathcal{H}$. The definition of $\widetilde{p}$ tells us that $\widetilde{q}(v) = \widetilde{p} \circ \widetilde{\phi}(v)$ for all $v \in \mathcal{C}^\infty(\mathbb{R})$ and $\widetilde{q}(w) = \widetilde{q}(\widetilde{\phi}(v)) = \widetilde{\phi}(\widetilde{q}(v)) = \widetilde{\phi} \circ \widetilde{p}(w)$ for v, $w \in \mathcal{C}^\infty(\mathbb{R})$ satisfying $\widetilde{\phi}(v) = w$. Hence it is indeed $\widetilde{q} = \widetilde{p} \circ \widetilde{\phi} = \widetilde{\phi} \circ \widetilde{p}$.

This shows that Definition 2.7(2) represents the unique extension of the algebra homomorphism T given in Remark 2.2 from $\mathcal{R}$ to the larger ring $\mathcal{H}$.

Let us illustrate the general delay-differential operator by the following example, which is in some sense the simplest nonordinary delay-differential operator.

Example 2.9. Let $p := (z-1)s^{-1} \in \mathbb{R}(s)[z]$. Then $p^*(s) = (e^{-s} - 1)s^{-1}$ is an entire function; thus $p \in \mathcal{H}$. The associated operator is given by

$$\widetilde{p}: \mathcal{C}^\infty(\mathbb{R}) \longrightarrow \mathcal{C}^\infty(\mathbb{R}),$$
$$w \longmapsto \sigma(v) - v, \text{ where } v^{(1)} = w.$$

Obviously, $\ker(\widetilde{z-1}) = \{v \in \mathcal{C}^\infty(\mathbb{R}) | v \text{ is of period } 1\}$; therefore,

$$\ker\widetilde{p} = \{w \in \mathcal{C}^\infty(\mathbb{R}) | \exists v \in \mathcal{C}^\infty(\mathbb{R}) \text{ of period 1 and with } v^{(1)} = w\},$$

which is a proper subspace of $\ker(\widetilde{z-1})$. Note that in the above case we have $\widetilde{p}(w) = \int_t^{t-1} w(\tau)d\tau$, which indicates that $\mathcal{H}$ includes not only point-delay operators but also distributed-delay operators.

As we will see in section 4, it is just the ring $\mathcal{H}$ which gives an algebraic description of the relation between behaviors of the type $\ker\widetilde{p} \subset \mathcal{C}^\infty(\mathbb{R})$: the lattice of kernels of operators $\widetilde{p}$ corresponds to the lattice of principal ideals in $\mathcal{H}$. Therefore, for the development of this correspondence it makes sense to consider also delay-differential operators in the generalized version of Definition 2.7. The ring $\mathcal{H}$ will be investigated in the next section.

We close the preliminaries with the following proposition.

PROPOSITION 2.10. *Let $p \in \mathcal{H}\backslash\{0\}$. Then*

(1) *The map $\widetilde{p} \in \mathrm{End}_{\mathbb{R}}(\mathcal{C}^\infty(\mathbb{R}))$ is surjective.*

(2) *Let $\deg_z p = L > 0$. If $w \in \mathcal{C}^\infty(\mathbb{R})$ satisfies $\widetilde{p}(w) = 0$ and $w|_{[k,k+L]} = 0$ for some $k \in \mathbb{Z}$, then $w = 0$.*

The result of part (1) can be found in [6, p. 697]. Since [6] uses rather difficult methods to also prove surjectivity for other (more general) operators, we present a complete and elementary proof of both parts of the proposition in the appendix. Of course, the surjectivity of $\widetilde{p}$ is well known if $p \in \mathbb{R}[s]$.

3. Properties of the ring $\mathcal{H}$. Two facts about the ring $\mathcal{H}$ will be important for what follows. One is that the division structure of $\mathcal{H}$ corresponds to the division properties of the associated entire functions in the full ring $H(\mathbb{C})$. This is made precise in Proposition 3.1(5). The other main fact about $\mathcal{H}$ is its advantageous ring structure. In Theorem 3.2 we will show that $\mathcal{H}$ is a *Bézout ring*, i.e., that every finitely generated ideal is principal. Stated in other words, finitely many elements $p_1, \dots, p_r \in \mathcal{H}$ have a greatest common divisor $d \in \mathcal{H}$ which fulfills a Bézout equation $d = \sum_{i=1}^r a_i p_i$ over $\mathcal{H}$. Furthermore, with Lemma 3.4 it will be proven that $\mathcal{H}$ is an *elementary divisor ring*, which means that matrices over $\mathcal{H}$ can be brought into diagonal form via multiplication with unimodular matrices from both sides. This is a very useful fact in order to handle the matrix case of delay-differential equations. One should note that both properties hold true also for the ring $H(\mathbb{C})$ (see, e.g., [17, Thm. 5, p. 136 and Thm. 8, p. 141]), but not for $\mathcal{R}$.

PROPOSITION 3.1.

(1) *If $p \in \mathcal{H}$ and $\alpha \in \mathbb{C}$, then $p^*(\bar{\alpha}) = \overline{p^*(\alpha)}$, where $^-$ denotes complex conjugation.*

(2) *Define $\mathcal{H}^\times := \{p \in \mathcal{H} \,|\, p \text{ is a unit}\}$. Then $\mathcal{H}^\times = \{az^k \,|\, a \in \mathbb{R}\backslash\{0\},\, k \in \mathbb{Z}\} = \{p \in \mathcal{H} \,|\, \mathcal{V}(p^*) = \emptyset\}$.*

(3) *$\mathcal{H}$ is not a unique factorization domain and not a Noetherian ring.*

(4) *For $p \in \mathcal{H}$ the following statements are equivalent:* (i) *p is irreducible,* (ii) *$p = \phi z^k$ for some irreducible $\phi \in \mathbb{R}[s]$ and $k \in \mathbb{Z}$, and* (iii) *p is prime.*

(5) *Let $p, q \in \mathcal{H}$. Then $p^* \,|\, q^*$ in $H(\mathbb{C}) \Longleftrightarrow p \,|\, q$ in $\mathcal{H}$.*

(6) *For $p, q \in \mathcal{H}$, not both zero, there exists a greatest common divisor (gcd) $d \in \mathcal{H}\backslash\{0\}$ of p, q which is unique up to multiplication by units in $\mathcal{H}$. Moreover, $\mathcal{V}(d^*) = \mathcal{V}(p^*, q^*)$. In particular, p and q are coprime in $\mathcal{H}$ iff $\mathcal{V}(p^*, q^*) = \emptyset$.*

(7) *Let $p = ad$, $q = bd \in \mathcal{H}\backslash\{0\}$, with d being a gcd of p, q and with $a, b \in \mathcal{H}$. Then $c := abd \in \mathcal{H}$ is a least common multiple (lcm) of p, q. An lcm is unique up to multiplication by units in $\mathcal{H}$.*

Proof. (1) This is obvious.

(2) Let $p \in \mathcal{H}^\times$. Then p is also a unit in $\mathbb{R}(s)[z, z^{-1}]$. Thus $p = az^k$ for some $a \in \mathbb{R}(s)$ and $k \in \mathbb{Z}$. Since $p^*(s) = a(s)e^{-ks}$ and $(p^{-1})^*(s) = a(s)^{-1}e^{ks}$ are both entire functions, it follows that $a \in \mathbb{R}\backslash\{0\}$. The last equality holds with Proposition 2.6.

(3) Consider $z - 1 \in \mathcal{H}$. Let $(\alpha_i)_{i\in\mathbb{N}} \subset \mathbb{C}\backslash\{0\}$ so that $e^{-\alpha_i} - 1 = 0$, $\alpha_i \neq \alpha_j$ for $i \neq j$ and $\alpha_{2i+1} = \overline{\alpha_{2i}}$ for $i \in \mathbb{N}$. Then $p_i := (s - \alpha_{2i})(s - \alpha_{2i+1}) \in \mathbb{R}[s]$ satisfies $z - 1 = \frac{z-1}{p_i} p_i = \frac{z-1}{\Pi_{i=1}^n p_i} \prod_{i=1}^n p_i$, and these are factorizations of $z - 1$ in $\mathcal{H}$. Moreover, the chain

$$\tfrac{z-1}{p_1}\mathcal{H} \subseteq \tfrac{z-1}{p_1 p_2}\mathcal{H} \subseteq \tfrac{z-1}{p_1 p_2 p_3}\mathcal{H} \subseteq \cdots$$

of ideals in $\mathcal{H}$ will not become stationary.

(4) "(i) $\Rightarrow$ (ii)" Let $p \in \mathcal{H}$ be irreducible. According to part (2) there exists $\alpha \in \mathbb{C}$ with $p^*(\alpha) = 0$. If $\alpha \in \mathbb{R}$, then $p = \frac{p}{s-\alpha}(s-\alpha)$ is a factorization in $\mathcal{H}$; thus $\frac{p}{s-\alpha}$ has to be a unit in $\mathcal{H}$. By (2) this yields $p = az^k(s-\alpha)$ for some nonzero $a \in \mathbb{R}$ and $k \in \mathbb{Z}$, which gives (ii). If $\alpha \notin \mathbb{R}$, then with part (1) one gets analogously $p = az^k(s-\alpha)(s-\bar{\alpha})$.

"(ii) $\Rightarrow$ (iii)" Let $\phi \in \mathbb{R}[s]$ be irreducible. Then ϕ is prime in $\mathbb{R}[s]$ and of the form $\phi = s-\alpha$ or $\phi = (s-\alpha)(s-\bar{\alpha})$. Suppose $p = \phi z^k$ and $p \,|\, fg$ in $\mathcal{H}$ for some $f, g \in \mathcal{H}$. Then $(fg)^* p^{*^{-1}} = (f^* g^*) p^{*^{-1}} \in H(\mathbb{C})$ and both cases for ϕ imply by use of (1): $p \,|\, f$ or $p \,|\, g$.

"(iii) $\Rightarrow$ (i)" holds true in every commutative domain.

(5) The direction "$\Leftarrow$" holds since $p \mapsto p^*$ is a ring homomorphism. "$\Rightarrow$" Let $q^*(p^*)^{-1} \in H(\mathbb{C})$. In the field $\mathbb{R}(s,z)$ we can write $qp^{-1} = ab^{-1}$ with coprime $a, b \in \mathbb{R}[s,z]$. The theorem of Bézout for algebraic curves implies

$$\#\{(\lambda,\mu) \in \mathbb{C}^2 \,|\, a(\lambda,\mu) = 0 = b(\lambda,\mu)\} < \infty.$$

Since $a^*(b^*)^{-1} = q^*(p^*)^{-1} \in H(\mathbb{C})$ yields $\mathcal{V}(b^*) \subseteq \mathcal{V}(a^*)$, we get $\#\mathcal{V}(b^*) < \infty$. By use of Proposition 2.6, this leads to $b = \phi z^k$ for some $\phi \in \mathbb{R}[s]\backslash\{0\}$ and $k \in \mathbb{Z}$. Hence $qp^{-1} = az^{-k}\phi^{-1} \in \mathcal{H}$.

(6) Since $\mathcal{H} \subset \mathbb{R}(s)[z,z^{-1}]$, there exists a gcd $d \in \mathbb{R}(s)[z,z^{-1}]$ of p, q. Thus $p = fd$, $q = gd$ with coprime f, $g \in \mathbb{R}(s)[z,z^{-1}]$.

In order to derive from this suitable factorizations in $\mathcal{H}$, we shall shift the poles of f^* or g^* and the common zeros of f^* and g^* within multiplicities into the factor d. To do so, let

$$\mathcal{P} = \{\alpha \in \mathbb{C} | \mu_\alpha(f^*) < 0 \text{ or } \mu_\alpha(g^*) < 0\}$$

be the set of poles of f or g. Then we have $\#\mathcal{P} < \infty$ as well as $\#\mathcal{V}(f^*, g^*) < \infty$ and $\mathcal{P} \cap \mathcal{V}(f^*, g^*) = \emptyset$. Put

$$\phi := \prod_{\alpha \in \mathcal{P}} (s-\alpha)^{\max\{-\mu_\alpha(f^*), -\mu_\alpha(g^*)\}} \in \mathbb{R}[s],$$
$$\psi := \prod_{\alpha \in \mathcal{V}(f^*,g^*)} (s-\alpha)^{\min\{\mu_\alpha(f^*), \mu_\alpha(g^*)\}} \in \mathbb{R}[s].$$

This leads to

$$(3.1)\quad p = \frac{f\phi}{\psi}\frac{\psi}{\phi}d,\ q = \frac{g\phi}{\psi}\frac{\psi}{\phi}d \ \text{ where } \frac{f\phi}{\psi},\ \frac{g\phi}{\psi} \in \mathcal{H} \ \text{ and } \mathcal{V}\left(\left(\frac{f\phi}{\psi}\right)^*, \left(\frac{g\phi}{\psi}\right)^*\right) = \emptyset.$$

Moreover, $\frac{\psi}{\phi}d \in \mathcal{H}$, for if $\alpha \in \mathbb{C}$ was a pole of $(\frac{\psi}{\phi}d)^*$, then it would follow that $\alpha \in \mathcal{V}((\frac{f\phi}{\psi})^*, (\frac{g\phi}{\psi})^*)$ since $p^*, q^* \in H(\mathbb{C})$. Hence we have a factorization $p = f'd'$, $q = g'd'$ in $\mathcal{H}$, and $\mathcal{V}((f')^*, (g')^*) = \emptyset$ implies that $(d')^*$ is a gcd of p^*, q^* in $H(\mathbb{C})$.

To show that d' is a gcd of p, q in $\mathcal{H}$, let $p = f''d''$, $q = g''d''$ with f'', g'', $d'' \in \mathcal{H}$. Then $p^* = (f'')^*(d'')^*$, $q^* = (g'')^*(d'')^*$, and thus $(d'')^* \,|\, (d')^*$ in $H(\mathbb{C})$. By part (5) this yields $ad'' = d'$ for some $a \in \mathcal{H}$ and therefore d' is a gcd of p, q in $\mathcal{H}$. This argument also implies the uniqueness property claimed for a gcd in $\mathcal{H}$.

The equality $\mathcal{V}(d^*) = \mathcal{V}(p^*, q^*)$ follows from (3.1), and the last claim of part (6) is an easy consequence of (2).

(7) Obviously $p \,|\, c$ and $q \,|\, c$ in $\mathcal{H}$. Let $c' \in \mathcal{H}$ be another common multiple of p and q; i.e., let there exist v, $w \in \mathcal{H}$ with $adv = c' = bdw$. Therefore $av = bw$

and $a^*v^* = b^*w^*$ in $H(\mathbb{C})$. This yields $w^* = (a^*v^*)(b^*)^{-1} \in H(\mathbb{C})$, and moreover $v^*(b^*)^{-1} \in H(\mathbb{C})$, since by (6) a^* and b^* have no common zeros. From (5) we get the existence of $b' \in \mathcal{H}$ with $bb' = v$ and thus $c' = adbb' = cb'$. □

Now we can prove the following.

THEOREM 3.2. *$\mathcal{H}$ is a* Bézout ring*, i.e., every finitely generated ideal is a principal ideal.*

Proof. We need to show that for $p, q \in \mathcal{H}$ and a gcd $d \in \mathcal{H}$ of p, q there exist $a, b \in \mathcal{H}$ so that $d = ap+bq$, for this implies $p\mathcal{H}+q\mathcal{H} = d\mathcal{H}$. Without loss of generality we can assume $d = 1$; hence by Proposition 3.1(6) we see that $\mathcal{V}(p^*, q^*) = \emptyset$.

Step 1. The elements p, q are coprime also in $\mathbb{R}(s)[z, z^{-1}]$. To see this, let $uv = p$, $uw = q$ with $u, v, w \in \mathbb{R}(s)[z, z^{-1}]$; then let $u = \widetilde{u}\phi^{-1}$, $v = \widetilde{v}\phi^{-1}, p = \tilde{p}\phi^{-1}, q = \tilde{q}\phi^{-1}$ with $\widetilde{u}, \widetilde{v}, \widetilde{p}, \widetilde{q} \in \mathcal{R}$, $\phi \in \mathbb{R}[s]$. Then $\widetilde{u}\widetilde{v} = p\phi$ and $\deg_z \widetilde{u} \geq 1$ would imply that all irreducible factors u_i of $\widetilde{u}$ with $\deg_z u_i \geq 1$ divide $\widetilde{p}$ in $\mathcal{R}$. Similarly $u_i \mid \widetilde{q}$ in $\mathcal{R}$ and thus u would be a nontrivial common factor of p, q in $\mathcal{H}$, which contradicts the coprimeness of p, q in $\mathcal{H}$. Thus $u \in \mathbb{R}(s)$ and is therefore a unit in $\mathbb{R}(s)[z, z^{-1}]$.

Hence there exists a Bézout equation in $\mathbb{R}(s)[z, z^{-1}]$; i.e.,

$$1 = ap + bq \text{ with suitable } a, b \in \mathbb{R}(s)[z, z^{-1}]. \tag{3.2}$$

Step 2. Next we will vary the coefficients a, b of (3.2) in such a way that we get a Bézout equation for p and q with coefficients in $\mathcal{H}$. More precisely, we will construct a rational function $v \in \mathbb{R}(s)$ so that

$$b + vp,\ a - vq \in \mathcal{H}. \tag{3.3}$$

Then (3.2) will imply the Bézout equation $1 = (a - vq)p + (b + vp)q$ in $\mathcal{H}$.

Step 2a. In order to achieve (3.3) we have to get rid of the poles of a^* and b^*. Therefore, write

$$a = \frac{\widetilde{a}}{\psi},\ b = \frac{\widetilde{b}}{\phi} \text{ with } \widetilde{a}, \widetilde{b} \in \mathcal{H},\ \psi, \phi \in \mathbb{R}[s] \text{ and } \mathcal{V}(\widetilde{a}^*, \psi) = \mathcal{V}(\widetilde{b}^*, \phi) = \emptyset. \tag{3.4}$$

Let $h \in \mathbb{R}[s]$ be a gcd of ψ, ϕ and $\psi = h\psi_1$, $\phi = h\phi_1$ with ψ_1, $\phi_1 \in \mathbb{R}[s]$. Then (3.2) becomes

$$h\psi_1\phi_1 = \phi_1\widetilde{a}p + \psi_1\widetilde{b}q, \tag{3.5}$$

where all elements are in $\mathcal{H}$. From $\psi_1(h\phi_1 - \widetilde{b}q) = \phi_1\widetilde{a}p$ and $\mathcal{V}(\psi_1, \phi_1) = \emptyset = \mathcal{V}(\widetilde{a}^*, \psi_1)$ it results with Proposition 3.1(5) $\psi_1 \mid p$ in $\mathcal{H}$. So let $p = p_1\psi_1$ with $p_1 \in \mathcal{H}$. Similarly, it is $q = q_1\phi_1$ with $q_1 \in \mathcal{H}$. Thus after cancellation of $\psi_1\phi_1$, (3.5) reads

$$h = \widetilde{a}p_1 + \widetilde{b}q_1. \tag{3.6}$$

Step 2b. Put $v = \frac{f}{h\psi_1\phi_1} \in \mathbb{R}(s)$, where $f \in \mathbb{R}[s]$ still has to be specified. Then (3.3) implies that we have to find $f \in \mathbb{R}[s]$ such that

$$\begin{cases} (b+vp)^* &= \big(\dfrac{\widetilde{b}}{h\phi_1} + \dfrac{f}{h\phi_1\psi_1}p_1\psi_1\big)^* = \dfrac{(\widetilde{b}+fp_1)^*}{h\phi_1} \in H(\mathbb{C}), \\ (a-vq)^* &= \big(\frac{\widetilde{a}}{h\psi_1} - \frac{f}{h\phi_1\psi_1}q_1\phi_1\big)^* = \frac{(\widetilde{a}-fq_1)^*}{h\psi_1} \in H(\mathbb{C}). \end{cases} \tag{3.7}$$

Hence we have to look for a polynomial $f \in \mathbb{R}[s]$ which places the zeros of $\widetilde{b}^* + fp_1^*$ and $\widetilde{a}^* - fq_1^*$ appropriately at the same time. In the rest of the proof we will show

that these are two interpolation problems for f which can in fact be solved with the same polynomial $f \in \mathbb{R}[s]$.

First, for $\alpha \in \mathcal{V}(\phi_1 h)$ one has $p_1^*(\alpha) \neq 0$, since (i) If $\alpha \in \mathcal{V}(\phi_1) \subset \mathcal{V}(q^*)$, then $\alpha \notin \mathcal{V}(p^*)$; hence $\alpha \notin \mathcal{V}(p_1^*)$. (ii) If $h(\alpha) = 0$, then by (3.6) and (3.4) it follows that $0 = \widetilde{a}^*(\alpha)p_1^*(\alpha) + \widetilde{b}^*(\alpha)q_1^*(\alpha)$ and $\widetilde{a}^*(\alpha) \neq 0 \neq \widetilde{b}^*(\alpha)$. Therefore, $\mathcal{V}(p^*, q^*) = \emptyset$ yields $p_1^*(\alpha) \neq 0 \neq q_1^*(\alpha)$.

For $\alpha \in \mathcal{V}(\phi_1 h)$ this leads to

$$\begin{aligned}
\mu_\alpha(\widetilde{b}^* + fp_1^*) \geq k &\iff (\widetilde{b}^* + fp_1^*)^{(\nu)}(\alpha) = 0,\ \nu = 0, \dots, k-1 \\
&\iff \widetilde{b}^{*^{(\nu)}}(\alpha) + \sum_{\mu=0}^{\nu} \tbinom{\nu}{\mu} f^{(\mu)}(\alpha) p_1^{*^{(\nu-\mu)}}(\alpha) = 0,\ \nu = 0, \dots, k-1 \\
&\iff f^{(\nu)}(\alpha) = -\frac{1}{p_1^*(\alpha)} \left[\widetilde{b}^{*^{(\nu)}}(\alpha) + \sum_{\mu=0}^{\nu-1} \tbinom{\nu}{\mu} p_1^{*^{(\nu-\mu)}}(\alpha) f^{(\mu)}(\alpha) \right] \\
&\qquad \text{for } \nu = 0, \dots, k-1.
\end{aligned}$$

A similar result holds for $\alpha \in \mathcal{V}(\psi_1 h)$. As a consequence $f \in \mathbb{R}[s]$ satisfies (3.7) iff

$$f^{(\nu)}(\alpha) = \begin{cases}
-\dfrac{1}{p_1^*(\alpha)} \left[\widetilde{b}^{*^{(\nu)}}(\alpha) + \displaystyle\sum_{\mu=0}^{\nu-1} \tbinom{\nu}{\mu} p_1^{*^{(\nu-\mu)}}(\alpha) f^{(\mu)}(\alpha) \right] & \\
\text{for } \nu = 0, \dots, \mu_\alpha(\phi_1 h) - 1 \qquad \text{if } \alpha \in \mathcal{V}(\phi_1 h) & \\
\dfrac{1}{q_1^*(\alpha)} \left[\widetilde{a}^{*^{(\nu)}}(\alpha) - \displaystyle\sum_{\mu=0}^{\nu-1} \tbinom{\nu}{\mu} q_1^{*^{(\nu-\mu)}}(\alpha) f^{(\mu)}(\alpha) \right] & \\
\text{for } \nu = 0, \dots, \mu_\alpha(\psi_1 h) - 1 \qquad \text{if } \alpha \in \mathcal{V}(\psi_1 h). &
\end{cases} \tag{3.8}$$

In particular, for $\alpha \in \mathcal{V}(\phi_1 h) \cap \mathcal{V}(\psi_1 h) = \mathcal{V}(h)$ and $\nu = 0, \dots, \mu_\alpha(h) - 1$ the derivative $f^{(\nu)}(\alpha)$ has to be equal to both expressions given in (3.8). Thus we can find such an f only if for those α and ν it is true that

$$\begin{aligned}
&-\frac{1}{p_1^*(\alpha)} \left[\widetilde{b}^{*^{(\nu)}}(\alpha) + \sum_{\mu=0}^{\nu-1} \tbinom{\nu}{\mu} p_1^{*^{(\nu-\mu)}}(\alpha) f^{(\mu)}(\alpha) \right] \\
&\qquad = \frac{1}{q_1^*(\alpha)} \left[\widetilde{a}^{*^{(\nu)}}(\alpha) - \sum_{\mu=0}^{\nu-1} \tbinom{\nu}{\mu} q_1^{*^{(\nu-\mu)}}(\alpha) f^{(\mu)}(\alpha) \right].
\end{aligned}$$

But this is indeed valid, since from (3.6) it follows that

$$\begin{aligned}
0 = h^{(\nu)}(\alpha) &= (\widetilde{a}^* p_1^* + \widetilde{b}^* q_1^*)^{(\nu)}(\alpha) \\
&= \sum_{\mu=0}^{\nu} \tbinom{\nu}{\mu} \widetilde{a}^{*^{(\mu)}}(\alpha) p_1^{*^{(\nu-\mu)}}(\alpha) + \sum_{\mu=0}^{\nu} \tbinom{\nu}{\mu} \widetilde{b}^{*^{(\mu)}}(\alpha) q_1^{*^{(\nu-\mu)}}(\alpha)
\end{aligned}$$

for $\nu = 0, \dots, \mu_\alpha(h) - 1$ and therefore one can apply Lemma A.2.

As $\mathcal{V}(\phi_1 \psi_1 h) \subseteq \mathbb{C}$ is symmetric with respect to complex conjugation, Proposition 3.1(1) and Lemma A.1 imply the existence of $f \in \mathbb{R}[s]$ with the properties required in (3.8). ∎

Example 3.3. Let $p = s^2$, $q = z - 1 \in \mathcal{H}$. Then $s \,|\, q^*$ but $s^2 \not| \, q^*$ in $H(\mathbb{C})$, thus $d = s$ is a gcd of p, q. A Bézout equation is given by

$$s = \frac{(1-s)z + 2s - 1}{s^2} s^2 + (s-1)(z-1).$$

Note also that $\ker \widetilde{p} = \{w \in \mathcal{C}^\infty(\mathbb{R}) \,|\, \exists\, \alpha, \beta \in \mathbb{R} \text{ for all } t \in \mathbb{R} : w(t) = \alpha + \beta t\}$ and $\ker \widetilde{q} = \{w \in \mathcal{C}^\infty(\mathbb{R}) \,|\, w \text{ is of period } 1\}$; hence $\ker \widetilde{p} \cap \ker \widetilde{q} = \{w \in \mathcal{C}^\infty(\mathbb{R}) \,|\, w \text{ constant}\} = \ker \widetilde{d}$.

It is a standing conjecture that every commutative Bézout domain is an *elementary divisor domain*, which means that matrices can be brought into diagonal form via left–right equivalence; see, e.g., [3, p. 92]. In the present case, one can in fact prove the elementary divisor property. To do so, we will show the following lemma, which states that $\mathcal{H}$ is a so-called *adequate ring*; see, e.g., [12, p. 473].

LEMMA 3.4. *Let $p, q \in \mathcal{H}$, $p \neq 0$. There exists a factorization $p = ab$ with $a, b \in \mathcal{H}$ such that a and q are coprime, whereas $\hat{b}$ and q are not coprime whenever $\hat{b} \in \mathcal{H} \backslash \mathcal{H}^\times$ is a divisor of b.*

Proof. The idea for the factorization is as follows: factorize $p = ab$ such that $\mathcal{V}(b^*) = \mathcal{V}(p^*, q^*)$ and $\mu_\lambda(b^*) = \mu_\lambda(p^*)$ for all $\lambda \in \mathcal{V}(b^*)$. This can easily be done if $\#\mathcal{V}(p^*, q^*) < \infty$. In the infinite case it needs an iterative procedure as described below.

Let $b_1 \in \mathcal{H}$ be a gcd of p and q and put $a_1 = \frac{p}{b_1}$ so that $p = a_1 b_1$. Define successively the following elements:

$$\text{let } c_i \in \mathcal{H} \text{ be a gcd of } a_i \text{ and } b_i; \text{ define } a_{i+1} = \frac{a_i}{c_i} \text{ and } b_{i+1} = c_i b_i. \tag{3.9}$$

Hence $p = a_i b_i = a_{i+1} c_i b_i = a_{i+1} b_{i+1}$. This gives a sequence of elements $a_i \in \mathcal{H}$ with the property that a_{i+1} divides a_i in $\mathcal{H}$. But then a_{i+1} divides a_i also in the principal ideal ring $\mathbb{R}(s)[z, z^{-1}]$ with the consequence that for some $k \in \mathbb{N}$ there exist $l \in \mathbb{Z}$ and $\phi \in \mathbb{R}[s] \backslash \{0\}$ such that $c_k = \phi z^l$ is a unit in $\mathbb{R}(s)[z, z^{-1}]$. Thus the procedure (3.9) yields the existence of a factorization:

$$p = a_k b_k \text{ with } \phi \in \mathbb{R}[s] \text{ as a gcd of } a_k \text{ and } b_k \text{ in } \mathcal{H}.$$

This implies that $\mathcal{V}(a_k^*, b_k^*)$ is finite, say $\mathcal{V}(a_k^*, b_k^*) = \{\lambda_1, \ldots, \lambda_n\}$, and we can define

$$f := \prod_{i=1}^{n} (s - \lambda_i)^{l_i} \in \mathbb{R}[s], \text{ where } l_i = \mu_{\lambda_i}(a_k^*).$$

With $a := a_k f^{-1} \in \mathcal{H}$ and $b := f b_k \in \mathcal{H}$ we get the factorization $p = ab$, which in fact satisfies the requirements of the lemma: (1) To establish the coprimeness of a and q, suppose $\mathcal{V}(a^*, q^*) \neq \emptyset$. Thus let $\lambda \in \mathcal{V}(a^*, q^*) \subseteq \mathcal{V}(p^*, q^*) = \mathcal{V}(b_1^*)$. Then $\lambda \in \mathcal{V}(b_1^*, a_k^*) \subseteq \mathcal{V}(a_k^*, b_k^*) = \{\lambda_1, \ldots, \lambda_n\}$. But for $\lambda = \lambda_j$ it is $\mu_\lambda(a^*) = \mu_{\lambda_j}(a_k^*) - \mu_{\lambda_j}(f) = 0$. Hence $\mathcal{V}(a^*, q^*) = \emptyset$ and from Proposition 3.1(6) we conclude the coprimeness of a and q.

(2) Let $\hat{b} \in \mathcal{H} \backslash \mathcal{H}^\times$ be a divisor of b and fix some $\lambda \in \mathcal{V}(b^*)$ with $\hat{b}^*(\lambda) = 0$. The construction (3.9) of the sequences (c_i) and (b_i) leads to the following equality of zero sets (note that we count zeros in $\mathcal{V}$ not with multiplicity):

$$\mathcal{V}(b^*) = \mathcal{V}(f^* b_k^*) = \mathcal{V}(b_k^*) = \mathcal{V}(c_{k-1}^* b_{k-1}^*) = \mathcal{V}(b_{k-1}^*) = \cdots = \mathcal{V}(b_1^*) = \mathcal{V}(p^*, q^*).$$

Thus $\lambda \in \mathcal{V}(q^*, \hat{b}^*)$ and therefore $\hat{b}$ and q are not coprime.

Note that when $\mathcal{V}(p^*, q^*) = \{\lambda_1, \ldots, \lambda_n\}$ is finite, the above construction leads to the factorization $p = \frac{p}{b} b$ with $b = \prod_{i=1}^{n} (s - \lambda_i)^{l_i}$ and $l_i = \mu_{\lambda_i}(p^*)$. □

Now we can summarize the properties for matrices over $\mathcal{H}$ as they follow from the above ring theory results.

THEOREM 3.5.

(1) *Let $a_1, \ldots, a_n \in \mathcal{H}$ and $d \in \mathcal{H}$ be a gcd of $a_1, \ldots, a_n$. Then there exists a matrix $A \in \mathcal{H}^{n \times n}$ with $[a_1, \ldots, a_n]$ as its first row and* $\det A = d$.

(2) *For $P \in \mathcal{H}^{n \times m}$ there exists $U \in Gl_n(\mathcal{H})$ so that $UP \in \mathcal{H}^{n \times m}$ has upper triangular form.*

(3) *Let $P \in \mathcal{H}^{n \times m}$ and $Q \in \mathcal{H}^{l \times m}$. There exists a greatest common right divisor (gcrd) $D \in \mathcal{H}^{m \times m}$ of P and Q and matrices $A \in \mathcal{H}^{m \times n}$, $B \in \mathcal{H}^{m \times l}$ with $D = AP + BQ$. If* $\mathrm{rk} D = m$, *then D is unique modulo multiplication from the left by unimodular matrices.*

(4) *Let $P, Q \in \mathcal{H}^{m \times m}$ with* $\mathrm{rk} P = \mathrm{rk} Q = m$. *Then there exists a least common left multiple (lclm) $M \in \mathcal{H}^{m \times m}$ of P and Q which is uniquely determined up to unimodular factors from the left.*

(5) *$\mathcal{H}$ is an* elementary divisor ring*; that is, for $P \in \mathcal{H}^{n \times m}$ with* $\mathrm{rk} P = r$ *there exist $U \in Gl_n(\mathcal{H})$ and $V \in Gl_m(\mathcal{H})$ such that*

$$UPV = \begin{bmatrix} P_1 & 0 \\ 0 & 0 \end{bmatrix} \in \mathcal{H}^{n \times m} \quad \text{with} \quad P_1 = \begin{bmatrix} p_1 & 0 & \cdots & 0 \\ 0 & \ddots & & \vdots \\ \vdots & & \ddots & 0 \\ 0 & \cdots & 0 & p_r \end{bmatrix} \in \mathcal{H}^{r \times r}, \tag{3.10}$$

where $p_i \neq 0$ for all i and $p_i \,|\, p_{i+1}$ for $i = 1, \ldots, r-1$.

Proof. Parts (1)–(4) hold in general for matrices over commutative Bézout domains. The proof of these parts is identical with that given for principal ideal domains in [13, pp. 31–36]. Part (5) follows from Lemma 3.4, as shown in [12, p. 473] for arbitrary adequate rings. □

The existence of an lcm for elements $p, q \in \mathcal{H}$ as we proved in Proposition 3.1(7) can also be concluded from part (1) of the above theorem (see, e.g., [4, Cor. 2, p. 126]).

4. Correspondence between behaviors and ideals in $\mathcal{H}$. The results in section 3 enable us to show a correspondence between the lattice of behaviors associated with delay-differential equations of the type (1.1) and the lattice of finitely generated ideals in $\mathcal{H}$. After introducing multivariable delay-differential operators, an analogous version of this correspondence will be shown also in that case.

Remember that, as outlined in Definition 2.7 and Remark 2.8, for $p \in \mathcal{H}$ the operator $\widetilde{p} \in \mathrm{End}_{\mathbb{R}}(\mathcal{C}^\infty(\mathbb{R}))$ exists. In particular, for $p \in \mathbb{R}[s, z] \subset \mathcal{H}$ this includes the classical case as in equation (1.1).

PROPOSITION 4.1. *For $p, q \in \mathcal{H} \backslash \{0\}$ let $d \in \mathcal{H}$ be a gcd of p, q, and $c \in \mathcal{H}$ be an lcm of p, q. Then*

(1) $\ker \widetilde{p} \subseteq \ker \widetilde{q} \iff p \,|\, q$;

(2) $\ker \widetilde{d} = \ker \widetilde{p} \cap \ker \widetilde{q}$;

(3) $\ker \widetilde{c} = \ker \widetilde{p} + \ker \widetilde{q}$;

(4) *if $d \in \mathcal{H}^\times$, then* $\ker \widetilde{p} + \ker \widetilde{q} = \ker \widetilde{pq} = \ker \widetilde{qp}$;

(5) *let $a \in \mathcal{H}$ be such that* $\ker \widetilde{p} \cap \ker \widetilde{q} \subseteq \ker \widetilde{a}$. *Then $a \in p\mathcal{H} + q\mathcal{H}$.*

Proof. (1)"⇒" Let $p = a\phi^{-1}$, $q = b\phi^{-1}$ with $a, b \in \mathcal{R}$ and $\phi \in \mathbb{R}[s]$. Then it is easy to see that $\ker \widetilde{p} \subseteq \ker \widetilde{q}$ implies $\ker \widetilde{a} \subseteq \ker \widetilde{b}$. Thus by (2.2) one has $b^*(a^*)^{-1} = q^*(p^*)^{-1} \in H(\mathbb{C})$ and with Proposition 3.1(5) it follows that $p \,|\, q$ in $\mathcal{H}$.

"$\Leftarrow$" If $q = ap$ with some $a \in \mathcal{H}$, then Remark 2.8(3) yields $\widetilde{q} = \widetilde{a} \circ \widetilde{p}$, and therefore $\ker \widetilde{p} \subseteq \ker \widetilde{q}$.

(2) This is a consequence of (1) and the existence of a Bézout equation $d = ap + bq$ in $\mathcal{H}$ together with Remark 2.8(3).

(3) "$\supseteq$" follows from (1).

"$\subseteq$" Let $p = ad$, $q = bd$ with $a, b \in \mathcal{H}$. Then, by Proposition 3.1(7) we can take $c = abd$ as an lcm of p, q. By coprimeness of a, b there exists $f, g \in \mathcal{H}$ with $1 = af + bg$. Hence $w \in \ker \widetilde{c}$ satisfies $w = \widetilde{fa}(w) + \widetilde{gb}(w) \in \ker \widetilde{q} + \ker \widetilde{p}$.

(4) If $d \in \mathcal{H}^\times$, then pq is an lcm of p, q; hence the claim holds by (3).

(5) This follows from (1) and (2) and the equality $d\mathcal{H} = p\mathcal{H} + q\mathcal{H}$. □

Notice that the Examples 2.9 and 3.3 correspond to the situation given in (1) and (2) of the above proposition.

Now we will come to the multivariable case. From Remark 2.8 we conclude that for a matrix $P = (p_{ij}) \in \mathcal{H}^{n\times m}$ the operator

$$\begin{aligned} \widetilde{P} : \quad & \mathcal{C}^\infty(\mathbb{R}^m) \longrightarrow \mathcal{C}^\infty(\mathbb{R}^n), \\ & (w_1, \dots, w_m)^{\mathtt{t}} \longmapsto \left(\sum_{j=1}^m \widetilde{p_{1j}}(w_j), \dots, \sum_{j=1}^m \widetilde{p_{nj}}(w_j) \right)^{\mathtt{t}} \end{aligned}$$

is well defined. Thus the behavior, defined by a system of delay-differential equations, can be described as $\ker \widetilde{P}$ with some $P \in \mathcal{R}^{n\times m}$, or in the more general case $P \in \mathcal{H}^{n\times m}$.

Remark 4.2.

(1) The map

$$\begin{aligned} \mathcal{H}^{n\times m} &\longrightarrow \mathrm{Hom}_{\mathbb{R}}(\mathcal{C}^\infty(\mathbb{R}^m), \mathcal{C}^\infty(\mathbb{R}^n)), \\ P &\longmapsto \widetilde{P} \end{aligned}$$

is $\mathbb{R}$ linear and injective and satisfies $\widetilde{PQ} = \widetilde{P} \circ \widetilde{Q}$ for $P \in \mathcal{H}^{n\times m}$, $Q \in \mathcal{H}^{m\times l}$.

(2) Analogous to the scalar case in Definition 2.4(1) the map

$$\begin{aligned} \mathcal{H}^{n\times m} &\longrightarrow H(\mathbb{C})^{n\times m}, \\ P &\longmapsto P^*(s) := P(s, e^{-s}) \end{aligned}$$

is a well-defined $\mathbb{R}$-linear map and fulfills $(PQ)^*(s) = P^*(s)Q^*(s)$ for $P \in \mathcal{H}^{n\times m}$, $Q \in \mathcal{H}^{m\times l}$.

Let us first list some properties of the operator $\widetilde{P}$.

PROPOSITION 4.3. *Let $P \in \mathcal{H}^{n\times m}$. Then*

(1) *if $n = m$ and $P \in Gl_n(\mathcal{H})$, then $\widetilde{P}$ is bijective and $P^*(s) \in Gl_n(\mathbb{C})$ for all $s \in \mathbb{C}$;*

(2) *$\widetilde{P}$ is surjective iff* $\mathrm{rk}P = n$;

(3) *the following properties are equivalent:*

(i) *$\widetilde{P}$ is injective,*

(ii) $\mathrm{rk}P^*(s) = m$ *for all $s \in \mathbb{C}$,*

(iii) *there exists $Q \in \mathcal{H}^{m\times n}$ with $QP = I_m$.*

Proof. (1) This follows from the existence of $Q \in \mathcal{H}^{n\times n}$ with $PQ = QP = I_n$ together with Remark 4.2.

(2) Let $\mathrm{rk}P = r \leq n$. By Theorem 3.5(5) there exist $U \in Gl_n(\mathcal{H})$ and $V \in Gl_m(\mathcal{H})$ so that UPV is as in (3.10). By (1) $\widetilde{P}$ is surjective iff $\widetilde{UPV}$ is surjective, and together with Proposition 2.10 this holds iff $r = n$.

(3) All three properties are invariant under multiplication with unimodular matrices from the left or from the right. Thus, using again Theorem 3.5(5), we can restrict ourselves to diagonal P. Since all three properties imply $\mathrm{rk}P = m$, we can assume

$$P = \begin{bmatrix} P_1 \\ 0 \end{bmatrix} \in \mathcal{H}^{n\times m} \text{ with } P_1 = \mathrm{diag}(p_1, \ldots, p_m) \in \mathcal{H}^{m\times m}.$$

Now (i) implies the injectivity of $\widetilde{p_i}$; thus, with (2.2) and Proposition 3.1(2), $p_i \in \mathcal{H}^\times$. This yields (ii). In the same way, (ii) leads to $p_i \in \mathcal{H}^\times$ for all i, and (iii) can be concluded. The implication "(iii) $\Rightarrow$ (i)" follows from Remark 4.2(1). □

Now we can generalize part of the results in Proposition 4.1 to the multivariable case.

PROPOSITION 4.4. *Let $P \in \mathcal{H}^{n\times m}$, $Q \in \mathcal{H}^{l\times m}$, and $D \in \mathcal{H}^{m\times m}$ be a gcrd of P, Q. Then*

(1) $\ker \widetilde{P} \cap \ker \widetilde{Q} = \ker \widetilde{D}$;

(2) *P is a right divisor of Q iff* $\ker \widetilde{P} \subseteq \ker \widetilde{Q}$;

(3) *under the condition* $\mathrm{rk}P = n$, $\mathrm{rk}Q = l$ *the following holds true:* $\ker \widetilde{P} = \ker \widetilde{Q}$ *iff $n = l$ and $P = UQ$ for some $U \in Gl_n(\mathcal{H})$.*

Proof. (1) Since "$\Rightarrow$" of (2) holds by Remark 4.2(1), part (1) follows from the existence of a Bézout equation for D (see Theorem 3.5(3)).

(2) It remains to prove "$\Leftarrow$." Let $r = \mathrm{rk}P$ and $U \in Gl_n(\mathcal{H})$, $V \in Gl_m(\mathcal{H})$ be such that $P' = UPV$ is as in (3.10). Denoting $Q' = UQV$, Proposition 4.3(1) implies $\ker \widetilde{P'} \subseteq \ker \widetilde{Q'}$. This yields $Q' = [R, 0]$ with $R \in \mathcal{H}^{l\times r}$, and, moreover, $\ker \widetilde{p_j} \subseteq \ker \widetilde{R_{ij}}$ for all $j = 1, \ldots, r$ and $i = 1, \ldots, l$. Hence, using Proposition 4.1(1) we get the existence of $A \in \mathcal{H}^{l\times n}$ such that $AP' = Q'$ and therefore $U^{-1}AUP = Q$.

(3) "$\Leftarrow$" is obvious.

"$\Rightarrow$" By (2) there exist $P = UQ$ and $Q = VP$ for some $U \in \mathcal{H}^{n\times l}$, $V \in \mathcal{H}^{l\times n}$. Then the full rank assumption implies $VU = I_l$ and $UV = I_n$, which leads to the desired result. □

5. Controllability. In this section we will generalize the well-known Hautus criterion for controllability to delay-differential systems. For time-delay state-space systems this criterion characterizes spectral controllability as it is known from, e.g., [18] and [2]. In the behavioral context this criterion is established for finite-dimensional discrete- or continuous-time AR-systems (see, e.g., [22, Prop. 4.3]) and, very recently, in [19] for exactly the same situation of delay-differential equations as presented in the paper at hand. However, the proof in [19] uses quite different methods than those developed in this paper.

Whereas controllability for state-space systems is formulated, of course, in terms of control functions and state trajectories, we do not have this possibility for behaviors. Hence we will use the notion of controllability as defined in [22]. For this we have to introduce first the concatenation of two functions.

DEFINITION 5.1. *Let $-\infty \le a_1 < a_2 \le b_1 < b_2 \le \infty$, and let $w_1 \colon (a_1, b_1) \to \mathbb{R}^m$ and $w_2 : [a_2, b_2) \to \mathbb{R}^m$ be two functions. For $t_0 \in [a_2, b_1]$ denote by $w_1 \wedge_{t_0} w_2 : (a_1, b_2) \to \mathbb{R}^m$ the following* concatenation of w_1 and w_2 at t_0*:*

$$(w_1 \wedge_{t_0} w_2)(t) := \begin{cases} w_1(t) & \text{for } a_1 < t < t_0, \\ w_2(t) & \text{for } t_0 \le t < b_2. \end{cases}$$

Using this definition, a behavior is called controllable if it is closed under concatenation in the sense given below. In [22, p. 186] one can find the system theory justification of this notion.

DEFINITION 5.2. *Let $\mathcal{B}$ be a shift-invariant subspace of $\mathcal{C}^\infty(\mathbb{R}^m)$. Then $\mathcal{B}$ is called* controllable *if it satisfies the following: for all w, $w' \in \mathcal{B}$ there exists $t_0 \geq 0$ and $c \in \mathcal{C}^\infty([0,t_0),\mathbb{R}^m)$ with $w\wedge_0 c\wedge_{t_0}\sigma^{t_0}w' \in \mathcal{B}$.*

The requirement $w\wedge_0 c\wedge_{t_0}\sigma^{t_0}w' \in \mathcal{B}$ yields in particular that the concatenation is in $\mathcal{C}^\infty(\mathbb{R}^m)$.

Note that $\mathcal{C}^\infty(\mathbb{R}^m)$ is controllable; even more, $\mathcal{C}^\infty(\mathbb{R}^m)$ is controllable in arbitrary short time: for all w, $w' \in \mathcal{C}^\infty(\mathbb{R}^m)$ and all $t_0 > 0$ there exists $c \in \mathcal{C}^\infty([0,t_0),\mathbb{R}^m)$ with $w\wedge_0 c\wedge_{t_0}\sigma^{t_0}w' \in \mathcal{C}^\infty(\mathbb{R}^m)$.

Since we introduce the concept of controllability only for shift-invariant subspaces, it makes sense to consider only controllability at time zero.

Whereas it is obvious that for $U \in \mathbb{R}[s]^{n\times m}$ and w, $w' \in \mathcal{C}^\infty(\mathbb{R}^m)$ it is $\widetilde{U}(w\wedge_0 w') = \widetilde{U}(w)\wedge_0\widetilde{U}(w')$ if $w\wedge_0 w'$ is sufficiently differentiable at $t_0 = 0$, it is a priori not clear that $\widetilde{U}(w\wedge_0 w')$ is a sort of concatenation of $\widetilde{U}(w)$ and $\widetilde{U}(w')$ if $U \in \mathbb{R}[s,z]^{n\times m}$ or even $U \in \mathcal{H}^{n\times m}$.

LEMMA 5.3. *Let $U = \sum_{j=0}^L U_j z^j \in \mathbb{R}[s,z]^{n\times m}$ with $U_j \in \mathbb{R}[s]^{n\times m}$. Further, let w, $w' \in \mathcal{C}^\infty(\mathbb{R}^m)$, $t_0 \in \mathbb{R}$ with $w\wedge_{t_0} w' \in \mathcal{C}^\infty(\mathbb{R}^m)$. Then there exists $c \in \mathcal{C}^\infty([t_0,t_0+L),\mathbb{R}^n)$ so that $\widetilde{U}(w\wedge_{t_0}w') = \widetilde{U}(w)\wedge_{t_0}c\wedge_{t_0+L}\widetilde{U}(w')$.*

Proof. A direct calculation shows

$$\begin{aligned}\widetilde{U}(w\wedge_{t_0}w')(t) &= \sum_{j=0}^L \widetilde{U}_j(w\wedge_{t_0}w')(t-j) = \sum_{j=0}^L (\widetilde{U}_j(w)\wedge_{t_0}\widetilde{U}_j(w'))(t-j)\\ &= \begin{cases} \sum_{j=0}^L \widetilde{U}_j(w')(t-j) = \widetilde{U}(w')(t) & \text{if } t \geq t_0+L,\\ c(t) & \text{if } t_0 \leq t < t_0+L,\\ \sum_{j=0}^L \widetilde{U}_j(w)(t-j) = \widetilde{U}(w)(t) & \text{if } t < t_0 \end{cases}\end{aligned}$$

for some function $c : [t_0, t_0+L) \to \mathbb{R}^n$. Hence $\widetilde{U}(w\wedge_{t_0}w') = \widetilde{U}(w)\wedge_{t_0}c\wedge_{t_0+L}\widetilde{U}(w')$. Since $\widetilde{U}(w\wedge_{t_0}w') \in \mathcal{C}^\infty(\mathbb{R}^n)$, we also get $c \in \mathcal{C}^\infty([t_0,t_0+L),\mathbb{R}^n)$. □

With this knowledge we can prove the following.

LEMMA 5.4. *Let $\mathcal{B}$ be a shift-invariant linear controllable subspace of $\mathcal{C}^\infty(\mathbb{R}^m)$ and let $U \in \mathcal{H}^{n\times m}$. Then $\widetilde{U}(\mathcal{B})$ is a shift-invariant linear controllable subspace of $\mathcal{C}^\infty(\mathbb{R}^n)$.*

Proof. Since $\mathcal{B}$ is shift invariant, it is enough to consider $U = \sum_{j=0}^L U_j z^j \in \mathbb{R}(s)[z]^{n\times m}$ with $U_j \in \mathbb{R}(s)^{n\times m}$.

Let w, $w' \in \mathcal{B}$. Then $\sigma^L w' \in \mathcal{B}$ and there exist $t_0 \geq 0$ and $c \in \mathcal{C}^\infty([0,t_0),\mathbb{R}^m)$ so that $\bar{w} := w\wedge_0 c\wedge_{t_0}\sigma^{t_0+L}w' \in \mathcal{B}$.

Case 1. Let $U_j \in \mathbb{R}[s]^{n\times m}$ for all j, thus $U \in \mathbb{R}[s,z]^{n\times m}$. Then by Lemma 5.3 we get the existence of $c' \in \mathcal{C}^\infty([0,t_0+L),\mathbb{R}^n)$, so that

$$\begin{aligned}\widetilde{U}(\bar{w}) &= \widetilde{U}(w\wedge_0 c\wedge_{t_0}\sigma^{t_0+L}w')\\ &= \widetilde{U}(w)\wedge_0 c'\wedge_{t_0+L}\widetilde{U}(\sigma^{t_0+L}w')\\ &= \widetilde{U}(w)\wedge_0 c'\wedge_{t_0+L}\sigma^{t_0+L}\widetilde{U}(w') \in \widetilde{U}(\mathcal{B}).\end{aligned}$$

Since w, $w' \in \mathcal{B}$ were arbitrary, this yields the controllability of $\widetilde{U}(\mathcal{B})$.

Case 2. Let $U_j = V_j\phi^{-1}$ with $V_j \in \mathbb{R}[s]^{n\times m}$. Put $V = \sum_{j=0}^L V_j z^j \in \mathbb{R}[s,z]^{n\times m}$. Then $U = V\phi^{-1}$ and by definition $\widetilde{U}(\bar{w}) = \widetilde{V}(v)$, if $v \in \mathcal{C}^\infty(\mathbb{R}^m)$ fulfills $\widetilde{\phi}(v) = \bar{w}$.

As in the first case, we shall show that $\widetilde{U}(\bar{w})$ is a concatenation of $\widetilde{U}(w)$ and $\sigma^{t_0+L}\widetilde{U}(w')$ so that $\widetilde{U}(\bar{w}) \in \widetilde{U}(\mathcal{B})$ implies the controllability of $\widetilde{U}(\mathcal{B})$. In order to do

so, we will construct a solution of $\widetilde{\phi}(v) = \bar{w}$ which corresponds to the special form of $\bar{w} = w \wedge_0 c \wedge_{t_0} \sigma^{t_0+L} w'$. For this let $c' \in \mathcal{C}^\infty([0, t_0), \mathbb{R}^m)$ be so that $\widetilde{\phi}(c') = c$. Then the solutions $v_i \in \mathcal{C}^\infty(\mathbb{R}^m)$, $i = 1, 2$, of

$$\begin{aligned} &\widetilde{\phi}(v_1) = w, && v_1^{(\nu)}(0) = c'^{(\nu)}(0) \text{ for } \nu = 0, \dots, \deg \phi - 1, \\ &\widetilde{\phi}(v_2) = \sigma^{t_0+L} w', && v_2^{(\nu)}(t_0) = c'^{(\nu)}(t_0) \text{ for } \nu = 0, \dots, \deg \phi - 1 \end{aligned}$$

satisfy $v := v_1 \wedge_0 c' \wedge_{t_0} v_2 \in \mathcal{C}^\infty(\mathbb{R}^m)$ and $\widetilde{\phi}(v) = \bar{w}$. Moreover, $\widetilde{V}(v_1) = \widetilde{U}(w)$, $\widetilde{V}(v_2) = \widetilde{U}(\sigma^{t_0+L} w')$. Now, by the first case of this proof there exists $c'' \in \mathcal{C}^\infty([0, t_0 + L), \mathbb{R}^n)$ so that

$$\begin{aligned} \widetilde{U}(\bar{w}) = \widetilde{V}(v) &= \widetilde{V}(v_1 \wedge_0 c' \wedge_{t_0} v_2) = \widetilde{V}(v_1) \wedge_0 c'' \wedge_{t_0+L} \widetilde{V}(v_2) \\ &= \widetilde{U}(w) \wedge_0 c'' \wedge_{t_0+L} \widetilde{U}(\sigma^{t_0+L} w') \\ &= \widetilde{U}(w) \wedge_0 c'' \wedge_{t_0+L} \sigma^{t_0+L} \widetilde{U}(w') \in \widetilde{U}(\mathcal{B}). \qquad \square \end{aligned}$$

Now we can prove the main part of this section

THEOREM 5.5. *Let $P \in \mathcal{H}^{n \times m}$. Then* $\ker \widetilde{P}$ *is controllable iff* $\mathrm{rk} P^*(s) = \mathrm{rk} P$ *for all $s \in \mathbb{C}$.*

Proof. (a) We first prove the scalar case $p \in \mathcal{H}$. If $p = 0$ then obviously $\ker \widetilde{p} = \mathcal{C}^\infty(\mathbb{R})$ is controllable. Let $p \neq 0$.

"$\Leftarrow$" This holds, since $\ker \widetilde{p} = \{0\}$ if $p \in \mathcal{H}^\times$.

"$\Rightarrow$" Let $w_1 \in \ker \widetilde{p}$. Then there exist $t_0 > 0$ and some $c \in \mathcal{C}^\infty([0, t_0), \mathbb{R})$ with $v := w_1 \wedge_0 c \wedge_{t_0} 0 \in \ker \widetilde{p}$ and Proposition 2.10(2) implies $v = 0$; hence, again by Proposition 2.10(2), $w_1 = 0$. Therefore controllability of $\ker \widetilde{p}$ implies $\ker \widetilde{p} = \{0\}$ and from Proposition 3.1(2) it follows that $p \in \mathcal{H}^\times$.

(b) Let $P \in \mathcal{H}^{n \times m}$. Using Theorem 3.5(5) and Lemma 5.4 we can restrict ourselves to the case of P being as in (3.10).

"$\Leftarrow$" The assumption on the rank implies that $p_j \in \mathcal{H}^\times$ for $j = 1, \dots, r$, and therefore $\ker \widetilde{P} = \{(0, \dots, 0, w_{r+1}, \dots, w_m)^\mathrm{t} \mid w_i \in \mathcal{C}^\infty(\mathbb{R}),\ i = r+1, \dots, m\}$, which is indeed controllable.

"$\Rightarrow$" The controllability of $\ker \widetilde{P}$ yields the controllability of $\ker \widetilde{p_j}$ for $j = 1, \dots, r$. Hence by the scalar case $p_j \in \mathcal{H}^\times$ and the desired conclusion follows. $\square$

Conclusions. As can be seen from sections 4 and 5, the ring $\mathcal{H}$ seems to be the adequate object for an algebraic treatment of delay-differential equations as (1.1) and (1.2). Once the algebraic properties of $\mathcal{H}$ are established, the translation into properties of the solution spaces are nearly straightforward.

In a forthcoming paper it will be shown how the existence of image representations for the systems under investigation can be characterized with the help of this algebraic framework. Moreover, the analytical meaning of the operators in $\mathcal{H}$ has to be clarified.

Appendix. We start with the following.

Proof of Proposition 2.10.

(1) Let $p \in \mathcal{H} \backslash \{0\}$ and $v \in \mathcal{C}^\infty(\mathbb{R})$. We have to find $w \in \mathcal{C}^\infty(\mathbb{R})$ fulfilling $\widetilde{p}(w) = v$.

First, it suffices to assume $p \in \mathcal{R}$ for let $p = q\phi^{-1}$ with $q \in \mathcal{R}$, $\phi \in \mathbb{R}[s]$. If we find $f \in \mathcal{C}^\infty(\mathbb{R})$ with $\widetilde{q}(f) = v$ and put $\widetilde{\phi}(f) = w$, then we have $\widetilde{p}(w) = v$. Hence we need to show the surjectivity of $\widetilde{q}$.

Thus let $p \in \mathcal{R}$ and, more precisely,

$$p = \sum_{j=0}^{L} p_j z^j \in \mathbb{R}[s,z] \text{ with } p_j \in \mathbb{R}[s] \text{ and } L \geq 1.$$

Put $p_0 = \sum_{i=0}^{l} a_i s^i$, $a_l = 1$, and $p_L = \sum_{i=0}^{r} b_i s^i$, $b_r \neq 0$.

We will construct piecewise a function $w \in \mathcal{C}^\infty(\mathbb{R})$ which fulfills for all $t \in \mathbb{R}$

$$\widetilde{p}(w)(t) = \sum_{j=0}^{L} \widetilde{p_j}(w)(t-j) = v(t). \tag{A.1}$$

The idea of the construction is as follows: start with a function $w_0 \in \mathcal{C}^\infty[0,L]$. In order to extend w_0 via concatenation (see Definition 5.1) to a solution of $\widetilde{p}(w) = v$ one has to solve successively ordinary inhomogeneous differential equations of the type

$$\widetilde{p_0}(\bar{w}_{k+1}) = v - \widetilde{(p-p_0)}(w_k) \text{ on the time interval } [L+k, L+k+1] \text{ for } k \geq 0,$$
$$\widetilde{p_L}(\bar{w}_k) = \sigma^{-L}\Big(v - \widetilde{(p-p_L)}(w_{k+1})\Big) \text{ on the time interval } [k, k+1] \text{ for } k \leq -1,$$

where the right-hand sides are determined successively by

$$w_k = w_0 \wedge_L \bar{w}_1 \wedge_{L+1} \ldots \wedge_{L+k-1} \bar{w}_k \text{ on } [0, L+k] \text{ for } k \geq 1,$$
$$w_{k+1} = \bar{w}_{k+1} \wedge_{k+2} \ldots \wedge_{-1} \bar{w}_{-1} \wedge_0 w_0 \text{ on } [k+1, L] \text{ for } k < -1.$$

The initial conditions at the points $L+k$ (for $k \geq 0$) and $k+1$ (for $k \leq -1$) have, of course, to be prescribed such that the concatenations are as smooth as possible. If one chooses the initial function $w_0 \in \mathcal{C}^\infty[0,L]$ appropriately, this procedure leads indeed to infinitely smooth concatenations and thus to a solution $w \in \mathcal{C}^\infty(\mathbb{R})$ of $\widetilde{p}(w) = v$.

The choice of the function w_0 is carried out in step (i) of the following elaboration. Steps (ii) and (iii) give the details of the extension of w_0 to a solution of $\widetilde{p}(w) = v$ on the positive real line, whereas step (iv) extends w_0 for negative time.

(i) Let $f \in \mathcal{C}^\infty[0,L]$ satisfy

$$\sum_{i=0}^{l} a_i f^{(i)}(t) = v(t), \qquad t \in [0,L],\ f^{(\nu)}(L) = 0 \text{ for } \nu = 0, \ldots, l-1.$$

(In the case $l = 0$, one has no freedom for the initial conditions. In this case the rest of the proof in (ii) and (iii) works analogously.) Let $g \in \mathcal{C}^\infty[0,L]$ be such that $g|_{[0,L-1]} = 0$ and $g|_{[L-0.5,L]} = 1$. Put $w_0 := fg \in \mathcal{C}^\infty[0,L]$. Then $w_0|_{[0,L-1]} = 0$ and

$$w_0^{(\nu)}(L) = f^{(\nu)}(L) = \begin{cases} 0 & \text{for } \nu = 0, \ldots, l-1, \\ v^{(\nu-l)}(L) - \sum_{i=0}^{l-1} a_i w_0^{(\nu-l+i)}(L) & \text{for } \nu \geq l. \end{cases}$$

(ii) Let $\bar{w}_1 \in \mathcal{C}^\infty[L, L+1]$ fulfill

$$\sum_{i=0}^{l} a_i \bar{w}_1^{(i)}(t) = v(t) - \sum_{j=1}^{L} \widetilde{p_j}(w_0)(t-j),\ t \in [L, L+1],$$

with initial conditions $\bar{w}_1^{(\nu)}(L) = 0$ for $\nu = 0, \ldots, l-1$. By differentiation one checks that $\bar{w}_1^{(\nu)}(L) = w_0^{(\nu)}(L)$ for all $\nu \in \mathbb{N}_0$ and thus $w_1 := w_0 \wedge_L \bar{w}_1 \in \mathcal{C}^\infty[0, L+1]$ fulfills $\sum_{j=0}^{L} \widetilde{p_j}(w_1)(t-j) = v(t)$ for $t \in [L, L+1]$.

(iii) Inductively, if $w_k \in \mathcal{C}^\infty[0, L+k]$ satisfies

$$\sum_{i=0}^{l} a_i w_k^{(i)}(t) = v(t) - \sum_{j=1}^{L} \widetilde{p_j}(w_k)(t-j)$$

for $t \in [L, L+k]$, then take the solution $\bar{w}_{k+1} \in \mathcal{C}^\infty[L+k, L+k+1]$ of the ODE

$$\sum_{i=0}^{l} a_i \bar{w}_{k+1}^{(i)}(t) = v(t) - \sum_{j=1}^{L} \widetilde{p_j}(w_k)(t-j),\ t \in [L+k, L+k+1],$$

with initial conditions $\bar{w}_{k+1}^{(\nu)}(L+k) = w_k^{(\nu)}(L+k)$ for $\nu = 0, \dots, l-1$. From this we obtain again by differentiation $\bar{w}_{k+1}^{(\nu)}(L+k) = w_k^{(\nu)}(L+k)$ for all $\nu \in \mathbb{N}_0$, and therefore we get a solution $w_{k+1} := w_k \wedge_{L+k} \bar{w}_{k+1} \in \mathcal{C}^\infty[0, L+k+1]$. Hence we can construct a function $w_+ \in \mathcal{C}^\infty[0, \infty)$ which satisfies (A.1) for $t \geq L$.

(iv) Let $\bar{w}_{-1} \in \mathcal{C}^\infty[-1, 0]$ satisfy

$$\sum_{i=0}^{r} b_i \bar{w}_{-1}^{(i)}(t) = v(t+L) - \sum_{j=0}^{L-1} \widetilde{p_j}(w_+)(t+L-j),\ t \in [-1, 0],$$

with the initial conditions $\bar{w}_{-1}^{(\nu)}(0) = 0$ for $\nu = 0, \dots, r-1$. Then $\bar{w}_{-1}^{(\nu)}(0) = 0$ for all $\nu \in \mathbb{N}_0$ and the function $w_{-1} := \bar{w}_{-1} \wedge_0 w_+ \in \mathcal{C}^\infty[-1, \infty)$ satisfies (A.1) for $t \geq L-1$. In a way analogous to (iii) we can proceed inductively and finally find a solution $w \in \mathcal{C}^\infty(\mathbb{R})$ for $\widetilde{p}(w) = v$.

(2) Put $p = q\phi^{-1}$ with $q \in \mathcal{R}$ and $\phi \in \mathbb{R}[s]$ and let $w \in \mathcal{C}^\infty(\mathbb{R})$ be given as in Proposition 2.10(2). It is easy to see that there exists $v \in \mathcal{C}^\infty(\mathbb{R})$ with $\widetilde{\phi}(v) = w$ and $v|_{[k, k+L]} = 0$. But then $0 = \widetilde{p}(w) = \widetilde{q}(v)$ and the proof of (1) shows by proceeding step by step on the intervals $[j, j+1]$ that $v = 0$ and thus $w = 0$. ∎

The following two lemmas are used in the proof of Theorem 3.2. The first one states the interpolation property for polynomials: given a finite set of points in the complex plane, there exists a polynomial $f \in \mathbb{C}[s]$, such that a specified number of derivatives $f^{(\nu)}$ take prescribed values at those points. If the required situation is symmetric with respect to complex conjugation, one can find a real interpolation polynomial.

LEMMA A.1. *Let $\alpha_1, \dots, \alpha_r \in \mathbb{C}\backslash\mathbb{R}$, $\alpha_{r+1}, \dots, \alpha_{r+t} \in \mathbb{R}$, $k_1, \dots, k_{r+t} \in \mathbb{N}_0$, $c_{j\nu} \in \mathbb{C}$ for $j = 1, \dots, r$ and $\nu = 0, \dots, k_j$, and $c_{j\nu} \in \mathbb{R}$ for $j = r+1, \dots, r+t$ and $\nu = 0, \dots, k_j$. Then there exists a unique $f \in \mathbb{R}[s]$ satisfying*

$$\begin{aligned}
&\deg f \leq N := 2\sum_{j=1}^{r}(k_j+1) + \sum_{j=r+1}^{r+t}(k_j+1) - 1,\\
&f^{(\nu)}(\alpha_j) = c_{j\nu} \text{ for } j = 1, \dots, r+t,\ \nu = 0, \dots, k_j,\\
&f^{(\nu)}(\overline{\alpha_j}) = \overline{c_{j\nu}} \text{ for } j = 1, \dots, r,\ \nu = 0, \dots, k_j.
\end{aligned}$$

Proof. The existence and uniqueness of $f \in \mathbb{C}[s]$ with the desired properties can be found, e.g., in [5, p. 37]. But this already implies $f \in \mathbb{R}[s]$, since with $f = \sum_{j=0}^{N} f_j s^j \in \mathbb{C}[s]$, $\bar{f} = \sum_{j=0}^{N} \bar{f}_j s^j$ also fulfills the above requirements. ∎

The second lemma is just a rather specific calculation. It is used to show that the interpolation requirements given in (3.8) can be satisfied by one polynomial, $f \in \mathbb{R}[s]$.

LEMMA A.2. *Let $K \in \mathbb{N}_0$ and a_j, b_j, p_j, $q_j \in \mathbb{C}$ for $j = 0, \dots, K$. Let $p_0 \neq 0 \neq q_0$ and*

$$\sum_{m=0}^{n} \binom{n}{m} b_m q_{n-m} = -\sum_{m=0}^{n} \binom{n}{m} a_m p_{n-m} \text{ for } n = 0, \dots, K. \tag{A.2}$$

Put $f_n := q_0^{-1}[a_n - \sum_{m=0}^{n-1} \binom{n}{m} q_{n-m} f_m]$ for $n = 0, \dots, K$. Then the recursion $f_n = -p_0^{-1}[b_n + \sum_{m=0}^{n-1} \binom{n}{m} p_{n-m} f_m]$ is also valid for $n = 0, \dots, K$.

Proof. For $n = 0$ it is $b_0 q_0 = -a_0 p_0$; hence $f_0 = \frac{a_0}{q_0} = -\frac{b_0}{p_0}$.

Suppose the claim holds true for $f_0, \dots, f_n$, $n < K$. Then one calculates

$$\begin{aligned}
q_0 f_{n+1} &= a_{n+1} - \sum_{m=0}^{n} \tbinom{n+1}{m} q_{n+1-m} f_m \\
&= a_{n+1} + \sum_{m=0}^{n} \tbinom{n+1}{m} q_{n+1-m} \left[\frac{1}{p_0} \left(b_m + \sum_{k=0}^{m-1} \tbinom{m}{k} p_{m-k} f_k \right) \right] \\
&= a_{n+1} + \frac{1}{p_0} \left[\sum_{m=0}^{n} \tbinom{n+1}{m} q_{n+1-m} b_m \right. \\
&\qquad \left. + \sum_{m=1}^{n} \sum_{k=0}^{m-1} \tbinom{n+1}{m} \tbinom{m}{k} q_{n+1-m} p_{m-k} f_k \right] \\
&= \frac{1}{p_0} \left[-b_{n+1} q_0 - \sum_{m=0}^{n} \tbinom{n+1}{m} a_m p_{n+1-m} \right. \\
&\qquad \left. + \sum_{k=0}^{n-1} \sum_{m=k+1}^{n} \tbinom{n+1}{m} \tbinom{m}{k} q_{n+1-m} p_{m-k} f_k \right] \\
&= -\frac{1}{p_0} \left[b_{n+1} q_0 + \sum_{m=0}^{n} \tbinom{n+1}{m} a_m p_{n+1-m} \right. \\
&\qquad \left. - \sum_{m=1}^{n} \sum_{k=0}^{m-1} \tbinom{n+1}{m} \tbinom{m}{k} p_{n+1-m} q_{m-k} f_k \right] \\
&= -\frac{1}{p_0} \left[b_{n+1} q_0 + q_0 \sum_{m=0}^{n} \tbinom{n+1}{m} p_{n+1-m} \frac{1}{q_0} \left(a_m - \sum_{k=0}^{m-1} \tbinom{m}{k} q_{m-k} f_k \right) \right] \\
&= -\frac{q_0}{p_0} \left[b_{n+1} + \sum_{m=0}^{n} \tbinom{n+1}{m} p_{n+1-m} f_m \right],
\end{aligned}$$

where the fourth equation follows from (A.2) and the fifth one from

$$\begin{aligned}
\sum_{m=k+1}^{n} \tbinom{n+1}{m} \tbinom{m}{k} q_{n+1-m} p_{m-k} &= \sum_{l=k+1}^{n} \tbinom{n+1}{n+1+k-l} \tbinom{n+1+k-l}{k} q_{l-k} p_{n+1-l} \\
&= \sum_{l=k+1}^{n} \tbinom{n+1}{l} \tbinom{l}{k} q_{l-k} p_{n+1-l}. \qquad \square
\end{aligned}$$

REFERENCES

[1] R. Bellman and K. L. Cooke, *Differential-Difference Equations*, Academic Press, New York, 1963.

[2] K. P. M. Bhat and H. N. Koivo, *Modal characterizations of controllability and observability in time delay systems*, IEEE Trans. Automat. Control, AC-21 (1976), pp. 292–293.

[3] J. W. Brewer, J. W. Bunce, and F. S. V. Vleck, *Linear Systems Over Commutative Rings*, Lecture Notes in Pure and Applied Mathematics 104, Marcel Dekker, New York, 1986.

[4] P. M. Cohn, *Free Rings and Their Relations*, Academic Press, London, 1971.

[5] P. J. Davis, *Interpolation and Approximation*, Dover, New York, 1975.

[6] L. Ehrenpreis, *Solutions of some problems of division. Part* III. *Division in the spaces* $\mathcal{D}'$, $\mathcal{H}$, $\mathcal{Q}_A$, $\mathcal{O}$, Amer. J. Math., 78 (1956), pp. 685–715.

[7] L. E. El'sgol'ts and S. B. Norkin, *Introduction to the Theory and Application of Differential Equations With Deviating Arguments*, Academic Press, New York, 1973.

[8] L. C. G. J. M. Habets, *Algebraic and Computational Aspects of Time-Delay Systems*, Ph.D. thesis, Department of Mathematics and Computing Science, Eindhoven University of Technology, Eindhoven, the Netherlands, 1994.

[9] A. S. B. Holland, *Introduction to the Theory of Entire Functions*, Academic Press, New York, 1973.

[10] E. W. Kamen, *On an algebraic theory of systems defined by convolution operators*, Math. Systems Theory, 9 (1975), pp. 57–74.

[11] E. W. Kamen, P. P. Khargonekar, and A. Tannenbaum, *Proper stable Bezout factorizations and feedback control of linear time-delay systems*, Internat. J. Control, 43 (1986), pp. 837–857.

[12] I. Kaplansky, *Elementary divisors and modules*, Trans. Amer. Math. Soc., 66 (1949), pp. 464–491.

[13] C. C. MacDuffee, *The Theory of Matrices*, Chelsea, New York, 1946.

[14] B. Malgrange, *Existence et approximation des solutions des équations aux dérivées partielles et des équations des convolution*, Ann. Inst. Fourier (Grenoble), 6 (1955/1956), pp. 271–355.

[15] A. Manitius and R. Triggiani, *Function space controllability of linear retarded systems: A derivation from abstract operator conditions*, SIAM J. Control Optim., 16 (1978), pp. 599–645.

[16] A. S. Morse, *Ring models for delay-differential systems*, Automatica, 12 (1976), pp. 529–531.

[17] R. Narasimhan, *Complex Analysis in One Variable*, Birkhäuser Boston, Cambridge, MA, 1985.

[18] L. Pandolfi, *On the infinite dimensional controllability of differential-difference control processes*, Boll. Un. Mat. Ital., 10 (1974), pp. 114–123.

[19] P. Rocha and J. C. Willems, *Behavioral controllability of delay-differential systems*, SIAM J. Control Optim., 35 (1997), pp. 254–264.

[20] H. Soethoudt, *Introduction to a Behavioral Approach for Continuous-Time Systems*, Ph.D. thesis, Department of Mathematics and Computing Science, Eindhoven University of Technology, Eindhoven, the Netherlands, 1993.

[21] E. D. Sontag, *Linear systems over commutative rings: A survey*, Ricerche di Automatica, 7 (1976), pp. 1–34.

[22] J. C. Willems, *Models for dynamics*, Dynam. Report., 2 (1989), pp. 171–269.

SIAM J. CONTROL OPTIM.
Vol. 35, No. 2, pp. 500–511, March 1997

007

FEEDBACK STABILIZATION OF AFFINE IN THE CONTROL STOCHASTIC DIFFERENTIAL SYSTEMS BY THE CONTROL LYAPUNOV FUNCTION METHOD*

PATRICK FLORCHINGER†

Abstract. The purpose of this paper is to study the asymptotic stability in probability of affine in the control stochastic differential systems. Sufficient conditions for the existence of control Lyapunov functions leading to the existence of stabilizing feedback laws which are smooth, except possibly at the equilibrium point of the system, are provided.

Key words. stochastic stability, control stochastic differential equation, control Lyapunov function, feedback law

AMS subject classifications. 60H10, 93C10, 93D05, 93D15, 93E15

PII. S0363012995279961

Introduction. The stabilization of nonlinear stochastic differential systems by means of state feedback laws is an important problem in control theory. The stochastic version of the Lyapunov theorem has been used to derive necessary and sufficient conditions for stabilization of stochastic differential systems at their equilibrium state. The stabilizability of various types of nonlinear stochastic differential systems has been studied for different notions of stochastic stability in the last past years (see, for instance, [8], [3], [4], [5], [7], or [6]).

The procedure used by Gao and Ahmed [8] relies on the stochastic Lyapunov theory and on the properties of the solution of a stochastic algebraic Riccati equation introduced by Wonham [16].

In [3], [4], and [6] the necessary and sufficient conditions for the asymptotic feedback stability in probability of the stochastic differential systems at their equilibrium state are of Lyapunov type, the stabilizers computed in these papers are smooth except possibly at the equilibrium state, and their construction is based on the knowledge of an appropriate control Lyapunov function. The aim of this paper is to study the asymptotic stability in probability for a wider class of affine in the control nonlinear stochastic differential systems than the one considered in [6]. This class of stochastic differential systems can be characterized in terms of *computable* control Lyapunov functions which depend on the system coefficients.

Note that a wide class of stochastic bilinear differential systems (like those used to model biological processes) as well as the equation of the angular velocity of a rigid body corrupted by noise (see [13] for the deterministic case) can be stabilized by using the results proved in the following. The main tools used in this paper are the stochastic Lyapunov theorem proved by Khasminskii [9] and the converse stability theorems of Kushner [10], Khasminskii [9], and Wilson [15]. In this paper, we extend some results proved by Tsinias [14] for deterministic control systems to control stochastic differential systems driven by a Wiener process. The analysis used in this paper is

*Received by the editors January 18, 1995; accepted for publication (in revised form) January 24, 1996. A version of this paper was presented at the 35th IEEE Conference on Decision and Control, Kobe, Japan, December 11–13, 1996.

http://www.siam.org/journals/sicon/35-2/27996.html

†Unité de Recherche Associée, Centre National de la Recherche Scientifique No. 399, Département de Mathématiques, Unité de Formation et de Recherche Mathématique Informatique Mécanique, Université de Metz, Ile du Saulcy, F 57045 Metz Cedex, France (florchin@ilm.loria.fr).

closely related to that of [14], taking into account that one needs differentiability of higher order than the one needed in the deterministic case. Furthermore, in order to use converse stochastic Lyapunov theorems, more restrictive assumptions on the system coefficients than those stated in [14] have to be assumed. This paper is divided in three sections organized as follows. In section 1, we introduce the class of affine in the control stochastic differential systems that we are dealing with in this paper, and we recall the stochastic version of Artstein's theorem proved in [6]. In section 2, we state and prove the main results of the paper on the existence of control Lyapunov functions (for the class of systems introduced in section 1) leading to the existence of stabilizing feedback laws which are smooth, except possibly at the equilibrium points of the systems. In section 3, we provide some numerical examples, for which the results proved in the previous sections allow, to compute stabilizing feedback laws. For a brief review of the Lyapunov machinery that we need in what follows to study the stochastic stability of the equilibrium solution of a stochastic differential equation, we refer the reader to [6], and for a detailed exposition of the stochastic stability theory we refer the reader to Khasminskii [9] or Arnold [1], for example.

1. Problem statement. The purpose of this section is to introduce the class of affine in the control stochastic differential systems that we are dealing with in this paper and to recall the stochastic version of Artstein's theorem proved in [6]. Denote by $(\Omega, \mathcal{F}, P)$ a complete probability space and by $w = \{w_t, t \in \mathbb{R}_+\}$ a standard $\mathbb{R}^m$-valued Wiener process defined on this space. Consider the multi-input stochastic differential system in $\mathbb{R}^n$ written in the Itô form,

$$x_t = x_0 + \int_0^t (f(x_s) + h(x_s)u)ds + \int_0^t g(x_s)dw_s, \tag{1}$$

where

1. x_0 is given in $\mathbb{R}^n$;
2. u is an $\mathbb{R}^p$-valued control law;
3. f, g, and h are C^∞ functionals mapping $\mathbb{R}^n$ into $\mathbb{R}^n$, $\mathbb{R}^{n\times m}$, and $\mathbb{R}^{n\times p}$, respectively, vanishing in the origin and such that there exists a nonnegative constant K such that for any $x \in \mathbb{R}^n$,

$$|f(x)| + |g(x)| + |h(x)| \leq K(1 + |x|).$$

The stochastic differential system (1) is said to be *asymptotically stabilizable in probability* (at the origin) if there exist a neighborhood D of the origin in $\mathbb{R}^n$ and a function k mapping D in $\mathbb{R}^p$, vanishing in the origin, such that

1. for every $x \in D$, the solution x_t of the closed-loop system

$$x_t = x + \int_0^t (f(x_s) + h(x_s)k(x_s))ds + \int_0^t g(x_s)dw_s \tag{2}$$

is uniquely defined;

2. the equilibrium solution $x_t \equiv 0$ of the resulting closed-loop system (2) is asymptotically stable in probability.

Denoting by L the infinitesimal generator of the stochastic process solution of the uncontrolled part of the stochastic differential system (1), that is, L is the second-order differential operator defined for any function Ψ in $C^2(\mathbb{R}^n; \mathbb{R})$ by

$$L\Psi(x) = \sum_{i=1}^n f_i(x)\frac{\partial \Psi}{\partial x_i}(x) + \frac{1}{2}\sum_{i,j=1}^n a_{ij}(x)\frac{\partial^2 \Psi}{\partial x_i \partial x_j}(x),$$

where $a_{ij}(x) = \sum_{k=1}^m g_k^i(x) g_k^j(x)$, $1 \le i, j \le n$, one can introduce the notion of *control Lyapunov function* as follows.

DEFINITION 1.1. *The stochastic differential system* (1) *satisfies a* stochastic Lyapunov condition *at the origin if there exist a neighborhood D of the origin in $\mathbb{R}^n$ and a Lyapunov function V defined on D such that for every $x \in D \setminus \{0\}$ the following condition holds:*

$$\sum_{j=0}^n h_i^j(x) \frac{\partial V}{\partial x_j}(x) = 0,\ i = 1, ..., p \ \Rightarrow\ LV(x) < 0.$$

A Lyapunov function V satisfying the above condition is called a control Lyapunov function for the stochastic differential system (1).

The control Lyapunov function V defined above is said to satisfy the *bounded control property* if there exists a positive real function d mapping D in $\mathbb{R}$ such that d is bounded on D and for every $x \in D \setminus \{0\}$ there exists a control $u \in \mathbb{R}^p$ such that

$$||u|| < d(x) \tag{3}$$

and

$$LV(x) + \sum_{i=1}^p \sum_{j=1}^n h_i^j(x) \frac{\partial V}{\partial x_j}(x) u^i < 0. \tag{4}$$

If in addition $\lim_{x \to 0} d(x) = 0$, then we say that the control Lyapunov function V satisfies the *small control property*. The following theorem, established in [6], is an extension of Artstein's theorem [2] to the feedback stabilization of stochastic differential systems. Another version of this result can be found in [3].

THEOREM 1.2. 1. *The stochastic differential system* (1) *satisfies a stochastic Lyapunov condition at the origin if and only if it is asymptotically stabilizable by means of a feedback law $u = k(x)$ which is smooth in a neighborhood of the origin except possibly in zero.*

2. *The corresponding control Lyapunov function V satisfies the bounded control property if and only if there exists a stabilizing feedback law $u = k(x)$ which is smooth in a neighborhood D of the origin, except possibly in zero, and such that*

$$||k(x)|| < d(x)$$

for every $x \in D \setminus \{0\}$ where d is defined in (3).

Furthermore, the function k is bounded in a neighborhood of the origin in $\mathbb{R}^n$, and if in addition the control Lyapunov function V satisfies the small control property, then $k(x)$ tends to zero as x tends to zero.

In the following section, we consider stochastic differential systems (1) of the form

$$\begin{aligned} d \begin{pmatrix} x_{1,t} \\ x_{2,t} \end{pmatrix} = \begin{pmatrix} f_1(x_{1,t}, x_{2,t}) \\ f_2(x_{1,t}, x_{2,t}) \end{pmatrix} dt + \begin{pmatrix} 0 \\ h_2(x_{1,t}, x_{2,t}) \end{pmatrix} u dt \\ + \begin{pmatrix} g_1(x_{1,t}) \\ g_2(x_{1,t}, x_{2,t}) \end{pmatrix} dw_t, \end{aligned} \tag{5}$$

where $x_t = \binom{x_{1,t}}{x_{2,t}} \in \mathbb{R}^{n_1} \times \mathbb{R}^{n_2}$, $n = n_1 + n_2$, and we derive sufficient conditions for the existence of control Lyapunov functions guaranteeing the existence of stabilizing feedback laws. A similar decomposition for the stochastic differential system (1) has

been considered by Mao [11], but his methodology is quite different from the one used in this work. To illustrate the method used in the following section, consider a single-input stochastic differential system in the form (5) with $n_2 = 1$. Assume that there exists a C^2 function ϕ mapping $\mathbb{R}^{n_1}$ into $\mathbb{R}$, vanishing in the origin, which is the unique solution of the equation $h_2(x_1, \nu) = 0$ and such that $\nu = \phi(x_1)$ is a stabilizing feedback law for the single-input stochastic differential system in $\mathbb{R}^{n_1}$,

$$dx_{1,t} = f_1(x_{1,t}, \nu)dt + g_1(x_{1,t})dw_t \tag{6}$$

(see Example 1). Then, the stochastic differential system (5) satisfies a stochastic Lyapunov condition at the origin, and so by application of Theorem 1.2 it is asymptotically stabilizable in probability. Indeed, if V is a Lyapunov function for the closed-loop system

$$dx_{1,t} = f_1(x_{1,t}, \phi(x_{1,t}))dt + g_1(x_{1,t})dw_t \tag{7}$$

deduced from (6) when the control law ν is given by $\nu = \phi(x_1)$, then the function Φ defined on $\mathbb{R}^n$ by

$$\Phi(x_1, x_2) = V(x_1) + \frac{1}{4}(x_2 - \phi(x_1))^4$$

is a control Lyapunov function for the stochastic differential system (5). It is obvious that the function Φ is twice continuously differentiable on $\mathbb{R}^n$ and positive definite, and for any $x \neq 0$ with

$$h_2(x)\frac{\partial \Phi}{\partial x_2}(x) = h_2(x_1, x_2)(x_2 - \phi(x_1))^3 = 0$$

one has $x_2 = \phi(x_1)$, and therefore, denoting by $\mathcal{L}_1$ the infinitesimal generator of the stochastic process solution of the stochastic differential equation (7), it yields

$$L\Phi(x_1, x_2)|_{x_2=\phi(x_1)} = \mathcal{L}_1 V(x_1) < 0.$$

Hence, the stochastic differential system (5) satisfies a stochastic Lyapunov condition at the origin, and, according to Theorem 1.2, it is asymptotically stabilizable in probability by means of a feedback law which is smooth in a neighborhood of the origin except possibly in zero. The method described above is applicable to stochastic differential systems in the form

$$d\begin{pmatrix} x_1 \\ . \\ . \\ . \\ x_n \end{pmatrix} = \begin{pmatrix} x_2 \\ . \\ . \\ x_n \\ f(x) \end{pmatrix} dt + u\begin{pmatrix} 0 \\ . \\ . \\ 0 \\ h(x) \end{pmatrix} dt + \begin{pmatrix} x_1 \\ . \\ . \\ x_{n-1} \\ g(x) \end{pmatrix} dw_t, \tag{8}$$

where f, g, and h are C^∞ functionals vanishing in the origin. The stochastic differential system (8) satisfies a stochastic Lyapunov condition at the origin provided that there exists a function ϕ mapping $\mathbb{R}^{n-1}$ into $\mathbb{R}$, vanishing in the origin, such that $x_n = \phi(x_1, ..., x_{n-1})$ is the unique solution of the equation $h(x) = 0$ and the equilibrium solution $x_t \equiv 0$ of the lower-dimensional stochastic differential system

$$d\begin{pmatrix} x_1 \\ . \\ . \\ . \\ x_{n-1} \end{pmatrix} = \begin{pmatrix} x_2 \\ . \\ . \\ x_{n-1} \\ \phi(x_1, ..., x_{n-1}) \end{pmatrix} dt + \begin{pmatrix} x_1 \\ . \\ . \\ . \\ x_{n-1} \end{pmatrix} dw_t$$

is asymptotically stable in probability. In the following section, we extend the above results to more general stochastic differential systems. A special emphasis is given to the asymptotic stabilization in probability of stochastic differential systems (5) in which the function h is constant and they are of the form

$$d\begin{pmatrix} x_{1,t} \\ x_{2,t} \end{pmatrix} = \begin{pmatrix} f_1(x_{1,t}, x_{2,t}) \\ f_2(x_{1,t}, x_{2,t}) \end{pmatrix} dt + \begin{pmatrix} 0 \\ u \end{pmatrix} dt + \begin{pmatrix} g_1(x_{1,t}) \\ g_2(x_{1,t}, x_{2,t}) \end{pmatrix} dw_t, \tag{9}$$

where $(x_1, x_2) \in \mathbb{R}^{n_1} \times \mathbb{R}^{n_2}$ and u is an $\mathbb{R}^{n_2}$-valued control law.

2. Main results. Consider the lower-dimensional subsystems of the stochastic differential system (5):

$$dx_{1,t} = f_1(x_{1,t}, \nu)dt + g_1(x_{1,t})dw_t, \tag{10}$$

$$dx_{2,t} = h_2(x_{1,t}, x_{2,t})udt, \tag{11}$$

where $\nu \in \mathbb{R}^{n_2}$ and u is an $\mathbb{R}^p$-valued control law. Then, one can prove the following result, where the stochastic Lyapunov condition for the stochastic differential system (5) is characterized in terms of suitable positive functions for the stochastic differential systems (10)–(11).

THEOREM 2.1. *Assume that there exist neighborhoods D and D_1 of the origin in $\mathbb{R}^n$ and $\mathbb{R}^{n_1}$, respectively, and mappings $r : D \to \mathbb{R}^p$, $\phi : D_1 \to \mathbb{R}^{n_2}$, and $W : D \to \mathbb{R}$ such that $\phi(0) = 0$, ϕ is continuous, and W is twice continuously differentiable.*

Denote by M, S_1, and S_2 the subsets of $\mathbb{R}^n$ defined by

$$M = \{x \in D \ / \ x_2 = \phi(x_1)\}, \quad S_1 = \{x \in D \ / \ \nabla^\alpha W(x) = 0, \ \alpha = 0, 1, 2\}$$

and

$$S_2 = \left\{ x \in D \ / \ \left(\frac{\partial W}{\partial x_2} h_2 r \right)(x) = 0 \right\},$$

and suppose that

1. *for any x in a neighborhood of the origin in $\mathbb{R}^n$, $W(x) \geq 0$ and $S_2 \subset S_1 = M$;*
2. *the stochastic differential system (10) is asymptotically stabilizable in probability by means of the feedback law $\nu = \phi(x_1)$.*

Then, the stochastic differential system (5) satisfies a stochastic Lyapunov condition at the origin. Furthermore, if Φ denotes the corresponding control Lyapunov function and if one assumes further that

3. *the function r is bounded, and the functions ϕ and $\frac{\partial W}{\partial x_2}$ are continuously differentiable;*
4. *there exist positive constants c and $a \leq 2$ such that*

$$\left(\frac{\partial W}{\partial x_2} h_2 r \right)(x) \leq -c||x_2 - \phi(x_1)||^a \tag{12}$$

for x in a neighborhood of the origin in $\mathbb{R}^n$,
then if $a \leq 1$ the function Φ satisfies the small control property, and if $1 < a \leq 2$ and the equilibrium solution $x_{1,t} \equiv 0$ of the stochastic differential system

$$dx_{1,t} = f_1(x_{1,t}, \phi(x_{1,t}))dt + g_1(x_{1,t})dw_t,$$

the coefficients of which are assumed to be C_b^2, is exponentially stable in mean square, the stochastic control Lyapunov function Φ satisfies the bounded control property.

Proof of Theorem 2.1. Since the equilibrium solution $x_{1,t} \equiv 0$ of the closed-loop system

$$dx_{1,t} = f_1(x_{1,t}, \phi(x_{1,t}))dt + g_1(x_{1,t})dw_t, \tag{13}$$

deduced from the stochastic differential system (10) when the control law ν is given by $\nu = \phi(x_1)$, is asymptotically stable in probability, the converse Lyapunov theorem for asymptotic stability in probability proved by Kushner [10] asserts that there exist a neighborhood $\mathcal{D}$ of the origin in $\mathbb{R}^{n_1}$ and a Lyapunov function V defined on $\mathcal{D}$ such that

$$L_1 V(x_1) < 0$$

for any $x_1 \in \mathcal{D}$, $x_1 \neq 0$, where L_1 denotes the infinitesimal generator of the stochastic process solution of the closed-loop system (13). On the other hand, the function Φ defined on $\mathbb{R}^n$ by

$$\Phi(x) = V(x_1) + W(x) \tag{14}$$

is positive definite, and for any $x \in (\mathcal{D} \times \mathbb{R}^{n_2}) \cap D$,

$$\nabla\Phi(x).\begin{pmatrix} 0 \\ h_2(x) \end{pmatrix} = \frac{\partial W}{\partial x_2}(x)h_2(x).$$

Hence for any $x \neq 0$ such that $\nabla\Phi(x).\binom{0}{h_2(x)} = 0$ one has, according to assumption 1 of Theorem 2.1, $x_2 = \phi(x_1)$, which implies that

$$\nabla W(x_1, \phi(x_1)) = 0 \text{ and } \nabla^2 W((x_1, \phi(x_1)) = 0.$$

Then, denoting by $\mathcal{L}$ the infinitesimal generator of the stochastic process solution of the uncontrolled part of the stochastic differential system (5) yields

$$\mathcal{L}\Phi(x)|_{x_2=\phi(x_1)} = L_1 V(x_1) < 0.$$

Therefore, the stochastic differential system (5) satisfies a stochastic Lyapunov condition at the origin, and the function Φ defined by (14) is a control Lyapunov function. Now, assume that (12) holds, and let $a \leq 1$. Since the function f is continuously differentiable, W is twice differentiable and for any $x_1 \in \mathcal{D}$, $\nabla W(x_1, \phi(x_1)) = 0$ and $\nabla^2 W(x_1, \phi(x_1)) = 0$, one can prove easily that there exist nonnegative constants c_1, c_2, and c_3 such that

$$||\nabla f(x)|| \leq c_1, \tag{15}$$

$$||\nabla W(x)|| \leq c_2||x_2 - \phi(x_1)||, \tag{16}$$

and

$$||\nabla^2 W(x)|| \leq c_3||x_2 - \phi(x_1)|| \tag{17}$$

for x in a neighborhood of the origin in $\mathbb{R}^n$. On the other hand, denoting by q the positive definite functional defined by

$$q(x) = -L_1 V(x_1) + ||x_2 - \phi(x_1)||^2$$

and denoting by b the functional defined by

$$b(x) = c_1||\nabla V(x_1)|| + c_2||f(x)|| + \frac{c_3}{2}||(gg^\star)(x)||$$

yield

$$\begin{aligned}|\mathcal{L}\Phi(x) + q(x)| &\leq b(x)||x_2 - \phi(x_1)|| \\ &\leq b(x)||x_2 - \phi(x_1)||^a.\end{aligned}$$

Then, taking into account inequality (12) one gets

$$\mathcal{L}\Phi(x) + \left(\frac{\partial \Phi}{\partial x_2} h_2\right)(x) \frac{b(x) r(x)}{c} \leq -q(x) < 0$$

for x in a neighborhood of the origin $x \neq 0$. Therefore, since r is bounded and b is continuous with $b(0) = 0$, it follows that the control Lyapunov function Φ satisfies the small control property. Finally, assume that (12) holds with $1 < a \leq 2$ and that the function ϕ is continuously differentiable. Since the function g is continuously differentiable there exists a nonnegative constant c_4 such that

$$||\nabla(gg^\star)(x)|| \leq c_4 \tag{18}$$

for x in a neighborhood of the origin. On the other hand, since $\phi(0) = 0$, there exists a nonnegative constant c_5 such that

$$||\phi(x_1)|| \leq c_5||x_1|| \tag{19}$$

for x_1 in a neighborhood of the origin in $\mathbb{R}^{n_1}$, and since $f(0) = 0$ and $g(0) = 0$, one can deduce from (15), (18), and (19) that

$$||f(x)|| \leq c_1\left((c_5 + 1)||x_1|| + ||x_2 - \phi(x_1)||\right)$$

and

$$||(gg^\star)(x)|| \leq c_4\left((c_5 + 1)||x_1|| + ||x_2 - \phi(x_1)||\right)$$

for all x in a neighborhood of the origin in $\mathbb{R}^n$. Furthermore, since the equilibrium solution $x_t \equiv 0$ of the stochastic differential equation (13) is exponentially stable in mean square, the converse Lyapunov theorem proved by Khasminskii [9] asserts that there exists a Lyapunov function V and positive constants β_1 and β_2 such that

$$L_1 V(x_1) \leq -\beta_1||x_1||^2$$

and

$$||\nabla V(x_1)|| \leq \beta_2||x_1||^2$$

for all x_1 in a neighborhood of the origin in $\mathbb{R}^{n_1}$. Then, defining the function q on $\mathbb{R}^n$ by

$$\begin{aligned}q(x) = -L_1 V(x_1) - \left(\nabla V \frac{\partial f_1}{\partial x_2}\right)(x_1, \phi(x_1))(x_2 - \phi(x_1)) - \mathcal{L}W(x) \\ + k||x_2 - \phi(x_1)||^2,\end{aligned}$$

where k is nonnegative, one can prove that there exist nonnegative constants M_1 and M_2 such that

$$q(x) \geq \beta_1 ||x_1||^2 - M_1 ||x_1||.||x_2 - \phi(x_1)|| + (k - M_2)||x_2 - \phi(x_1)||^2,$$

and so q is positive definite provided that k is large enough. On the other hand, denoting by K a nonnegative constant such that

$$||f_1(x) - f_1(x_1, \phi(x_1)) - \frac{\partial f_1}{\partial x_2}(x_1, \phi(x_1))(x_2 - \phi(x_1))|| \leq K||x_2 - \phi(x_1)||,$$

one can prove that there exists a constant $s > 0$ such that for all x in a neighborhood of the origin in $\mathbb{R}^n$,

$$\begin{aligned} |\mathcal{L}\Phi(x) + q(x)| &\leq s||x_2 - \phi(x_1)||^a \\ &\leq -\frac{s}{c}\left(\frac{\partial W}{\partial x_2} h_2 r\right)(x). \end{aligned}$$

Therefore,

$$\mathcal{L}\Phi(x) + \left(\frac{\partial \Phi}{\partial x_2} h_2\right)(x)\frac{rs}{c} \leq -q(x)$$

for all x in a neighborhood of the origin in $\mathbb{R}^n$, $x \neq 0$, and since $\frac{rs}{c}$ is bounded, the control Lyapunov function Φ satisfies the bounded control property. This concludes the proof of Theorem 2.1.

The following result, which is an immediate consequence of the previous theorem, provides a control Lyapunov function for the stochastic differential system (5) that depends directly on the dynamics of the stochastic differential systems (10) and (11).

THEOREM 2.2. *Assume that there exist neighborhoods D and D_1 of the origin in $\mathbb{R}^n$ and $\mathbb{R}^{n_1}$, respectively, and mappings $r : D \to \mathbb{R}^p$ and $\phi : D_1 \to \mathbb{R}^{n_2}$ such that r is Lipschitz continuous, $\phi(0) = 0$, ϕ is continuous, the equilibrium solution $x_{1,t} \equiv 0$ of the stochastic differential system*

$$dx_{1,t} = f_1(x_{1,t}, \phi(x_{1,t}))dt + g_1(x_{1,t})dw_t \tag{20}$$

is asymptotically stable in probability, and the set

$$M = \{x \in D \ / \ x_2 = \phi(x_1), \ x_1 \in D_1\}$$

is asymptotically stable with respect to the differential system

$$\dot{x} = \begin{pmatrix} 0 \\ (h_2 r)(x) \end{pmatrix}. \tag{21}$$

Then, the stochastic differential system (5) *satisfies a stochastic Lyapunov condition at the origin.*

Proof of Theorem 2.2. Let N be the subset of $\mathbb{R}^n$ defined by

$$N = \{x \in \mathbb{R}^n \ / \ x_1 \in D_1\}.$$

Then the region N is positively invariant for the ordinary differential system (21) (cf. [12]), and since $M \subset N$ is asymptotically stable with respect to (21), one can

deduce from [15] that there exists a smooth Lyapunov function W defined on N such that $W(x) > 0$ and $(\frac{\partial W}{\partial x_2} h_2 r)(x) < 0$ for $x \notin M$, whereas $W(x) = 0$ for $x \in M$. Therefore, $(\frac{\partial W}{\partial x_2} h_2 r)(x) \leq 0$ in a neighborhood of the origin and $(\frac{\partial W}{\partial x_2} h_2 r)(x) = 0$ if and only if $x_2 = \phi(x_1)$. Then, assumptions 1 and 2 of Theorem 2.1 are satisfied, and so the stochastic differential system (5) satisfies a stochastic Lyapunov condition at the origin. This concludes the proof of Theorem 2.2.

In the following, we study the particular case of stochastic differential systems in the form (9). Note that one can assume without loss of generality that $f_2 \equiv 0$. Otherwise, it suffices to apply in (9) the smooth feedback law $u \to -f_2 + u$, and the stochastic differential system becomes

$$d \begin{pmatrix} x_{1,t} \\ x_{2,t} \end{pmatrix} = \begin{pmatrix} f_1(x_{1,t}, x_{2,t}) \\ 0 \end{pmatrix} dt + \begin{pmatrix} 0 \\ u \end{pmatrix} dt + \begin{pmatrix} g_1(x_{1,t}) \\ g_2(x_{1,t}, x_{2,t}) \end{pmatrix} dw_t. \tag{22}$$

The stabilizability of such stochastic differential systems has already been studied in [6]; however, the following result asserts that the existence of stabilizing feedback laws for the stochastic differential system (22) is a consequence of Theorem 2.1.

PROPOSITION 2.3. *Assume that the stochastic differential system* (10) *is asymptotically stabilizable in probability by means of a feedback law* $\nu = \phi(x_1)$ *which is continuously differentiable in a neighborhood of the origin in* $\mathbb{R}^{n_1}$. *Then, the stochastic differential system* (22) *satisfies a stochastic Lyapunov condition at the origin, and the corresponding control Lyapunov function satisfies the small control property.*

Proof of Proposition 2.3. Since the equilibrium solution $x_{1,t} \equiv 0$ of the stochastic differential equation

$$dx_{1,t} = f_1(x_{1,t}, \phi(x_{1,t}))dt + g_1(x_{1,t})dw_t$$

is asymptotically stable in probability, the converse Lyapunov theorem proved by Kushner [10] asserts that there exists a Lyapunov function V defined in a neighborhood $\mathcal{D}$ of the origin in $\mathbb{R}^{n_1}$ such that

$$L_1 V(x_1) < 0$$

for any $x_1 \in \mathcal{D}$, $x_1 \neq 0$. On the other hand, let W be the functional defined on $\mathbb{R}^n$ by

$$W(x) = \frac{1}{2} ||x_2 - \phi(x_1)||^2.$$

Then, W is semipositive definite and $\frac{\partial W}{\partial x_2}$ is continuously differentiable in a neighborhood of the origin. Furthermore, if r denotes the functional defined on $\mathbb{R}^n$ by

$$r_i(x) = -\operatorname{sgn}\left((x_2 - \phi(x_1))_i\right), \qquad 1 \leq i \leq n_2,$$

it is obvious that r is uniformly bounded on $\mathbb{R}^n$ and so W satisfies the hypotheses of Theorem 2.1.

In particular, note that (12) is fulfilled with $a = 1$, which implies that the control Lyapunov function Φ defined on $\mathbb{R}^n$ by

$$\Phi(x) = V(x_1) + W(x)$$

satisfies the small control property. This completes the proof of Proposition 2.3.

3. Numerical examples.

Example 1. Let x_0 be given in $\mathbb{R}^2$, and denote by $x_t \in \mathbb{R}^2$, the solution of the stochastic differential system

$$d\begin{pmatrix} x_{1,t} \\ x_{2,t} \end{pmatrix} = \begin{pmatrix} -\frac{1}{2}x_{1,t} + x_{2,t} \\ \phi_1(x_{1,t}, x_{2,t}) \end{pmatrix} dt + u \begin{pmatrix} 0 \\ x_{2,t} + \phi_2(x_{1,t}) \end{pmatrix} dt + \begin{pmatrix} x_{1,t} \\ x_{2,t} \end{pmatrix} dw_t, \tag{23}$$

where ϕ_2 is a C^2 functional on $\mathbb{R}$ such that

$$-2x_1\phi_2(x_1) + (x_1\nabla\phi_2(x_1) - \phi_2(x_1))^2 < 0 \tag{24}$$

for any x_1 in a neighborhood of the origin in $\mathbb{R}$, $x_1 \neq 0$. (Note that $\phi_2(x_1) = x_1$ satisfies inequality (24).) Then one can prove easily that

- the function Φ defined on $\mathbb{R}^2$ by

$$\Phi(x) = x_1^2 + (x_2 + \phi_2(x_1))^2$$

is a control Lyapunov function for the stochastic differential system (23);

- the feedback law $\nu = -\phi_2(x_1)$ renders the control stochastic differential system

$$dx_{1,t} = -\frac{1}{2}x_{1,t}dt + \nu dt + x_{1,t}dw_t$$

asymptotically stable in probability.

Therefore, according to Theorems 1.2 and 2.1 the stochastic differential system (23) is asymptotically stabilizable in probability by means of a feedback law which is smooth in a neighborhood of the origin in $\mathbb{R}^2$, except possibly at the origin. On the other hand, if W denotes the function defined on $\mathbb{R}^2$ by

$$W(x) = (x_2 + \phi(x_1))^2,$$

one can prove that inequality (12) is fulfilled with $a = 2$ and $r = -1$; therefore, the function Φ satisfies the bounded control property provided that $\nabla\phi_2(0) = 0$.

Furthermore, in this case the equilibrium solution $x_{1,t} \equiv 0$ of the stochastic differential equation

$$dx_{1,t} = \left(-\frac{1}{2}x_{1,t} - \phi_2(x_{1,t})\right) dt + x_{1,t}dw_t$$

is exponentially stable in mean square, which implies, according to Theorem 2.1, that the stochastic differential system (23) is asymptotically stabilizable in probability by a bounded feedback law.

Example 2. Consider the stochastic differential system in $\mathbb{R}^3$ defined by

$$dy_t = \begin{pmatrix} y_{1,t}^3 - y_{1,t}^2(y_{2,t} + y_{3,t}) \\ y_{2,t} \\ 0 \end{pmatrix} dt + \begin{pmatrix} 0 \\ (y_{1,t} - y_{2,t})^3 \\ 0 \end{pmatrix} u_1 dt + \begin{pmatrix} 0 \\ 0 \\ 1 \end{pmatrix} u_2 dt + \begin{pmatrix} y_{1,t}^2 \\ y_{2,t} \\ y_{3,t} \end{pmatrix} dw_t, \tag{25}$$

where y_0 is given in $\mathbb{R}^3$. The stochastic differential system (25) has the form (5), where

$$x_{1,t} = y_{1,t}, x_{2,t} = \begin{pmatrix} y_{2,t} \\ y_{3,t} \end{pmatrix}, \qquad x_t = \begin{pmatrix} x_{1,t} \\ x_{2,t} \end{pmatrix},$$
$$f_1(x) = y_{1,t}^3 - y_{1,t}^2(y_{2,t} + y_{3,t}), \qquad f_2(x) = \begin{pmatrix} y_{2,t} \\ 0 \end{pmatrix},$$
$$g_1(x_{1,t}) = x_{1,t}^2, \qquad g_2(x) = x_{2,t},$$
$$u = \begin{pmatrix} u_1 \\ u_2 \end{pmatrix},$$

and

$$h_2(x) = \begin{pmatrix} (y_{1,t} - y_{2,t})^3 & 0 \\ 0 & 1 \end{pmatrix}.$$

Furthermore, if r denotes the functional defined on $\mathbb{R}^3$ by

$$r(x) = \begin{pmatrix} 1 \\ y_{1,t} - y_{3,t} \end{pmatrix}$$

and if ϕ denotes the functional defined on $\mathbb{R}$ by

$$\phi(x_1) = \begin{pmatrix} x_1 \\ x_1 \end{pmatrix}, \tag{26}$$

then the set

$$M = \left\{ x \in \mathbb{R}^3 \ / \ x_2 = \phi(x_1) \right\}$$

is asymptotically stable with respect to the ordinary differential system (21). This result is easily proved by evaluating the derivative of the Lyapunov function W defined on $\mathbb{R}^3$ by

$$W(x) = \frac{1}{2} ||x_2 - \phi(x_1)||^2$$

along the trajectories of (21). On the other hand, the feedback law ν defined on $\mathbb{R}$ by

$$\nu = \phi(x_1)$$

asymptotically stabilizes in probability the stochastic differential system (10). Indeed, if V denotes the Lyapunov function defined on $\mathbb{R}$ by

$$V(x_1) = \frac{1}{2} x_1^2,$$

one has

$$L_1 V(x_1) = -\frac{1}{2} x_1^4,$$

which implies, according to the stochastic Lyapunov theorem (Theorem 1.3 in [6]), that the equilibrium solution $x_{1,t} \equiv 0$ of the stochastic differential system (20) is asymptotically stable in probability. Therefore, the stochastic differential system (25) satisfies the hypotheses of Theorem 2.2 and so is asymptotically stabilizable in probability.

REFERENCES

[1] L. Arnold, *Stochastic Differential Equations: Theory and Applications*, John Wiley, New York, 1974.

[2] Z. Artstein, *Stabilization with relaxed controls*, Nonlinear Anal.,7 (1983), pp. 1163–1173.

[3] P. Florchinger, *A universal formula for the stabilization of control stochastic differential equations*, Stochastic Anal. Appl., 11 (1993), pp. 155–162.

[4] P. Florchinger, *A stochastic version of Jurdjevic–Quinn theorem*, Stochastic Anal., 12 (1994), pp. 473–480.

[5] P. Florchinger, *On the stabilization of homogeneous control stochastic systems*, in Proc. 32nd IEEE Conference on Decision and Control, San Antonio, TX, 1993, pp. 855–856.

[6] P. Florchinger, *Lyapunov-like techniques for stochastic stability*, SIAM J. Control Optim., 33 (1995), pp. 1151–1169.

[7] P. Florchinger, A. Iggidr, and G. Sallet, *Stabilization of a class of nonlinear stochastic systems*, Stochastic Process. Appl., 50 (1994), pp. 235–243.

[8] Z. Y. Gao and N. U. Ahmed, *Feedback stabilizability of nonlinear stochastic systems with state–dependent noise*, Internat. J. Control, 45 (1987), pp. 729–737.

[9] R. Z. Khasminskii, *Stochastic Stability of Differential Equations*, Sijthoff & Noordhoff, Alphen aan den Rijn, the Netherlands, 1980.

[10] H. J. Kushner, *Converse theorems for stochastic Liapunov functions*, SIAM J. Control, 5 (1967), pp. 228–233.

[11] X. Mao, *Exponential stability of large-scale stochastic differential equations*, Systems Control Lett., 19 (1992), pp. 71–81.

[12] J. L. Massera, *Contributions to stability theory*, Ann. of Math., 64 (1956), pp. 182–206; erratum in Ann. of Math., 68 (1958), p. 202.

[13] E. D. Sontag and H. J. Sussmann, *Further comments on the stabilizability of the angular velocity of a rigid body*, Systems Control Lett., 12 (1989), pp. 213–217.

[14] J. Tsinias, *Existence of control Lyapunov functions and applications to state feedback stabilizability of nonlinear systems*, SIAM J. Control Optim., 29 (1991), pp. 457–473.

[15] F. W. Wilson, *The structure of the level surfaces of a Lyapunov function*, J. Differential Equations, 3 (1967), pp. 323–329.

[16] W. M. Wonham, *On a matrix Riccati equation of stochastic control*, SIAM J. Control Optim., 6 (1968), pp. 681–697.

SIAM J. CONTROL OPTIM.
Vol. 35, No. 2, pp. 512–531, March 1997

008

OPTIMAL STRATEGIES FOR BILEVEL DYNAMIC PROBLEMS*

JANE J. YE†

Abstract. In this paper we study the bilevel dynamic problem, which is a hierarchy of two dynamic optimization problems, where the constraint region of the upper level problem is determined implicitly by the solutions to the lower level optimal control problem. To obtain optimality conditions, we reformulate the bilevel dynamic problem as a single level optimal control problem that involves the value function of the lower-level problem. Sensitivity analysis of the lower-level problem with respect to the perturbation in the upper-level decision variable is given and first-order necessary optimality conditions are derived by using nonsmooth analysis. A constraint qualification of calmness type and a sufficient condition for the calmness are also given.

Key words. necessary conditions, bilevel dynamic problems, sensitivity analysis, nonsmooth analysis, value function, constraint qualification, calmness condition

AMS subject classifications. 90D65, 49K4

PII. S0363012993256150

1. Introduction.

Let us consider a two-level hierarchical system where two decision makers try to find best decisions with respect to certain, but generally different, goals. Moreover, assume that these decision makers cannot act independently of each other but only according to a certain hierarchy whereby the optimal strategy chosen by the lower level (hereafter the "follower") depends on the strategy selected by the upper level (hereafter the "leader"). On the other hand, let the objective function of the leader depend not only on his own decision but also on the reaction of the follower. Then while having the first choice, the leader is able to evaluate the true value of his own selection only after knowing the follower's possible reactions. Assume that the game is cooperative; i.e., if the follower's problem has several optimal decisions for a given leader's decision, then the follower allows the leader to choose which of them is actually used. Thus the leader will choose his optimal decision among all decisions available and the follower's optimal decision to minimize his objective. In particular, we consider a hierarchical dynamical system, where the state $x(t) \in R^d$ is influenced by the decisions of both leader and follower $u(\cdot)$ and $v(\cdot)$. The state $x(t) \in R^d$ is described by

$$\begin{aligned} &\dot{x}(t) = \phi(t, x(t), u(t), v(t)) \qquad \text{almost everywhere (a.e.) } t \in [t_0, t_1], \\ &x(t_0) = x_0, \end{aligned}$$

where $u(t) \in U$, a closed subset of R^n and $v(t) \in W(t) \subset R^m$ for almost all $t \in [t_0, t_1]$. In mathematical terms, given any control function $u(\cdot)$ selected by the leader, the follower faces the ordinary (single-level) optimal control problem involving a parameter u,

$$P_2(u) \qquad \min J_2(x, u, v) = \int_{t_0}^{t_1} G(t, x(t), u(t), v(t))dt + g(x(t_1))$$

*Received by the editors September 24, 1993; accepted for publication (in revised form) January 30, 1996. This research was supported in part by the National Sciences and Engineering Research Council of Canada.

http://www.siam.org/journals/sicon/35-2/25615.html

†Department of Mathematics and Statistics, University of Victoria, Victoria, BC, Canada V8W 3P4 (janeye@uvaix.uvic.ca).

$$\begin{aligned} \text{subject to (s.t.) } & \dot{x}(t) = \phi(t, x(t), u(t), v(t)) \qquad \text{a.e.}, \\ & x(t_0) = x_0, \\ & v(t) \in W(t) \qquad \text{a.e.}, \end{aligned}$$

while the leader faces the *bilevel dynamic problem*,

$$\begin{aligned} P_1 \qquad & \min\ J_1(x,u,v) = \int_{t_0}^{t_1} F(t, x(t), u(t), v(t))dt + f(x(t_1)) \\ & \text{over } u \in L^2([t_0,t_1], U) \text{ and all solutions } (x,v) \text{ of } P_2(u). \end{aligned}$$

The bilevel static problem, where both the leader's and the follower's decisions are vectors instead of control functions, was first introduced by von Stackelberg [14] for an economic model. The bilevel dynamic problems were first considered by Chen and Cruz in [2]. Most of the bilevel (static or dynamic) problems are attacked by reducing the bilevel problem to a single-level problem with the first-order necessary optimality conditions for the lower-level problem as additional constraints (cf. Bard and Falk [1] and Zhang [20], [21] for bilevel static problems, Chen and Cruz [2] and Zhang [20] for bilevel dynamic problems). The reduction is equivalent provided the lower-level optimal control problem is convex, since in this case the first-order necessary optimality condition is also sufficient. Apart from the strong convexity assumption, the resulting optimality conditions of the above approach involve second-order derivatives and a larger system, since the reduced problem minimizes over the set of original decision variables as well as the set of multipliers of the lower-level problem.

To our knowledge, there is no optimality condition for a general bilevel dynamic problem to date. The necessary condition obtained by Chen and Cruz in [2] holds in the case where Pontryagin's maximum principle for the lower-level optimal control problem is sufficient for optimality and no bounds are allowed for the control functions. The necessary condition was stated in a normal form (i.e., the multiplier for the objective function of the upper-level problem is 1) that holds only when the reduced single-level optimal control problem is calm (see [3] for definition). The necessary condition obtained by Zhang in [20] is only for a bilevel dynamic problem in which the dynamics are linear in the state and control variables and require convexity assumptions on the objective function of the lower-level problem. The purpose of this paper is to provide first-order necessary optimality conditions for problem P_1 under *very general* assumptions (in particular, without convexity assumptions and with bounds on the control functions).

Define the *value function of the lower-level optimal control problem* as an extended-valued functional $V(u) : L^2([t_0,t_1], U) \to \bar{R}$ defined by

$$V(u) := \inf \left\{ \int_{t_0}^{t_1} G(t, x(t), u(t), v(t))dt + g(x(t_1)) : \begin{array}{ll} \dot{x}(t) = \phi(t, x(t), u(t), v(t)) & \text{a.e.} \\ v(t) \in W(t) \qquad \text{a.e.} & \\ x(t_0) = x_0 & \end{array} \right\},$$

where $\bar{R} := R \cup \{-\infty\} \cup \{+\infty\}$ is the extended real line and $\inf \emptyset = +\infty$ by convention. Our approach is to reformulate P_1 as the following single-level optimal control problem:

$$\widetilde{P}_1 \qquad \min\ J_1(u,v) = \int_{t_0}^{t_1} F(t, x(t), u(t), v(t))dt + f(x(t_1))$$

$$\begin{aligned} \text{s.t. } & \dot{x}(t) = \phi(t, x(t), u(t), v(t)) \qquad \text{a.e.}, \\ & x(t_0) = x_0, \\ & u(\cdot) \in L^2([t_0, t_1], U), v(t) \in W(t) \qquad \text{a.e.}, \\ & \int_{t_0}^{t_1} G(t, x(t), u(t), v(t))dt + g(x(t_1)) - V(u) = 0. \end{aligned} \tag{1}$$

The above problem is obviously equivalent to the original bilevel dynamic problem P_1 and is a *nonstandard optimal control problem* since the constraint (1) involves a functional defined by the value function $V(u)$ of the lower-level optimal control problem. In general $V(u)$ is not an explicit function of the problem data and is nonsmooth even in the case where all problem data are smooth functions. To derive a necessary condition for optimality for problem P_1, one needs to study Lipschitz continuity and generalized gradients of the value function $V(u)$ and develop a necessary optimality condition for the nonstandard optimal control problem with functional constraints (1). Recent developments in nonsmooth analysis allow us to study Lipschitz continuity and generalized gradients of the value function $V(u)$ with respect to a *nonadditive* infinite-dimensional perturbation u. We then reformulate the nonstandard optimal control problem as an infinite-dimensional optimization problem and use a result due to Ioffe [8] to derive a necessary optimality condition for the nonstandard optimal control problem with functional constraints.

The approach of reducing a bilevel problem to a single-level problem using the value function was used in the literature (see [11], [12]) for numerical purposes and for deriving first-order necessary conditions for the static bilevel optimization problem [17], [18]. The essential issue in the static case is the constraint qualification since the generalized differentiability of the value function in the finite-dimensional case is well known and the resulting equivalent single-level problem is an ordinary mathematical programming problem. It was shown in [17] and [18] that bilevel problems always have abnormal multipliers, and the right constraint qualification for ensuring the existence of a normal multiplier is the calmness condition. In Ye [16], a bilevel dynamic optimization problem where the lower level is an optimal control problem while the upper-level decision variable is a vector is considered. Although the bilevel dynamic optimization problem considered in [16] is a special case of the problem we study in this paper, it deserves special attention since it reduces to a single-level optimal control problem with end point constraints involving a value function that is a function of the upper-level decision vector. Fritz John–type necessary optimality conditions were derived under more general assumptions.

The following basic assumptions are in force throughout this paper:

(A1) $W(t) : [t_0, t_1] \to R^m$ is a nonempty, compact-valued, set-valued map. The graph of $W(t)$ (i.e., the set $\{(s, r) : s \in [t_0, t_1], r \in W(s)\}$), denoted by Gr$W$, is $\mathcal{L} \times \mathcal{B}$ measurable, where $\mathcal{L} \times \mathcal{B}$ denotes the σ-algebra of subsets of $[t_0, t_1] \times R^m$ generated by product sets $M \times N$ where M is a Lebesgue measurable subset of $[t_0, t_1]$ and N is a Borel subset of R^m.

(A2) The function $F(t, x, u, v) : [t_0, t_1] \times R^d \times R^n \times R^m \to R$ is $\mathcal{L} \times \mathcal{B}$ measurable in (t, v) and continuously differentiable in x and u. The functions $\phi(t, x, u, v) : [t_0, t_1] \times R^d \times R^n \times R^m \to R^d$, $G(t, x, u, v) : [t_0, t_1] \times R^d \times R^n \times R^m \to R$ are measurable in t, continuously differentiable in x and u, and lower semicontinuous in v.

(A3) There exists an integrable function $\psi : [t_0, t_1] \to R$ such that

$$|\nabla_{(x,u)} F| + |\nabla_{(x,u)} G| + |\nabla_{(x,u)} \phi| \le \psi(t) \quad \forall (t, x, u, v) \in [t_0, t_1] \times R^d \times U \times W(t).$$

(A4) The function $f(x) : \mathbb{R}^d \to R$ is locally Lipschitz continuous, and the function $g(x) : R^d \to R$ is Lipschitz continuous of rank $L_g \geq 0$.

(A5) For any $u \in L^2([t_0, t_1], U)$, $P_2(u)$ has an admissible pair (whose definition is given below).

A *control function* for $P_2(u)$ is a (Lebesgue) measurable selection $v(\cdot)$ for $W(\cdot)$, that is, a measurable function satisfying $v(t) \in W(t)$ a.e. $t \in [t_0, t_1]$. An *arc* is an absolutely continuous function. An *admissible pair* for $P_2(u)$ is a pair of functions $(x(\cdot), v(\cdot))$ on $[t_0, t_1]$ of which $v(\cdot)$ is a control function for $P_2(u)$ and $x(\cdot) : [t_0, t_1] \to R^d$ is an arc that satisfies the differential equation $\dot{x}(t) = \phi(t, x(t), u(t), v(t))$ a.e., together with the initial condition $x(t_0) = x_0$. The first and the second components of an admissible pair are called an *admissible trajectory* and *admissible control*, respectively. A *solution* to problem $P_2(u)$ is an admissible pair for $P_2(u)$ that minimizes the value of the cost functional $J_2(x, u, v)$ over all admissible pairs for $P_2(u)$. An *admissible strategy* for P_1 includes $u \in L^2([t_0, t_1], U)$ and an optimal control v for $P_2(u)$. The strategy (u, v) and the corresponding trajectory x are *optimal* for the bilevel dynamic problem P_1 if (x, u, v) minimizes the value of the cost functional $J_1(x, u, v)$ among all admissible strategies and the corresponding trajectories for P_1.

The plan of the paper is as follows. In section 2, we study generalized differentiability of the value function $V(u)$. In section 3, under a calmness-type constraint qualification, we derive a Kuhn–Tucker–type necessary optimality condition for the bilevel dynamic problem. It is also shown that the existence of a uniformly weak sharp minimum is a sufficient condition for the calmness, and a sufficient condition for existence of a weak sharp minimum is given. Finally, three examples are given in section 3 to illustrate applications of the constraint qualification and the necessary optimality conditions.

2. Differentiability of the value function. Let X be a Hilbert space. Consider a lower semicontinuous functional $\phi : X \to R \cup \{+\infty\}$ and a point $\bar{x} \in X$, where ϕ is finite. A vector $\zeta \in X$ is called a *proximal subgradient* of $\phi(\cdot)$ at $\bar{x}$ provided that there exist $M > 0, \delta > 0$ such that

$$\phi(x') - \phi(\bar{x}) + M\|x' - \bar{x}\|^2 \geq \langle \zeta, x' - \bar{x} \rangle, \qquad x' \in \bar{x} + \delta B.$$

The set of all proximal subgradients of $\phi(\cdot)$ at $\bar{x}$ is denoted $\partial^\pi \phi(\bar{x})$. A *limiting subgradient* of ϕ at $\bar{x}$ is the set

$$\hat{\partial}\phi(\bar{x}) := \{\text{weak} \lim_{k \to \infty} \zeta_k : \zeta_k \in \partial^\pi \phi(x_k), x_k \to \bar{x}, \phi(x_k) \to \phi(\bar{x})\}.$$

The limiting subgradient is a smaller object than the *Clarke generalized gradient* (see Clarke [3] for definition). In fact, if ϕ is Lipschitz continuous near $\bar{x}$, we have $\partial\phi(\bar{x}) = \text{clco}\hat{\partial}\phi(\bar{x})$, where ∂ and clcoA denote the Clarke generalized gradient and closed convex hull of set A, respectively. For the definition and more details of the precise relation between the limiting subgradient and the Clarke generalized gradient, the reader is referred to Clarke [4] and Rockafellar [13].

The following result concerning the compactness of trajectories of a differential inclusion is slightly different from [3, Theorem 3.1.7] and will be used repeatedly. We omit the proof here since it can be proved similarly to [3, Theorem 3.1.7].

PROPOSITION 2.1. *Let $\Gamma : [t_0, t_1] \times R^d \times R^n \to R^n \times R^n$ be a set-valued map. We suppose that Γ is integrably bounded (i.e., there exists an integrable function $k(t)$ such that $|v| \leq k(t) \forall v \in \Gamma(t, x, u)$) and that Γ is nonempty, compact, and convex. We suppose that for every $(t, x, u) \in [t_0, t_1] \times R^d \times R^n$ the set-valued map $t' \to \Gamma(t', x, u)$*

is measurable and $\forall [t_0,t_1]\times R^d\times R^n$*, the set-valued map* $(x',u')\to\Gamma(t,x',u')$ *is upper semicontinuous. Let* Γ *be* $\mathcal{L}\times\mathcal{B}$ *measurable, where* $\mathcal{L}\times\mathcal{B}$ *denotes the* σ*-algebra of subsets of* $[t_0,t_1]\times R^d\times R^n$ *generated by product sets* $M\times N$*, where* M *is a Lebesque measurable subset of* $[t_0,t_1]$ *and* N *is a Borel subset of* $R^d\times R^n$*.*

Let $\{x_i\}$ *be a sequence of arcs on* $[t_0,t_1]$ *and* $\{\zeta_i\}$ *be a sequence of functions in* $L^2([t_0,t_1],R^n)$ *satisfying*

(i) $(\dot{x}_i(t),\zeta_i(t))\in\Gamma(t,x_i(t),u_i(t))$ *a.e.* $t\in[t_0,t_1]$,

(ii) $\zeta_i\to\zeta$ *weakly in* L^2,

(iii) $u_i\to u$ *in* L^2,

(iv) $\{x_i(t_0)\}$ *is bounded.*

Then there exists a subsequence of $\{x_i\}$ *that converges uniformly to an arc* x *such that*

$$(\dot{x}(t),\zeta(t))\in\Gamma(t,x(t),u(t))\qquad \text{a.e. } t\in[t_0,t_1].$$

To discuss generalized differentiability of the value function $V(u)$, we will need the following assumptions:

(A6) There exists $k(t)\in L^2([t_0,t_1],R)$ such that

$$|\phi|+|\nabla_{(x,u)}\phi|+|G|+|\nabla_{(x,u)}G|\le k(t)\quad\forall(t,x,u,v)\in[t_0,t_1]\times R^d\times U\times W(t).$$

(A7) For any $(t,x,u)\in[t_0,t_1]\times R^d\times R^n$ the set

$$\{(\phi(t,x,u,v),G(t,x,u,v)+r):v\in W(t),r\ge 0\}$$

is convex.

(A8) $|\nabla_u\phi|\le M\quad\forall(t,x,u,v)\in[t_0,t_1]\times R^d\times U\times W(t)$, where $M>0$ is a constant.

Remark 2.2. Assumption (A7) is standard in control theory to ensure the existence of an optimal control for the lower-level problem. In the case where this assumption is not satisfied, the standard procedure is to go for the relaxed control (see, e.g., [19] and [22]).

Let the Hamiltonian for $P_2(u)$ be the function defined by

$$H_2(t,x,u,p_2):=\sup\{p_2\cdot\phi(t,x,u,v)-G(t,x,u,v):v\in W(t)\}$$

and Y_u be the set of all optimal trajectories x to problem $P_2(u)$.

The following result gives the Lipschitz continuity of the value function and characterizes the generalized gradient of the value function. It extends the result of Clarke [5] to allow general *nonadditive* perturbations in both the dynamics and the objective function.

THEOREM 2.3. *Suppose that assumptions* (A1)–(A8) *hold. Then* V *is Lipschitz continuous near* u *and*

$$\partial V(u)\subset clco\cup_{x\in Y_u}\{\zeta:\exists\ arc\ p_2\ s.t.\ (-\dot{p_2},-\zeta,\dot{x})\in\partial H_2(t,x,u,p_2)\ a.e.\ -p_2(t_1)\in\hat{\partial}g(x(t_1))\},$$

where ∂H_2 *denotes the Clarke generalized gradient with respect to* (x,u,p_2)*.*

Before proving Theorem 2.3, we first give the following result.

LEMMA 2.4. *Let* u_i *be a sequence converging (in* L^2*) to* u *and let* (x_i,v_i) *be an admissible pair for* $P_2(u_i)$*. Then there exist a subsequence of* $\{x_i\}$ *converging*

uniformly to an arc x and a control v with (x, v) being an admissible pair for $P_2(u)$ such that

$$J_2(x, u, v) \le \liminf J_2(x_i, u_i, v_i).$$

Proof. Let

$$\dot{y}_i(t) := G(t, x_i(t), u_i(t), v_i(t)).$$

Then

$$(\dot{x}_i(t), \dot{y}_i(t)) \in \Gamma(t, x_i(t), y_i(t), u_i(t)), \tag{2}$$

where

$$\Gamma(t, x, y, u) := \{(\phi(t, x, u, v), r) : G(t, x, u, v) \le r \le k(t) + 1, v \in V(t)\}.$$

The proof can be reduced to an application of Proposition 2.1 by studying the differential inclusion (2). The essential fact in the reduction is Fillipov's lemma: an (extended) arc (x, y) satisfies the differential inclusion iff there is a control function v for x such that (x, v) is feasible for $P_2(u)$ and y satisfies $G(t, x, u, v) \le \dot{y} \le k(t) + 1$. □

We now turn to the proof of the theorem. By (A5), $P_2(u)$ has an admissible pair. So $V(u)$ is finite. By Lemma 2.4, V is (strongly) lower semicontinuous.

Step 1. Let $u \in L^2([t_0, t_1], U)$ and $\zeta \in \partial^\pi V(u)$. Let (x, v) be a solution of $P_2(u)$ that exists by virtue of Lemma 2.4. Then by definition, for some $M > 0$ and $\forall\ u'$ near u (in the L^2 norm), we have

$$\begin{aligned} V(u') - \langle \zeta, u' \rangle + M\|u' - u\|_2^2 &\ge V(u) - \langle \zeta, u \rangle \\ &= \int_{t_0}^{t_1} G(t, x(t), u(t), v(t))dt + g(x(t_1)) - \int_{t_0}^{t_1} \langle \zeta(t), u(t) \rangle dt. \end{aligned}$$

Let (x', v') be an admissible pair for $P_2(u')$. Then

$$\begin{aligned} &\int_{t_0}^{t_1} G(t, x'(t), u'(t), v'(t))dt + g(x'(t_1)) - \int_{t_0}^{t_1} \langle \zeta(t), u'(t) \rangle dt + M\|u' - u\|_2^2 \\ &\qquad \ge \int_{t_0}^{t_1} G(t, x(t), u(t), v(t))dt + g(x(t_1)) - \int_{t_0}^{t_1} \langle \zeta(t), u(t) \rangle dt. \end{aligned}$$

Hence (x, u, v) is a solution of the following optimal control problem:

$$\begin{aligned} \min & \int_{t_0}^{t_1} [G(t, x'(t), u'(t), v'(t)) - \langle \zeta(t), u'(t) \rangle]dt + g(x'(t_1)) + M\|u' - u\|_2^2 \\ \text{s.t. } & \dot{x'}(t) = \phi(t, x'(t), u'(t), v'(t)) \qquad \text{a.e.}, \\ & x'(t_0) = x_0, \\ & v'(t) \in W(t) \qquad \text{a.e.}, \\ & u'(t) \in U(t) := \{u' \in R^n : |u' - u(t)| \le \epsilon\}. \end{aligned}$$

Applying Theorem 5.2.1 of Clarke [3] with the Clarke generalized gradient replaced by the limiting subgradient in the transversality conditions (cf. [4, 10, 9]) to the above

optimal control problem with free end points leads to the existence of an arc p_2 such that

$$-\dot{p_2}(t) = \nabla_x \phi(t, x(t), u(t), v(t))^\top p_2(t) - \nabla_x G(t, x(t), u(t), v(t)) \quad \text{a.e.,} \tag{3}$$
$$\max_{u \in U(t), v \in W(t)} \{p_2(t) \cdot \phi(t, x(t), u, v) - G(t, x(t), u, v) + \langle \zeta(t), u \rangle\}$$

$$= p_2(t) \cdot \phi(t, x(t), u(t), v(t)) - G(t, x(t), u(t), v(t)) + \langle \zeta(t), u(t) \rangle \text{ a.e.,} \tag{4}$$

$$-p_2(t_1) \in \hat{\partial} g(x(t_1)), \tag{5}$$

where $^\top$ denotes the transpose. Equation (4) implies that

$$\max_{v \in W(t)} \{p_2(t) \cdot \phi(t, x(t), u(t), v) - G(t, x(t), u(t), v)\}$$
$$= p_2(t) \cdot \phi(t, x(t), u(t), v(t)) - G(t, x(t), u(t), v(t)) \text{ a.e.}$$

and

$$-\zeta(t) = \nabla_u \phi(t, x(t), u(t), v(t))^\top p_2(t) - \nabla_u G(t, x(t), u(t), v(t)). \tag{6}$$

Step 2. For any $\zeta \in \hat{\partial} V(u)$ by definition $\zeta = \text{weak} \lim_{i \to \infty} \zeta_i$, where $\zeta_i \in \partial^\pi V(u_i)$, $u_i \to u$ in L^2 and $V(u_i) \to V(u)$. By Step 1, for each u_i there exists an arc p_2^i and an arc x_i that solves $P_2(u_i)$ (along with v_i) such that

$$-\dot{p_2}^i(t) = \nabla_x \phi(t, x_i(t), u_i(t), v_i(t))^\top p_2^i(t) - \nabla_x G(t, x_i(t), u_i(t), v_i(t)) \quad \text{a.e.,} \tag{7}$$
$$\max_{v \in W(t)} \{p_2^i(t) \cdot \phi(t, x_i(t), u_i(t), v) - G(t, x_i(t), u_i(t), v)\}$$

$$= p_2^i(t) \cdot \phi(t, x_i(t), u_i(t), v_i(t)) - G(t, x_i(t), u_i(t), v_i(t)) \text{ a.e.,} \tag{8}$$

$$-\zeta_i(t) = \nabla_u \phi(t, x_i(t), u_i(t), v_i(t))^\top p_2^i(t) - \nabla_u G(t, x_i(t), u_i(t), v_i(t)), \tag{9}$$

$$-p_2^i(t_1) \in \hat{\partial} g(x_i(t_1)). \tag{10}$$

By [3, Theorem 2.8.2], (7), (8), and (9) imply that

$$(-\dot{p_2^i}(t), -\zeta_i(t), \dot{x_i}(t)) \in \partial H_2(t, x_i(t), u_i(t), p_2^i(t)) \quad \text{a.e.} \tag{11}$$

From (7)

$$p_2^i(t) = p_2^i(t_1) - \int_{t_1}^{t} [\nabla_x \phi(s, x_i(s), u_i(s), v_i(s))^\top p_2^i(s) - \nabla_x G(s, x_i(s), u_i(s), v_i(s))] ds.$$

By assumption (A4) and inclusion (10), the norm of $p_2^i(t_1)$ is bounded by L_g. Assumption (A3) implies that the norms of $\nabla_x \phi$ and $\nabla_x G$ are bounded by the integrable function ψ. Thus

$$|p_2^i(t)| \le (L_g + \int_{t_0}^{t_1} \psi(s) ds) + \int_t^{t_1} \psi(s) |p_2^i(s)| ds$$
$$= K + \int_t^{t_1} \psi(s) |p_2^i(s)| ds,$$

where $K := L_g + \int_{t_0}^{t_1} \psi(s)ds$. Invoking Gronwall's inequality, we conclude that

$$|p_2^i(t)| \le Ke^{\int_t^{t_1} \psi(s)ds},$$

which implies that $\|p_2^i\|_\infty$ is bounded. It follows that the set-valued map ∂H_2 is integrably bounded. Applying Proposition 2.1 to differential inclusion (11) with boundary condition (10), we conclude that there exists a convergent subsequence of $\{x_i, p_2^i\}$ that converges to the arcs x, p_2 such that

$$(-\dot{p_2}(t), -\zeta(t), \dot{x}(t)) \in \partial H_2(t, x(t), u(t), p_2(t)) \qquad \text{a.e.}$$

Note that by Lemma 2.4 we may suppose $x \in Y_u$ since x_i is an optimal trajectory of $P_2(u_i)$. From the upper semicontinuity of the limiting subgradients

$$-p_2(t_1) \in \hat{\partial} g(x(t_1)).$$

Therefore we conclude that

$$\hat{\partial} V(u) \subset \cup_{x \in Y_u} \{\zeta : \exists \text{ arc } p_2 \text{ s.t. } (-\dot{p_2}, -\zeta, \dot{x}) \in \partial H_2(t, x, u, p_2) \text{ a.e.}, -p_2(t_1) \in \hat{\partial} g(x(t_1))\}.$$

Step 3. To complete the proof of the theorem, we only have to show that V is Lipschitz near u. By [6, Theorem 3.6], V is Lipschitz near u of rank C iff

$$\sup\{\|\zeta\|_2 : \zeta \in \partial^\pi V(u')\} \le C \qquad \forall u' \text{ in a neighborhood of } u.$$

Indeed, by Step 1, for any u and any $\zeta \in \partial^\pi V(u)$ there exists an arc p_2 along with a solution (x, v) of $P_2(u)$ such that (3), (5), and (6) hold. Therefore

$$|\zeta(t)| \le M(|p_2(t)| + |\nabla_u G|). \tag{12}$$

Since $\forall$ such p_2, $\|p_2\|_\infty \le Ke^{\int_{t_0}^{t_1} \psi(s)ds}$, it then follows from (12) that all $\zeta \in \partial^\pi V(u), \forall u \in L^2([t_0, t_1], U)$ are bounded in L^2. Hence V is Lipschitz continuous, and the proof of Theorem 2.3 is now complete. □

3. Necessary conditions for optimality. As in the static case (cf. [17, 18]), it is easy to show that the equivalent single-level optimal control problem $\widetilde{P}_1$ always has a nontrivial abnormal multiplier; i.e., there always exists (λ, r, p_1) not all equal to zero with $\lambda = 0$ satisfying (13), (14), (15), and (16). Hence the traditional technique of concluding the existence of a normal multiplier from the nonexistence of a nontrivial abnormal multiplier will not work for the bilevel dynamic problem, and the calmness is the right constraint qualification (see more discussion in [17, 18]). The purpose of this section is to derive a Kuhn–Tucker–type necessary optimality condition for the bilevel dynamic problem under a *calmness*-type constraint qualification. Our approach is to reformulate the original problem as an infinite-dimensional optimization problem and derive the desired result from the necessary optimality condition for this infinite-dimensional optimization problem. Formulation as an infinite-dimensional optimization problem takes care of the functional constraints. However, the usual Lagrange multiplier rule for infinite-dimensional optimization problems cannot be used here since the problem data are not Lipschitz in the control variable in the lower-level optimal control problem. Ioffe [8] derived a very general maximum principle for the standard optimal control problem by reduction to an infinite-dimensional optimization problem. We will use the result and approach of Ioffe to derive the necessary optimality condition of the maximum principle type for the bilevel dynamic problem.

DEFINITION 3.1. *Let (u^*, v^*) be an optimal strategy of P_1 (equivalently $\widetilde{P}_1$) and x^* the corresponding trajectory. $\widetilde{P}_1$ is said to be* partially calm *at (x^*, u^*, v^*) with modulus $\mu \geq 0$ if $\forall (x, u, v)$ satisfying*

$$\begin{aligned} &\dot{x}(t) = \phi(t, x(t), u(t), v(t)) \quad \text{a.e.}, \\ &x(t_0) = x_0, \\ &u(\cdot) \in L^2([t_0, t_1], U), v(\cdot) \in \mathcal{V}, \end{aligned}$$

we have

$$J_1(x, u, v) - J_1(x^*, u^*, v^*) + \mu(J_2(x, u, v) - V(u)) \geq 0,$$

where $\mathcal{V}$ denotes the collection of all admissible control functions for $P_2(u)$.

Define the *pseudo Hamiltonian* for problem $(\widetilde{P}_1)$ as

$$H_1(t, x, u, v, p_1; \lambda, r) := p_1 \cdot \phi(t, x, u, v) - rG(t, x, u, v) - \lambda F(t, x, u, v),$$

for $t \in [t_0, t_1]$, $x, p_1 \in \mathbb{R}^d$, $u \in R^n$, $v \in R^m$, $\lambda, r \in \mathbb{R}$.

THEOREM 3.2. *Assume that* (A1)–(A5) *hold. Let (x^*, u^*, v^*) be an optimal solution of P_1. Suppose that $\widetilde{P}_1$ is* partially calm *at (x^*, u^*, v^*) with modulus $\mu \geq 0$. Assume that the value function for the lower-level problem V is locally Lipschitz continuous near u^*. Then there exist $\lambda > 0, r = \lambda\mu$, and an arc p_1 such that*

$$\begin{aligned} -\dot{p_1}(t) &= \nabla_x H_1(t, x^*(t), u^*(t), v^*(t), p_1(t); \lambda, r) \qquad \text{a.e.}, \\ &\max_{v \in W(t)} H_1(t, x^*(t), u^*(t), v, p_1(t); \lambda, r) \end{aligned} \tag{13}$$

$$= H_1(t, x^*(t), u^*(t), v^*(t), p_1(t); \lambda, r) \quad \text{a.e.}, \tag{14}$$

$$-p_1(t_1) \in \lambda \partial f(x^*(t_1)) + r\partial g(x^*(t_1)), \tag{15}$$

$$\nabla_u H_1(\cdot, x^*(\cdot), u^*(\cdot), v^*(\cdot), p_1(\cdot); \lambda, r) \in -r\partial V(u^*) + N_{L^2([t_0, t_1], U)}(u^*). \tag{16}$$

Proof. Since $\widetilde{P}_1$ is partially calm at (x^*, u^*, v^*) with modulus μ, it is easy to see that (x^*, u^*, v^*) is also optimal for the following penalized problem:

$$\begin{aligned} P(\mu) \qquad \min \;& J_1(x, u, v) + \mu(J_2(x, u, v) - V(u)) \\ \text{s.t.} \;& \dot{x}(t) = \phi(t, x(t), u(t), v(t)) \qquad \text{a.e.}, \\ & x(t_0) = x_0, \\ & u(\cdot) \in L^2([t_0, t_1], U), \quad v(t) \in W(t) \qquad \text{a.e.}, \end{aligned}$$

which can be equivalently posed as the following problem:

$$\begin{aligned} \widehat{P}_1 \qquad \min \;& f(x(t_1)) + z(t_1) + \mu(g(x(t_1)) + y(t_1) - V(u)) \\ \text{s.t.} \;& \dot{x}(t) = \phi(t, x(t), u(t), v(t)) \qquad \text{a.e.}, \\ & \dot{y}(t) = G(t, x(t), u(t), v(t)) \qquad \text{a.e.}, \\ & \dot{z}(t) = F(t, x(t), u(t), v(t)) \qquad \text{a.e.}, \\ & v(t) \in W(t) \qquad \text{a.e.}, \\ & (x, y, z)(t_0) \in \{x_0\} \times \{0\} \times \{0\}. \end{aligned}$$

We now reformulate the above problem as an infinite-dimensional optimization problem. Let $C([t_0,t_1],R^n)$ be the space of continuous mappings from $[t_0,t_1]$ into R^n with the usual supremum norm. Set

$$\widetilde{x} := (x,y,z), \qquad \widetilde{\phi} := (\phi,G,F).$$

For $v(\cdot)\in\mathcal{V}$, the mapping $(\widetilde{x}(\cdot),u(\cdot))\to F_0(\widetilde{x}(\cdot),u(\cdot),v(\cdot))$ from $X := C([t_0,t_1],R^{d+2})$ $\times L^2([t_0,t_1],U)$ into $Y := C([t_0,t_1],R^{d+2})$:

$$F_0(\widetilde{x}(\cdot),u(\cdot),v(\cdot))(t) := \widetilde{x}(t)-\widetilde{x}(t_0)+\int_{t_0}^{t}\widetilde{\phi}(s,\widetilde{x}(s),u(s),v(s))ds$$

is well defined, continuously differentiable in $\widetilde{x}(\cdot)$, and Lipschitz continuous in $u(\cdot)$. Finally, let

$$f_0(\widetilde{x}(\cdot)) := f(x(t_1))+z(t_1), \tag{17}$$

$$G_0(\widetilde{x}(\cdot),u(\cdot)) := y(t_1)+g(x(t_1))-V(u), \tag{18}$$

$$S := \{\widetilde{x}\subset Y : x(t_0)=x_0, y(t_0)=0, z(t_0)=0\}.$$

Then problem $\widehat{P}_1$ is equivalent to the following infinite-dimensional optimization problem:

$$\begin{aligned} P_1{}' \qquad \min\ & f_0(\widetilde{x})+\mu G_0(\widetilde{x},u)\\ \text{s.t.}\ & F_0(\widetilde{x},u,v)=0,\\ & (\widetilde{x},u)\in S\times L^2([t_0,t_1],U),\\ & v\in\mathcal{V}. \end{aligned}$$

The above problem is in the form of a very general problem in section 4 of Ioffe [8]. Let the Lagrangian of the above problem be

$$L(\lambda,\alpha,\widetilde{x},u,v) := \lambda(f_0(\widetilde{x})+\mu G_0(\widetilde{x},u))+\langle\alpha,F_0(\widetilde{x},u,v)\rangle.$$

As in section 5 of Ioffe [8], the assumptions for [8, Theorem 2] can be verified. By [8, Theorem 2], if (x^*,u^*,v^*) is a local solution to $P_1{}'$, then there exist Lagrange multipliers $\lambda\ge 0$, $\alpha\in Y^*$ not all equal to zero such that

$$0\in\partial_{(\widetilde{x},u)}L(\lambda,\alpha,\widetilde{x}^*,u^*,v^*)+N_S(\widetilde{x}^*)\times N_{L^2([t_0,t_1],U)}(u^*), \tag{19}$$

$$L(\lambda,\alpha,\widetilde{x}^*,u^*,v^*)=\min_{v\in\mathcal{V}}L(\lambda,\alpha,\widetilde{x}^*,u^*,v), \tag{20}$$

where Y^* denotes the space of continuous linear functions on Y. Since f_0,G_0 are separable functions of $(\widetilde{x},u)$ (f_0 is independent of u and G_0 is the sum of a function independent of $\widetilde{x}$ and a function independent of u), by [15, Proposition 1.8], (19) implies that

$$\begin{aligned} 0\in{}&\lambda\partial f_0(\widetilde{x}^*)\times\{0\}+(\lambda\mu\partial_{\widetilde{x}}G_0(\widetilde{x}^*,u^*))\times(-\lambda\mu\partial V(u^*))\\ &+\partial_{(\widetilde{x},u)}\langle\alpha,F_0(\widetilde{x}^*,u^*,v^*)\rangle+N_S(\widetilde{x}^*)\times N_{L^2([t_0,t_1],U)}(u^*). \end{aligned} \tag{21}$$

Notice that $\langle \alpha, F_0(\widetilde{x}, u, v)\rangle$ can be represented as an integral functional on $X \times L^2$ by

$$\begin{aligned}&\langle \alpha, F_0(\widetilde{x}, u, v)\rangle\\ &= \int_{t_0}^{t_1} \langle \widetilde{x}(s) - \widetilde{x}(t_0), \xi(s)\rangle d\mu - \int_{t_0}^{t_1} \left\langle \int_t^{t_1} \xi(\tau) d\mu, \widetilde{\phi}(t, \widetilde{x}(t), u(t), v(t))\right\rangle dt,\end{aligned}$$

where the pair $(\mu, \xi(\cdot))$ represents the functional $\alpha \in Y^*$ (μ being a nonnegative Radon measure on $[t_0, t_1]$ and $\xi(\cdot) : [t_0, t_1] \to R^{d+2}$, μ-integrable); i.e.,

$$\int_{t_0}^{t_1} \langle \xi(t), y(t)\rangle d\mu = \langle \alpha, y(\cdot)\rangle \qquad \forall y(\cdot) \in Y.$$

Hence by Theorems 2.7.4 and 2.7.5 of [3] it is regular. Therefore, by [3, Proposition 2.3.15], (21) implies that

$$0 \in \lambda \partial f_0(\widetilde{x}^*) + \lambda\mu \partial_{\widetilde{x}} G_0(\widetilde{x}^*, u^*) + D_{\widetilde{x}} \langle \alpha, F_0(\widetilde{x}^*, u^*, v^*)\rangle + N_S(\widetilde{x}^*), \tag{22}$$

$$0 \in -\lambda\mu \partial V(u^*) + \partial_u \langle \alpha, F_0(\widetilde{x}^*, u^*, v^*)\rangle + N_{L^2([t_0,t_1],U)}(u^*), \tag{23}$$

where $D_{\widetilde{x}} \langle \alpha, F_0(\widetilde{x}, u, v)\rangle$ denotes the Gâteaux derivative of the functional $\langle \alpha, F_0(\widetilde{x}, u, v)\rangle$ with respect to $\widetilde{x}$.

Now let us analyze (22). We have that $\partial f_0(\widetilde{x}(\cdot))$ contains those $\beta \in Y^*$ that can be represented in the form

$$\langle \beta, h(\cdot)\rangle = \langle a, h(t_1)\rangle$$

for some $a \in \partial f(x(t_1)) \times \{0\} \times \{1\}$.

Similarly, $\partial_{\widetilde{x}} G_0(\widetilde{x}, u)$ contains those $\beta \in Y^*$ that can be represented in the form

$$\langle \beta, h(\cdot)\rangle = \langle b, h(t_1)\rangle$$

for some $b \in \partial g(x(t_1)) \times \{1\} \times \{0\}$.

Let $p(t) := \int_t^{t_1} \xi(\tau) d\mu$. Then p is an arc. For any $h \in X$,

$$\begin{aligned}\langle D_{\widetilde{x}} \langle \alpha, F_0(\widetilde{x}, u, v)\rangle, h(\cdot)\rangle &= \int_{t_0}^{t_1} \langle h(t) - h(t_0), \xi(t)\rangle d\mu\\ &\quad - \int_{t_0}^{t_1} \langle \nabla_{\widetilde{x}} \widetilde{\phi}(t, \widetilde{x}(t), u(t), v(t))^\top p(t), h(t)\rangle dt.\end{aligned}$$

$N_S(\widetilde{x})$ contains those $\beta \in Y^*$ that can be represented in the form

$$\langle \beta, h(\cdot)\rangle = \langle c, h(t_0)\rangle$$

for some $c \in N_{\{x_0\}\times\{0\}\times\{0\}}(\widetilde{x}(t_0))$.

Inclusion (22) yields the existence of

$$a \in \partial f(x^*(t_1)) \times \{0\} \times \{1\}, b \in \partial g(x^*(t_1)) \times \{1\} \times \{0\}, c \in N_{\{x_0\}\times\{0\}\times\{0\}}(\widetilde{x}^*(t_0))$$

such that

$$\begin{aligned}0 &= \lambda \langle a, h(t_1)\rangle + \lambda\mu \langle b, h(t_1)\rangle + \int_{t_0}^{t_1} \langle h(t) - h(t_0), \xi(t)\rangle d\mu\\ &\quad - \int_{t_0}^{t_1} \langle \nabla_{\widetilde{x}} \widetilde{\phi}(t, \widetilde{x}^*(t), u^*(t), v^*(t))^\top p(t), h(t)\rangle dt + \langle c, h(t_0)\rangle \qquad \forall h \in X.\end{aligned}$$

Let us denote $h = (h_1, h_2, h_3)$, $\xi = (\xi_1, \xi_2, \xi_3)$, $p = (p_1, p_2, p_3)$, where subscript 1 corresponds to vectors in R^d and subscripts $2, 3$ to vectors in R. In particular, if we choose $h(\cdot)$ that are absolutely continuous with $h(t_0) = 0, h_i(\cdot) = 0$ for $i = 1, 3$, we have

$$0 = \lambda\mu h_2(t_1) + \int_{t_0}^{t_1} h_2(t)\xi_2(t)d\mu,$$

which is equal to

$$0 = \int_{t_0}^{t_1} \left(\int_t^{t_1} \xi_2(s)d\mu + \lambda\mu \right) dh_2(t),$$

which implies that $p_2(t) = -\lambda\mu$.

Similarly, if we choose $h(\cdot)$ that are absolutely continuous with $h(t_0) = 0, h_i(\cdot) = 0$ for $i = 1, 2$, we have

$$0 = \lambda h_3(t_1) + \int_{t_0}^{t_1} h_3(t)\xi_3(t)d\mu,$$

which implies that $p_3(t) = -\lambda$.

If we choose $h(\cdot)$ that are absolutely continuous with $h(t_0) = 0, h_i(\cdot) = 0$ for $i = 2, 3$, we have

$$\begin{aligned} 0 = \lambda\langle a_1, h_1(t_1)\rangle + \lambda\mu\langle b_1, h_1(t_1)\rangle + \int_{t_0}^{t_1} \langle h_1(t), \xi_1(t)\rangle d\mu \\ - \int_{t_0}^{t_1} \langle \nabla_x \widetilde{\phi}(t, \widetilde{x}^*(t), u^*(t), v^*(t))^\top p(t), h_1(t)\rangle dt. \end{aligned}$$

Setting $-q = \lambda a_1 + \lambda\mu b_1$ and changing the order of integration, we obtain

$$\begin{aligned} 0 = \int_{t_0}^{t_1} \Big\langle \int_t^{t_1} \xi_1(t)d\mu + \nabla_x \phi(t, x^*(t), u^*(t), v^*(t))^\top p_1(t) \\ - \lambda\mu\nabla_x G(t, x^*(t), u^*(t), v^*(t)) - \lambda\nabla_x F(t, x^*(t), u^*(t), v^*(t)) - q, k(t) \Big\rangle dt, \end{aligned}$$

where $k(t) = \dot{h}(t)$ is an arbitrary integrable mapping. In view of the definition of $p_1(t)$, this implies

$$\begin{aligned} p_1(t) - q = -\int_t^{t_1} (\nabla_x \phi(s, x^*(s), u^*(s), v^*(s))^\top p_1(s) \\ + \lambda\mu\nabla_x G(s, x^*(s), u^*(s), v^*(s)) + \lambda\nabla_x F(s, x^*(s), u^*(s), v^*(s)))ds, \end{aligned}$$

from which we derive (13).

Let us now analyze (23). Since $\langle \alpha, F_0(\widetilde{x}, u, v)\rangle$ is an integral functional of u on L^2, it is not Gâteaux differentiable. However, under our assumptions, [3, Theorem 2.7.5] applies. Therefore, for $\beta \in \partial_u \langle \alpha, F_0(\widetilde{x}^*, u^*, v^*)\rangle$,

$$\langle \beta, h(\cdot)\rangle = -\int_{t_0}^{t_1} \langle \nabla_u \widetilde{\phi}(t, \widetilde{x}^*(t), u^*(t), v^*(t))^\top p(t), h(t)\rangle dt$$

for any $h \in L^2([t_0, t_1], R^n)$. Hence (23) implies (16).

We also have

$$p(t_1) = q \in -\lambda \partial f(x^*(t_1)) - \lambda\mu \partial g(x^*(t_1)).$$

That is (15).

Equation (20) implies that

$$-\int_{t_0}^{t_1} \langle p(t), \widetilde{\phi}(t, \widetilde{x}^*(t), u^*(t), v^*(t))\rangle dt \le -\int_{t_0}^{t_1} \langle p(t), \widetilde{\phi}(t, \widetilde{x}^*(t), u^*(t), v(t))\rangle dt.$$

Since $-\lambda = p_2(t), \lambda\mu = -p_3(t)$, the above inequality implies that

$$\int_{t_0}^{t_1} H_1(t, x^*(t), u^*(t), v^*(t), p_1(t); \lambda, \lambda\mu)dt \ge \int_{t_0}^{t_1} H_1(t, x^*(t), u^*(t), v(t), p_1(t); \lambda, \lambda\mu)dt$$

for any $v(\cdot) \in \mathcal{V}$. Since for any measurable set $E \subset [t_0, t_1]$,

$$v(\cdot) = \chi_E(\cdot)v(\cdot) + (1 - \chi_E(\cdot))v^*(\cdot),$$

where χ_E denotes the characteristic function of E, and belongs to $\mathcal{V}$ whenever $v(\cdot) \in \mathcal{V}$, it follows that

$$H_1(t, x^*(t), u^*(t), v^*(t), p_1(t); \lambda, \lambda\mu) \ge H_1(t, x^*(t), u^*(t), v(t), p_1(t); \lambda, \lambda\mu) \text{ a.e.}$$

for any $v(\cdot) \in \mathcal{V}$. From measurable selection theory, (14) follows.

Now we need to show that $\lambda \ne 0$. From the fact that λ and α are not all equal to zero, it follows easily that

$$\|p_1\|_\infty + \lambda > 0. \tag{24}$$

This condition prevents λ becoming zero. Indeed if $\lambda = 0$, then the transversality condition (15) would imply that $p_1(t_1) = 0$. This in turn implies that $p_1 \equiv 0$, which contradicts (24). The proof of the theorem is now complete. □

Combining Theorems 3.2 and 2.3, the following Kuhn–Tucker–type necessary optimality condition for the general bilevel dynamic problem is obtained.

THEOREM 3.3. *Assume* (A1)–(A8) *hold. Let* $(u^*(t), v^*(t))$ *be an optimal strategy of the bilevel dynamic problem* P_1 *and* $x^*(t)$ *the corresponding optimal trajectory. Suppose that* $\widetilde{P}_1$ *is partially calm at* (x^*, u^*, v^*) *with modulus* $\mu \ge 0$. *Then there exists arc* p_1 *such that*

$$-\dot{p_1}(t) = \nabla_x H_1(t, x^*(t), u^*(t), v^*(t), p_1(t); 1, \mu), \tag{25}$$

$$\max_{v \in W(t)} H_1(t, x^*(t), u^*(t), v, p_1(t); 1, \mu)$$

$$= H_1(t, x^*(t), u^*(t), v^*(t), p_1(t); 1, \mu) \text{ a.e.}, \tag{26}$$

$$-p_1(t_1) \in \partial f(x^*(t_1)) + \mu \partial g(x^*(t_1)), \tag{27}$$

$$\nabla_u H_1(\cdot, x^*(\cdot), u^*(\cdot), v^*(\cdot), p_1(\cdot); 1, \mu)$$

$$\in \mu \, clco \cup_{x \in Y_{u^*}} \{\zeta : \exists \text{ arc } p_2 \text{ s.t. } (-\dot{p_2}, \zeta, \dot{x}) \in \partial H_2(t, x, u^*, p_2) \quad \text{a.e.},$$

$$-p_2(t_1) \in \hat{\partial} g(x(t_1))\}$$

$$+ N_{L^2([t_0,t_1],U)}(u^*). \tag{28}$$

It is clear that in the minimax case (i.e., when $J_1 = -J_2$) and the trivial case (i.e., when $J_1 = J_2$), the calmness condition always holds with $\mu = 1$ and $\mu = 0$, respectively. We now give an example that satisfies the partial calmness condition.

Example 1. Consider the following bilevel dynamic problem:

$$\begin{aligned} &\min\ (x_1(1))^2 + (x_2(1))^2 \\ &\text{s.t. } u(t) \geq 0, (x, v) \in S(u), \end{aligned}$$

where $S(u)$ is the solution set of

$$\begin{aligned} &\min\ (x_1(1) + x_2(1))^3 \\ &\text{s.t. } \dot{x}_1(t) = u(t), \\ &\qquad \dot{x}_2(t) = v(t),\ v(t) \geq 0, \\ &\qquad x_1(0) = x_2(0) = 0. \end{aligned}$$

The solution of the above problem is $x^* = 0, u^* = 0, v^* = 0$. Since $J_1(x, u, v) \geq 0$ for all (x, u, v) that are admissible for $P(\mu)$ and $J(x^*, u^*, v^*) = 0$, it is easy to see that the above problem is partially calm.

As seen in Example 1, the calmness condition depends on knowledge of the optimal value of the dynamic bilevel problem. It is therefore important to find sufficient conditions for the calmness condition. For the static case, [17] identifies the existence of a uniformly weak sharp minimum as a sufficient condition for the calmness. It is shown in that paper that the bilevel programming problem in which the lower-level problem is linear is always calm, and sufficient conditions for the calmness of the bilevel problem where the lower-level problem is a linear quadratic problem are given.

To extend the definition of a uniform weak sharp minimum to our dynamic setting, we introduce the following notation. Given u, a control function for the upper level, let $\Omega(u)$ denote

$$\Omega(u) = \{(x, v) \in C([t_0, t_1], R^d) \times \mathcal{V} : \dot{x} = \phi(t, x, u, v), x(t_0) = x_0\}.$$

Let $S(u)$ denote the set of all solutions to problem $P_2(u)$. We say that the family of optimal control problems $\{P_2(u) : u \in L^2([t_0, t_1], U)\}$ has a uniformly weak sharp minimum with modulus $\alpha > 0$ if

$$d_{S(u)}(x, v) \leq \alpha(J_2(x, u, v) - V(u)) \qquad \forall (x, v) \in \Omega(u), u \in L^2([t_0, t_1], U),$$

where $d_{S(u)}(x, v)$ denotes the distance from (x, v) to the set $S(u)$. As in [17], we can show that a uniformly weak sharp minimum is a sufficient condition for partial calmness.

PROPOSITION 3.4. *In addition to* (A1) and (A7), *assume that for any* $u(\cdot)$ *there exists* $k(\cdot) \in L^1([t_0, t_1])$ *such that*

$$|F(t, x', u(t), v') - F(t, x'', u(t), v'')| \leq k(t)\|(x', v') - (x'', v'')\| \qquad \forall t \in [t_0, t_1], x', x'' \in R^d, v', v'' \in R^m$$

and that f *is Lipschitz continuous with constant* $L_f > 0$. *That* $\{P_2(u) : u \in L^2([t_0, t_1], U)\}$ *has a uniformly weak sharp minimum with modulus* α *implies that* $\widetilde{P}_1$ *is partially calm with modulus* $\mu \geq \alpha(\|k\|_1 + L_f)$ *at any solution of the problem.*

Proof. By the definition of a uniformly weak sharp minimum, there exists $\alpha > 0$ such that $\forall\ (x,v) \in \Omega(u), u \in L^2([t_0,t_1],U)$,

$$\begin{aligned} J_2(x,u,v) - V(u) &\geq (1/\alpha) d_{S(u)}(x,v) \\ &= (1/\alpha)|(x,v) - (x(u),v(u))|, \end{aligned}$$

where $(x(u),v(u))$ is the metric projection of (x,v) onto the set $S(u)$. Let (x^*,u^*,v^*) be any solution of the problem P_1. The assumptions imply that $J_1(x,u,v)$ is Lipschitz continuous in (x,v) uniformly in u with constant $L_1 = \|k\|_1 + L_f$. It follows that

$$\begin{aligned} J_2(x,u,v) - V(u) &\geq \frac{1}{\alpha} d_{S(u)}(x,v) \\ &= \frac{1}{\alpha}|(x,v) - (x(u),v(u))| \\ &\geq \frac{1}{\alpha L_1}(J_1(x,u,v) - J_1(x(u),u,v(u))) \\ &\geq \frac{1}{\alpha L_1}(J_1(x,u,v) - J_1(x^*,u^*,v^*)) \\ &\geq \frac{1}{\mu}(J_1(x,u,v) - J_1(x^*,u^*,v^*)). \end{aligned}$$

Therefore, we see that $\widetilde{P}_1$ is partially calm at any solution of the problem with modulus $\mu \geq \alpha L_1$. □

The following result is a sufficient condition for a uniformly weak sharp minimum. The proof technique follows from a result about regular points due to Ioffe (Theorem 1 and Corollary 1.1 of [7]).

PROPOSITION 3.5. *Suppose that $J_2(x,u,v)$ is Lipschitz continuous in (x,v) uniformly in u with constant $L > 0$. If there exists a constant $c > 0$ such that $\|\xi + \eta\| \geq c$ whenever $\xi \in \partial_{(x,v)} J_2(x,u,v)$, $\eta \in (L+1)\partial d_{\Omega(u)}(x,v)$ (or $\eta \in N_{\Omega(u)}(x,v)$)$\forall (x,v) \in \Omega(u)$ such that $(x,v) \notin S(u)\ \forall$ admissible controls u, then*

$$d_{S(u)}(x,v) \leq (1/c)(J_2(x,u,v) - V(u)) \forall (x,v) \in \Omega(u).$$

Proof. Assume that the statement is false. Then there is $u \in L^2([t_0.t_1],U)$ and $(x,v) \in \Omega(u)$ such that

$$d_{S(u)}(x,v) > \frac{1}{c}(J_2(x,u,v) - V(u)).$$

We can obviously choose $\delta > 1$ to make the following inequality valid:

$$d_{S(u)}(x,v) > \frac{\delta}{c}(J_2(x,u,v) - V(u)) := \gamma. \tag{29}$$

It is also obvious that

$$J_2(x,u,v) - V(u) \leq \inf_{(x,v)\in\Omega(u)} (J_2(x,u,v) - V(u)) + \frac{c\gamma}{\delta}.$$

Let δ_S denote the indicator function of set S. Applying the Ekeland variational principle [3, Theorem 7.5.1] with $F(x', v') := J_2(x', u, v') - V(u) + \delta_{\Omega(u)}(x', v')$, $\epsilon = \gamma c/\delta$, and $\lambda = \gamma$, we find $(\widetilde{x}, \widetilde{v}) \in \Omega(u)$ such that

$$\|(\widetilde{x}, \widetilde{v}) - (x, v)\| \leq \gamma \tag{30}$$

and

$$\phi(x', v') := J_2(x', u, v') - V(u) + (c/\delta)\|(x', v') - (\widetilde{x}, \widetilde{v})\|$$

attains its minimum on $\Omega(u)$ at $(\widetilde{x}, \widetilde{v})$. It follows that

$$\begin{aligned} 0 &\in \partial\phi(\widetilde{x}, \widetilde{v}) + (L+1)\partial d_{S(u)}(\widetilde{x}, \widetilde{v}) \\ &\subset \partial_{(x,v)} J_2(\widetilde{x}, u, \widetilde{v}) + (c/\delta)B + (L+1)\partial d_{S(u)}(\widetilde{x}, \widetilde{v}). \end{aligned}$$

Thus there exist

$$\xi \in \partial_{(x,v)} J_2(\widetilde{x}, u, \widetilde{v}), \quad \eta \in (L+1)\partial d_{S(u)}(\widetilde{x}, \widetilde{v})$$

such that

$$\|\xi + \eta\| \leq c/\delta < c. \tag{31}$$

According to (29), (30), and $(\widetilde{x}, \widetilde{v}) \in \Omega(u)$, we have that

$$(\widetilde{x}, \widetilde{v}) \notin S(u).$$

Therefore (31) contradicts the assumption. The proof of the proposition is then complete. □

We now use an example to illustrate the application of the above result. It is different from Example 1 only in the lower-level objective function.

Example 2. Consider the following bilevel dynamic problem:

$$\begin{aligned} &\min\ (x_1(1))^2 + (x_2(1))^2 \\ &\text{s.t. } u(t) \geq 0, (x, v) \in S(u), \end{aligned}$$

where $S(u)$ is the solution set of

$$\begin{aligned} &\min\ x_1(1) + x_2(1) + (x_1(1) + x_2(1))^3 \\ &\text{s.t. } \dot{x}_1(t) = u(t), \\ &\qquad \dot{x}_2(t) = v(t),\ v(t) \geq 0, \\ &\qquad x_1(0) = x_2(0) = 0. \end{aligned}$$

It is easy to see that $\Omega(u) = \{x_1 : x_1(t) = \int_0^t u(s)ds\} \times \{x_2 : x_2(t) \geq 0\} \times \{v : v(t) \geq 0\}$ and $S(u) = \{(x, v) : x_1(t) = \int_0^t u(s)ds, x_2 \equiv 0, v \equiv 0\}\ \forall\ u(t) \geq 0$. Since

$$\begin{aligned} \partial_{(x,v)} J_2(x, u, v) = \{(\xi_1, \xi_2, 0) :& \langle \xi_1, h(\cdot)\rangle = ((1 + 3(x_1(1) + x_2(1))^2)h(1), \\ & \langle \xi_2, h(\cdot)\rangle = (1 + 3(x_1(1) + x_2(1))^2)h(1)\ \forall\ h \in C[0,1]\}, \end{aligned}$$

and $N_{\Omega(u)}(x, v) = N_{\{x_1 : x_1(t) = \int_0^t u(s)ds\}}(x_1) \times \{0\} \times N_{\{v : v(t) \geq 0\}}(v)$ for any $(x_1, x_2, v) \notin S(u)$, it is easy to see that the assumptions in Proposition 3.5 are satisfied.

We now calculate that

$$H_1(t,x,u,v,p_1;1,\mu) = p_1^1 u + p_1^2 v, \qquad H_2(t,x,u,v,p_2) = \sup\{p_2^1 u + p_2^2 v : v \geq 0\}.$$

Since H_2 is independent of x, (28) implies that there exists an arc p_2 such that

$$\dot{p}_2(t) = 0, \tag{32}$$

$$-p_2(1) = (1 + 3(x_1^*(1) + x_2^*(1))^2, 1 + 3(x_1^*(1) + x_2^*(1))^2), \tag{33}$$

$$p_1^1 - \mu p_2^1 \in N_{\{u \in C[0,1]: u \geq 0\}}(u^*). \tag{34}$$

Observing that $x_1(0) = x_2(0) = 0$ and both $x_1(t)$ and $x_2(t)$ are nondecreasing, we derive from (32) and (33) that $p_2^1 = p_2^2 \leq -1$ are constants. Hence $H_2 = p_2^1 u$, which implies from (28) that $\dot{x}_2^*(t) = 0$. That is $x_2^* \equiv 0$. If $u^* \not\equiv 0$ then (34) implies that

$$p_1^1 = \mu p_2^1.$$

But by (25) and (27), p_1^1 and p_1^2 are nonpositive constants and

$$-2x_1^*(1) - \mu(1 + 3[x_1^*(1) + x_2^*(1)]^2) = -\mu(1 + 3[x_1^*(1) + x_2^*(1)]^2),$$

which implies that $x_1^*(1) = 0$. But this is a contradiction. Therefore $u^* \equiv 0, v^* \equiv 0$ is a candidate for an optimal solution since $x_1(0) = x_2(0) = 0$ and both $x_1(t)$ and $x_2(t)$ are nondecreasing. It is not hard to check that it is indeed a solution. Notice that in Example 2 the lower-level problem is not convex and hence is out of the scope of any currently available control theory.

Finally, we use another example to illustrate applications of Theorem 3.2 in solving bilevel dynamic problems in the absence of the calmness condition. Example 3 shows that even without the calmness condition, the necessary condition that we derived may be used to find condition for the existence of a normal multiplier.

Example 3. Consider the following bilevel dynamic problem with linear-quadratic cost functions on the interval $[0,1]$, where

$$F(t,x,u,v) = \frac{1}{2}[x^\top Q_1 x + u^\top R_{11} u + v^\top R_{12} v],$$

$$f(x) = \frac{1}{2} x^\top K_1 x,$$

$$G(t,x,u,v) = \frac{1}{2}[x^\top Q_2 x + u^\top R_{21} u + v^\top R_{22} v],$$

$$g(x) = \frac{1}{2} x^\top K_2 x,$$

$$\phi(t,x,u,v) = A(t)x + B(t)u + C(t)v,$$

where $x \in R^d$, $u \in R^n$, $v \in R^m$, Q_1, Q_2, K_1, K_2 are positive semidefinite matrices and $R_{22}, R_{11}, rR_{22} + R_{12}$, where $r \geq 0$ is any constant, are positive definite matrices with appropriate order; R_{21} is a $n \times n$ matrix; $A(t), B(t)$, and $C(t)$ are matrices with continuous components.

We can calculate

$$\begin{aligned}
H_1(t,x,u,v,p_1;1,\mu) &= p_1^\top \phi - \mu G - F \\
&= p_1^\top (A(t)x + B(t)u + C(t)v) \\
&\quad -\frac{1}{2}\mu[x^\top Q_2 x + u^\top R_{21} u + v^\top R_{22} v] \\
&\quad -\frac{1}{2}[x^\top Q_1 x + u^\top R_{11} u + v^\top R_{12} v], \\
H_2(t,x,u,p_2) &= \sup_v \{p_2^\top \phi - G\} \\
&= \sup_v \{p_2^\top (A(t)x + B(t)u + C(t)v) - \frac{1}{2}[x^\top Q_2 x + u^\top R_{21} u + v^\top R_{22} v]\} \\
&= p_2^\top (A(t)x + B(t)u + C(t)R_{22}^{-1} C(t)^\top p_2) \\
&\quad -\frac{1}{2}[x^\top Q_2 x + u^\top R_{21} u + p_2^\top C(t) R_{22}^{-1} C(t)^\top p_2].
\end{aligned}$$

Suppose that (u^*, v^*) is an optimal control pair and x^* is the corresponding trajectory. If the conclusion of Theorem 3.3 holds, then there exist adjoint arcs p_1, p_2 and constant $\mu \geq 0$ such that

$$\begin{aligned}
-\dot{p}_1 &= A(t)^\top p_1 - [\mu Q_2 + Q_1] x^*, \\
-p_1(1) &= [\mu K_2 + K_1] x^*(1), \\
-\dot{p}_2 &= A(t)^\top p_2 - Q_2^\top x^*, \\
-p_2(1) &= K_2 x^*(1),
\end{aligned}$$

$$\dot{x}^* = A(t)x^* + B(t)u^* + C(t)v^*, \tag{35}$$

$$B(t)^\top p_1 - R_{11} u^* = \mu B(t)^\top p_2, \tag{36}$$

$$v^*(t) = R_{22}^{-1} C(t)^\top p_2 = [\mu R_{22} + R_{12}]^{-1} C(t)^\top p_1. \tag{37}$$

Equation (36) implies that

$$u^*(t) = R_{11}^{-1} B(t)^\top (p_1 - \mu p_2). \tag{38}$$

Substituting (37) and (38) into (35) yields

$$\dot{x}^* = A(t)x^* + B(t)R_{11}^{-1} B(t)^\top (p_1 - \mu p_2) + C(t) R_{22}^{-1} C(t)^\top p_2. \tag{39}$$

Thus, for any $\mu \geq 0$ that satisfies (37), i.e.,

$$R_{22}^{-1} C(t)^\top p_2 = [\mu R_{22} + R_{12}]^{-1} C(t)^\top p_1, \tag{40}$$

we obtain a set of $3d$ equations with equal numbers of unknowns.

$$\begin{aligned}
\dot{x}^* &= A(t)x^* + B(t)R_{11}^{-1} B(t)^\top (p_1 - \mu p_2) + C(t) R_{22}^{-1} C(t)^\top p_2, \\
-\dot{p}_1 &= A(t)^\top p_1 - [\mu Q_2^\top + Q_1^\top] x^*, \\
-\dot{p}_2 &= A(t)^\top p_2 - Q_2^\top x^*.
\end{aligned}$$

Assume that

$$p_1(t) = \psi_1(t)x^*(t),$$
$$p_2(t) = \psi_2(t)x^*(t),$$

where ψ_i are matrices that satisfy the end point conditions

$$\psi_1(1) = -[K_1 + \mu K_2], \quad \psi_2(1) = -K_2$$

to be determined. Differentiating them with respect to t gives

$$\dot{p}_1 = \dot{\psi}_1 x^*(t) + \psi_1 \dot{x}^*(t),$$
$$\dot{p}_2 = \dot{\psi}_2 x^*(t) + \psi_2 \dot{x}^*(t).$$

Substituting for $\dot{x}^*, \dot{p}_1$, and $\dot{p}_2$ gives

$$\dot{\psi}_1 = -A(t)^\top \psi_1 - \psi_1 A(t) + \mu Q_2 + Q_1 - \psi_1 B(t) R_{11}^{-1} B(t)^\top (\psi_1 - \mu\psi_2) - \psi_1 C(t) R_{22}^{-1} C(t)^\top \psi_2,$$
$$\dot{\psi}_2 = -A(t)^\top \psi_2 - \psi_2 A(t) + Q_2 - \psi_2 B(t) R_{11}^{-1} B(t)^\top (\psi_1 - \mu\psi_2) - \psi_2 C(t) R_{22}^{-1} C(t)^\top \psi_2.$$

Moreover,

$$u^*(t) = R_{11}^{-1} B(t)^\top (\psi_1(t) - \mu\psi_2(t))x(t),$$
$$v^*(t) = R_{22}^{-1} C(t)^\top \psi_2(t) x(t).$$

Let $\psi_3 = \psi_1 - \mu\psi_2$. Then provided that there exists $\mu \geq 0$ that satisfies

$$R_{22}^{-1} C(t)^\top \psi_2 = [\mu R_{22} + R_{12}]^{-1} C(t)^\top (\psi_3 + \mu\psi_2) \tag{41}$$

we obtain

$$u^*(t) = R_{11}^{-1} B(t)^\top \psi_3(t) x(t), \tag{42}$$
$$v^*(t) = R_{22}^{-1} C(t)^\top \psi_2(t) x(t), \tag{43}$$

where ψ_3 and ψ_2 are solutions to

$$\dot{\psi}_3 = -A(t)^\top \psi_3 - \psi_3 A(t) + Q_1 - \psi_3 B(t) R_{11}^{-1} B(t)^\top \psi_3 - \psi_3 C(t) R_{22}^{-1} C(t)^\top \psi_2,$$
$$\dot{\psi}_2 = -A(t)^\top \psi_2 - \psi_2 A(t) + Q_2 - \psi_2 B(t) R_{11}^{-1} B(t)^\top \psi_3 - \psi_2 C(t) R_{22}^{-1} C(t)^\top \psi_2,$$

with end point conditions

$$\psi_3(1) = -K_1, \quad \psi_2(1) = -K_2.$$

It is clear that the existence of $\mu \geq 0$ that satisfies the equality (41) is a constraint qualification for ensuring the existence of normal multipliers for the class of linear-quadratic bilevel problems. Such $\mu \geq 0$ exists, for example, when

$$K_1 = 0, \quad Q_1 = 0, \quad R_{12} = 0$$

or

$$K_1 = K_2 = 0, \qquad Q_1 = Q_2 = 0.$$

Acknowledgments. The author would like to thank Qiji Zhu for suggestions which led to improvements in Theorem 2.3 of this paper. The author would also like to thank anonymous referees for comments on an early version of this paper that helped to improve the exposition.

REFERENCES

[1] J. F. Bard and J. E. Falk, *An explicit solution to the multi-level programming problem*, Oper. Res., 9 (1982), pp. 77–100.
[2] C. I. Chen and J. B. Cruz Jr., *Stackelberg solution for two-person games with biased information patterns*, IEEE Trans. Automat. Control, 6 (1972), pp. 791–798.
[3] F. H. Clarke, *Optimization and Nonsmooth Analysis*. Wiley-Interscience, New York, 1983.
[4] F. H. Clarke, *Methods of Dynamic and Nonsmooth Optimization*, NSF-CBMS Regional Conference Series in Applied Mathematics 57, SIAM, Philadelphia, PA, 1989.
[5] F. H. Clarke, *Perturbed optimal control problems*, IEEE Trans. Automat. Control, 6 (1986), pp 535–542.
[6] F. H. Clarke, R. J. Stern, and P. R. Wolenski, *Subgradient criteria for monotonicity, the Lipschitz condition and convexity*, Canad. J. Math., 45 (1993), pp. 1167–1183.
[7] A. D. Ioffe, *Regular points of Lipschitz functions*, Trans. Amer. Math. Soc., 251 (1979), pp. 61–69.
[8] A. D. Ioffe, *Necessary conditions in nonsmooth optimization*, Math. Oper. Res., 9 (1984), pp. 159–189.
[9] A. Y. Kruger and B. S. Mordukhovich, *Minimization of nonsmooth functionals in optimal control problems*, Engrg. Cybernetics, 16 (1978), pp. 126–133.
[10] B. S. Mordukhovich, *Maximum principle in problems of time optimal control with nonsmooth constraints*, J. Appl. Math. Mech., 40 (1976), pp. 960–969.
[11] J. Outrata, *A note on the usage of nondifferentiable exact penalties in some special optimization problems*, Kybernetika, 24 (1988), pp. 251–258.
[12] J. Outrata, *On the numerical solution of a class of Stackelberg problems*, Z. Oper. Res., 34 (1990), pp. 255–277.
[13] R. T. Rockafellar, *Extensions of subgradient calculus with applications to optimization*, Nonlinear Anal., 9 (1985), pp. 665–698.
[14] H. von Stackelberg, *The Theory of the Market Economy*, Oxford University Press, Oxford, UK, 1952.
[15] J. J. Ye, *Optimal Control of Piecewise Deterministic Markov Processes*, Ph.D. thesis, Department of Mathematics and Statistics, Dalhousie University, Halifax, Canada, 1990.
[16] J. J. Ye, *Necessary conditions for bilevel dynamic optimization problems*, SIAM J. Control Optim., 33 (1995), pp. 1208–1223.
[17] J. J. Ye and D. L. Zhu, *Optimality conditions for bilevel programming problems*, Optimization, 33 (1995), pp. 9–27.
[18] J. J. Ye, D. L. Zhu, and Q. J. Zhu, *Exact penalization and necessary optimality conditions for generalized bilevel programming problems*, SIAM J. Optim., 7 (1997), to appear.
[19] L. C. Young, *Lectures on the Calculus of Variations and Optimal Control Theory*, W. B. Saunders, Philadelphia, PA, 1969.
[20] R. Zhang, *Problems of Hierarchical Optimization: Nonsmoothness and Analysis of Solutions*, Ph.D. thesis, Department of Applied Mathematics, University of Washington, Seattle, WA, 1990.
[21] R. Zhang, *Problems of hierarchical optimization in finite dimensions*, SIAM J. Optim., 4 (1994), pp. 521–536.
[22] J. Warga, *Optimal Control of Differential and Functional Equations*, Academic Press, New York, 1972.

SIAM J. CONTROL OPTIM.
Vol. 35, No. 2, pp. 532–551, March 1997

009

A PREDICTOR-CORRECTOR ALGORITHM FOR A CLASS OF NONLINEAR SADDLE POINT PROBLEMS*

JIE SUN†, JISHAN ZHU‡, AND GONGYUN ZHAO§

Abstract. An interior path-following algorithm is proposed for solving the nonlinear saddle point problem

$$\text{minimax } c^T x + \phi(x) + b^T y - \psi(y) - y^T A x$$
$$\text{subject to } (x, y) \in \mathcal{X} \times \mathcal{Y} \subset R^n \times R^m,$$

where $\phi(x)$ and $\psi(y)$ are smooth convex functions and $\mathcal{X}$ and $\mathcal{Y}$ are boxes (hyperrectangles). This problem is closely related to the models in stochastic programming and optimal control studied by Rockafellar and Wets (*Math. Programming Studies*, 28 (1986), pp. 63–93; *SIAM J. Control Optim.*, 28 (1990), pp. 810–822). Existence and error-bound results on a central path are derived. Starting from an initial solution near the central path with duality gap $O(\mu)$, the algorithm finds an ϵ-optimal solution of the problem in $O(\sqrt{m+n}\,|\log \mu/\epsilon|)$ iterations if both $\phi(x)$ and $\psi(y)$ satisfy a scaled Lipschitz condition.

Key words. interior point methods, nonlinear complementarity problem, optimal control, saddle point problem, stochastic programming

AMS subject classifications. 49J35, 65K10, 90C06, 90C15, 90C33

PII. S0363012994276111

1. Introduction.

We discuss an interior path-following scheme for solving a class of nonlinear saddle point problems in the following form:

$$\text{(1.1)} \quad \begin{cases} \text{Find a saddle point for} & l(x,y) \equiv c^T x + \phi(x) + b^T y - \psi(y) - y^T A x \\ \text{subject to} & x \in \mathcal{X} \subset R^n, y \in \mathcal{Y} \subset R^m, \end{cases}$$

where $\phi(x)$ and $\psi(y)$ are C^2-convex functions, $c \in R^n$, $b \in R^m$, $A \in R^{m\times n}$, and the superscript T represents transpose. The sets of $\mathcal{X}$ and $\mathcal{Y}$ are boxes (hyperrectangles).

According to convex analysis [17], problem (1.1) has a pair of associated optimization problems: the primal problem

$$\text{(1.2)} \qquad \text{minimize}_{x\in\mathcal{X}}\ f(x) \equiv c^T x + \phi(x) + \sup_{y\in\mathcal{Y}}\{(b - Ax)^T y - \psi(y)\}$$

and the dual problem

$$\text{(1.3)} \qquad \text{maximize}_{y\in\mathcal{Y}}\ g(y) \equiv b^T y - \psi(y) - \sup_{x\in\mathcal{X}}\{(A^T y - c)^T x - \phi(x)\}.$$

Note that the function $f(x)$ (called the primal objective function) is convex, the function $g(y)$ (called the dual objective function) is concave, and both are nondifferentiable in general. (For a detailed analysis for the case that both $\phi(x)$ and $\psi(y)$ are quadratic,

*Received by the editors October 24, 1994; accepted for publication (in revised form) February 2, 1996. This research was supported in part by grants RP-920068 and RP-930033 from the National University of Singapore.

http://www.siam.org/journals/sicon/35-2/27611.html

†Department of Decision Sciences, National University of Singapore, Singapore 119260 (fbasunj@nus.sg).

‡Newbury College, Brookline, MA 02146.

§Department of Mathematics, National University of Singapore, Singapore 119260 (matzgy@nus.sg).

see [18], where $f(x)$ and $-g(y)$ turn out to be "piecewise quadratic" convex functions.) Therefore, problems (1.1)–(1.3) can be categorized as nonsmooth convex programming problems. Some fundamental duality relationships among (1.1), (1.2), and (1.3) have been established in [17] which include existence results on the saddle point(s) of (1.1) and the saddle point value—the common optimal value of (1.2) and (1.3).

Problems (1.1)–(1.3) stem from a development beyond the conventional formulation of optimization problems. These models provide a framework that allows penalty representations of constraints as well as accommodates other sources of nonsmoothness such as objectives produced by multistage optimization problems. For instance, linearly constrained convex optimization problems are usually posed in the form

$$\text{minimize } \ \phi(x) \ \text{ subject to } Ax \geq b, x \geq 0.$$

The corresponding Lagrangian saddle point problem is

$$\text{minimax}_{x\geq 0, y\geq 0} \ \phi(x) + y^T(b - Ax),$$

which is a special case of (1.1). Now let $\psi(y)$ be a convex function such that $\psi(0) = 0$ and $\psi(y) \geq 0$ and add $-\psi(y)$ to the Lagrangian function. Then the corresponding primal program becomes

$$\text{minimize } \ \phi(x) + \sup_{y\geq 0}\{y^T(b - Ax) - \psi(y)\} \ \text{ subject to } x \geq 0,$$

where the function $\sup_{y\geq 0}\{y^T(b-Ax)-\psi(y)\}$ is equal to zero for x satisfying $Ax \geq b$ and is greater or equal to zero for all x. Thus in this formulation the exact constraint $Ax \geq b$ is replaced by a penalty representation that allows the modeler to deal with $Ax \geq b$ more flexibly by selecting suitable $\psi(y)$ and $\mathcal{Y}$.

As an example of how multistage optimization could fit into the form of (1.1), we consider a two-stage stochastic programming model studied by Rockafellar and Wets [23]. At the first (current) stage, a decision $x \in R^n$ has to be made, incurring a direct cost $c^T x + \phi(x)$, subject to $x \in \mathcal{X}$, where $\mathcal{X}$ is a box. At the second (future) stage, a random event is observed with outcome $\omega \in \Omega$, where Ω is a probability space. The decision x and the outcome ω then determine an additional "recourse cost" $\rho_\omega(x)$. Under the practical circumstances such as soft constraints [18, 20] and simple recourse [31], the function $\rho_\omega(x)$ is expressed by an optimal value function as follows:

$$\rho_\omega(x) = \sup_{y_\omega \in \mathcal{Y}_\omega} \{y_\omega^T(b_\omega - A_\omega x) - \psi_\omega(y_\omega)\}. \tag{1.4}$$

In principle, the set $\mathcal{Y}_\omega$, the vector b_ω, the matrix A_ω, and the function ψ_ω are allowed to be random. The objective in this model is to make the best decision x with respect to the present cost and constraints as well as the expected cost $E_\omega[\rho_\omega(x)]$ and certain induced constraints. Assuming finite discrete distribution for ω, the decision problem can be described by

$$\text{minimize } \ c^T x + \phi(x) + \sum_{\omega\in\Omega} \pi_\omega \rho_\omega(x) \text{ subject to } x \in \mathcal{X}, \tag{1.5}$$

where π_ω is the probability of the random event ω. Now let

$$\mathcal{Y} = \prod_{\omega\in\Omega} \mathcal{Y}_\omega, \ b = \begin{bmatrix} \cdot \\ \cdot \\ \pi_\omega b_\omega \\ \cdot \\ \cdot \end{bmatrix}, \ A = \begin{bmatrix} \cdot \\ \cdot \\ \pi_\omega A_\omega \\ \cdot \\ \cdot \end{bmatrix}, \text{ and } \psi(y) = \sum_{\omega\in\Omega} \pi_\omega \psi_\omega(y_\omega).$$

Then problem (1.5) takes the form of (1.2).

More about the motivation of models (1.1)–(1.3) and their applications in stochastic programming and optimal control can be found in a series of pioneer papers of Rockafellar and Wets [18, 19, 20, 21, 23, 24].

Algorithms for linear-quadratic cases of problem (1.1) in which both $\phi(x)$ and $\psi(y)$ are linear or quadratic have been studied extensively. Among them are the L-shaped method [30], the decomposition methods [1, 5], the finite generation method [23], the projected gradient method [37], the steepest descent method [36], the sequential quadratic programming method [16], and some interior point methods [2, 3, 27, 32, 33]. For the special case where both $\mathcal{X}$ and $\mathcal{Y}$ are boxes and both $\phi(x)$ and $\psi(y)$ are separable quadratic functions, a simplex–active-set method has been developed [22]. The more general convex case of (1.1), however, has not yet received enough attention in algorithmic development, although the problem can be traced back to the extended Fenchel duality model in the 1970s [17].

The predictor-corrector algorithm studied in this paper is rooted in the primal-dual path-following algorithm of Kojima, Mizuno, and Yoshise [10] for linear complementarity problems and the predictor-corrector method of Mizuno, Todd, and Ye [14] for linear programming. The algorithms in [10] and [14] and their variants have been extensively studied on their global, local, and computational behaviors in the context of linear complementarity problems. It is not possible for us to point out all references on this method; the interested reader may look into the tutorial paper of Gonzaga [6] and recent papers such as [9, 11, 12, 25, 34, 35] and the references therein.

The purpose of writing this paper is threefold.

(i) We develop a predictor-corrector algorithm for problem (1.1) together with existence and error-bound results on a central path. In particular we show that if functions $\phi(x)$ and $\psi(y)$ satisfy a scaled Lipschitz condition, then starting from an initial solution near the central path, the algorithm converges to an ϵ-optimal solution of the problem in $O\left(\sqrt{m+n}\,|\log \mu_0/\epsilon|\right)$ iterations, where μ_0 is a number related to the initial duality gap.

(ii) To the research community in stochastic programming and optimal control, we demonstrate that the interior point method can be used as one of the theoretically efficient methods for convex-concave saddle point problems. We also suggest ways to take computational advantage of the special structure of problem (1.1) arising from stochastic programming and optimal control.

(iii) To the research community of interior point methods, this paper contributes a polynomial algorithm for a class of monotone complementarity problems (MCP). To our knowledge, only few algorithms for nonlinear MCP have been shown to have polynomial complexity (Nesterov and Nemirovskii [15] and Tseng [28]) under different conditions. The scaled Lipschitz condition used in this paper, as shown in [38], is satisfied by a fairly large class of convex functions.

This paper is organized as follows. In section 2 we discuss the conditions for the existence of the optimal solution(s) and the central path. We establish an estimate on the duality gap if (x, y) is near the central path. These results form a foundation for the interior path-following method for problem (1.1). Section 3 is devoted to the proof of polynomial convergence of the predictor-corrector algorithm. Finally, section 4 contains some concluding remarks.

2. Results on feasibility, existence, and error bounds related to the central path. In the following analysis, we assume that both $\mathcal{X}$ and $\mathcal{Y}$ are nonnegative orthants in order to simplify the statements. This is not an essential change from assuming that $\mathcal{X}$ and $\mathcal{Y}$ are boxes. We will elaborate this point at the end of section 3.

Suppose that a saddle point (x^*, y^*) of (1.1) exists; that is,

$$l(x^*, y) \le l(x^*, y^*) \le l(x, y^*) \text{ for all } x \in \mathcal{X} \text{ and } y \in \mathcal{Y}.$$

It can be shown [17] that x^* and y^* must be optimal solutions of (1.2) and (1.3), respectively, and that $f(x^*) = l(x^*, y^*) = g(y^*)$ and vice versa. In addition, the saddle point condition is equivalent to the following variational relationships:

$$-\nabla_x l(x^*, y^*) \in N_{\mathcal{X}}(x^*) \text{ and } \nabla_y l(x^*, y^*) \in N_{\mathcal{Y}}(y^*), \tag{2.1}$$

where $N_{\mathcal{X}}(x^*)$ stands for the normal cone of $\mathcal{X}$ at x^*. The notation $N_{\mathcal{Y}}(y^*)$ has a similar meaning. We assume that both $\phi(x)$ and $\psi(y)$ are finite and twice continuously differentiable on $\mathcal{X}$ and $\mathcal{Y}$, respectively. By introducing auxiliary vectors w^* and s^*, conditions (2.1) can be written in an explicit form as follows:

$$\begin{cases} \nabla\phi(x^*) - A^T y^* - w^* = -c, \\ Ax^* + \nabla\psi(y^*) - s^* = b, \\ (x^*)^T w^* = (y^*)^T s^* = 0, \\ x^*, y^*, w^*, s^* \ge 0. \end{cases} \tag{2.2}$$

The proposed algorithm finds a point $(x, y) \in \mathcal{X} \times \mathcal{Y}$ such that $f(x) - g(y) \le \epsilon$ by finding a sequence of approximate solutions of the following system as $\mu \downarrow 0$:

$$\begin{cases} \nabla\phi(x) - A^T y - w = -c, \\ Ax + \nabla\psi(y) - s = b, \\ x_j w_j = \mu, \quad j = 1, ..., n, \\ y_i s_i = \mu, \quad i = 1, ..., m, \\ x, w, y, s \ge 0. \end{cases} \tag{2.3}$$

Similar to the deduction of system (2.2), it can be shown that (x, y, w, s) is a solution to (2.3) if and only if (x, y) is a saddle point of

$$l_\mu(x, y) \equiv c^T x + \phi(x) + b^T y - \psi(y) - y^T A x - \mu \sum_{j=1}^{n} \log x_j + \mu \sum_{i=1}^{m} \log y_i$$

over $\mathcal{X} \times \mathcal{Y}$.

Naturally, as an interior point method, the proposed algorithm needs some kind of interior points from which to start. We make the following assumption.

Assumption 2.1. There is a quadruple $(x, y, w, s) > 0$ such that the first two equations of system (2.3) are satisfied.

Under this assumption, we can prove that the iterates generated by our algorithm are feasible solutions to problems (1.2) and (1.3) and that the solutions to problem (2.3) exist for all $\mu \ge 0$.

We first discuss the feasibility problem. An x is *feasible* to (1.2) if $x \ge 0$ and $f(x) < \infty$; a y is *feasible* to (1.3) if $y \ge 0$ and $g(y) > -\infty$. We have the following result under an assumption weaker than Assumption 2.1.

PROPOSITION 2.2. *Suppose that the following relations are valid:*

$$\begin{cases} \nabla\phi(x^k) - A^T y^k - w^k = -c, \\ Ax^k + \nabla\psi(y^k) - s^k = b, \\ x^k, w^k, y^k, s^k \ge 0. \end{cases} \tag{2.4}$$

Then x^k is a feasible solution to problem (1.2) *and y^k is a feasible solution to problem* (1.3).

Proof. The second equation of (2.4) and the convexity of $\psi(y)$ imply that for any $y \in \mathcal{Y}$,

$$\begin{aligned}(b - Ax^k)^T y - \psi(y) &= [\nabla\psi(y^k) - s^k]^T y - \psi(y) \\ &= \nabla\psi(y^k)^T(y - y^k) + \nabla\psi(y^k)^T y^k - y^T s^k - \psi(y) \\ &\le \nabla\psi(y^k)^T y^k - \psi(y^k) - y^T s^k.\end{aligned}$$

Since $y \ge 0$ and $s^k \ge 0$, we have

$$\begin{aligned}f(x^k) &= c^T x^k + \phi(x^k) + \sup_{y\ge 0}\{(b - Ax^k)^T y - \psi(y)\} \\ &\le c^T x^k + \phi(x^k) + \sup_{y\ge 0}\{\nabla\psi(y^k)^T y^k - \psi(y^k) - y^T s^k\} \\ &\le c^T x^k + \phi(x^k) + \nabla\psi(y^k)^T y^k - \psi(y^k) < \infty.\end{aligned}$$

Thus the second equation of (2.4), $s^k \ge 0$, and $x^k \ge 0$ imply that x^k is feasible to (1.2). Similarly, the first equation of (2.4), $w^k \ge 0$, and $y^k \ge 0$ imply that y^k is feasible to (1.3). □

Because our algorithm will ensure that (2.4) is valid for all iterates (see details below), the algorithm will generate a feasible sequence $\{(x^k, y^k)\}$ to problems (1.2) and (1.3) according to Proposition 2.2.

Now we discuss the existence of the saddle points of $l_\mu(x, y)$. Unlike the linear-quadratic case, the feasibility of both primal and dual problems is not enough for the existence of a saddle point. However, we will show that under Assumption 2.1 for any $\mu \ge 0$ a saddle point of $l_\mu(x, y)$ exists. We first prove a lemma.

LEMMA 2.3. *Suppose that the following relations are valid:*

$$\mathcal{P} \equiv \cap_{y>0}\{p \in R^n_+ | c^T p + (\phi 0^+)(p) - y^T A p \le 0\} = \{0\}, \tag{2.5}$$

$$\mathcal{Q} \equiv \cap_{x>0}\{q \in R^m_+ | - b^T q + (\psi 0^+)(q) + x^T A^T q \le 0\} = \{0\}, \tag{2.6}$$

where $(\phi 0^+)(p)$ and $(\psi 0^+)(q)$ are the recession functions of $\phi(x)$ and $\psi(y)$, respectively:

$$(\phi 0^+)(p) \equiv \lim_{\lambda\to 0^+} \lambda\phi(x + \lambda^{-1} p),$$

$$(\psi 0^+)(q) \equiv \lim_{\lambda\to 0^+} \lambda\psi(y + \lambda^{-1} q).$$

(The two recession functions are invariant regardless of the choice of $x \in \mathcal{X}$ and $y \in \mathcal{Y}$.) Then $l_\mu(x, y)$ has a saddle point on $\mathcal{X} \times \mathcal{Y}$.

Proof. According to convex analysis [17, Theorem 37.6], a sufficient condition for $l_\mu(x, y)$ to have a saddle point on $\mathcal{X} \times \mathcal{Y}$ is that the convex functions $l_\mu(\cdot, y)$ have no common direction of recession for $y \in \operatorname{ri}\mathcal{Y}$ and that the convex functions $-l_\mu(x, \cdot)$ have no common direction of recession for $x \in \operatorname{ri}\mathcal{X}$, where "ri" stands for the relative interior. Denote by R^n_+ and R^m_+ the nonnegative orthants of R^n and R^m, respectively. Now for fixed $y \in \operatorname{ri}\mathcal{Y}$, the set of directions of recession of $l_\mu(\cdot, y)$ is given by

$$\begin{aligned}\mathcal{P}_y &\equiv \{p \in R^n_+ | \lim_{\lambda\to 0^+} \lambda l_\mu(x+\lambda^{-1}p, y) \le 0\} \\ &= \left\{ p \in R^n_+ | \lim_{\lambda\to 0^+} \lambda\phi(\lambda^{-1}p) - \lim_{\lambda\to 0^+} \lambda\mu \sum_{j=1}^n \log(x_j + \lambda^{-1}p_j) + (c - A^T y)^T p \le 0 \right\} \\ &= \{p \in R^n_+ | (\phi 0^+)(p) + (c - A^T y)^T p \le 0\}.\end{aligned}$$

Similarly, for fixed $x \in \operatorname{ri}\mathcal{X}$, the set of directions of recession of $-l(x,\cdot)$ is

$$\mathcal{Q}_x \equiv \{q \in R^m_+ | (\psi 0^+)(q) - (b - Ax)^T q \le 0\}.$$

The statement "for $y \in \operatorname{ri}\mathcal{Y}$ there is no common direction of recession" is then interpreted as (2.5). We also get (2.6) in the same fashion. □

PROPOSITION 2.4. *Under Assumption* 2.1 *for any* $\mu \ge 0$, $l_\mu(x,y)$ *has a saddle point on* $\mathcal{X}\times\mathcal{Y}$.

Proof. Suppose that Assumption 2.1 is satisfied by a quadruple $(x^0, y^0, w^0, s^0) > 0$. Note that

$$(\phi 0^+)(p) = \lim_{\lambda\to 0^+} \lambda\phi(x^0 + \lambda^{-1}p) = \lim_{\lambda\to 0^+} \lambda[\phi(x^0 + \lambda^{-1}p) - \phi(x^0)] \ge \nabla\phi(x^0)^T p.$$

We have, for any $p \in \mathcal{P}_{y^0}$,

$$(w^0)^T p = c^T p + \nabla\phi(x^0)^T p - (y^0)^T A p \le c^T p + (\phi 0^+)(p) - (y^0)^T A p \le 0.$$

Since $w^0 > 0$ and $p \ge 0$, the above inequality implies $p = 0$. Therefore we have $\mathcal{P}_{y^0} = \{0\}$. Analogously we have $\mathcal{Q}_{x^0} = \{0\}$. By Lemma 2.3, $l_\mu(x,y)$ has a saddle point on $\mathcal{X}\times\mathcal{Y}$. □

Proposition 2.4 says that Assumption 2.1 is sufficient for the existence of optimal solutions of problems (1.1)–(1.3) as well as for the existence of the solutions of (2.3) for any $\mu > 0$. From the strict convexity of $\sup_{y\ge 0} l_\mu(x,y)$ and $-\inf_{x\ge 0} l_\mu(x,y)$, it can be seen that $(x(\mu), y(\mu))$ is unique for $\mu > 0$. We call the set $\{(x(\mu), y(\mu)) | \mu > 0\}$ the *central path* of problem (1.1) if $(x(\mu), y(\mu))$ is a saddle point of $l_\mu(x,y)$ on $\mathcal{X}\times\mathcal{Y}$.

It should be noted that system (2.2) can be viewed as an MCP of the mapping

$$F : R^{m+n} \to R^{m+n}, \quad F(x,y) = \begin{bmatrix} \nabla\phi(x) - A^T y + c \\ Ax + \nabla\psi(y) - b \end{bmatrix}.$$

Several conditions have been discussed in the literature for the existence of the central path under various situations. For example Güler [7] studies conditions for MCPs. It can be shown that his conditions are equivalent to Assumption 2.1 in the context of mapping F. The conditions involving recession functions stated in Lemma 2.3 appear to be new in the literature.

We now estimate the error if the solution of system (2.3) is used as an approximate solution to the saddle point problem (1.1). We prove that the duality gap of the solutions on the central path converges to zero as μ goes to zero. This is the basic fact that justifies interior path-following algorithms.

PROPOSITION 2.5. *Under Assumption* 2.1, *given any* $\mu \ge 0$ *we have*

$$0 \le f(x(\mu)) - g(y(\mu)) \le (m+n)\mu. \tag{2.7}$$

Proof. The assumption implies the existence of $(x(\mu), y(\mu))$, which together with a certain $(w(\mu), s(\mu))$ satisfies system (2.4). By definitions of $f(x)$ and $g(y)$, we always have $f(x) \geq g(y)$ for all $(x, y) \in \mathcal{X} \times \mathcal{Y}$ (the weak duality). Therefore we only need to prove the second inequality. Analogous to the proof of Proposition 2.2, we have

$$\begin{aligned} f(x(\mu)) &\leq c^T x(\mu) + \phi(x(\mu)) - \psi(y(\mu)) + \nabla\psi(y(\mu))^T y(\mu) \\ &= c^T x(\mu) + \phi(x(\mu)) - \psi(y(\mu)) + [b - Ax(\mu)]^T y(\mu) + s(\mu)^T y(\mu). \end{aligned} \tag{2.8}$$

The last equality above is based on the second equation of (2.4). A symmetric argument for the dual problem implies

$$g(y(\mu)) \geq b^T y(\mu) - \psi(y(\mu)) + \phi(x(\mu)) - [A^T y(\mu) - c]^T x(\mu) - w(\mu)^T x(\mu). \tag{2.9}$$

Proposition 2.5 is proved by subtracting (2.9) from (2.8):

$$f(x(\mu)) - g(y(\mu)) \leq w(\mu)^T x(\mu) + s(\mu)^T y(\mu) = (m+n)\mu. \qquad \square$$

Proposition 2.5 provides the estimation on the duality gap for the points on the central path. Denote the positive diagonal matrices $\mathrm{diag}(x_1, \ldots, x_n)$, $\mathrm{diag}(y_1, \ldots, y_m)$, $\mathrm{diag}(w_1, \ldots, w_n)$, and $\mathrm{diag}(s_1, \ldots, s_m)$ by X, Y, W, and S, respectively. For any $\mu > 0$, we do not have to obtain $x(\mu)$ and $y(\mu)$ exactly. In practice a path-following algorithm generates a sequence of (x^k, y^k) close to $(x(\mu^k), y(\mu^k))$. The closeness is defined by a proximity function

$$\delta(x, y, w, s, \mu) = \left(\left\| \frac{Wx}{\mu} - e \right\|^2 + \left\| \frac{Sy}{\mu} - e \right\|^2 \right)^{1/2}, \tag{2.10}$$

where e is a vector of ones of compatible dimension and $\|\cdot\|$ is the Euclidean norm. Notice that $\delta(x, y, w, s, \mu) = 0$ implies that (x, y) is on the central path; i.e., $(x, y) = (x(\mu), y(\mu))$. With a little abuse of the notations, the same e is used in (2.10) and below regardless of the dimension. The following result provides an error bound for an approximate solution of (2.3) which satisfies (2.4) but may not satisfy the other equations of (2.3).

PROPOSITION 2.6. *If (x, y, w, s) satisfies (2.4) and $\delta(x, y, w, s, \mu) \leq \alpha$, then*

$$0 \leq f(x) - g(y) \leq (1 + \alpha/\sqrt{n+m})(n+m)\mu.$$

Proof. Notice that the proof of inequalities (2.8) and (2.9) uses only the relationships in (2.4). Thus by following the proof of Proposition 2.5, we have

$$f(x) - g(y) \leq w^T x + s^T y.$$

On the other hand,

$$w^T x + s^T y = e^T W x + e^T S y = e^T \begin{bmatrix} Wx - \mu e \\ Sy - \mu e \end{bmatrix} + (m+n)\mu$$

$$\leq \|e\| \cdot \left\| \begin{bmatrix} Wx - \mu e \\ Sy - \mu e \end{bmatrix} \right\| + (m+n)\mu \leq \alpha\mu\sqrt{n+m} + (n+m)\mu.$$

Hence

$$f(x) - g(y) \leq (1 + \alpha/\sqrt{n+m})(n+m)\mu. \qquad \square$$

Proposition 2.6 provides an estimation of the duality gap for the solutions in a neighborhood of the central path. With this estimation, in order to find a pair of ϵ-optimal solution to (1.1) in the following sense

$$x \text{ is feasible to (1.2), } y \text{ is feasible to (1.3), and } 0 \le f(x) - g(y) \le \epsilon,$$

we only need to find a pair of primal and dual feasible solutions in a neighborhood of the central path satisfying $\delta(x, y, w, s, \mu) \le \alpha$, where $\mu = \epsilon/[(1+\alpha/\sqrt{n+m})(n+m)]$.

3. Convergence analysis of the predictor-corrector algorithm. Given any point $(x, y, w, s) > 0$ and any number $\mu > 0$ satisfying (2.4) and $\delta(x, y, w, s, \mu) \le \alpha$, we consider a path-following method for problem (1.1). A typical iteration in the algorithm applies one step of Newton's method to the system

$$\begin{cases} \nabla\phi(x) - A^T y - w = -c, \\ Ax + \nabla\psi(y) - s = b, \\ Wx = \lambda\mu e, \\ Sy = \lambda\mu e, \end{cases} \tag{3.1}$$

where λ is a certain constant. From the Newton approximation of system (3.1), the improving direction $(\Delta x, \Delta y, \Delta w, \Delta s)$ is determined by

$$\begin{cases} \Delta w = \nabla^2\phi(x)\Delta x - A^T\Delta y, \\ \Delta s = A\Delta x + \nabla^2\psi(y)\Delta y, \\ W\Delta x + X\Delta w = -Xw + \lambda\mu e, \\ S\Delta y + Y\Delta s = -Sy + \lambda\mu e. \end{cases} \tag{3.2}$$

The associated direction is called the predictor (affine-scaling) direction if $\lambda = 0$ and is called the corrector (centering) direction if $\lambda = 1$. The following algorithm moves a solution from a tight neighborhood of the central path to a loose one in each predictor step in order to reduce the central path parameter μ. It then draws a solution from the loose neighborhood back to the tight one in each corrector step. The algorithm terminates when $\mu \le \epsilon/[(1+\alpha/\sqrt{n+m})(n+m)]$. We now present the algorithm.

ALGORITHM 3.1 (a predictor-corrector algorithm for problem (1.1)).

Step 0 *(Initialization) Let* $k = 0$. *Choose* $(x^0, y^0, w^0, s^0) > 0$, $\mu_0 > 0$, $0 < \alpha < 1 < t$, *where* $\alpha t < 1$, *such that the first two equations of* (3.1) *are satisfied by* (x^0, y^0, w^0, s^0) *and such that* $\delta(x^0, y^0, w^0, s^0, \mu_0) \le \alpha$.

Step 1 **For** $k = 0, 1, \ldots,$ **until** $\mu_k \le \epsilon/[(1+\alpha/\sqrt{n+m})(n+m)]$ *(where* ϵ *is the user assigned tolerance),* **do**

Step 1.1 *Solve* (3.2) *with* $x = x^k$, $y = y^k$, $w = w^k$, $s = s^k$, $\mu = \mu_k$, *and* $\lambda = 0$. *Denote by* Δx^p, Δy^p, Δw^p, *and* Δs^p *the resulting directions. Let* θ *be a suitable positive number such that*

$$\delta(x(\theta), y(\theta), w(\theta), s(\theta), \mu(\theta)) \le t\alpha,$$

where

$$x(\theta) = x^k + \theta\Delta x^p, \quad y(\theta) = y^k + \theta\Delta y^p,$$
$$w(\theta) = w^k + \theta\Delta w^p + \nabla\phi(x^k + \theta\Delta x^p) - \nabla\phi(x^k) - \theta\nabla^2\phi(x^k)\Delta x^p,$$
$$s(\theta) = s^k + \theta\Delta s^p + \nabla\psi(y^k + \theta\Delta y^p) - \nabla\psi(y^k) - \theta\nabla^2\psi(y^k)\Delta y^p,$$

and

$$\mu(\theta) = (1-\theta)\mu_k.$$

This is the predictor step.

Step 1.2 *Solve* (3.2) *with* $x = x(\theta), y = y(\theta), w = w(\theta), s = s(\theta), \mu = \mu(\theta)$, *and* $\lambda = 1$, *resulting in* $\Delta x^c,\ \Delta y^c,\ \Delta w^c$, *and* Δs^c. *Let*

$$x^{k+1} = x(\theta) + \Delta x^c, \quad y^{k+1} = y(\theta) + \Delta y^c,$$
$$w^{k+1} = w(\theta) + \Delta w^c + \nabla\phi(x(\theta) + \Delta x^c) - \nabla\phi(x(\theta)) - \nabla^2\phi(x(\theta))\Delta x^c,$$
$$s^{k+1} = s(\theta) + \Delta s^c + \nabla\psi(y(\theta) + \Delta y^c) - \nabla\psi(y(\theta)) - \nabla^2\psi(y(\theta))\Delta y^c,$$

and

$$\mu_{k+1} = \mu(\theta).$$

This is the corrector step. Update k *and go to next iteration of Step* 1.

The steplength θ in Step 1.1 can be computed by an explicit formula which will be discussed in Proposition 3.6 below. It will become apparent at the end of our analysis that all (x^k, y^k) are feasible to (1.2) and (1.3).

Algorithm 3.1 can be generalized to find an approximate solution (u^*, v^*) of the following MCP:

$$\text{(MCP)} \qquad u, v \in R^{m+n},\ u \geq 0,\ v = F(u) \geq 0,\ u^T v = 0$$

in the sense of

$$u^*, v^* \in R^{m+n},\ u^* \geq 0,\ v^* = F(u^*) \geq 0,\ \delta(u^*, v^*, \mu) \leq \alpha,$$

where $0 < \alpha < 1$, μ is any preassigned small positive number, and $\delta(u, v, \mu)$ is defined similarly to $\delta(x, y, w, s, \mu)$. In order to facilitate comparisons of Algorithm 3.1 with existing interior point algorithms for nonlinear complementarity problems, we will prove our convergence result for the following algorithm for (MCP) which generalizes Algorithm 3.1.

ALGORITHM 3.2 (a predictor-corrector algorithm for problem (MCP)).

Step 0 *(Initialization) Let* $k = 0$. *Choose* $(u^0, v^0) > 0$, $\mu_0 > 0$, $0 < \alpha < 1 < t$, *where* $\alpha t < 1$, *such that the first equation of*

$$\begin{cases} F(u) - v = 0, \\ Vu = \lambda\mu e \end{cases} \tag{3.3}$$

is satisfied by (u^0, v^0) *and such that* $\delta(u^0, v^0, \mu_0) \leq \alpha$.

Step 1 **For** $k = 0, 1, \ldots,$ **until** $\mu_k \leq \epsilon/[(1 + \alpha/\sqrt{n+m})(n+m)]$ *(where* ϵ *is the user-assigned tolerance),* **do**

Step 1.1 *Solve*

$$\begin{cases} \Delta v = F'(u)\Delta u, \\ V\Delta u + U\Delta v = -Vu + \lambda\mu e, \end{cases} \tag{3.4}$$

with $u = u^k$, $v = v^k$, $\mu = \mu_k$, *and* $\lambda = 0$, *where* $F'(u)$ *represents the Jacobian of* F *at* u. *Denote by* Δu^p *and* Δv^p *the resulting directions. Let* θ *be a suitable positive number such that*

$$\delta(u(\theta), v(\theta), \mu(\theta)) \leq t\alpha,$$

where

$$u(\theta) = u^k + \theta \Delta u^p,$$

$$v(\theta) = v^k + \theta \Delta v^p + F(u^k + \theta \Delta u^p) - F(u^k) - \theta F'(u^k)\Delta u^p,$$

and

$$\mu(\theta) = (1-\theta)\mu_k.$$

This is the predictor step.

Step 1.2 *Solve* (3.4) *with* $u = u(\theta), v = v(\theta), \mu = \mu(\theta)$, *and* $\lambda = 1$, *resulting in* Δu^c *and* Δv^c. *Let*

$$u^{k+1} = u(\theta) + \Delta u^c,$$

$$v^{k+1} = v(\theta) + \Delta v^c + F(u(\theta) + \Delta u^c) - F(u(\theta)) - F'(u(\theta))\Delta u^c,$$

and

$$\mu_{k+1} = \mu(\theta).$$

This is the corrector step. Update k *and go to next iteration of Step* 1.

It is not hard to verify that by assigning

$$u = \begin{bmatrix} x \\ y \end{bmatrix}, \quad v = \begin{bmatrix} w \\ s \end{bmatrix}, \quad \text{and } F(u) = \begin{bmatrix} \nabla\phi(x) - A^T y + c \\ Ax + \nabla\psi(y) - b, \end{bmatrix}$$

Algorithm 3.2 specializes to Algorithm 3.1.

We now proceed to show that Algorithm 3.2 is of polynomial complexity. We take notations such as U and V in the same way as X, Y, W, and S. The following assumption is used in our proof.

The scaled Lipschitz condition (SLC) for $F(u)$.

Given $0 < \beta < 1$, *there exists* $M > 0$ *such that*

$$\|U[F(u+\Delta u) - F(u) - F'(u)\Delta u]\| \le M\Delta u^T F'(u)\Delta u$$

holds for any $u > 0$ *and* Δu *satisfying* $\|U^{-1}\Delta u\| \le \beta$.

Back to the saddle point problem, this condition is equivalent to

$$\left\|X[\nabla\phi(x+\Delta x) - \nabla\phi(x) - \nabla^2\phi(x)\Delta x]\right\| \le M\Delta x^T\nabla^2\phi(x)\Delta x$$

and

$$\left\|Y[\nabla\psi(y+\Delta y) - \nabla\psi(y) - \nabla^2\psi(y)\Delta y]\right\| \le M\Delta y^T\nabla^2\psi(y)\Delta y$$

for any $x > 0, \Delta x, y > 0$, and Δy satisfying $\left\|X^{-1}\Delta x\right\| \le \beta$ and $\left\|Y^{-1}\Delta y\right\| \le \beta$.

SLC has been employed in the analysis of interior point methods for convex programs [38]. Functions satisfying this condition include many useful functions such as $\phi(x) = \sum \phi_j(x_j)$ (similarly for $\psi(y) = \sum \psi_i(y_i)$), where $\phi_j(x_j)$ could be

$$-\log x_j,\ x_j \log x_j,\quad x_j^\alpha (\alpha < 0 \text{ or } \alpha > 1).$$

Note that $\phi(x)$ is not necessarily separable in general. For instance, the SLC is satisfied by the quadratic functions $\phi(x) = x^T Q x$ with Q being positive semidefinite. The same can be said about $\psi(y)$.

Denote the diagonal matrices

$$\mathrm{diag}(\Delta u_1, \ldots, \Delta u_{m+n}) \text{ and } \mathrm{diag}(\Delta v_1, \ldots, \Delta v_{m+n})$$

by ΔU and ΔV, respectively. We use the conventional notations such as $U^\nu = \mathrm{diag}(u_1^\nu, \ldots, u_{m+n}^\nu)$, $\Delta U^\nu = \mathrm{diag}(\Delta u_1^\nu, \ldots, \Delta u_{m+n}^\nu)$ and $u^\nu = (u_1^\nu, \ldots, u_{m+n}^\nu)^T$ for any real number ν. In the derivations below, we frequently use relationships defined by (3.4) and the following simple inequality:

For all u, v such that $\delta(u, v, \mu) \le \tau$ there holds

$$(1-\tau)\mu \le u_j v_j \le (1+\tau)\mu, \quad j = 1, \ldots, m+n. \tag{3.5}$$

Analysis on the predictor step. Given

$$\delta(u, v, \mu) \le \alpha$$

(the superscript k is omitted), we want to know how to choose θ such that

$$\delta(u(\theta), v(\theta), \mu(\theta)) \le t\alpha.$$

This will give us important information on the complexity of the algorithm because the algorithm sets $\mu_{k+1} = (1-\theta)\mu_k$ and stops when $\mu_k \le \epsilon/[(1+\alpha/\sqrt{n+m})(n+m)]$. An explicit formula for θ will be given following our analysis. Let

$$\xi = (I + \theta U^{-1}\Delta U)U[F(u + \theta\Delta u) - F(u) - \theta F'(u)\Delta u],$$

where I stands for the identity matrix. Note that

$$\begin{aligned}
&\|V(\theta)u(\theta) - (1-\theta)\mu e\| \\
&= \left\|Vu + \theta(V\Delta u + U\Delta v) + \theta^2 \Delta U\Delta v - (1-\theta)\mu e + \xi\right\| \\
&= \left\|(1-\theta)Vu - (1-\theta)\mu e + \theta^2\Delta U\Delta v + \xi\right\| \\
&\le (1-\theta)\,\|Vu - \mu e\| + \theta^2\,\|\Delta U\Delta v\| + \|\xi\| \\
&\le (1-\theta)\alpha\mu + \theta^2\,\|\Delta U\Delta v\| + \|\xi\| \qquad (3.6)
\end{aligned}$$

and that

$$\|\xi\| \le 2M\theta^2 \Delta u^T F'(u)\Delta u,$$

if $\|\theta U^{-1}\Delta u\| \le \beta < 1$ is satisfied. Let

$$\eta_1 \equiv \|\Delta U\Delta v\| \quad \text{and} \quad \eta_2 \equiv \Delta u^T F'(u)\Delta u.$$

Now we estimate η_1 and η_2 separately.

LEMMA 3.3.

$$\eta_1 \equiv \|\Delta U\Delta v\| \le (1+\alpha)(m+n)\mu.$$

Proof. Let

$$r_1 = (UV)^{-1/2} V\Delta u,$$

$$r_2 = (UV)^{-1/2} U\Delta v,$$

and

$$R_1 = (UV)^{-1/2} V\Delta U.$$

Then by (3.4) we have

$$r_1 + r_2 = -(Uv)^{1/2}$$

and

$$r_1{}^T r_2 = \Delta u^T \Delta v = \Delta u^T F'(u)\Delta u \geq 0.$$

Therefore we get

$$\eta_1 = \|\Delta U \Delta v\| = \|R_1 r_2\| \leq \max\{\|r_1\|^2, \|r_2\|^2\}$$
$$\leq \|r_1 + r_2\|^2 = \left\|(Uv)^{1/2}\right\|^2 \leq (1+\alpha)(m+n)\mu \qquad \text{(by (3.5))}. \qquad \square$$

Since we will apply the SLC in the estimate of η_2, we need to prove the following lemma.

LEMMA 3.4. *In the predictor step, if*

$$\theta \leq \sqrt{\frac{(1-\alpha)\beta}{(1+\alpha)(m+n)}},$$

then $\left\|\theta U^{-1}\Delta u\right\| \leq \beta$.

Proof. Multiplying both sides of

$$V\Delta u + U\Delta v = -Uv$$

by $U^{-1/2}V^{-1/2}$, we have

$$U^{-1/2}V^{1/2}\Delta u + U^{1/2}V^{-1/2}\Delta v = -U^{1/2}v^{1/2}.$$

Therefore by using (3.5) and noting that the inner product of the two terms on the left-hand side is nonnegative, we have

$$\left\|U^{-1/2}V^{1/2}\Delta u\right\|^2 \leq \left\|U^{1/2}v^{1/2}\right\|^2 \leq (1+\alpha)(m+n)\mu.$$

It in turn implies

$$(1+\alpha)(m+n)\mu \geq \left\|U^{-1/2}V^{1/2}\Delta u\right\|^2 = \left\|U^{1/2}V^{1/2}U^{-1}\Delta u\right\|^2 \geq (1-\alpha)\mu\left\|U^{-1}\Delta u\right\|^2.$$

Thus we have

$$\left\|\theta U^{-1}\Delta u\right\|^2 \leq (1+\alpha)(m+n)\theta^2/(1-\alpha) \leq \beta$$

as long as

$$\theta \leq \sqrt{\frac{(1-\alpha)\beta}{(1+\alpha)(m+n)}}. \qquad \square$$

The estimate related to η_2 is given in the following lemma.

LEMMA 3.5. *In the predictor step,*

$$\eta_2 \equiv \Delta u^T F'(u)\Delta u \le (1+\alpha)(m+n)\mu/4.$$

Proof. Setting $\lambda = 0$ in (3.4), we have

$$V\Delta u + U(F'(u)\Delta u) = -Uv.$$

Multiply both sides of the equation by $\Delta u^T U^{-1}$, and we get

$$\Delta u^T U^{-1} V\Delta u + \Delta u^T F'(u)\Delta u = -\Delta u^T v.$$

Hence

$$\begin{aligned}
&\Delta u^T F'(u)\Delta u \\
&= -\Delta u^T v - \Delta u^T U^{-1} V\Delta u \\
&= -\Delta u^T (U^{-1/2}V^{1/2})(U^{1/2}v^{1/2}) - \left\|U^{-1/2}V^{1/2}\Delta u\right\|^2 \\
&\le \left\|U^{1/2}v^{1/2}\right\| \cdot \left\|U^{-1/2}V^{1/2}\Delta u\right\| - \left\|U^{-1/2}V^{1/2}\Delta u\right\|^2 \\
&\le \left\|U^{1/2}v^{1/2}\right\|^2/4 \\
&\le (1+\alpha)(m+n)\mu/4.
\end{aligned}$$

The second to last inequality above is due to the fact that

$$bt - t^2 \le b^2/4$$

for any real numbers b and t. □

PROPOSITION 3.6. *Taking*

$$\theta = \min\left\{\sqrt{\frac{(1-\alpha)\beta}{(1+\alpha)(m+n)}}, \frac{2}{1+\left[1+\frac{(4+2M)(1+\alpha)(m+n)}{(t-1)\alpha}\right]^{1/2}}\right\} = O\left[(m+n)^{-1/2}\right],$$

we have

$$\|V(\theta)u(\theta) - (1-\theta)\mu e\| \le t\alpha(1-\theta)\mu.$$

Proof. From (3.6), Lemmas 3.3, 3.4, and 3.5 we have

$$\begin{aligned}
&\|V(\theta)u(\theta) - (1-\theta)\mu e\| \\
&\le (1-\theta)\alpha\mu + \theta^2\|\Delta U\Delta v\| + \|\xi\| \\
&\le (1-\theta)\alpha\mu + \eta_1\theta^2 + 2M\theta^2\eta_2.
\end{aligned}$$

This quantity will be not greater than $t\alpha(1-\theta)\mu$ as long as

$$0 \le \theta \le \frac{2}{1+\left[1+\frac{4(\eta_1+2M\eta_2)}{(t-1)\alpha\mu}\right]^{1/2}}$$

by simply solving a quadratic equation and picking the large root.

Replacing η_1 and η_2 by the estimates in Lemmas 3.3 and 3.5, we obtain an explicit upper bound for θ:

$$\theta \le \frac{2}{1+\left[1+\frac{(4+2M)(1+\alpha)(m+n)}{(t-1)\alpha}\right]^{1/2}}.$$

This, together with the condition of Lemma 3.4, results in the proposition. ∎

COROLLARY 3.7. *In each predictor step the central path parameter μ can be reduced at least by a factor of $1-\tau/\sqrt{m+n}$, where τ is a positive constant depending on the choices of α, β, t, and the smooth coefficient M in the SLC.*

Analysis on the corrector step. We begin the corrector step with $\delta(u(\theta), v(\theta), \mu(\theta)) \le t\alpha$. Our target is to show that after the step we will have

$$\delta(u^{k+1}, v^{k+1}, \mu_{k+1}) \le \alpha,$$

where

$$u^{k+1} = u(\theta) + \Delta u^c, \quad \mu_{k+1} = \mu(\theta),$$

and

$$v^{k+1} = v(\theta) + \Delta v^c + F(u(\theta) + \Delta u^c) - F(u(\theta)) - F'(u(\theta))\Delta u^c.$$

We will follow the same clue used in the analysis of the predictor step. Omitting the θ in the parentheses, we estimate

$$\eta_3 \equiv \|\Delta U \Delta v\| \quad \text{and} \quad \eta_4 \equiv \Delta u^T F'(u)\Delta u$$

in Lemmas 3.8 and 3.10, while Lemma 3.9 will guarantee the use of the SLC.

LEMMA 3.8. *If $\delta(u, v, \mu) \le t\alpha$ then*

$$\eta_3 \equiv \|\Delta U \Delta v\| \le \frac{(t\alpha)^2}{1-t\alpha}\mu.$$

Proof. Let

$$r_3 = (UV)^{-1/2} V \Delta u$$

and

$$r_4 = (UV)^{-1/2} U \Delta v.$$

Then

$$r_3 + r_4 = (UV)^{-1/2}(-Uv + \mu e).$$

Since $r_3^T r_4 = \Delta u^T \Delta v = \Delta u^T F'(u)\Delta u \ge 0$, similar to the proof of Lemma 3.3, we have

$$\|\Delta U \Delta v\| \le \|r_3 + r_4\|^2 \le \frac{(t\alpha\mu)^2}{(1-t\alpha)\mu} = \frac{(t\alpha)^2}{1-t\alpha}\mu. \qquad \blacksquare$$

LEMMA 3.9. *In the corrector step if $t\alpha \le \beta/(1+\beta)$, then*

$$\|U^{-1}\Delta u\| \le \beta.$$

Proof. Set $\lambda = 1$ in (3.4), and we get

$$V\Delta u + U\Delta v = -Uv + \mu e. \tag{3.7}$$

Let

$$D = U^{1/2}V^{-1/2}.$$

Multiply both sides of (3.7) by $U^{-1/2}V^{-1/2}$, and we obtain

$$D^{-1}\Delta u + D\Delta v = -U^{-1/2}V^{-1/2}\left(Uv - \mu e\right).$$

Since $\Delta u^T \Delta v = \Delta u^T F'(u)\Delta u \geq 0$, we have

$$\left(D^{-1}\Delta u\right)^T \left(D\Delta v\right) \geq 0$$

and

$$\begin{aligned}\|D^{-1}\Delta u\|^2 &\leq \|D^{-1}\Delta u + D\Delta v\|^2 \\ &\leq \|U^{-1/2}V^{-1/2}\|^2\|Uv - \mu e\|^2 \\ &\leq \left[((1-t\alpha)\mu)^{-1/2}\right]^2 (t\alpha\mu)^2 = \frac{(t\alpha)^2\mu}{1-t\alpha}.\end{aligned}$$

On the other hand,

$$\begin{aligned}\|D^{-1}\Delta u\|^2 &= \|U^{1/2}V^{1/2}U^{-1}\Delta u\|^2 \\ &\geq (1-t\alpha)\mu\|U^{-1}\Delta u\|^2.\end{aligned}$$

Hence we get

$$\|U^{-1}\Delta u\| \leq \frac{t\alpha}{1-t\alpha} \leq \beta. \qquad \square$$

LEMMA 3.10. *If $\delta(u, v, \mu) \leq t\alpha$, then*

$$\eta_4 \equiv \Delta u^T F'(u)\Delta u \leq \frac{(t\alpha)^2}{4(1-t\alpha)}\mu.$$

Proof. The proof is very similar to the proof of Lemma 3.5, so we omit the details. The only changes we have to make are to replace Uv and $U^{1/2}v^{1/2}$ with $Uv - \mu e$ and $\left(U^{-1/2}V^{-1/2}\right)(Uv - \mu e)$, respectively, and to use the fact that

$$\frac{\left(U^{-1/2}V^{-1/2}\right)(Uv - \mu e)}{4} \leq \frac{\|Uv - \mu e\|^2}{4(1-t\alpha)\mu} \leq \frac{(t\alpha)^2}{4(1-t\alpha)}\mu. \qquad \square$$

PROPOSITION 3.11. *If the parameters α, t, and M satisfy*

$$\frac{(t\alpha)^2}{1-t\alpha} + \frac{M(t\alpha)^2}{2(1-t\alpha)} \leq \alpha,$$

for example, we can choose

$$t = 2 \text{ and } 0 < \alpha \leq \frac{1}{6+2M};$$

then we have $(u^{k+1}, v^{k+1}) > 0$, and $\delta(u^{k+1}, v^{k+1}, \mu_{k+1}) \leq \alpha$.

Proof. Using similar derivations to that of (3.6) and invoking Lemmas 3.8–3.10, we have

$$\begin{aligned}
&\left\|V_{k+1}u^{k+1} - \mu_{k+1}e\right\| \\
&\le \|\Delta U \Delta v\| + 2M\left[\Delta u F'(u)\Delta u\right] \\
&= \eta_3 + 2M\eta_4 \\
&\le \frac{(t\alpha)^2\mu}{1-t\alpha} + \frac{M(t\alpha)^2\mu}{2(1-t\alpha)} \le \alpha\mu = \alpha\mu_{k+1}.
\end{aligned}$$

Thus we have $\delta(u^{k+1}, v^{k+1}, \mu_{k+1}) \le \alpha$, which further implies $(u^{k+1}, v^{k+1}) > 0$. ∎

Back to the saddle point algorithm 3.1, it is easy to see that $(x^k, y^k, w^k, s^k) > 0$ for all k because $\delta(x^k, y^k, w^k, s^k, \mu_k) < 1$. In addition, the choices of $(x^{k+1}, y^{k+1}, w^{k+1}, s^{k+1})$ in the algorithm ensure that condition (2.4) is satisfied for all k. Thus, by Proposition 2.2, x^k and y^k are feasible to problems (1.1) and (1.2), respectively, for all k.

In summary, the algorithm keeps $(x^k, y^k, w^k, s^k) > 0$ and $\delta(x^k, y^k, w^k, s^k, \mu^k) \le \alpha$ for each k while reducing μ at a linear rate of $1 - O[(m+n)^{-1/2}]$. From Proposition 2.6, it is readily to see that, in order to find an ϵ-optimal solution (x^k, y^k) of problem (1.1) satisfying $f(x^k) - g(y^k) \le \epsilon$, only $O\left[\sqrt{m+n}\,|\log(m+n)\mu_0/\epsilon|\right]$ iterations are necessary. Since μ_0/ϵ is usually larger than $m+n$, the algorithm has polynomial complexity of $O\left(\sqrt{m+n}|\log \mu_0/\epsilon|\right)$.

Remarks on taking advantage of the problem structure. Practical problems formulated as problem (1.1) often have special structures. For instance, in the two-stage model of stochastic programming mentioned in section 1, the matrix $\nabla^2\psi(y)$ is very large and block diagonal, whereas the matrix $\nabla^2\phi(x)$ is of ordinary size. So the key point is how to reduce the amount of work for solving system (3.2). The special feature of system (3.2) allows us to reduce the amount of computations substantially. Let us elucidate this point in some detail.

Problems with large, block-diagonal $\nabla^2\psi(y)$ and small $\nabla^2\phi(x)$. Substituting Δw and Δs into the last two equations and canceling Δy, we obtain an equivalent system to (3.2):

$$\begin{cases}
\left\{\nabla^2\phi(x) + X^{-1}W + A^T\left[\nabla^2\psi(y) + Y^{-1}S\right]^{-1}A\right\}\Delta x \\
\quad = A^T\left[\nabla^2\psi(y) + Y^{-1}S\right]^{-1}\left(-s + \lambda\mu Y^{-1}e\right) - w + \lambda\mu X^{-1}e, \\
\left[\nabla^2\psi(y) + Y^{-1}S\right]\Delta y = -A\Delta x - s + \lambda\mu Y^{-1}e, \\
\Delta w = \nabla^2\phi(x)\Delta x - A^T\Delta y, \\
\Delta s = A\Delta x + \nabla^2\psi(y)\Delta y.
\end{cases} \tag{3.8}$$

The principal work is to solve the first equation in (3.8). Since the dimension of x is low (less than 100, say), we can solve system (3.8) exactly even if the dimension of y is high. Notice that the matrix $(\nabla^2\psi(y) + Y^{-1}S)$ has a block-diagonal structure. Hence $A^T[\nabla^2\psi(y) + Y^{-1}S]^{-1}A$ can be computed blockwise. In particular the computation can proceed in parallel. The solution of the first equation in (3.8) can be obtained by a factorization of $\nabla^2\phi(x) + X^{-1}W + A^T[\nabla^2\psi(y) + Y^{-1}S]^{-1}A$, which might be a dense matrix but must be a low-dimensional and symmetric positive definite matrix. After Δx is obtained from the first equation, Δy can be calculated through the second

equation of (3.8) using $[\nabla^2\psi(y)+Y^{-1}S]^{-1}$ computed in the first step. As a matter of fact, we may store the block factorizations of $\nabla^2\psi(y)+Y^{-1}S$ in the first step for later use in the second step. Subsequently, we get Δw and Δs.

Problems with large, block-diagonal $\nabla^2\psi(y)$, $\nabla^2\phi(x)$, and block-banded A. Discretized optimal control problems often have large, block-diagonal $\nabla^2\psi(y)$ and $\nabla^2\phi(x)$ as described in [20] and [24]. In addition, the matrix A in problem (1.1) can be partitioned into blocks such that the resulting matrix of blocks is banded [32], [33]. For instance, the dynamic constraint system

$$x_t = A_t x_{t-1} + B_t u_t + b_t \text{ for } t = 0,\dots,T \text{ with } A_0 = 0 \tag{3.9}$$

can be expressed as an additional term in the saddle function (1.1):

$$-\sum_{t=0}^{T}(y_t - z_t)^T(-x_t + A_t x_{t-1} + B_t u_t + b_t), \tag{3.10}$$

where $y_t \ge 0,\ z_t \ge 0$ are some additional dual vectors. By doing this, we generate an infinite penalty for the violation of constraints (3.9) in the primal problem. Let $u = (u_0,\dots,u_T)$ be the primal control vector and $x = (x_0,\dots,x_T)$ be the primal state vector. It can be seen that the submatrix of A in problem (1.1) corresponding to the primal vector (u,x) and the dual vector $(y,z) = (y_0,\dots,y_T,z_0,\dots,z_T)$ is

$$\bar{A} = \begin{bmatrix} B & F \\ -B & -F \end{bmatrix}, \text{ where } B = \begin{bmatrix} B_0 & & \\ & \ddots & \\ & & B_T \end{bmatrix}, \quad F = \begin{bmatrix} -I_0 & & & \\ A_1 & -I_1 & & \\ & \ddots & \ddots & \\ & & A_T & -I_T \end{bmatrix}.$$

The computation of $A^T(\nabla^2\psi(y)+Y^{-1}S)^{-1}A$ in (3.8) will require us to compute the submatrices

$$M_1 \equiv \bar{Y}^{-1}\bar{S}_y,\ M_2 \equiv \bar{Z}^{-1}\bar{S}_z, \text{ and } \bar{A}^T\begin{bmatrix} M_1 & \\ & M_2 \end{bmatrix}^{-1}\bar{A},$$

where $\bar{Y},\ \bar{S}_y,\ \bar{Z}$, and $\bar{S}_z$ are some diagonal matrices. Note that

$$\begin{aligned} \bar{A}^T\begin{bmatrix} M_1 & \\ & M_2 \end{bmatrix}^{-1}\bar{A} &= \begin{bmatrix} B & F \\ -B & -F \end{bmatrix}^T\begin{bmatrix} M_1 & \\ & M_2 \end{bmatrix}^{-1}\begin{bmatrix} B & F \\ -B & -F \end{bmatrix} \\ &= \begin{bmatrix} B^T(M_1^{-1}+M_2^{-1})B & B^T(M_1^{-1}+M_2^{-1})F \\ F^T(M_1^{-1}+M_2^{-1})B & F^T(M_1^{-1}+M_2^{-1})F \end{bmatrix} \\ &\equiv \begin{bmatrix} H_{11} & H_{12} \\ H_{21} & H_{22} \end{bmatrix}. \end{aligned}$$

According to the structures of M_1, M_2, B, and F, matrix H_{11} is a block-diagonal matrix, matrix $H_{12} = H_{21}^T$ is a block-band matrix with bandwidth 2, and matrix H_{22} is also a block-band one with bandwidth 3. Therefore it is possible to solve equation system (3.8) by using block operations applied to a banded system.

The case of $\mathcal{X}$ and $\mathcal{Y}$ being boxes. Now we explain why changing $\mathcal{X}$ and $\mathcal{Y}$ into boxes will not complicate the computation. Consider a simple case that both $\mathcal{X}$ and $\mathcal{Y}$ are closed-end boxes:

$$\mathcal{X} = \{x \in R^n | 0 \le x \le u < \infty\} \text{ and } \mathcal{Y} = \{y \in R^m | 0 \le y \le v < \infty\}.$$

By introducing auxiliary variables $x^1 = u - x$, $y^1 = v - y$, s, s^1, w, and w^1 and using the variational relationships, the saddle point condition can be written as follows:

$$\begin{cases} \nabla\phi(x) - A^T y - w + w^1 = -c, \\ Ax + \nabla\psi(y) - s + s^1 = b, \\ x + x^1 = u, \\ y + y^1 = v, \\ x^T w = (x^1)^T w^1 = y^T s = (y^1)^T s^1 = 0, \\ x, x^1, y, y^1, w, w^1, s, s^1 \geq 0. \end{cases}$$

The corresponding Newton equations become

$$\begin{cases} \nabla^2\phi(x)\Delta x - A^T\Delta y + \Delta w^1 - \Delta w = a^1, \\ A\Delta x + \nabla^2\psi(y)\Delta y + \Delta s^1 - \Delta s = a^2, \\ \Delta x + \Delta x^1 = a^3, \\ \Delta y + \Delta y^1 = a^4, \\ X\Delta w + W\Delta x = a^5, \\ Y\Delta s + S\Delta y = a^6, \\ W_1\Delta x^1 + X_1\Delta w^1 = a^7, \\ S_1\Delta y^1 + Y_1\Delta s^1 = a^8, \end{cases} \tag{3.11}$$

where $a^1, \ldots, a^8$ are certain fixed vectors. After canceling all variables with superscript 1, we get a system whose major equations are

$$\begin{cases} \left[\nabla^2\phi(x) + X^{-1}W + X_1^{-1}W_1\right]\Delta x - A^T\Delta y = b^1, \\ A\Delta x + \left[\nabla^2\psi(y) + Y^{-1}S + Y_1^{-1}S_1\right]\Delta y = b^2, \end{cases} \tag{3.12}$$

where b^1 and b^2 are certain fixed vectors. After canceling Δy, system (3.12) becomes a similar system to (3.8). Therefore the computations needed for solving system (3.11) are roughly the same as solving system (3.2). Hence there are no significant changes in the algorithm if $\mathcal{X}$ and $\mathcal{Y}$ are changed into boxes. The case that $\mathcal{X}$ and $\mathcal{Y}$ are half-open-end and half-closed-end boxes can be treated analogously.

4. Conclusions and final remarks. We have shown the polynomiality of a predictor-corrector algorithm for a class of nonlinear saddle point problems. The model is an extension of the traditional Lagrange multiplier model in optimization. The results on existence of the central path and the relationship between the central path and the duality gap established in section 2 are useful in developing other interior path-following algorithms.

Several issues deserve further investigation.

(i) *The smooth condition.* We have proposed the SLC as the smooth condition. In the context of convex programming, various smooth conditions were proposed to characterize the problems which will have polynomial interior point algorithms. We refer the reader to references [4] and [15] for more details. A direction of future studies is to identify different classes of practical saddle point problems which will have polynomial interior point algorithms under smooth conditions other than the SLC.

(ii) *Infeasible starting point.* Algorithm 3.1 requires a starting point near the central path. It could be an obstacle if such a starting point is not provided. Recently,

much research has been done on interior point algorithms starting from an *infeasible* interior point; see [8, 12, 13, 29, 33, 35] for examples. The algorithms reduce both infeasibility and duality gap simultaneously. Some of the algorithms use large neighborhoods of the central path and have been tested, e.g., Wright and Ralph [33]. It makes practical sense to develop infeasible interior point methods with large stepsizes for saddle point problems like (1.1).

(iii) *Computation-related developments.* Primary studies [26, 32, 33] show that interior point methods might be fairly effective in solving linear-quadratic problems, but comparison studies between interior point methods and other existing methods have not been conducted fully. In addition, for large-scale problems, it is crucial to reduce the amount of computations for matrix inverse and Hessian evaluation. It is meaningful to develop algorithms that, for instance, can use inexact directions, inaccurate Hessians, and parallelizations in predictor and corrector steps.

Acknowledgment. The authors would like to thank three anonymous referees and the associate editor for their very valuable and detailed suggestions on improving this paper.

REFERENCES

[1] J. R. Birge, *Decomposition and partitioning methods for multistage stochastic linear programs*, Oper. Res., 33 (1985), pp. 989–1007.
[2] J. R. Birge and D. F. Holmes, *Efficient solution of two-stage stochastic linear programs using interior point methods*, Comput. Optim. Appl., 1 (1992), pp. 245–276.
[3] J. R. Birge and L. Qi, *Computing block-angular Karmarkar projections with application to stochastic programming*, Management Sci., 34 (1988), pp. 1472–1479.
[4] D. den Hertog, F. Jarre, C. Roos, and T. Terlaky, *A sufficient condition for self-concordance, with application to some classes of structured convex programming problems*, Math. Programming, 69 (1995), pp. 75–88.
[5] G. Dantzig and A. Madansky, *On the solution of two-stage linear programs under uncertainty*, in Proc. Fourth Berkeley Symposium on Math. and Probability, Vol. 1, University of California Press, Berkeley, CA, 1961.
[6] C. C. Gonzaga, *Interior point path following algorithms*, in Mathematical Programming: State of the Art 1994, J. Berge and K. Murty, eds., University of Michigan Press, Ann Arbor, MI, 1994.
[7] O. Güler, *Existence of interior point and interior paths in nonlinear monotone complementarity problems*, Math. Oper. Res., 18 (1993), pp. 128–147.
[8] F. Jarre and M. A. Saunders, *A practical interior-point method for convex programming*, SIAM J. Optim., 5 (1995), pp. 129–148.
[9] J. Ji, F. Potra, and S. Huang, *A predictor-corrector method for linear complementarity problems with polynomial complexity and superlinear convergence*, J. Optim. Theory Appl., 85 (1995), pp. 187–199.
[10] M. Kojima, S. Mizuno, and A. Yoshise, *A primal-dual algorithm for a class of linear complementarity problems*, Math. Programming, 44 (1989), pp. 1–26.
[11] M. Kojima, M. Megiddo, T. Noma, and A. Yoshise, *A Unified Approach to Interior Point Algorithms for Linear Complementarity Problems*, Lecture Notes in Computer Science, Springer-Verlag, Berlin, 1991.
[12] M. Kojima, T. Noma, and A. Yoshise, *Global convergence in infeasible-interior-point algorithms*, Math. Programming, 65 (1994), pp. 43–72.
[13] S. Mizuno, *A superlinearly convergent infeasible-interior-point algorithm for geometrical LCPs without a strictly complementary condition*, Math. Oper. Res., 21 (1996), pp. 382–400.
[14] S. Mizuno, M. Todd, and Y. Ye, *On adaptive step primal-dual interior-point algorithm for linear programming*, Math. Oper. Res., 18 (1993), pp. 964–981.
[15] Y. Nesterov and A. Nemirovskii, *Interior Point Polynomial Algorithms in Convex Programming*, SIAM, Philadelphia, PA, 1994.
[16] L. Qi and R. S. Womersley, *An SQP algorithm for extended linear-quadratic problems in stochastic programming*, Ann. Oper. Res., 56 (1995), pp. 251-285.
[17] R. T. Rockafellar, *Convex Analysis*, Princeton University Press, Princeton, NJ, 1970.

[18] R. T. Rockafellar, *Linear-quadratic programming and optimal control*, SIAM J. Control Optim., 25 (1987), pp. 781–814.
[19] R. T. Rockafellar, *Computational schemes for large-scale problems in extended linear-quadratic programming*, Math. Programming, 48 (1990), pp. 447–474.
[20] R. T. Rockafellar, *Large-scale extended linear-quadratic programming and multistage optimization*, in Proc. Fifth Mexico–U.S. Workshop on Numerical Analysis, S. Gomez, J. .P. Hennat, and R. Tapia, eds., SIAM, Philadelphia, PA, 1990.
[21] R. T. Rockafellar, *Nonsmooth optimization*, in Mathematical Programming: State of the Art 1994, J. Berge and K. Murty, eds., University of Michigan Press, Ann Arbor, MI, 1994.
[22] R. T. Rockafellar and J. Sun, *A finite simplex-active-set method for monotropic piecewise quadratic programming*, in Advances in Optimization and Approximation, D. Du and J. Sun, eds., Kluwer Academic Publishers, Dordrecht, the Netherlands, 1994.
[23] R. T. Rockafellar and R. J.-B. Wets, *A Lagrangian finite generation technique for solving linear-quadratic problems in stochastic programming*, Math. Programming Studies, 28 (1986), pp. 63–93.
[24] R. T. Rockafellar and R. J.-B. Wets, *Generalized linear-quadratic problems of deterministic and stochastic optimal control in discrete time*, SIAM J. Control Optim., 28 (1990), pp. 810–822.
[25] G. Sonnevend, J. Stoer, and G. Zhao, *On the complexity of following the central path of linear programs by linear extrapolation* II, Math. Programming, 52 (1991), pp. 527–553.
[26] J. Sun, K. Wee, and J. Zhu, *An interior point method for solving a class of linear-quadratic stochastic programming problems*, in Recent Advances in Nonsmooth Optimization, D. Du, L. Qi, and R. S. Womersley, eds., World Scientific, Singapore, 1995.
[27] J. Sun and J. Zhu, *A predictor-corrector method for extended linear-quadratic programming*, Comput. Oper. Res., 23 (1996), pp. 755–767.
[28] P. Tseng, *Global linear convergence of a path-following algorithm for some monotone variational inequality problems*, J. Optim. Theory Appl., 75 (1992), pp. 265–279.
[29] P. Tseng, *An Infeasible Path-Following Method for Monotone Complementarity Problem*, Department of Mathematics, University of Washington, Seattle, WA, 1994, preprint.
[30] R. Van Slyke and R. J.-B. Wets, *L-shaped linear programs with applications to optimal control and stochastic linear programs*, SIAM J. Appl. Math., 17 (1969), pp. 638–663.
[31] R. J.-B. Wets, *Solving stochastic programs with simple recourse*, Stochastics, 10 (1983), pp. 219–242.
[32] S. J. Wright, *Interior point methods for optimal control of discrete-time systems*, J. Optim. Theory Appl., 77 (1993), pp. 161–187.
[33] S. J. Wright and D. Ralph, *A superlinear infeasible-interior-point algorithm for monotone complementarity problems*, Math. Oper. Res., 21 (1996).
[34] Y. Ye and K. Anstreicher, *On quadratic and $O(\sqrt{n}L)$ convergence of a predictor-corrector algorithm for LCP*, Math. Programming, 62 (1993), pp. 537–552.
[35] Y. Zhang, *On the convergence of a class of infeasible-interior-point methods for the horizontal linear complementarity problem*, SIAM J. Optim., 4 (1994), pp. 208–227.
[36] C. Zhu, *On the primal-dual steepest descent algorithm for extended linear-quadratic programming*, SIAM J. Optim., 5 (1995), pp. 114–128.
[37] C. Zhu and R. T. Rockafellar, *Primal-dual projected gradient algorithms for extended linear-quadratic programming*, SIAM J. Optim., 3 (1993), pp. 751–783.
[38] J. Zhu, *A path-following algorithm for a class of convex programming problems*, Z. Oper. Res., 36 (1992), pp. 359–377.

SIAM J. CONTROL OPTIM.
Vol. 35, No. 2, pp. 552–561, March 1997

010

THE SYNTHESIS OF UNIVERSAL FEEDBACK PURSUIT STRATEGIES IN DIFFERENTIAL GAMES*

F. H. CLARKE†, YU. S. LEDYAEV‡, AND A. I. SUBBOTIN§

Abstract. We show how any (generalized) supersolution of the Hamilton–Jacobi equation can be used to construct a feedback pursuit strategy which guarantees (to any given tolerance) a capture time not exceeding the solution's value. If the supersolution is the value function, then a near-optimal pursuit strategy is obtained in this way. An important feature of the construction is its "universal" nature, i.e., the fact that the feedback law is uniformly effective on compact sets of initial conditions. This implies in particular that the feedback construction is one that exploits nonoptimal behavior on the part of the evader.

Key words. pursuit, differential game, feedback, synthesis, proximal analysis, Hamilton–Jacobi equation

AMS subject classification. 93B52

PII. S0363012995283972

Introduction. The differential equation bearing the names of Hamilton and Jacobi and of Bellman and Isaacs has been the essential ingredient of the dynamic programming approach to solving differential games. When the value function is smooth, it constitutes a classical solution of that equation; further, its derivatives give rise easily to optimal strategies in feedback form (we recall this familiar argument in section 1). Of course, it is now understood that it is essential to consider the value function as a solution in some generalized sense (minimax or viscosity, for example) via constructs of nonsmooth analysis [3]: derivates, generalized gradients, subdifferentials, etc. (see [5] for an overview of this topic). In this more realistic setting, the issue of effective feedback synthesis is a more subtle one; it is the one we focus upon in this article.

The particular tool of nonsmooth analysis used here is the *proximal subdifferential* $\partial_P f$ of a lower semicontinuous function f mapping $\mathbb{R}^n$ to $(-\infty, \infty]$. An element ζ of $\mathbb{R}^n$ belongs to $\partial_P f(x)$ (for a given x at which f is finite) iff there exists $\sigma \geq 0$ and a neighborhood $N(x)$ of x such that

$$f(y) - f(x) + \sigma \|y - x\|^2 \geq \langle \zeta, y - x \rangle \quad \forall y \in N(x).$$

Supersolutions of the Hamilton–Jacobi equation can be defined in terms of the proximal subdifferential in a manner akin to the well-known approaches of Subbotin [14] and of Crandall and Lions [6] (see section 2 and [4] for a comparison of these solution concepts). As shown in [5], the proximal approach is particularly well suited to

*Received by the editors April 3, 1995; accepted for publication (in revised form) February 2, 1996. The research of the second and third authors was supported in part by Russian Fund for Fundamental Research grant 93–011–16032. A version of this paper was presented at the 35th IEEE Conference on Decision and Control, Kobe, Japan, December 11–13, 1996.

http://www.siam.org/journals/sicon/35-2/28397.html

†Centre de recherches mathématiques, Université de Montréal, C.P. 6128, Succ. Centre-ville, Montréal, PQ, Canada H3C 3J7 (clarke@crm.umontreal.ca), and Institut Desargues, Université de Lyon I, Villeurbanne, France. The research of this author was supported by the Natural Sciences and Engineering Research Council of Canada and le Fonds FCAR du Québec.

‡Steklov Institute of Mathematics, Moscow 117966, Russia (ledyaev@mian.su).

§Institute of Mathematics and Mechanics, Ekaterinburg 620219, Russia (subbotin@imm.e-burg.su).

the construction of (set) invariant feedbacks; that same property is what lies at the heart of the present article. Experts will recognize our basic constructs in part as especially simple yet powerful variants of the extremal aiming techniques introduced by Krasovskii and Subbotin [10]. The simplicity of the approach is underlined by the fact that the main result (Theorem 2.1) is completely self-contained, although we do cite a few other results in the next section for the purpose of situating the present contribution.

The context of this article is that of a general differential game of pursuit and evasion. Starting with any supersolution of the associated Hamilton–Jacobi equation, Theorem 2.1 constructs an explicit feedback pursuit strategy which guarantees (to any specified tolerance) a capture time no greater than the value of the supersolution at the starting point x_0. When the supersolution is the value function, the ε-optimal feedback strategies are obtained.

An important feature of this feedback is its "universality," i.e., the property of remaining uniformly effective for all starting positions x_0 in a given compact set. In practical terms, this implies that the feedback strategy is one that exploits nonoptimal behavior: if the evader makes a bad move, this is equivalent to starting over from a new position from which the capture time is less, but then the given feedback pursuit strategy is effective from this new starting point. Similarly, a universal strategy is one that has robustness properties in the presence of jumps or uncertainties in state position.

1. The time-optimal differential game. We proceed to formulate precisely a time-optimal differential game whose payoff is the elapsed time before the state reaches a prescribed terminal set. The value function of the game may be discontinuous and extended valued.

Main assumptions. Let the motion of the controlled system be described by the equation

$$\dot{x}(t) = f(x(t), p(t), q(t)), \tag{1.1}$$

where $x(t) \in \mathbb{R}^n$ is the state, $p(t) \in P$ and $q(t) \in Q$ are the controls of the pursuer and the evader, and P and Q are compact sets. It is assumed that the function $f : \mathbb{R}^n \times P \times Q \mapsto \mathbb{R}^n$ is continuous and satisfies the Lipschitz condition

$$\|f(x,p,q) - f(y,p,q)\| \le \lambda \|x - y\| \tag{1.2}$$

for all $x, y \in \mathbb{R}^n$, $p \in P$, $q \in Q$. Suppose for the moment that

$$\min_{p\in P}\max_{q\in Q}\langle s, f(x,p,q)\rangle = \max_{q\in Q}\min_{p\in P}\langle s, f(x,p,q)\rangle =: H(x,s) \tag{1.3}$$

for all $s \in \mathbb{R}^n$ and $x \in \mathbb{R}^n$. This local saddle point condition is a familiar one frequently referred to as the *Isaacs condition*. The function $H(x,s)$ defined by the equality (1.3) is called the Hamiltonian.

Under the above assumptions, for any initial point $x_0 \in \mathbb{R}^n$ and any choice of measurable controls $p(\cdot) : \mathbb{R}^+ \mapsto P$, $q(\cdot) : \mathbb{R}^+ \mapsto Q$, the corresponding solution of equation (1.1) exists, is unique, and can be extended over the whole semiaxis $\mathbb{R}^+ := [0, \infty)$.

Let a closed set $M \subset \mathbb{R}^n$ be given. We define τ the payoff functional on the space $\mathcal{C}(\mathbb{R}^+; \mathbb{R}^n)$ of continuous functions $x(\cdot) : \mathbb{R}^+ \mapsto \mathbb{R}^n$ by the equality

$$\tau(x(\cdot)) := \min\{t \in \mathbb{R}^+ : x(t) \in M\}. \tag{1.4}$$

If $x(t) \notin M$ for all $t \in \mathbb{R}^+$, then we put $\tau(x(\cdot)) = \infty$.

Feedback strategies and value. An arbitrary function $U : \mathbb{R}^n \mapsto P$ (a function $V : \mathbb{R}^n \mapsto Q$) will be called a feedback strategy of the pursuer (the evader). Let an initial point $x_0 \in \mathbb{R}^n$ be given. Let a feedback strategy U and a partition Δ of $[0, \infty)$,

$$\Delta := \{0 = t_0 < t_1 < \cdots\}, \quad t_i \to \infty \quad \text{as } i \to \infty,$$

be chosen by the pursuer. Denote by $\mathbf{X}(x_0, U, \Delta)$ the set of "step-by-step" trajectories of the differential inclusion

$$\dot{x}(t) \in \operatorname{co}\{f(x(t), U(x(t_i)), q) : q \in Q\}, \tag{1.5}$$

$t_i \le t < t_{i+1}$, $i = 0, 1, 2, \ldots$. Specifically, the elements of the set $\mathbf{X}(x_0, U, \Delta)$ are continuous functions $x(\cdot)\colon \mathbb{R}^+ \mapsto \mathbb{R}^n$ which satisfy the initial condition $x(0) = x_0$; for any $\theta > 0$ their restrictions to $[0, \theta]$ are absolutely continuous and satisfy the differential inclusion (1.5) for almost all $t \in [0, \theta]$.

Similarly, let a feedback strategy $V : \mathbb{R}^n \mapsto Q$ and a partition Δ be chosen by the evader. By the symbol $\mathbf{X}(x_0, V, \Delta)$ we denote the set of step-by-step trajectories of the differential inclusion

$$\dot{x}(t) \in \operatorname{co}\{f(x(t), p, V(x(t_i)))\colon p \in P\}, \tag{1.6}$$

$t_i \le t < t_{i+1}$, $i = 0, 1, 2, \ldots$. It is obvious that any pair (U, V) of feedback strategies of the pursuer and the evader is compatible in the following sense: for arbitrary partitions $\Delta^{(1)}$ and $\Delta^{(2)}$ chosen by the pursuer and the evader there exists a unique trajectory of (1.1) corresponding to the controls given by

$$p(t) = U(x(t_i^{(1)})), \quad t_i^{(1)} \le t < t_{i+1}^{(1)}, \quad t_i^{(1)} \in \Delta^{(1)},\ i = 0, 1, 2, \ldots,$$
$$q(t) = V(x(t_j^{(2)})), \quad t_j^{(2)} \le t < t_{j+1}^{(2)},\ t_j^{(2)} \in \Delta^{(2)},\ j = 0, 1, 2, \ldots.$$

Therefore $\mathbf{X}(x_0, U, \Delta^{(1)}) \cap \mathbf{X}(x_0, V, \Delta^{(2)}) \ne \emptyset$.

Below we will use the notation

$$\operatorname{diam}(\Delta) := \sup_i (t_{i+1} - t_i) \quad \text{for } i = 0, 1, 2, \ldots. \tag{1.7}$$

Since the payoff functional (1.4) may in general be discontinuous, we need some additional constructions in order to define the value. Consider the functional

$$\tau_\varepsilon(x(\cdot)) := \min\{t \in \mathbb{R}^+ : x(t) \in M^\varepsilon\}. \tag{1.8}$$

If $x(t) \notin M^\varepsilon$ for all $t \in \mathbb{R}^+$, then we put $\tau_\varepsilon(x(\cdot)) := \infty$. Here ε is a positive number and M^ε is the ε-neighborhood of the terminal set M; that is,

$$M^\varepsilon := \{x + y : x \in M,\ \|y\| \le \varepsilon\}.$$

Further we introduce the values

$$T_1^\varepsilon(x_0, U, \Delta) := \sup\{\tau_\varepsilon(x(\cdot)) : x(\cdot) \in \mathbf{X}(x_0, U, \Delta)\},$$
$$T_1^\varepsilon(x_0, U) := \limsup_{\operatorname{diam}(\Delta) \downarrow 0} T_1^\varepsilon(x_0, U, \Delta),$$
$$T_1^\varepsilon(x_0) := \inf_U T_1^\varepsilon(x_0, U),$$
$$T_1^0(x_0) := \lim_{\varepsilon \downarrow 0} T_1^\varepsilon(x_0).$$

It is evident that $T_1^{\varepsilon'}(x_0) \geq T_1^{\varepsilon''}(x_0)$ for $\varepsilon' \leq \varepsilon''$. From this monotonicity we have that the last limit exists.

The quantity $T_1^0(x_0)$ is called the optimal result for the pursuer in the class of feedback strategies. Assume that $T_1^0(x_0) < \infty$. From the definition we obtain immediately that for any $\varepsilon > 0$ and $\theta > T_1^0(x_0)$ a feedback strategy U_* and a number $\delta > 0$ can be found such that for any step-by-step motion $x(\cdot) \in \mathbf{X}(x_0, U_*, \Delta)$, where $\operatorname{diam}(\Delta) \leq \delta$ the point $x(t)$ reaches the ε-neighborhood of the set M no later than at time θ.

Similarly we define the optimal result for the evader. We set

$$\begin{aligned} T_2^\varepsilon(x_0, V, \Delta) &:= \inf\{\tau_\varepsilon(x(\cdot)) : x(\cdot) \in \mathbf{X}(x_0, V, \Delta)\}, \\ T_2^\varepsilon(x_0, V) &:= \liminf_{\operatorname{diam}(\Delta)\downarrow 0} T_2^\varepsilon(x_0, V, \Delta), \\ T_2^\varepsilon(x_0) &:= \sup_V T_2^\varepsilon(x_0, V), \\ T_2^0(x_0) &:= \lim_{\varepsilon\downarrow 0} T_2^\varepsilon(x_0). \end{aligned}$$

The quantity $T_2^0(x_0)$ is called the optimal result for the evader in the class of feedback strategies. From the definitions we have that for any $\theta < T_2^0(x_0)$ there exist an $\varepsilon > 0$, a feedback strategy V_*, and a $\delta > 0$ such that any step-by-step motion $x(\cdot) \in \mathbf{X}(x_0, V_*, \Delta)$ on the time interval $[0, \theta]$ will avoid the ε-neighborhood of the set M, provided $\operatorname{diam}(\Delta) \leq \delta$.

For arbitrary step-by-step control choices $(U, \Delta^{(1)})$ and $(V, \Delta^{(2)})$ of the pursuer and the evader we have the inequalities

$$T_1^\varepsilon(x_0, U, \Delta^{(1)}) \geq \tau_\varepsilon(x_*(\cdot)) \geq T_2^\varepsilon(x_0, V, \Delta^{(2)}),$$

where

$$x_*(\cdot) \in \mathbf{X}(x_0, U, \Delta^{(1)}) \cap \mathbf{X}(x_0, V, \Delta^{(2)}).$$

Therefore $T_1^0(x_0) \geq T_2^0(x_0)$.

It is known [10] that under our assumptions the value $\operatorname{Val}(x_0)$ of the time-optimal differential game exists; that is, the equality

$$\operatorname{Val}(x_0) := T_1^0(x_0) = T_2^0(x_0) \tag{1.9}$$

holds.

The framework presented here is that of Krasovskii and Subbotin [10]. Of course, other well-known approaches have been developed; see, for example, [1], [2], [7], [8], [12].

The Bellman–Isaacs equation. Consider the following boundary value problem:

$$H(x, Dv(x)) + 1 = 0, \qquad x \in G := \mathbb{R}^n \setminus M, \tag{1.10}$$

$$v(x) = 0, \qquad x \in \partial G, \tag{1.11}$$

where $H(x, s)$ is the Hamiltonian defined by (1.3). For the differential game under consideration, equation (1.10) is called the Bellman–Isaacs equation. Let us recall some well-known facts.

Let a continuous function $v : \overline{G} \mapsto \mathbb{R}^+$ satisfy the boundary condition (1.11). Let this function be continuously differentiable in the domain G and satisfy equation (1.10). Then the function $v(\cdot)$ coincides with the value function $\mathrm{Val}(\cdot)$ of the time-optimal differential game. Moreover, in this case optimal feedback strategies U_0 and V_0 of the pursuer and the evader can be constructed with the help of v as follows. Introduce *extremal prestrategies* [10] which are defined by the relations

$$p_0(x,s) \in \operatorname{Arg}\min_{p\in P}[\max_{q\in Q}\langle s, f(x,p,q)\rangle], \tag{1.12}$$

$$q_0(x,s) \in \operatorname{Arg}\max_{q\in Q}[\min_{p\in P}\langle s, f(x,p,q)\rangle]. \tag{1.13}$$

We construct feedback strategies U_0 and V_0 as compositions of the prestrategies and the gradient Dv; that is,

$$U_0(x) := p_0(x, Dv(x)), \quad V_0(x) := q_0(x, Dv(x)). \tag{1.14}$$

It is well known, however, that the value function is differentiable only in rare situations. We will show that in the general case, optimal or suboptimal strategies of the pursuer can still be defined via v by relations of the form (1.14), but instead of the gradient $Dv(x)$, proximal subgradients of v are utilized.

2. The main result. Since it is unrealistic to suppose that v is smooth, a generalized solution concept is required. For our present purposes, we consider *proximal solutions*. We stress that as a solution concept, this is entirely equivalent (see [4]) to the earlier formulations of minimax [13], [14] or viscosity solutions [6], but certain advantages derive from using the proximal formulation: the calculus is more natural than that of Dini subderivates (used in minimax solutions) or that of comparison functions (used in viscosity solutions), and the geometric interpretation of a proximal subgradient lends itself to the dualization of results in either set terms or functional terms. To put this another way, the proximal formulation leads to a generalized solution concept that is well integrated with both the theory of nonsmooth analysis and a body of results in control theory bearing on many other issues, as opposed to standing alone. The survey paper [5] is in part a demonstration of this point.

A lower semicontinuous function $v\colon \overline{G} \to [0,\infty]$ is called a *proximal supersolution* of (1.10) provided that

$$H(x,\zeta) + 1 \le 0 \quad \forall x \in G, \qquad \forall \zeta \in \partial_P v(x). \tag{2.1}$$

Note that the inequality (2.1) holds vacuously at points x for which $\partial_P v(x)$ is empty (among which are the points x at which $v(x) = \infty$).

Given any nonnegative lower semicontinuous proximal supersolution v, we associate with it for any $\alpha \in (0, 1/2)$ a certain feedback strategy U_α^P of the pursuer, as follows. For $x \in \overline{G}$, select any minimizer $y_\alpha(x)$ over $\overline{G}$ of the function

$$y \mapsto \frac{|x-y|^2}{2\alpha^2} - \exp\{-2\lambda v(y)\}.$$

This minimum over $\overline{G}$ is attained, as is easily seen. Thus we have

$$y_\alpha(x) \in \operatorname{Arg}\min_{y\in\overline{G}}\left\{\frac{|x-y|^2}{2\alpha^2} - \exp(-2\lambda v(y))\right\}. \tag{2.2}$$

Subsequently, set

$$\zeta_\alpha(x) := \frac{x - y_\alpha(x)}{\alpha^2} \tag{2.3}$$

and define

$$U_\alpha^P(x) := p_0(x, \zeta_\alpha(x)), \tag{2.4}$$

where p_0 is an extremal prestrategy (1.12). Observe that the specification of U_α^P is quite explicit.

In what follows, we posit the hypotheses of the previous section on the data of the problem, except that the Isaacs condition (1.3) is *not* required. In its absence, and without having recourse to relaxed (mixed) control strategies or counterstrategies of the evader, the existence of the value cannot be asserted. However, this has no bearing on the theorem below, which simply provides a feedback strategy for one player (the pursuer) furnishing (up to a specified tolerance) a guaranteed result corresponding to any supersolution. In this one-sided setting, the Hamiltonian H appearing in (2.1) is given by

$$H(x, \zeta) := \min_{p \in P} \max_{q \in Q} \langle \zeta, f(x, p, q) \rangle.$$

THEOREM 2.1. *Let $v: \overline{G} \to [0, \infty]$ be a lower semicontinuous proximal supersolution of the Bellman–Isaacs equation (i.e., v satisfies (2.1)). Let D be a compact subset of G upon which v is bounded. Then for any $\varepsilon > 0$, there exist $\alpha > 0$ and $\delta > 0$ such that*

$$x_0 \in D,\ \operatorname{diam}(\Delta) < \delta \Rightarrow T_1^\varepsilon(x_0, U_\alpha^P, \Delta) \le v(x_0) + \varepsilon, \tag{2.5}$$

where U_α^P is the feedback strategy constructed from v as described above.

Remark 2.2.

(a) The feedback strategy U_α^P is defined independently of D, and the theorem asserts that it is "universal"; i.e., that within the specified tolerance and for α small enough (this does depend on D), it provides the guaranteed upper bound associated with v uniformly for all initial points in D.

(b) It follows from the results in [14] that in the presence of the Isaacs condition the value function $\operatorname{Val}(\cdot)$ is a supersolution of the Bellman–Isaacs equation. In fact, it is the minimal supersolution satisfying the boundary condition $v(x) = 0$, $x \in \partial G$. With $v(\cdot) = \operatorname{Val}(\cdot)$, the theorem then provides a universal ε-optimal feedback strategy. Note, however, that we do not suppose here that v coincides with Val and that (surprisingly) no boundary condition on ∂G enters into the theorem. The theorem evidently implies $T_1^0 \le v$.

(c) The construction of U_α^P corresponds to a functional version of the geometric approach called "proximal aiming" in [5]; it can also be viewed as a variant of the extremal aiming method of [9], [10]. We stress that the proof of the theorem is completely self-contained and elementary and that specific criteria for α and $\operatorname{diam}(\Delta)$ to be "small enough" can be inferred from it. Thus the feedback U_α^P is implementable at least conceptually. Let us stress that "small enough" here corresponds to picking α small *first* and *then* $\delta = \operatorname{diam}(\Delta)$ small in accordance with the order of the iterated limit appearing in (2.12) below.

(d) We remark that the construction of an explicit universal feedback strategy for the evader appears to be an essentially different problem, one that we do not address

here. It is known [10] that in the presence of the Isaacs condition, an appropriate pair of pursuit/evasion strategies would give rise to an approximate saddle point.

Proof of Theorem 2.1. We define a new lower semicontinuous $u\colon \overline{G} \to [0,1]$ via the Kruzhkov transformation:

$$u(x) := 1 - \exp(-2\lambda v(x)). \tag{2.6}$$

Then we claim that $u(\cdot)$ satisfies

$$H(x,\zeta) - 2\lambda(u(x)-1) \le 0 \quad \forall x \in G, \qquad \forall \zeta \in \partial_P u(x). \tag{2.7}$$

This is straightforward from (2.1), (2.6) when $u(x) < 1$, since then $v(x) < \infty$ and $\partial_P u(x)$ and $\partial_P v(x)$ are related in an evident way. When $u(x) = 1$, the only possible element ζ of $\partial_P u(x)$ is $\zeta = 0$, as follows readily from the definition of $\partial_P u(x)$ and the fact that $u \le 1$ everywhere. But then (2.7) holds, since $H(x,0) = 0$.

We define u_α on $\mathbb{R}^n$ by

$$u_\alpha(x) := \min_{y \in \overline{G}} \left\{ u(y) + \frac{|x-y|^2}{2\alpha^2} \right\}, \tag{2.8}$$

an inf-convolution familiar in many settings, notably convex and functional analysis (Moreau–Yosida approximations) and partial differential equations (see, for example, [11], [15]).

Note that by (2.2) the minimum is attained at $y_\alpha(x)$ and that we have $0 \le u_\alpha(x) \le u(x) \le 1$. In addition, the minimum must be attained at points y for which $|x-y|^2/(2\alpha^2)$ is no greater than 1, whence

$$|x - y_\alpha(x)| \le 2\alpha. \tag{2.9}$$

Finally, when $y_\alpha(x) \in G$ (open), then the stationarity condition holds at $y_\alpha(x)$: zero is a proximal subgradient of the function of y appearing in (2.8); i.e., we have

$$y_\alpha(x) \in G \Rightarrow \zeta_\alpha(x) \in \partial_P u(y_\alpha(x)) \tag{2.10}$$

(recall that ζ_α is defined by (2.3)).

Denote by $X(x_0)$ the family of all trajectories $x(\cdot)$ on $[0,\infty)$ of the differential inclusion

$$\dot{x}(t) \in \operatorname{co}\{f(x(t),p,q)\colon p \in P, q \in Q\}$$

satisfying the initial condition $x(0) = x_0$. Let T_0 be an upper bound for $v(\cdot)$ on D, and define

$$K := \{x(t) \in \mathbb{R}^n\colon x(\cdot) \in X(x_0), t \in [0, T_0+1]\},$$
$$m := \sup\{|f(x,p,q)|\colon x \in K, p \in P, q \in Q\}.$$

Note that K is bounded and $m < \infty$. Choose numbers $\alpha \in (0,1/2)$ and $\delta_0 > 0$ such that

$$3\alpha \le \varepsilon. \tag{2.11}$$

Choose any $x_0 \in D$ and set $\theta := v(x_0) + \varepsilon$. Without loss of generality we may assume $\varepsilon \le 1$, so that $\theta \le T_0 + 1$.

We now construct a certain function h which will provide the estimate (2.5) of the theorem. □

LEMMA 2.3. *There is a nonnegative function $h(\alpha,\delta)$ depending only on α and δ (and not on $x_0 \in D$ nor on $x(\cdot) \in X(x_0, U_\alpha^P, \Delta)$) such that*

$$\lim_{\alpha\downarrow 0}(\lim_{\delta\downarrow 0} h(\alpha,\delta)) = 0 \tag{2.12}$$

and having the following property: for any $x(\cdot) \in X(x_0, U_\alpha^P, \Delta)$ let $t_i \in \Delta$, $t_i < \theta$ and suppose $\operatorname{dist}(x(t_i), M) > 3\alpha$. *Then for any $\tau \in [t_i, t_{i+1}] \cap [0,\theta]$ we have*

$$u_\alpha(x(\tau)) \le 1 + e^{2\lambda(\tau-t_i)}[u_\alpha(x(t_i)) - 1] + (\tau - t_i)e^{2\lambda(\tau-t_i)}h(\alpha, \operatorname{diam}(\Delta)). \tag{2.13}$$

To prove this, let f^* be defined by

$$f^* := \frac{x(\tau) - x(t_i)}{\tau - t_i} = \frac{1}{\tau - t_i}\int_{t_i}^{\tau} \dot{x}(t)\,dt.$$

Note that $\dot{x}(t) \in \operatorname{co}\{f(x(t), U_\alpha^P(x(t_i)), q)\colon q \in Q\}$ and $|f^*| \le m$. Below we use the notation

$$\xi := x(t_i), \qquad \zeta_\alpha := \zeta_\alpha(\xi), \qquad \eta_\alpha := y_\alpha(x(t_i)), \qquad \mu := \tau - t_i.$$

Note that $\zeta_\alpha = (\xi - \eta_\alpha)/\alpha^2$ and $u_\alpha(x(\tau)) = u_\alpha(\xi + \mu f^*)$. We observe from (2.8) and the definition of $\eta_\alpha = y_\alpha(\xi)$ that

$$\begin{aligned} u_\alpha(\xi + \mu f^*) &\le u(\eta_\alpha) + |\xi + \mu f^* - \eta_\alpha|^2/(2\alpha^2) \\ &= u(\eta_\alpha) + |\xi - \eta_\alpha|^2/(2\alpha^2) + \langle \xi - \eta_\alpha, \mu f^*\rangle/\alpha^2 + \frac{\mu^2}{2\alpha^2}|f^*|^2 \\ &= u_\alpha(\xi) + \mu\langle \zeta_\alpha, f^*\rangle + \frac{\mu^2}{2\alpha^2}|f^*|^2. \end{aligned} \tag{2.14}$$

We now require a bound for the middle term in this last expression.

By definition (i.e., (2.4)) we have

$$\max_{q\in Q}\langle \zeta_\alpha, f(\xi, U_\alpha^P(\xi), q)\rangle = \min_{p\in P}\max_{q\in Q}\langle \zeta_\alpha, f(\xi, p, q)\rangle = H(\xi, \zeta_\alpha),$$

whence

$$\langle \zeta_\alpha, f\rangle \le H(\xi, \zeta_\alpha) \quad \forall f \in \operatorname{co}\{f(\xi, U_\alpha^P(\xi), q)\colon q \in Q\}.$$

Also, we have $|\xi - \eta_\alpha| \le 2\alpha$ and $|\zeta_\alpha| \le 2/\alpha$ by (2.9) and (2.3). Invoking the positive homogeneity of $H(x,s)$ in s and its Lipschitz continuity in x, we derive

$$\begin{aligned} H(\xi, \zeta_\alpha) &\le H(\eta_\alpha, \zeta_\alpha) + \lambda|\xi - \eta_\alpha|\,|\zeta_\alpha| \\ &= H(\eta_\alpha, \zeta_\alpha) + \lambda|\xi - \eta_\alpha|^2/\alpha^2. \end{aligned} \tag{2.15}$$

We note the existence of $\tilde{f} \in \operatorname{co}\{f(\xi, U_\alpha^P(\xi), q)\colon q \in Q\}$ such that $|f^* - \tilde{f}| \le \gamma(\mu)$, where $\gamma(\mu) \to 0$ as $\mu \downarrow 0$. Hence

$$\langle \zeta_\alpha, f^*\rangle \le \langle \zeta_\alpha, \tilde{f}\rangle + \tilde{h}(\alpha, \mu), \tag{2.16}$$

where $\tilde{h}(\alpha,\mu) := 2\gamma(\mu)/\alpha$. Combining (2.15), (2.16) gives

$$\langle \zeta_\alpha, f^* \rangle \le H(\eta_\alpha, \zeta_\alpha) + \lambda|\xi - \eta_\alpha|^2/\alpha^2 + \tilde{h}(\alpha,\mu),$$

where $\lim_{\alpha\downarrow 0}(\lim_{\mu\downarrow 0} \tilde{h}(\alpha,\mu)) = 0$.

We now invoke the above estimates together with (2.14) to deduce

$$u_\alpha(\xi + \mu f^*) \le u_\alpha(\xi) + \mu[H(\eta_\alpha,\zeta_\alpha) + \lambda|\xi-\eta_\alpha|^2/\alpha^2] + \frac{\mu^2}{2\alpha^2}|f^*|^2 + \mu\tilde{h}(\alpha,\mu).$$

Since $\operatorname{dist}(\xi, M) > 3\alpha$ by assumption and since $|\xi - \eta_\alpha| \le 2\alpha$, it follows that η_α lies in G. Thus $\zeta_\alpha \in \partial_P u_\alpha(\eta_\alpha)$ by (2.10). We may therefore use (2.7) to write

$$\begin{aligned} H(\eta_\alpha,\zeta_\alpha) + \lambda|\xi-\eta_\alpha|^2/\alpha^2 &\le 2\lambda u(\eta_\alpha) - 2\lambda + \lambda|\xi-\eta_\alpha|^2/\alpha^2 \\ &= 2\lambda\left\{u(\eta_\alpha) + \frac{|\xi-\eta_\alpha|^2}{2\alpha^2}\right\} - 2\lambda = 2\lambda u_\alpha(\xi) - 2\lambda. \end{aligned}$$

Substituting in the previous estimate gives

$$\begin{aligned} u_\alpha(\xi+\mu f^*) &\le u_\alpha(\xi) + 2\lambda\mu[u_\alpha(\xi) - 1] + \frac{\mu^2}{2\alpha^2}|f^*|^2 + \mu\tilde{h}(\alpha,\mu) \\ &= (1+2\lambda\mu)u_\alpha(\xi) - 2\lambda\mu + \frac{\mu^2}{2\alpha^2}|f^*|^2 + \mu\tilde{h}(\alpha,\mu) \\ &\le e^{2\lambda\mu}u_\alpha(\xi) - 2\lambda\mu + \frac{\mu^2}{2\alpha^2}m^2 + \mu\tilde{h}(\alpha,\mu) \\ &= 1 + e^{2\lambda\mu}[u_\alpha(\xi) - 1] + \mu e^{2\lambda\mu}h(\alpha,\mu), \end{aligned}$$

where

$$h(\alpha,\mu) := \left[\tilde{h}(\alpha,\mu) + \frac{\mu}{2\alpha^2}m^2 - 2\lambda + \frac{e^{2\lambda\mu}-1}{\mu}\right]e^{-2\lambda\mu}$$

is nonnegative and has the required limiting property. The lemma is proved.

We now require of the parameters α and δ $(= \operatorname{diam}(\Delta))$ that they satisfy

$$\theta e^{2\lambda\theta}h(\alpha,\delta) < e^{2\lambda\varepsilon} - 1. \tag{2.17}$$

If for every such Δ and $x(\cdot)$ as in the lemma there exists $t_i \in \Delta$ such that $t_i < \theta$ and $\operatorname{dist}(x(t_i), M) \le 3\alpha$, then (since $3\alpha \le \varepsilon$ by (2.11)) the conclusion (2.5) of Theorem 2.1 follows immediately. Let us suppose the contrary (for some such Δ and x) and derive a contradiction.

By iterating (2.13) as a recurrent inequality for $u_\alpha - 1$ we easily derive

$$u_\alpha(x(\theta)) - 1 \le [u_\alpha(x_0) - 1]e^{2\lambda\theta} + \theta e^{2\lambda\theta}h(\alpha,\delta).$$

It is clear that $e^{2\lambda\theta} = e^{2\lambda\varepsilon}(1 - u(x_0))^{-1}$, and we have $u_\alpha(x_0) \le u(x_0)$. This yields

$$u_\alpha(x(\theta)) \le 1 - e^{2\lambda\varepsilon} + \theta e^{2\lambda\theta}h(\alpha,\delta) < 0 \qquad \text{(by (2.17))},$$

which cannot be since u_α is nonnegative by construction. This completes the proof of Theorem 2.1.

Remark 2.4. It is a familiar observation that the Lipschitz condition on $f(x,p,q)$ in the x variable can be weakened by requiring it to hold for x in any given compact set K (with the Lipschitz constant λ depending on K), if one postulates the linear growth (in x) condition on f.

REFERENCES

[1] L. D. Berkovitz, *Differential games of generalized pursuit and evasion*, SIAM J. Control, 24 (1986), pp. 361–373.

[2] A. Blaquiere and L. G. Leitmann, *Jeux quantitatifs*, Gauthier-Villars, Paris, 1969.

[3] F. H. Clarke, *Optimization and Nonsmooth Analysis*, Classics in Applied Mathematics 5, SIAM, Philadelphia, PA, 1990 (reprinted). (Original edition: Wiley, New York, 1983.)

[4] F. H. Clarke and Yu. S. Ledyaev, *Mean value inequalities in Hilbert space*, Trans. Amer. Math. Soc., 344 (1994) pp. 307–324.

[5] F. H. Clarke, Yu. S. Ledyaev, R. J. Stern, and P. R. Wolenski, *Qualitative properties of trajectories of control systems: A survey*, J. Dynamical and Control Systems, 1 (1995), pp. 1–48.

[6] M. G. Crandall and P.-L. Lions, *Viscosity solutions of Hamilton-Jacobi equations*, Trans. Amer. Math. Soc., 277 (1983), pp. 1–42.

[7] R. J. Elliott and N. J. Kalton, *The existence of value in differential games of pursuit and evasion*, J. Differential Equations, 126 (1972), pp. 504–523.

[8] A. Friedman, *Differential Games*, Wiley-Interscience, New York, 1971.

[9] G. G. Garnysheva and A. I. Subbotin, *Strategies of minimax aiming in the direction of quasigradient*, Prikl. Mat. Mekh., 58 (1994), pp. 72–78 (in Russian).

[10] N. N. Krasovskii and A. I. Subbotin, *Positional Differential Games*, Nauka, Moscow, 1974 (in Russian); French translation: *Jeux Différentiels*, editions Mir, Moscou, 1979; revised English translation: *Game-Theoretical Control Problems*, Springer-Verlag, New York, 1988.

[11] R. R. Phelps, *Convex Functions, Monotone Operators and Differentiability*, 2nd ed., Lecture Notes in Mathematics 1364, Springer-Verlag, New York, 1993.

[12] P. Soravia, *Pursuit evasion problems and viscosity solutions of Isaacs' equations*, SIAM J. Control Optim., 31 (1993), pp. 604–623.

[13] A. I. Subbotin, *Minimax Inequalities and Hamilton-Jacobi Equations*, Nauka, Moscow, 1991 (in Russian).

[14] A. I. Subbotin, *Generalized Solutions of First Order PDEs: The Dynamic Optimization Perspective*, Birkhäuser Boston, Cambridge, MA, 1994.

[15] E. Zeidler, *Nonlinear Functional Analysis and Its Applications* III, Springer-Verlag, New York, 1984.

SIAM J. CONTROL OPTIM.
Vol. 35, No. 2, pp. 562–584, March 1997

011

BRACHISTOCHRONE WITH COULOMB FRICTION*

STEPHEN C. LIPP†

Abstract. This paper formulates and solves in closed form the problem of finding the minimum-time path of a particle between two points in a uniform gravitational field when motion of the particle is resisted by a force proportional to the normal force exerted on the particle by the path. This resistance to motion is the common mathematical form for Coulomb friction. The problem solution involves the reformulation of the classical brachistochrone of Bernoulli in terms of a singular control problem in which the time derivative of the heading angle of the particle is the control parameter. As such, this solution provides a unique approach to the solution of minimum-time path problems.

Key words. brachistochrone, singular control, extremal control, minimum-time control, corner conditions, optimal control

AMS subject classifications. Primary, 49K15, 49K30; Secondary, 49J15, 49J30

PII. S0363012995287957

1. Introduction.

Minimum-time optimal control problems are a common research area in analytical and numerical control system synthesis. Current research in robotics and automation is fraught with minimum-time optimization problems [4, 8, 16, 20]. Some of the earliest work in optimal control theory was concerned with minimum-time optimization [3]. Perhaps the earliest problem proposed in minimum-time optimization was the brachistochrone of Bernoulli [1]. This problem may be stated as follows: A bead slides on a frictionless wire between points A and B in a constant-gravity field. The bead has an initial speed V_0 at point A. What is the shape of the wire that will produce a minimum-time path between the two points? Assuming the positive y axis points upward with $A = (x_0, y_0)$ and $B = (0, 0)$, the optimal path is a portion of a cycloid [2], the path generated by a point on the circumference of a circle as it rolls in the direction of the x axis. The minimum-time solution in this case requires that the time rate of change of heading angle of the bead is constant [2].

Euler proposed an extension to this problem. As quoted from Goldstine [7], Euler "... takes up the elegant problem of finding the shape of the brachystochrone [sic] curve in case the medium through which the heavy particle falls resists the motion depending only on the velocity v. Here he assumes that the resistance function R is proportional to v^{2n}." Euler's solution to this problem is an implicit solution [5]. Due to the fact that the friction is proportional to a function of the velocity only, this problem may be solved using the calculus of variations.

Current research into the brachistochrone has been primarily focused on different approaches to solving the classical brachistochrone problem. For example, Roomany [19] uses a graph theoretical approach to solve the classical brachistochrone problem. This technique is efficient and convergent but is only useful if the field through which the particle moves is derivable from a potential. In a similar manner, Razzaghi and Elnagar solve the classical brachistochrone problem using interpolating polynomials of appropriate degree [18]. Szarkowicz approaches the solution of the brachistochrone

*Received by the editors June 19, 1995; accepted for publication (in revised form) February 4, 1996. A version of this paper was presented at the 35th IEEE Conference on Decision and Control, Kobe, Japan, December 11–13, 1996.

http://www.siam.org/journals/sicon/35-2/28795.html

†Department of Mechanical Engineering, University of New Orleans, New Orleans, LA 70148-2230 (sclme@jazz.ucc.uno.edu).

problem with use of multistage Monte Carlo methods [22]. Some extensions of the classical brachistochrone problem have been discussed in the current literature as well. Hoskins extended and solved the classical brachistochrone problem, where the change in direction is allowed only along circles of a given radius R [9]. Perlick solved the classical brachistochrone problem extended to a stationary Lorentzian space-time [14].

The present paper examines the brachistochrone problem in which the friction force on the particle is a resistance proportional to the force exerted by the particle by its path (the normal force). This formulation is the equivalent of determining minimum-time paths in a uniform gravitational field with Coulomb friction resisting the motion. With this choice of motion resistance, the penalty for curvature in the path is high. In fact, the dynamics to be derived will demand that there be no abrupt changes in curvature. Therefore, if the optimal control problem is reformulated with an intermediate point as the initial point and the corresponding speed at that point as the initial speed, a *different* optimal path will result. This again differs from the classical brachistochrone result in that the position and speed at any point on an optimal cycloidal path determine the cycloid. It is the position and *velocity* which uniquely determine the time-optimal path when Coulomb friction is present in the brachistochrone problem.

A discretization of this problem was formulated and solved using a Davidon–Fletcher–Powell algorithm in 1990 [17]. A number of important points in this result were left unsolved, including the question of whether a solution exists and how close a solution is to the actual minimum-time solution. Some interesting features of the minimum-time path were not illustrated in the discretized solution. This paper will present the closed-form solution for the minimum-time path. Therefore, it will demonstrate the validity of the discretized approach and provide further insight into the limitations of that generic approach in its application to this problem.

2. Optimal control formulation. The classical brachistochrone problem may be formulated as an optimal control problem with the heading angle θ as the control variable. In the formulation of any optimal control problem, the first step is the formation of the scalar Hamiltonian [15]. The Hamiltonian is defined by

$$\mathcal{H} = \langle \boldsymbol{\lambda}, \boldsymbol{f}(\boldsymbol{x}, u) \rangle + L(\boldsymbol{x}, u), \tag{2.1}$$

where $\dot{\boldsymbol{x}} = \boldsymbol{f}(\boldsymbol{x}, u)$ are the state dynamics, $\dot{\boldsymbol{\lambda}} = -\frac{\partial \mathcal{H}}{\partial \boldsymbol{x}}$ are the "costate" dynamics, $L(\boldsymbol{x}, u)$ is the scalar function whose time integral is to be minimized, u is the control variable, and $\langle \cdot, \cdot \rangle$ denotes the scalar inner (dot) product. For minimum-time control problems, the scalar cost function $L(\boldsymbol{x}, u) \equiv 1$, since the minimization of the time integral of one corresponds to the minimization of the final time.

Pontryagin's "minimum" principle [15] states that the optimal control is the control u which minimizes the Hamiltonian almost everywhere. When the Hamiltonian is convex in the control variable u, the optimal control necessarily satisfies Euler's condition (the first variation) $\frac{\partial \mathcal{H}}{\partial u} = 0$ and sufficiency is guaranteed with the "strict" Legendre–Clebsch condition (the second variation) $\frac{\partial^2 \mathcal{H}}{\partial u^2} > 0$. The strict Legendre–Clebsch condition is guaranteed in the case of the classical brachistochrone.

The state dynamics for the standard brachistochrone may be derived in x and y coordinates using Newton's laws. Assume that positive x is to the right and positive y is up. Additionally, assume that a first quadrant heading angle θ corresponds to negative x-speed and y-speed. The particle speed is denoted by V. A diagram of the geometry is shown in Figure 2.1.

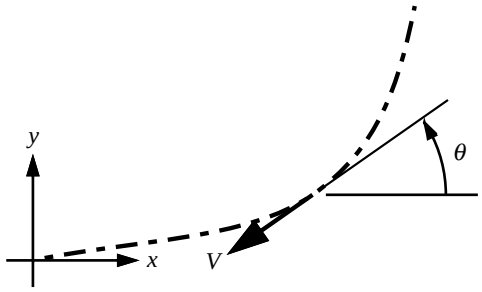


FIG. 2.1. *State-variable geometry.*

Using dimensionless variables

$$\begin{aligned} \dot{x} &= -V\cos\theta, \\ \dot{y} &= -V\sin\theta, \\ \dot{V} &= \frac{1}{2}\sin\theta\,. \end{aligned} \tag{2.2}$$

The dimensionless variables chosen were

$$x = \frac{x^*}{x^*(0)}, \qquad y = \frac{y^*}{x^*(0)}, \qquad V = \frac{V^*}{\sqrt{2gx^*(0)}}, \qquad t = t^*\sqrt{\frac{2g}{x^*(0)}}\,, \tag{2.3}$$

where the superscript $*$ indicates dimensional variables. Treating θ as the control variable, the initial and final conditions on the states are

$$x(0) = 1, \quad y(0) = y_0, \quad V(0) = V_0, \quad x(t_f) = 0, \quad y(t_f) = 0\,. \tag{2.4}$$

Setting

$$\boldsymbol{\lambda} = \left\{ \begin{array}{c} \lambda_x \\ \lambda_y \\ \lambda_V \end{array} \right\}, \tag{2.5}$$

the Hamiltonian may be written as

$$\mathcal{H} = -\lambda_x V\cos\theta - \lambda_y V\sin\theta + \frac{1}{2}\lambda_V\sin\theta + 1\,. \tag{2.6}$$

The costate dynamics as obtained from $\dot{\boldsymbol{\lambda}} = -\frac{\partial\mathcal{H}}{\partial \boldsymbol{x}}$ are

$$\begin{aligned} \dot{\lambda}_x &= \dot{\lambda}_y = 0, \\ \dot{\lambda}_V &= \lambda_x\cos\theta + \lambda_y\sin\theta\,. \end{aligned} \tag{2.7}$$

With the state boundary conditions given, the costates $\boldsymbol{\lambda}$ have free (unknown) boundary conditions except at the boundary where $\boldsymbol{x}$ is unconstrained. Thus the only boundary condition satisfied by $\boldsymbol{\lambda}$ is

$$\lambda_V(t_f) = 0\,. \tag{2.8}$$

This means that there are six differential equations and seven unknowns ($V(t_f)$, $\lambda_x(0)$, $\lambda_y(0)$, $\lambda_V(0)$, $\lambda_x(t_f)$, $\lambda_y(t_f)$, and t_f).

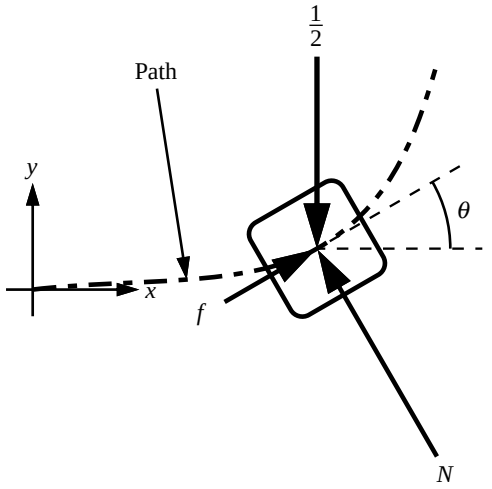


FIG. 2.2. *Free-body diagram of bead on path with Coulomb friction resisting motion.*

The final condition for determination of the optimal control is the transversality condition [15], which states that

$$\mathcal{H}(t_f) = \frac{\partial L(\boldsymbol{x}, u)}{\partial t} = 0\,. \tag{2.9}$$

With the Hamiltonian having no explicit time dependence, this means that $\mathcal{H}(t) \equiv 0$. Simple integration demonstrates that the first and second variation conditions hold for a cycloidal path. Therefore, the cycloidal path is the optimal minimum-time path for control.

When Coulomb friction is added as a resisting force to the motion, the dimensionless equations of motion may be determined from a free-body diagram of the particle (see Figure 2.2). In the figure, the dimensionless normal force N and dimensionless friction force f are given by

$$N = \frac{N^*}{2mg} \qquad \text{and} \qquad f = \frac{f^*}{2mg}\,. \tag{2.10}$$

Coulomb friction resistance has the constitutive law $f = \mu|N|$, where μ is the coefficient of friction. The normal force exerted by the path on the particle is equal to the component of the weight of the particle (1/2 in dimensionless terms) in the direction of the normal force minus the mass of the particle multiplied by the centripetal acceleration caused by the particle's speed and the curvature of the path. In dimensionless terms, this is $V\dot{\theta}$. Therefore, the dimensionless equations of motion are

$$\begin{aligned} \dot{x} &= -V\cos\theta, \\ \dot{y} &= -V\sin\theta, \\ \dot{V} &= \frac{1}{2}\sin\theta - \mu\left|\frac{1}{2}\cos\theta - V\dot{\theta}\right|\,. \end{aligned} \tag{2.11}$$

Unlike the classical or Euler's brachistochrone problem, $\dot{\theta}$ has entered the dynamics. In order to formulate the problem as an optimal control problem, θ will be adjoined as an extra state with $\dot{\theta} = \Omega$ as the control variable.

This formulation is troublesome due to the fact that the control objective and the state dynamics are not convex in the control variable. The second derivative of the Hamiltonian with respect to the control being zero is not sufficient for optimality. In fact, Euler's necessary condition is not satisfied initially for nonzero initial velocity. Recourse to Pontryagin's minimum principle yields extremal control initially but yields no information about the control when the system equations simplify with Euler's condition being satisfied. This problem requires the theory of singular control [10, 11].

When the partial derivative of the Hamiltonian with respect to the control variable equals zero but yields a relation in which there is no explicit control variable, then a singular control solution may exist for the system. Singular control methods yield the following constraint equations for determining the control:

$$\begin{aligned} &\frac{d^i}{dt^i}\left[\frac{\partial \mathcal{H}}{\partial u}\right] = 0, \qquad i = 0, 1, \ldots, m-1, \\ &\frac{d^m}{dt^m}\left[\frac{\partial \mathcal{H}}{\partial u}\right] = g(\boldsymbol{x}, \boldsymbol{\lambda}, u) = 0. \end{aligned} \tag{2.12}$$

The first $m-1$ derivatives are independent of the control u. The mth derivative determines the control u as a function of the states/costates. A result obtainable from this theory is [11]: (1) m is always even; (2) if the number of states is n, the singular control restricts state/costate space to dimension $2n-m$ (the singular hypersurface); and (3) it is necessary that

$$(-1)^{\frac{m}{2}} \frac{\partial}{\partial u}\left(\frac{d^m}{dt^m}\left[\frac{\partial \mathcal{H}}{\partial u}\right]\right) \geq 0\,.$$

The final condition is a generalized Legendre–Clebsch condition, a *necessary* condition for optimality. In general there are no sufficiency conditions for optimality of a singular control arc. Even the strict inequality does not guarantee optimality. Nevertheless, this method will prove invaluable in generating control trajectories for the brachistochrone with Coulomb friction. Furthermore, since the classical brachistochrone can be derived as a singular control problem and the Coulomb friction brachistochrone simplifies to the classical brachistochrone when the coefficient of friction $\mu \to 0$, the necessarily optimal path derived by singular control methods is sufficient in the limit.

3. Reformulation of the classical brachistochrone problem. The classical brachistochrone problem can also be formulated as a singular control problem where the control variable is Ω, the time rate of change of the slope of the time-optimal path. In this case, the state equations are

$$\begin{aligned} \dot{x} &= -V\cos\theta, \\ \dot{y} &= -V\sin\theta, \\ \dot{V} &= \frac{1}{2}\sin\theta, \\ \dot{\theta} &= \Omega\,. \end{aligned} \tag{3.1}$$

The dimensionless Hamiltonian is thus

$$\mathcal{H} = -\lambda_x V\cos\theta - \lambda_y V\sin\theta + \frac{1}{2}\lambda_V \sin\theta + \lambda_\theta \Omega + 1\,. \tag{3.2}$$

The costate equations are

$$\begin{aligned}
\dot{\lambda}_x &= -\frac{\partial \mathcal{H}}{\partial x} = 0,\\
\dot{\lambda}_y &= -\frac{\partial \mathcal{H}}{\partial y} = 0,\\
\dot{\lambda}_V &= -\frac{\partial \mathcal{H}}{\partial V} = \lambda_x \cos\theta + \lambda_y \sin\theta,\\
\dot{\lambda}_\theta &= -\frac{\partial \mathcal{H}}{\partial \theta} = V(\lambda_y \cos\theta - \lambda_x \sin\theta) - \frac{1}{2}\,\lambda_V \cos\theta
\end{aligned} \tag{3.3}$$

with the boundary conditions

$$\lambda_V(t_f) = 0, \qquad \lambda_\theta(0) = \lambda_\theta(t_f) = 0, \qquad \mathcal{H}(t_f) = 0\,. \tag{3.4}$$

With the Hamiltonian being linear in the control variable Ω, the optimal control will be singular control. The singular control is determined with the following conditions:

$$\begin{aligned}
\frac{\partial \mathcal{H}}{\partial \Omega} &= \lambda_\theta = 0,\\
\frac{d}{dt}\,\frac{\partial \mathcal{H}}{\partial \Omega} &= V(\lambda_y \cos\theta - \lambda_x \sin\theta) - \frac{1}{2}\,\lambda_V \cos\theta = 0,\\
\frac{d^2}{dt^2}\,\frac{\partial \mathcal{H}}{\partial \Omega} &= -\frac{1}{2}\,\lambda_x + \left(\frac{1}{2}\,\lambda_V \sin\theta - V(\lambda_x \cos\theta + \lambda_y \sin\theta)\right)\Omega = 0\,.
\end{aligned} \tag{3.5}$$

These conditions restrict state/costate space to dimension six. This restriction in the dimensionality of the state/costate dynamics is known as the singular surface; the trajectory defined on this surface is known as the singular arc. The generalized Legendre–Clebsch condition for this singular control problem is

$$-\frac{\partial}{\partial \Omega}\left[\frac{d^2}{dt^2}\left(\frac{\partial \mathcal{H}}{\partial \Omega}\right)\right] = V(\lambda_x \cos\theta + \lambda_y \sin\theta) - \frac{1}{2}\,\lambda_V \sin\theta \geq 0\,. \tag{3.6}$$

Solving the differential equations for the states/costates and choosing θ as the independent parameter yields (see section 4.2 for more details)

$$\begin{aligned}
x(\theta) &= \frac{V_f^2}{2\cos^2\theta_f}\,(2(\theta - \theta_f) + \sin 2\theta - \sin 2\theta_f), & \lambda_x(\theta) &= \frac{\cos\theta_f}{V_f},\\
y(\theta) &= \frac{V_f^2}{2\cos^2\theta_f}\,(\cos 2\theta_f - \cos 2\theta), & \lambda_y(\theta) &= \frac{\sin\theta_f}{V_f},\\
V(\theta) &= V_f\,\frac{\cos\theta}{\cos\theta_f}, & \lambda_V(\theta) &= -\frac{2\sin(\theta - \theta_f)}{\cos\theta_f},\\
t(\theta) &= t_f - \frac{2V_f}{\cos\theta_f}\,(\theta - \theta_f), & \lambda_\theta(\theta) &= 0\,.
\end{aligned} \tag{3.7}$$

The values of V_f, θ_f, and θ_0 are determined from the initial conditions on x, y, and V. The fact that $\lambda_\theta(\theta) \equiv 0$ indicates the singular solution for the control exists throughout the control trajectory. The optimal control is

$$\Omega = -\frac{\cos\theta_f}{2V_f}, \tag{3.8}$$

satisfying the necessary generalized Legendre–Clebsch condition

$$V(\lambda_x \cos\theta + \lambda_y \sin\theta) - \frac{1}{2}\,\lambda_V \sin\theta = 1 > 0\,.$$

These results will serve as a check for the brachistochrone problem with Coulomb friction. When $\mu = 0$, the Coulomb friction results should reduce to the above results.

4. Brachistochrone with Coulomb friction.

4.1. The equations of motion. The dimensionless equations of motion for the brachistochrone with Coulomb friction are obtained from (2.11) with $\dot\theta = \Omega$ appended. The boundary conditions for the states are

$$(x(\,0\,), y(\,0\,), V(\,0\,), \theta(\,0\,)) = (1, y_0, V_0, \theta_0)\,, \tag{4.1}$$

$$(x(t_f), y(t_f), V(t_f), \theta(t_f)) = (0,\ 0\,, V_f, \theta_f)\,, \tag{4.2}$$

where θ_0, V_f, and θ_f are unknown parameters.

The Hamiltonian for the system is

$$\mathcal{H} = -\lambda_x V\cos\theta - \lambda_y V\sin\theta + \lambda_V\left(\frac{1}{2}\,\sin\theta - \mu\left|\frac{1}{2}\,\cos\theta - V\Omega\right|\right) + \lambda_\theta\Omega + 1\,. \tag{4.3}$$

Therefore, the equations for the costates are

$$\begin{aligned}
\dot\lambda_x &= 0,\\
\dot\lambda_y &= 0,\\
\dot\lambda_V &= \lambda_x\cos\theta + \lambda_y\sin\theta - \mu s\lambda_V\Omega,\\
\dot\lambda_\theta &= -\lambda_x V\sin\theta + \lambda_y V\cos\theta - \frac{1}{2}\,\lambda_V(\cos\theta + \mu s\sin\theta)
\end{aligned} \tag{4.4}$$

with the boundary conditions

$$(\lambda_x(\,0\,), \lambda_y(\,0\,), \lambda_V(\,0\,), \lambda_\theta(\,0\,)) = (\lambda_{x_0}, \lambda_{y_0}, \lambda_{V_0}, 0)\,, \tag{4.5}$$

$$(\lambda_x(t_f), \lambda_y(t_f), \lambda_V(t_f), \lambda_\theta(t_f)) = (\lambda_{x_f}, \lambda_{y_f},\ \ 0\ \ , 0)\,, \tag{4.6}$$

where λ_{x_0}, λ_{y_0}, λ_{V_0}, λ_{x_f}, and λ_{y_f} are unknown. In the above equations

$$s = \operatorname{sgn}\left(\frac{1}{2}\,\cos\theta - V\Omega\right) = \operatorname{sgn}(N)\,. \tag{4.7}$$

Note that for $N = 0$, the above differential equations for λ_V and λ_θ are not properly defined. This case will be considered in section 4.3.

4.2. The singular control arc. The first step in synthesizing the singular control arc for this problem is to derive the constraint equations. To simplify the following analysis, three parameters will be introduced to replace three parameters in the state/costate equations. These parameters are defined as follows:

$$\sin\theta_\mu = \frac{s\mu}{\sqrt{1+\mu^2}},\quad \cos\theta_\mu = \frac{1}{\sqrt{1+\mu^2}},\quad \lambda_x = L\cos\theta_\lambda,\quad \lambda_y = L\sin\theta_\lambda \tag{4.8}$$

with $L > 0$. Noting that the time derivatives of λ_x and λ_y are both zero, the reduced set of six state/costate equations is

$$\begin{aligned}
\dot{x} &= -V\cos\theta,\\
\dot{y} &= -V\sin\theta,\\
\dot{V} &= \frac{1}{2\cos\theta_\mu}\sin(\theta-\theta_\mu) + \tan\theta_\mu V\Omega,\\
\dot{\theta} &= \Omega,\\
\dot{\lambda}_V &= L\cos(\theta-\theta_\lambda) - \tan\theta_\mu\lambda_V\Omega,\\
\dot{\lambda}_\theta &= -LV\sin(\theta-\theta_\lambda) - \frac{1}{2\cos\theta_\mu}\lambda_V\cos(\theta-\theta_\mu)\,.
\end{aligned} \tag{4.9}$$

The seven unknown parameters for these six equations are θ_0, θ_f, V_f, L, θ_λ, λ_{V_0}, and t_f. There is also θ_μ, which is unknown because of s being unknown. In forthcoming equations, s is assumed constant; its sign will be determined from the generalized Legendre–Clebsch condition. The seventh condition is the transversality condition which, for minimum-time optimization, states that the Hamiltonian is zero at time $t = t_f$. Using the final conditions of the states and costates, equations (4.2) and (4.6), respectively, in the Hamiltonian, (4.3), with use of the parameter equations (4.8), the transversality condition may be written as

$$LV_f\cos(\theta_f-\theta_\lambda) = 1\,. \tag{4.10}$$

For this problem, the constraints for singular control may be written as

$$\begin{aligned}
\frac{\partial\mathcal{H}}{\partial\Omega} &= \lambda_\theta + \tan\theta_\mu V\lambda_V = 0,\\
\cos\theta_\mu\frac{d}{dt}\frac{\partial\mathcal{H}}{\partial\Omega} &= -LV\sin(\theta-\theta_\mu-\theta_\lambda) - \frac{1}{2\cos\theta_\mu}\lambda_V\cos\theta = 0,\\
\cos^2\theta_\mu\frac{d^2}{dt^2}\frac{\partial\mathcal{H}}{\partial\Omega} &= \frac{L}{2}(\sin\theta_\mu\sin(2\theta-\theta_\mu-\theta_\lambda) - \cos\theta_\lambda)\\
&\quad - LV\Omega\cos(\theta-2\theta_\mu-\theta_\lambda) + \frac{1}{2\cos\theta_\mu}\lambda_V\Omega\sin(\theta+\theta_\mu) = 0\,.
\end{aligned} \tag{4.11}$$

With the condition that $\lambda_V(t_f) = \lambda_\theta(t_f) = 0$, it is evident that singular control will lead the system to the final state.

The generalized Legendre–Clebsch condition for the existence of singular arcs is

$$-\frac{\partial}{\partial\Omega}\left[\frac{d^2}{dt^2}\left(\frac{\partial\mathcal{H}}{\partial\Omega}\right)\right] > 0\,.$$

For this problem, this is equivalent to the condition

$$-\frac{1}{2\cos\theta_\mu}\lambda_V\sin(\theta+\theta_\mu) + LV\cos(\theta-2\theta_\mu-\theta_\lambda) > 0\,. \tag{4.12}$$

The assumptions upon which the singular control derivation is based are that (1) the speed, including the parameter V_f, is positive; (2) the parameter L is positive; (3) the heading angle θ lies in the interval $(-\pi/2, \pi/2)$; (4) the variable s as manifested in the parameter θ_μ is constant; and (5) the generalized Legendre–Clebsch condition holds.

The derivation will proceed assuming these conditions hold. When the singular-control state trajectories are finally derived, these assumptions will be verified.

The control as determined from the final condition is

$$\Omega = \frac{\frac{L}{2}(\sin\theta_\mu \sin(2\theta - \theta_\mu - \theta_\lambda) - \cos\theta_\lambda)}{LV\cos(\theta - 2\theta_\mu - \theta_\lambda) - \frac{1}{2\cos\theta_\mu}\lambda_V \sin(\theta + \theta_\mu)}\,. \tag{4.13}$$

If the second equation is solved for λ_V, the result is

$$\lambda_V = -\frac{2V}{\cos\theta}\, L\cos\theta_\mu \sin(\theta - \theta_\mu - \theta_\lambda)\,. \tag{4.14}$$

Substituting this result into the control equation,

$$\Omega = \frac{\cos\theta}{2V}\,\frac{\sin\theta_\mu \sin(2\theta - \theta_\mu - \theta_\lambda) - \cos\theta_\lambda}{\sin\theta_\mu \sin(2\theta - \theta_\mu - \theta_\lambda) + \cos(2\theta_\mu + \theta_\lambda)}\,. \tag{4.15}$$

Notice the equations are identical to the singular control equations for the classical brachistochrone when $\theta_\mu = 0$.

With λ_V obtained from (4.14), the generalized Legendre–Clebsch condition becomes

$$\frac{LV}{\cos\theta}\,(\sin\theta_\mu \sin(2\theta - \theta_\mu - \theta_\lambda) + \cos(2\theta_\mu + \theta_\lambda)) > 0\,. \tag{4.16}$$

With $-\pi/2 < \theta < \pi/2$, the constraint that $\lambda_V(t_f) = 0$ implies that if (4.14) is evaluated at t_f then

$$\theta_f = \theta_\mu + \theta_\lambda + k\pi,$$

where k is an integer. Substituting $\theta = \theta_f$ in the generalized Legendre–Clebsch condition yields

$$\frac{LV}{\cos\theta}\,(\sin\theta_\mu \sin(2\theta - \theta_\mu - \theta_\lambda) + \cos(2\theta_\mu + \theta_\lambda))\Big|_{t=t_f} = (-1)^k LV_f \cos\theta_\mu > 0\,.$$

Therefore, k must be an even integer. Without loss of generality, $k = 0$ and

$$\theta_\lambda = \theta_f - \theta_\mu\,. \tag{4.17}$$

Hereafter, the variable θ_λ will be replaced by the right-hand side of (4.17). Rewriting the generalized Legendre–Clebsch condition in terms of θ_f,

$$\frac{LV}{\cos\theta}\,\{\sin\theta_\mu \sin(2\theta - \theta_f) + \cos(\theta_\mu + \theta_f)\} > 0\,. \tag{4.18}$$

With L and V positive, and $-\pi/2 < \theta < \pi/2$, the generalized Legendre–Clebsch condition holds if the term in the braces is positive.

The dimensionless normal force in the system is

$$N = \frac{1}{2}\,\cos\theta - V\Omega = \cos\theta\,\frac{\cos\theta_\mu \cos\theta_f}{\sin\theta_\mu \sin(2\theta - \theta_f) + \cos(\theta_\mu + \theta_f)}\,. \tag{4.19}$$

The generalized Legendre–Clebsch condition explicitly requires $s = \mathrm{sgn}(N) = +1$. In other words, there is no way of picking an optimal singular control for which $N < 0$ at

some point in the singular arc. Furthermore, the assumption of s constant is justified. Hereafter, the parameter θ_μ will be defined by the relations (4.8) with $s \equiv +1$.

It will be helpful to express the generalized Legendre–Clebsch condition, and other relationships in the equations, in a different form. First,

$$\begin{aligned}(4.20)\qquad &\frac{LV}{\cos\theta}\left(\sin\theta_\mu \sin(2\theta-\theta_f)+\cos(\theta_\mu+\theta_f)\right)\\ &= LV\cos\theta\cos\theta_f\cos\theta_\mu((\tan\theta+\tan\theta_\mu)^2-(\tan\theta_f+\tan\theta_\mu)^2+\sec^2\theta_f)>0\,.\end{aligned}$$

The dimensionless normal force may be written as

$$(4.21)\qquad N=\frac{\sec\theta}{(\tan\theta+\tan\theta_\mu)^2-(\tan\theta_f+\tan\theta_\mu)^2+\sec^2\theta_f}\,.$$

Similarly, (4.15) may be written as

$$(4.22)\qquad \Omega=-\frac{\cos\theta}{2V}\,\frac{(\tan\theta-\tan\theta_\mu)^2-(\tan\theta_f-\tan\theta_\mu)^2+\sec^2\theta_f}{(\tan\theta+\tan\theta_\mu)^2-(\tan\theta_f+\tan\theta_\mu)^2+\sec^2\theta_f}\,.$$

Using this last relationship and writing the differential equation for $V(\theta)$,

$$(4.23)\qquad \frac{dV}{d\theta}=-V\,\frac{\tan^3\theta+(1-2\tan\theta_f\tan\theta_\mu)\tan\theta-2\tan\theta_\mu}{(\tan\theta-\tan\theta_\mu)^2-(\tan\theta_f-\tan\theta_\mu)^2+\sec^2\theta_f}\,.$$

The solution of this differential equation is

$$(4.24)\qquad V_+(\theta)=\frac{V_f\sec\theta_f\sec\theta}{(\tan\theta-\tan\theta_\mu)^2-(\tan\theta_f-\tan\theta_\mu)^2+\sec^2\theta_f}\,.$$

The subscript + indicates that this and subsequent formulas are valid for θ near θ_f (singular control). The beauty of this result is its simplicity. With $\theta_\mu = 0$,

$$V(\theta)=V_f\,\frac{\sec\theta_f}{\sec\theta}=V_f\,\frac{\cos\theta}{\cos\theta_f},$$

which is the result for the classical brachistochrone.

With $V_+ > 0$, $\cos\theta > 0$, and the generalized Legendre–Clebsch condition, $\Omega(\theta) < 0$, which implies that θ decreases as a function of time. This means that $\theta \geq \theta_f$, where θ represents the heading angle at time $t \leq t_f$. The implicit assumption in the analysis to this point has been that $V_+ > 0$. In order for this to be true,

$$\begin{aligned}V_+(\theta)&=\frac{V_f\sec\theta_f\sec\theta}{(\tan\theta-\tan\theta_\mu)^2-(\tan\theta_f-\tan\theta_\mu)^2+\sec^2\theta_f}\\ &=V_f\,\frac{\cos\theta_f}{\cos\theta}\,\frac{1}{f^2(\theta)-f^2(\theta_f)+1}>0,\end{aligned}$$

where

$$(4.25)\qquad f(x)=\frac{\tan x-\tan\theta_\mu}{\sec\theta_f}=\cos\theta_f(\tan x-\mu)\,.$$

Given that θ and θ_f are each bounded between $-\pi/2$ and $\pi/2$, this implies that $f^2(\theta)-f^2(\theta_f)+1>0$. If $1-f^2(\theta_f)>0$, then the constraint holds regardless of the value of θ. If $1-f^2(\theta_f)<0$, then θ must be constrained. The fact that $V_+ > 0$ and

$\Omega(\theta) < 0$ translates to the following constraints on θ and θ_f:

$$\begin{aligned} &\theta_f > 2\theta_\mu - \frac{\pi}{2}\,, && \theta_f < \theta < \frac{\pi}{2}\,, \\ &\theta_f \le 2\theta_\mu - \frac{\pi}{2}\,, && \theta_f < \theta < \tan^{-1}\left(\tan\theta_\mu - \sqrt{(\tan\theta_f - \tan\theta_\mu)^2 - \sec^2\theta_f}\right). \end{aligned} \tag{4.26}$$

This constraint on the range of θ also implies the generalized Legendre–Clebsch condition is satisfied in general for the state trajectories to be computed.

The differential equations satisfied by x, y, and t are as follows:

$$\begin{aligned} \frac{dx}{d\theta} &= 2V_f^2\cos\theta_f\,\frac{f^2(\theta) - 4f_0 f(\theta) - f^2(\theta_f) + 4f_0 f(\theta_f) + 1}{(f^2(\theta) - f^2(\theta_f) + 1)^3}\,f'(\theta)\,, \\ \frac{dy}{d\theta} &= 2V_f^2\,\frac{(f(\theta) - f_0)(f^2(\theta) - 4f_0 f(\theta) - f^2(\theta_f) + 4f_0 f(\theta_f) + 1)}{(f^2(\theta) - f^2(\theta_f) + 1)^3}\,f'(\theta)\,, \end{aligned} \tag{4.27}$$

and

$$\frac{dt}{d\theta} = -2V_f\,\frac{f^2(\theta) - 4f_0 f(\theta) - f^2(\theta_f) + 4f_0 f(\theta_f) + 1}{(f^2(\theta) - f^2(\theta_f) + 1)^2}\,f'(\theta),$$

where $f_0 = f(0) = -\mu\cos\theta_f$.

The solution of each differential equation differs with θ_f greater than, equal to, or less than $2\theta_\mu - \pi/2$. The value for $\theta_f > 2\theta_\mu - \pi/2$ is presented first. The value for $\theta_f < 2\theta_\mu - \pi/2$ is obtained from this result by replacing each term $\sqrt{1 - f^2(\theta_f)}$ with $i\sqrt{f^2(\theta_f) - 1}$, and $\tan^{-1}(ix)$ with $i\coth^{-1}(x)$. For each solution the corresponding formula for $\theta_f = 2\theta_\mu - \pi/2$ is also presented. The solutions of the differential equations are

$$\begin{aligned} \frac{x_+(\theta)}{V_f^2} = \cos\theta_f \Bigg[&\frac{1 - f^2(\theta_f) + 3f(\theta_f)f_0}{(1 - f^2(\theta_f))^2}\left(\frac{f(\theta)}{f^2(\theta) - f^2(\theta_f) + 1} - f(\theta_f)\right. \\ &\left. + \frac{1}{\sqrt{1 - f^2(\theta_f)}}\left(\tan^{-1}\left(\frac{f(\theta)}{\sqrt{1 - f^2(\theta_f)}}\right) - \tan^{-1}\left(\frac{f(\theta_f)}{\sqrt{1 - f^2(\theta_f)}}\right)\right)\right) \\ &+ \frac{2f_0 f(\theta_f)}{1 - f^2(\theta_f)}\left(\frac{f(\theta)}{(f^2(\theta) - f^2(\theta_f) + 1)^2} - f(\theta_f)\right) \\ &- 2f_0\left(1 - \frac{1}{(f^2(\theta) - f^2(\theta_f) + 1)^2}\right)\Bigg], \end{aligned} \tag{4.28}$$

$$\begin{aligned} \frac{y_+(\theta)}{V_f^2} = 1 &- \frac{1}{f^2(\theta) - f^2(\theta_f) + 1} + 2f_0(f(\theta_f) + f_0)\left(1 - \frac{1}{(f^2(\theta) - f^2(\theta_f) + 1)^2}\right) \\ &+ \frac{2f_0(1 - f^2(\theta_f) - f(\theta_f)f_0)}{1 - f^2(\theta_f)}\left(\frac{f(\theta)}{(f^2(\theta) - f^2(\theta_f) + 1)^2} - f(\theta_f)\right) \\ &- \frac{f_0(2 - 2f^2(\theta_f) + 3f(\theta_f)f_0)}{(1 - f^2(\theta_f))^2}\left(\frac{f(\theta)}{f^2(\theta) - f^2(\theta_f) + 1} - f(\theta_f)\right. \\ &\left. + \frac{1}{\sqrt{1 - f^2(\theta_f)}}\left(\tan^{-1}\left(\frac{f(\theta)}{\sqrt{1 - f^2(\theta_f)}}\right) - \tan^{-1}\left(\frac{f(\theta_f)}{\sqrt{1 - f^2(\theta_f)}}\right)\right)\right), \end{aligned} \tag{4.29}$$

and

$$\frac{t_+(\theta)}{2V_f} = \frac{t_f}{2V_f} - \frac{2f_0 f(\theta_f)}{1-f^2(\theta_f)}\left(\frac{f(\theta)}{f^2(\theta)-f^2(\theta_f)+1} - f(\theta_f)\right)$$
$$+\ 2f_0\left(1 - \frac{1}{f^2(\theta)-f^2(\theta_f)+1}) - (1 + \frac{2f_0 f(\theta_f)}{1-f^2(\theta_f)}\right)$$
$$\cdot \frac{1}{\sqrt{1-f^2(\theta_f)}}\left(\tan^{-1}\left(\frac{f(\theta)}{\sqrt{1-f^2(\theta_f)}}\right) - \tan^{-1}\left(\frac{f(\theta_f)}{\sqrt{1-f^2(\theta_f)}}\right)\right). \tag{4.30}$$

The solutions of the differential equations for $f(\theta_f) = -1$ ($\theta_f = 2\theta_\mu - \pi/2$) are

$$\frac{x_+(\theta)}{V_f^2} = \sin 2\theta_\mu\left[-\frac{2}{3}\left(\frac{1}{f^3(\theta)}+1\right) + 2f_0\left(\frac{1}{f^4(\theta)}-1\right) + \frac{8f_0}{5}\left(\frac{1}{f^5(\theta)}+1\right)\right], \tag{4.31}$$

$$\frac{y_+(\theta)}{V_f^2} = 1 - \frac{1}{f^2(\theta)} + 2f_0(1-f_0)\left(\frac{1}{f^4(\theta)}-1\right)$$
$$+\ \frac{10f_0}{3}\left(\frac{1}{f^3(\theta)}+1\right) - \frac{8f_0^2}{5}\left(\frac{1}{f^5(\theta)}+1\right), \tag{4.32}$$

and

$$\frac{t_+(\theta)}{2V_f} = \frac{t_f}{2V_f} + \frac{1}{f(\theta)} + 1 + 2f_0\left(1 - \frac{1}{f^2(\theta)}\right) - \frac{4f_0}{3}\left(\frac{1}{f^3(\theta)}+1\right). \tag{4.33}$$

When $\theta_\mu = 0$, the equations simplify to

$$x(\theta) = \frac{V_f^2}{2\cos^2\theta_f}\left(2(\theta-\theta_f) + \sin 2\theta - \sin 2\theta_f\right),$$

$$y(\theta) = \frac{V_f^2}{2\cos^2\theta_f}\left(\cos 2\theta_f - \cos 2\theta\right),$$

and

$$t(\theta) = t_f - \frac{2V_f}{\cos\theta_f}\left(\theta - \theta_f\right).$$

Admittedly, the results were presented without exposition, but the fact that the results yielded are identical to classical brachistochrone results when $\mu = 0$ is at least reassuring.

4.3. The extremal control arc. With the derivation of the singular control arc, the optimal control has yet to be completely described. For one, each of the expressions derived contains two unknowns, V_f and θ_f. There must be relations connecting the states in the singular arc with states starting at time $t = 0$.

Initially the control will be extremal control, that is, control which lies at the boundary of the admissible domain of control. As demonstrated earlier, the optimal control for initial conditions in the first or fourth quadrant (x_0 positive) is control in which the normal force is nonnegative. Therefore

$$\frac{1}{2}\cos\theta - V\Omega \geq 0. \tag{4.34}$$

This constraint will yield the initial extremal control.

The partial derivative of the Hamiltonian with respect to the control variable Ω is given by

$$\frac{\partial \mathcal{H}}{\partial \Omega} = \lambda_\theta + s\mu V \lambda_V \,. \tag{4.35}$$

Examining this term at time $t = 0$, if $V_0 = V(0) \neq 0$ then this term is nonzero. Furthermore, the time derivative of the Hamiltonian is

$$\frac{d\mathcal{H}}{dt} = (\lambda_\theta + s\mu V \lambda_V)\dot{\Omega} - \mu \lambda_V \left(\frac{1}{2} \cos\theta - V\Omega \right) \dot{s} = 0 \,. \tag{4.36}$$

There are thus two options for the initial control Ω if $V_0 \neq 0$: (1) choose Ω constant with s constant (either + or − 1), or (2) choose $\Omega = \frac{\cos\theta}{2V}$ such that the normal force is zero and define $s = -\frac{\lambda_\theta}{\mu V \lambda_V}$ so that the coefficient of $\dot{\Omega}$ in the time derivative of the Hamiltonian is also zero.

Choosing case (1), the *constant* extremal control which "minimizes" the Hamiltonian must be $\Omega = \pm\infty$; that is, the initial control will be a jump change in the direction of the particle. By Pontryagin's principle [15], the optimal control is the control which extremizes the Hamiltonian. In this case, extremization of the Hamiltonian occurs with $\Omega = \pm\infty$ if and only if

$$\left| \frac{\lambda_\theta}{\mu V \lambda_V} \right| \geq 1 \,. \tag{4.37}$$

This constraint follows from the fact that λ_V, from (4.14), is less than zero and that λ_θ, from the first equation of (4.11), is greater than zero at the corner between singular control and extremal control. Thus, in order for constant control to be extremal control, the control reduces to a jump change in slope. After the jump in the control trajectory, the velocity is now zero and singular control yields the constraints. Intuitively, this possibility is trivial as setting the initial normal force infinite makes this problem identical to a zero-initial-velocity problem. All that is gained from having nonzero initial energy is lost immediately.

Therefore, the extremal control will be to let

$$\Omega = \frac{\cos\theta}{2V} \,. \tag{4.38}$$

This control minimizes the Hamiltonian if and only if

$$s = -\frac{\lambda_\theta}{\mu V \lambda_V} \qquad \text{with} \qquad |s| \leq 1 \,. \tag{4.39}$$

This initial extremal control is different from the singular control derived earlier. Due to the order of the singularity, the control *must* be discontinuous [21] at the corner between extremal control and singular control. Setting the normal force equal to zero means that the curve described by the motion of the particle is that of a particle in "free fall."

Starting with the free-fall equation for velocity,

$$\frac{dV}{d\theta} = V \tan\theta \,.$$

Integrating this result yields

$$V_-(\theta) = V_0 \frac{\cos\theta_0}{\cos\theta}\,. \tag{4.40}$$

The subscript $-$ indicates that this and subsequent formulas are valid for θ near θ_0 (extremal control). Using this result in the differential equations for x, y, and t yields

$$\begin{aligned} \frac{dx}{d\theta} &= -2V_0^2 \cos^2\theta_0 \sec^2\theta\,, \\ \frac{dy}{d\theta} &= -2V_0^2 \cos^2\theta_0 \sec^2\theta \tan\theta\,, \end{aligned} \tag{4.41}$$

and

$$\frac{dt}{d\theta} = 2V_0 \cos\theta_0 \sec^2\theta\,.$$

The solution of this set of differential equations is

$$\begin{aligned} x_-(\theta) &= 1 - 2V_0^2 \cos^2\theta_0(\tan\theta - \tan\theta_0) \\ &= 1 - 2V_0^2 \frac{\cos^2\theta_0}{\cos^2\theta_f} \cos\theta_f(f(\theta) - f(\theta_0))\,, \end{aligned} \tag{4.42}$$

$$\begin{aligned} y_-(\theta) &= y_0 - V_0^2 \cos^2\theta_0(\tan^2\theta - \tan^2\theta_0) \\ &= y_0 - V_0^2 \frac{\cos^2\theta_0}{\cos^2\theta_f}(f^2(\theta) - 2f(\theta)f_0 - f^2(\theta_0) + 2f(\theta_0)f_0), \end{aligned} \tag{4.43}$$

and

$$t_-(\theta) = 2V_0 \cos\theta_0(\tan\theta - \tan\theta_0) = 2V_0 \frac{\cos\theta_0}{\cos\theta_f}(f(\theta) - f(\theta_0))\,. \tag{4.44}$$

If θ is the heading angle at the time at which the control switches, then letting $x_+(\theta)$ from (4.28) equal $x_-(\theta)$ from (4.42), $y_+(\theta)$ from (4.29) equal $y_-(\theta)$ from (4.43), and $V_+(\theta)$ from (4.24) equal $V_-(\theta)$ from (4.40) yields three equations in unknowns θ_0, θ_f, V_f, and θ. Therefore it is necessary to derive a fourth equation in these unknowns. The fourth equation will come from meeting the first two conditions for the existence of a singular arc at heading angle θ.

In order to derive the two necessary conditions for a singular arc, the differential equations for λ_V and λ_θ need to be solved. Using the costate time differential equations (4.4) with $s = -\frac{\lambda_\theta}{\mu V \lambda_V}$ yields

$$\begin{aligned} \dot\lambda_V &= L\cos(\theta - \theta_f + \theta_\mu) + \frac{1}{2}\frac{\lambda_\theta \cos\theta}{V_-^2(\theta)}, \\ \dot\lambda_\theta &= -LV_-(\theta)\sin(\theta - \theta_f + \theta_\mu) - \frac{1}{2}\lambda_V \cos\theta + \frac{1}{2}\frac{\lambda_\theta \sin\theta}{V_-(\theta)}\,. \end{aligned} \tag{4.45}$$

Rewriting these time differential equations, with a change of dependent variable, in terms of θ

$$\begin{aligned} \frac{d\lambda_V}{d\theta} &= \frac{2LV_0 \cos\theta_0 \cos(\theta - \theta_f + \theta_\mu)}{\cos^2\theta} + \frac{\lambda_\theta}{V_-}, \\ \frac{d}{d\theta}\frac{\lambda_\theta}{V_-} &= -\frac{2LV_0 \cos\theta_0 \sin(\theta - \theta_f + \theta_\mu)}{\cos^2\theta} - \lambda_V\,. \end{aligned} \tag{4.46}$$

Furthermore, the first two necessary conditions for a singular solution may be written as

$$\lambda_\theta + \tan\theta_\mu V_-(\theta)\lambda_V = \frac{V_-(\theta)}{\cos\theta_\mu}\left(\frac{\lambda_\theta}{V_-(\theta)}\cos\theta_\mu + V_-(\theta)\lambda_V \sin\theta_\mu\right) = 0\,, \tag{4.47}$$

$$-LV_-(\theta)\sin(\theta-\theta_f) - \frac{1}{2\cos\theta_\mu}\lambda_V\cos\theta = 0\,. \tag{4.48}$$

Solving the differential equations for λ_V and λ_θ/V_- yields

$$\lambda_V(\theta) = \lambda_{V_0}\cos(\theta-\theta_0) + 2L(V_0\sin(\theta-\theta_0-\theta_f+\theta_\mu) + V_-(\theta)\sin(\theta_f-\theta_\mu))$$

and

$$\frac{\lambda_\theta(\theta)}{V_-(\theta)} = -\lambda_{V_0}\sin(\theta-\theta_0) + 2L(V_0\cos(\theta-\theta_0-\theta_f+\theta_\mu) - V_-(\theta)\cos(\theta_f-\theta_\mu)),$$

where $\lambda_{V_0} = \lambda_V(0)$.

With these two quantities determined, the value of the switching function given by (4.47) and the constraint given by (4.48) is

$$-\frac{V_-(\theta)}{\cos\theta_\mu}\left(\lambda_{V_0}\sin(\theta-\theta_0-\theta_\mu) + 2LV_0\,\frac{\sin(\theta-\theta_0)\sin(\theta-\theta_f)}{\cos\theta}\right) = 0\,, \tag{4.49}$$

$$-\frac{1}{2\cos\theta_\mu}\left[\lambda_{V_0}\cos(\theta-\theta_0)\cos\theta + L\left(\frac{\cos\theta_f\cos\theta_\mu + \cos\theta\cos(\theta-\theta_f+\theta_\mu)}{\cos\theta}\sin(\theta-\theta_0)\right.\right.$$

$$\left.\left. + \frac{\cos\theta_0\cos\theta_\mu + \cos\theta\cos(\theta-\theta_0+\theta_\mu)}{\cos\theta}\sin(\theta-\theta_f)\right)V_0\right] = 0\,. \tag{4.50}$$

These two linear equations have a nontrivial solution in λ_{V_0} and L if and only if the determinant of the coefficient matrix is zero. The coefficients of λ_{V_0} and L in the second equation are *not* (scaled) time derivatives of the corresponding coefficients in the first equation, despite the derivation given for the conditions in the singular solution. This is due to the fact that the costates on the (constrained) extremal trajectory have different governing equations than the singular trajectory paths (s is *not* constant here). Expanding and simplifying the determinant of the coefficient matrix yield

$$\frac{V_0^2\cos\theta_0}{\cos^2\theta}[\sin(\theta-\theta_0)(\sin(\theta-\theta_0)-\mu\cos(\theta-\theta_0))\cos\theta_f - \mu\sin(\theta-\theta_f)\cos\theta] = 0\,. \tag{4.51}$$

Provided that the (dimensionless) initial velocity V_0 is not zero, the term in brackets being zero is the fourth equation that must hold. The constraints on these four equations come from the fact that $\theta > \theta_0$, $\theta > \theta_f$, and $V_f > 0$ and the restrictions on θ given with the states, inequality (4.26).

4.4. Abnormal trajectories. The derivation up to this point has implicitly assumed "normality" of the minimum-time trajectories. The complete formulation of an optimal control problem in the calculus of variations requires the formation of the Hamiltonian as

$$\mathcal{H} = \langle \boldsymbol{\lambda}, \boldsymbol{f}(\boldsymbol{x}, u)\rangle + \lambda_0 L(\boldsymbol{x}, u), \tag{4.52}$$

where λ_0 is a nonnegative constant. To this point the assumption has been that λ_0 was strictly positive. Thus equation (2.1) is obtained by scaling the remaining costates by the positive constant λ_0.

If, however, λ_0 is zero, then a degenerate condition exists which may yield optimal trajectories. For example, as noted in a novel case in [13], a completely controllable optimal control system with k linear controls and n states, $k < n$, and a cost function L which is quadratic positive-definite in the controls yields an optimal solution which is abnormal. The full definition of the optimal control problem is as follows [12, 23]:

(1) For any t, $\boldsymbol{\lambda}(t)$ and λ_0 are not both zero.

(2) At $(t, \boldsymbol{x}(t), \boldsymbol{\lambda}(t), \lambda_0)$, the function of u,

$$\mathcal{H} = \langle \boldsymbol{\lambda}, \boldsymbol{f}(\boldsymbol{x}, u) \rangle + \lambda_0 L(\boldsymbol{x}, u),$$

assumes for $u = u(t)$ its minimum. With the cost and the state dynamics having no explicit time dependence, the minimum value of the function $\mathcal{H}$ is constant.

(3) For the arguments $t, \boldsymbol{x}(t), \boldsymbol{\lambda}(t), \lambda_0$ the canonical Euler equations

$$\dot{\boldsymbol{x}} = \frac{\partial \mathcal{H}}{\partial \boldsymbol{\lambda}}, \qquad \dot{\boldsymbol{\lambda}} = -\frac{\partial \mathcal{H}}{\partial \boldsymbol{x}}$$

hold.

(4) If T_0 and T_f are the tangent spaces to the state manifold at $t = t_0$ and $t = t_f$, respectively, then

$$\boldsymbol{\lambda}(t_0) \perp T_0 \qquad \text{and} \qquad \boldsymbol{\lambda}(t_f) \perp T_f \,.$$

This is the *transversality* condition.

The abnormal case corresponds to choosing $\lambda_0 = 0$ and finding nontrivial $\boldsymbol{\lambda}(t)$ which satisfy the Euler equations everywhere except possibly at points where the state dynamics $\boldsymbol{f}$ are discontinuous [6]. An equivalent definition of abnormality [23] is that for an n-dimensional system $\boldsymbol{x}(t)$, there do not exist $2n$ linearly independent permissible variations $\delta P, \delta Q$, where

$$(P, Q) = ((t_0, \boldsymbol{x}(t_0)), (t_f, \boldsymbol{x}(t_f))) \,.$$

This second definition indicates the nature of abnormal trajectories; they are trajectories where the states/control cannot be admissibly varied about a given trajectory.

The singular control trajectory or the extremal control trajectory limits the admissible variations of the states and control about the given trajectory. For example, the extremal control trajectory requires that the normal force, given by

$$\frac{1}{2} \cos\theta - V\Omega,$$

be zero along the entire path. This specifies the control Ω as a function of the states V and θ. If V or θ is varied, the control varies. The control itself does *not* restrict the variation of V and θ. If it did, this would indicate an abnormal trajectory.

Similarly, the singular control trajectory requires the states/costates/control to satisfy the three constraints (4.11), thus restricting the state/costate manifold to dimension six, rather than the unconstrained eight. Thus in the singular trajectory, there are only six admissible variations in the states/costates, rather than the eight

of the unconstrained system. Nevertheless, the derivation of the singular control relies upon the cost function L in that the generalized Legendre–Clebsch condition (4.18) implicitly uses the transversality condition for admissible variations of t_f to ensure positive definiteness. With $\lambda_0 = 0$, there are no admissible variations of t_f and the generalized Legendre–Clebsch condition is trivial (zero). The singular paths therefore require "normal" trajectories, that is, those in which all seven unconstrained parameters are allowed to vary in order to realize a minimum.

The question then remains: is there a control strategy such that the states/costates may not be admissibly varied to the full dimension of the space. This question may be answered by examining the "smoothness" of the state manifold. The state manifold consists of two smooth submanifolds joined at the corner where

$$\frac{1}{2}\cos\theta - V\Omega = 0\,.$$

The corner is defined by the control Ω. Furthermore, there are no stationary points in the state-space for $V > 0$. Thus with $V > 0$, the states and the cost may be varied on the corner and on both sides of the corner due to the definition of the corner in terms of the control Ω. The minimization of the final time t_f trivially requires $V > 0$ at all interior points in a state trajectory. Thus at every point in the interior of a state trajectory, there exist admissible variations in the states/costates allowing the normal minimization.

5. The brachistochrone when the initial velocity is zero. If $V_0 = 0$, then the solution of the optimization problem is much simpler than the general case. In this case $\theta_0 = \theta = \pi/2$. The optimal control becomes the solution of the two equations

$$x_+(\pi/2) = 1 \qquad \text{and} \qquad y_+(\pi/2) = y_0\,.$$

The velocity equation and (4.51) both hold. Furthermore,

$$f(\theta) = f(\pi/2) \to \infty \qquad \text{and} \qquad \theta_f > 2\theta_\mu - \pi/2\,.$$

Taking the ratio of y_+ to x_+ will eliminate V_f from the equations, yielding one equation in one unknown, θ_f. When this is solved, the corresponding value of V_f can be solved for in either of the two equations as a positive scaling factor.

The one equation in θ_f can be used to obtain θ_f as a function of y_0 and μ. Expressing this relationship in one equation,

$$\begin{aligned} &1 + \frac{f_0(2f(\theta_f) - f_0)}{1 - f^2(\theta_f)} + \frac{3f_0^2}{(1 - f^2(\theta_f))^2} - f_0\left(\frac{2}{1 - f^2(\theta_f)}\right. \\ &\quad \left. + \frac{3f(\theta_f)f_0}{(1 - f^2(\theta_f))^2}\right)\frac{1}{\sqrt{1 - f^2(\theta_f)}}\cos^{-1}(f(\theta_f)) = y_0\cos\theta_f\left[\frac{f_0 - f(\theta_f)}{1 - f^2(\theta_f)}\right. \\ &\quad \left. - \frac{3f_0}{(1 - f^2(\theta_f))^2} + \left(\frac{1}{1 - f^2(\theta_f)} + \frac{3f(\theta_f)f_0}{(1 - f^2(\theta_f))^2}\right)\frac{1}{\sqrt{1 - f^2(\theta_f)}}\cos^{-1}(f(\theta_f))\right]. \end{aligned} \tag{5.1}$$

Expanding this equation in a power series in θ_f about $\theta_f = 2\theta_\mu - \pi/2$ will yield y_0 as a function of θ_f. Then, reverting this power series will yield θ_f as a function of y_0.

The result of this analysis is

$$\theta_f = 2\theta_\mu - \frac{\pi}{2} + 3\,\frac{y_0-\mu}{1+\mu^2} + \frac{3}{\mu}\,(1-2\mu^2)\left(\frac{y_0-\mu}{1+\mu^2}\right)^2 - \frac{3\sqrt{6}}{\pi\sqrt{\mu}}\,(3-\mu^2)\left(\frac{y_0-\mu}{1+\mu^2}\right)^{\frac{5}{2}} + O((y_0-\mu)^3). \tag{5.2}$$

Immediately apparent from this series expansion is that $y_0 = \mu$ corresponds to a final angle $\theta_f = 2\theta_\mu - \pi/2$. If this angle is substituted into equation (4.30), it is evident that the time-to-go $t_f - t(\theta)$ is infinite. Furthermore, the gradient $\frac{\partial \theta_f}{\partial y_0}$ is positive. Thus initial conditions for y_0 which are less than or equal to μ will not yield the origin in any time. This may also be argued on the grounds of the energy content of the system.

A further point to be noted is that the formal power series derived for θ_f does not converge if $\mu = 0$. This is due to the fact that the power series was expanded about $\theta_f = 2\theta_\mu - \pi/2$. In order to obtain a convergent power series for $\mu = 0$, (5.1) would need to be solved about $\mu = 0$. In this case a power series expansion for θ_f in terms of μ with y_0 as a parameter would be obtained. The initial value for θ_f would be the solution for θ_f in the classical frictionless brachistochrone. Due to the difficulty of the solution of these implicit transcendental equations, this power series solution will not be attempted.

An examination of the brachistochrones computed for $V_0 = 0$, $y_0 = 1$, and varying values of μ (Figure 5.1) demonstrates that the curves cross. The crossing occurs because near the starting point $(x, y) = (1, 1)$, the curvature of the brachistochrones is an increasing function of μ, with the curvature being infinite for $\theta = \pi/2$ and $\mu = 1$. With the increasing curvature, the brachistochrones lie closer to the line $y = y_0 x$, which, in the limit, is the "curve" for $\mu = y_0 = 1$. However, at the final point $(x, y) = (0, 0)$ of each brachistochrone, the slope $\tan\theta_f$ is a *decreasing* function of μ. This observation was not predicted correctly by Ramamani, Lu, and Tabarrok [17] in their discretization of this problem.

The final heading θ_f, the final velocity V_f, and the time-to-go t_f as a function of the coefficient of friction μ are plotted in Figure 5.2. The figure uses an initial condition of $y_0 = 1$. Although the trends displayed in the figures are consistent with different values of y_0, there is no linear relationship between figures with different values of y_0. Demonstrated in the figure is that θ_f is a decreasing function of μ. Furthermore, this function also has negative concavity. This negative concavity, which becomes infinite as $\mu \to y_0^-$, means that $t_f \to \infty$ "slowly" as $\mu \to y_0^-$. The slowness of this approach to infinity can be determined by substituting the power series for θ_f into the relationship for t_f, and by noting the implicit relationship for V_f in this expression. The result is

$$t_f = 4\sqrt{\frac{\pi}{3}}\left(\frac{6(y_0-\mu)}{\mu(1+\mu^2)}\right)^{-\frac{1}{4}} + O((y_0-\mu)^{\frac{3}{4}}). \tag{5.3}$$

This slow divergence is demonstrated in Figure 5.2.

A point not immediately apparent in Figure 5.2 is that although V_f decreases with increasing μ, its concavity has an inflection point around $\mu = 0.65$. Referring to the differential equation for $V(\theta)$, (4.23), for a given value of y_0, there is a value

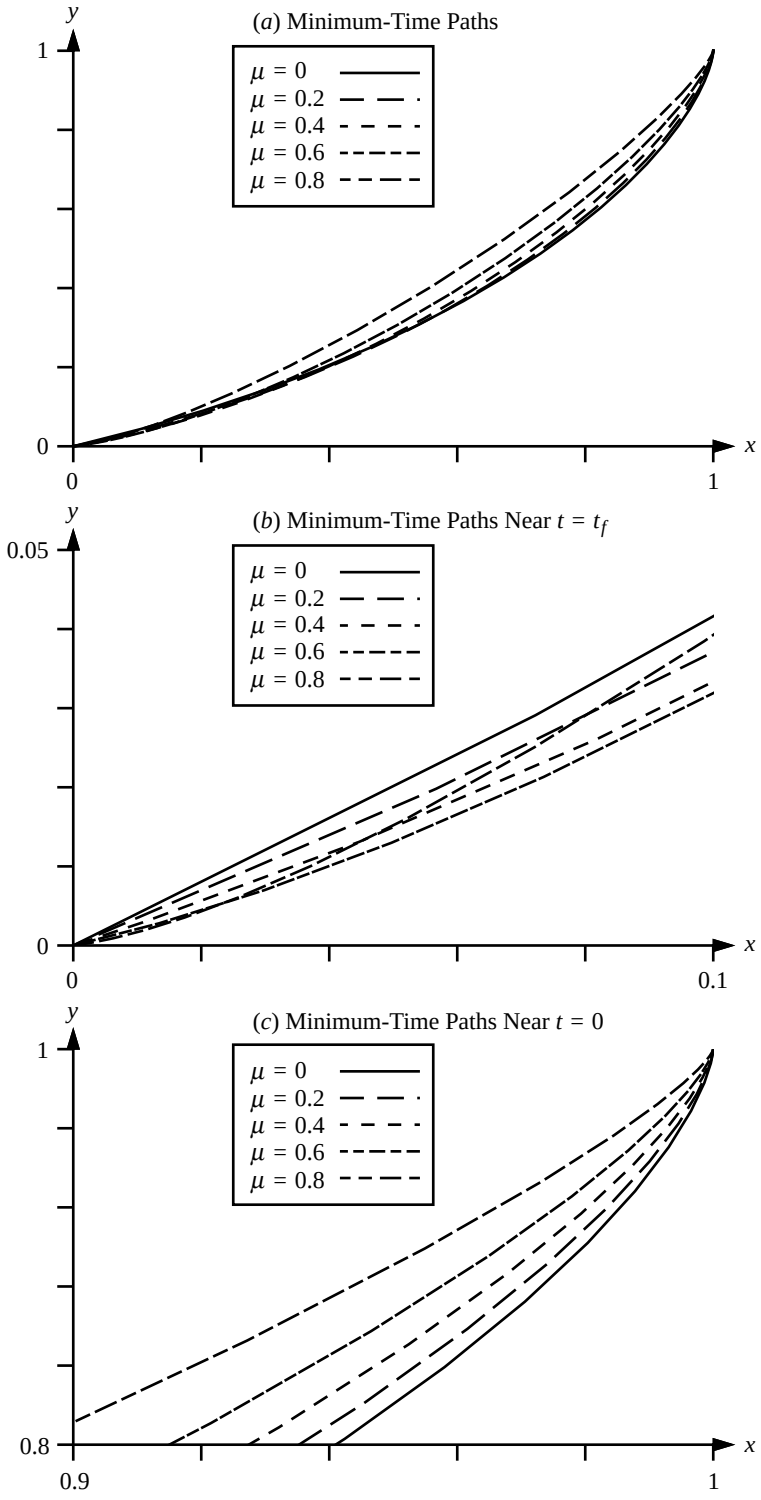


FIG. 5.1. *Coulomb friction brachistochrones with* $V_0 = 0$, $y_0 = 1$, *and differing values for* μ.

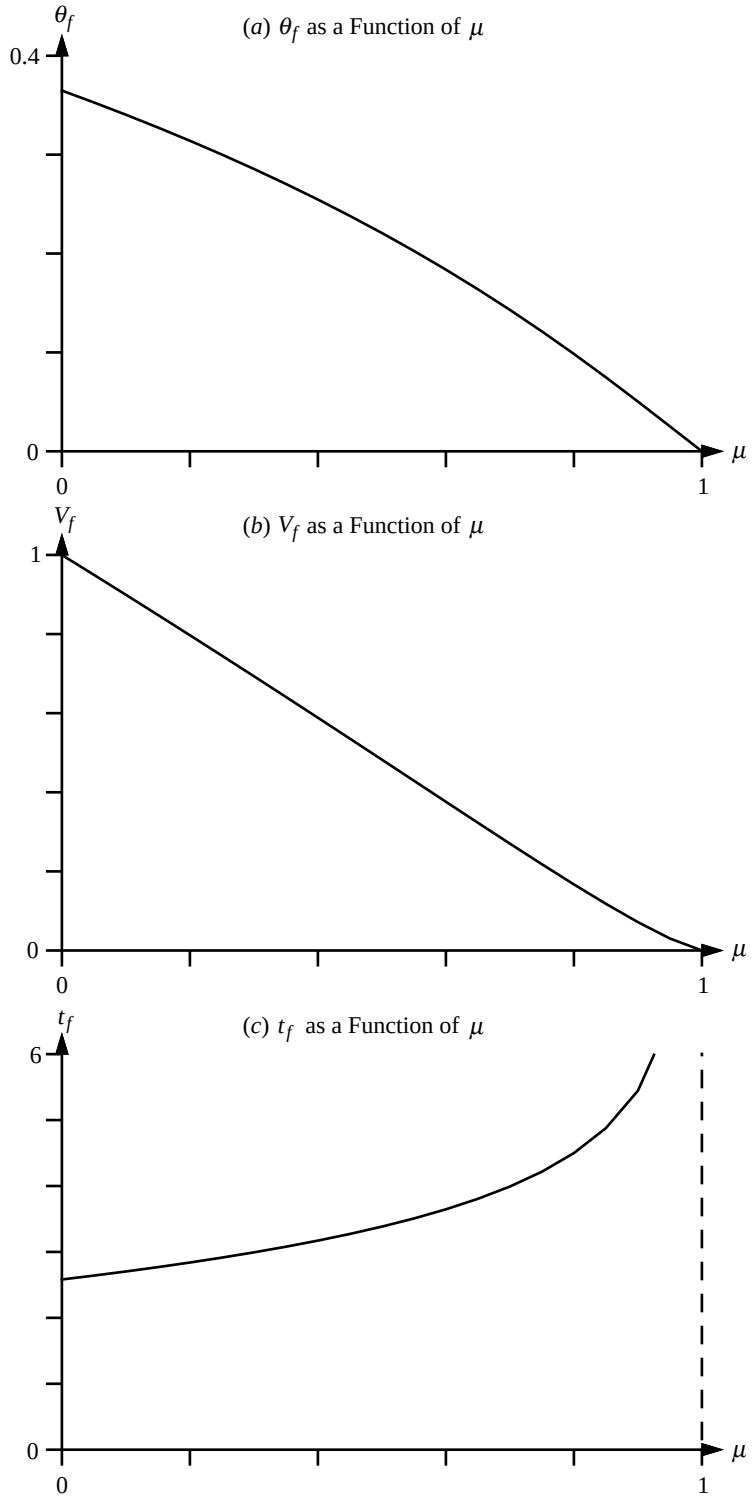


FIG. 5.2. (a) *Final heading angle* θ_f, (b) *final speed* V_f, *and* (c) *minimum time* t_f *for* $V_0 = 0$, $y_0 = 1$, *and variable* μ.

of μ at which the derivative $\frac{dV(\theta_f)}{d\theta} = 0$. This occurs when $\tan\theta_f = 2\mu$. For $y_0 = 1$, this corresponds to $\mu = 0.162\ldots$. Thus, for $y_0 = 1$ and $\mu > 0.162\ldots$, the value of V_f is no longer the largest value of V for the brachistochrone; that is, the speed of the particle initially increases but then decreases as the origin is approached. As the value of μ increases, the point of maximum velocity moves toward $\theta = \pi/2$. Due to the larger difference in angle between $V_{\max}$ and V_f for increasing μ, the plot of V_f as a function of μ will begin concave downward, inflect, and finish concave upward.

6. Some observations about minimum-time paths with nonzero initial velocity. As derived in section 4.3, the minimum-time path for a particle with nonzero initial velocity in a uniform gravitational field with Coulomb friction resisting motion consists of a "free fall" segment followed by a singular arc. The corner conditions joining the two segments are satisfied when (4.51) is satisfied. Setting $\theta = \theta_0$ in (4.51) and solving for θ_0 (and V_0) yields the minimum-time solution for which the entire path is a singular trajectory. The result of this substitution is the zero initial velocity paths presented in the previous section. These paths require $\theta = \theta_0 = \pi/2$.

Setting $\theta = \theta_f$ in (4.51) will yield the solution for a minimum-time path in which the optimal trajectory is entirely a "free fall" trajectory, that is, a trajectory in which the normal force exerted by the path on the particle is zero. Solving for θ_f yields

$$\theta_f = \theta_0 + \theta_\mu \,. \tag{6.1}$$

Substituting this result for θ in (4.42) and (4.43) and solving for θ_0 and V_0 yields

$$\theta_0 = \tan^{-1}\left(y_0 + \mu^{-1} - \sqrt{1 + y_0^2 + \mu^{-2}}\right) \tag{6.2}$$

and

$$V_0 = \sqrt{\frac{\sqrt{1 + y_0^2 + \mu^{-2}} - y_0}{2}} \,. \tag{6.3}$$

The question then is whether velocities greater than this value will yield optimal free-fall trajectories. The answer to this question comes in the solution of $s = -\frac{\lambda_\theta}{\mu V \lambda_V}$ when λ_θ, V, and λ_V are chosen to satisfy the boundary conditions at $t = 0$ and $t = t_f$. The solutions to the differential equations (4.46) with boundary conditions (4.6) are

$$\lambda_V(\theta) = -2\lambda_x V_0 \cos\theta_0 \cos\theta(\tan\theta_f - \tan\theta)(1 + \tan\theta\tan\theta_0) \,, \tag{6.4}$$

$$\frac{\lambda_\theta(\theta)}{V(\theta)} = 2\lambda_x V_0 \cos\theta_0 \cos\theta(\tan\theta_f - \tan\theta)(\tan\theta - \tan\theta_0) \,. \tag{6.5}$$

Thus

$$s = -\frac{\lambda_\theta}{\mu V \lambda_V} = \frac{\tan(\theta - \theta_0)}{\mu} \qquad \text{and} \qquad s_{\max} = \frac{\tan(\theta_f - \theta_0)}{\mu} \,. \tag{6.6}$$

Hence, if a free-fall path were time optimal, $s_{\max} \le 1$. Solving equation (4.42) equals zero and (4.43) equals zero for $\tan\theta_0$ and $\tan\theta_f$,

$$\tan\theta_0 = \sqrt{(2V_0^2 + y_0)^2 - 1 - y_0^2} - 2V_0^2, \quad \tan\theta_f = 2(V_0^2 + y_0) - \sqrt{(2V_0^2 + y_0)^2 - 1 - y_0^2} \,.$$

This means that

$$s_{\max} = \frac{\tan(\theta_f - \theta_0)}{\mu} = \frac{1}{\mu\sqrt{(2V_0^2 + y_0)^2 - 1 - y_0^2}} \le 1 \,. \tag{6.7}$$

Thus for V_0 greater than the critical V_0 given by equation (6.3), the minimum-time path is exclusively a free-fall path.

7. Conclusions. This paper has presented the time-optimal path of a particle between two points in a uniform gravitational field when motion of the particle is resisted by a force proportional to the normal force exerted on the particle by the path. This problem is equivalent to finding a brachistochrone in a uniform gravitational field with Coulomb friction resisting motion. The construction of the brachistochrone in the general case of nonzero initial velocity less than the critical initial velocity given by equation (6.3) becomes the solution of four nonlinear equations in four unknowns. For velocities greater than or equal to the critical initial velocity, the minimum-time path to the origin is a free-fall path. The special case of zero initial velocity is discussed in detail. This case simplifies the analysis to the solution of two transcendental equations in two unknowns. With a simplification, the solution becomes (5.1), which is a single transcendental equation in the unknown θ_f, the final heading angle.

A set of Coulomb friction brachistochrones with zero initial velocity are presented. The features displayed for the specific case plotted are general features of all zero initial velocity Coulomb friction brachistochrones, namely: (1) the (x, y) path is concave upward, (2) the initial heading is always directly downward, (3) the initial curvature is an increasing function of μ for a fixed initial position, (4) the final heading is a concave-downward decreasing function of the coefficient of friction μ for a fixed initial position, (5) the final velocity is a decreasing function of μ with a value of zero at $\mu = y_0$, and (6) the time-to-go is an increasing function of μ with $t_f \to +\infty$ for $\mu \to y_0^-$. Points (3) and (4) above indicate that if two brachistochrones between the same initial and final positions are drawn with different values of μ, the paths cross. This is not a common occurrence in time-optimal path problems. Typically, a change in a parameter in a time-optimal control problem does not produce a path which intersects the original path. The crossing of the paths with a change in the coefficient of friction μ is due to the lack of smoothness in the dynamics of the problem.

REFERENCES

[1] J. Bernoulli, *Jacobi Bernoulli solutio problematum fraternorum*, Acta Eruditorum, Leipzig, May 1697, p. 214.

[2] A. E. Bryson, Jr., and Y.-C. Ho, *Applied Optimal Control: Optimization, Estimation, and Control*, rev. ed., Hemisphere Publishing, Washington, 1975.

[3] D. Bushaw, *Optimal discontinuous forcing terms*, in Contributions to the Theory of Nonlinear Oscillations, vol. 4, S. Lefschetz, ed., Princeton University Press, Princeton, NJ, 1958, pp. 29–52.

[4] R. N. Danbury, *Time-optimal servomechanisms—discrete-time estimation of the optimum switching function*, IEEE Trans. Industry Appl., 30 (1994), pp. 333–340.

[5] L. Euler, *Methodus Inveniendi Lineas Curvas Maximi Minimive Proprietate Guadentes sive Solutio Problematis Isoperimetrici Latissimo Sensu Accepti*, Lausanne, Geneva, 1744.

[6] G. M. Ewing, *Calculus of Variations with Applications*, Dover Publications, New York, 1985.

[7] H. H. Goldstine, *A History of the Calculus of Variations from the* 17*th through the* 19*th Century*, Springer-Verlag, New York, 1980.

[8] T. Heckenthaler and S. Engell, *Approximately time-optimal fuzzy control of a two-tank system*, IEEE Control Systems, 14 (1994), pp. 24–30.

[9] J. A. Hoskins, W. D. Hoskins, and R. G. Stanton, *The cardioid and a variation of the brachistochrone problem*, Utilitas Math., 40 (1991), pp. 65–70.

[10] C. D. Johnson, *Singular solutions in problems of optimal control*, in Advances in Control Systems: Theory and Applications, vol. 2, C. T. Leondes, ed., Academic Press, New York, London, 1965, pp. 209–267.

[11] H. J. Kelley, R. E. Kopp, and H. G. Moyer, *Singular extremals*, in Topics in Optimization, G. Leitmann, ed., Academic Press, New York, London, 1967, pp. 63–101.

[12] G. Knowles, *An Introduction to Applied Optimal Control*, Academic Press, New York, 1981.
[13] R. Montgomery, *Abnormal minimizers*, SIAM J. Control Optim., 32 (1994), pp. 1605–1620.
[14] V. Perlick, *The brachistochrone problem in a stationary space-time*, J. Math. Phys., 32 (1991), pp. 3148–3157.
[15] L. S. Pontryagin, V. G. Boltyanskii, R. V. Gamkrelidze, and E. F. Mishchenko, *The Mathematical Theory of Optimal Processes*, Pergamon Press Reprint (UK) dist. by Franklin Book Co., Sausalito, CA, 1964.
[16] R. L. Racicot, *Limiting servo motor torque gradients with near minimum time repositioning*, IEEE Trans. Control Systems Tech., 1 (1993), pp. 284–289.
[17] S. Ramamani, W.-S. Lu, and B. Tabarrok, *Optimization of a discretized brachistochrone problem*, Internat. J. Mech. Engr. Education, 19 (1990), pp. 79–87.
[18] M. Razzaghi and G. N. Elnagar, *A pseudospectral collocation method for the brachistochrone problem*, Math. Comput. Simulation, 36 (1994), pp. 241–246.
[19] H. H. Roomany, *A graph theoretic approach to the brachistochrone problem*, Comput. Phys., 4 (1990), pp. 303–306.
[20] Z. Shiller, *On singular time-optimal control along specific paths*, IEEE Trans. Robotics Automat., 10 (1994), pp. 561–566.
[21] J. L. Speyer, *Necessary conditions for optimality for paths lying on a corner*, Management Sci., 19 (1973), pp. 1257–1270.
[22] D. S. Szarkowicz, *Investigating the brachistochrone with a multistage Monte Carlo method*, Internat. J. Systems Sci., 26 (1995), pp. 233–243.
[23] L. C. Young, *Lectures on the Calculus of Variations and Optimal Control Theory*, W. B. Saunders Co., Philadelphia, PA, 1969.

SIAM J. CONTROL OPTIM.
Vol. 35, No. 2, pp. 585–603, March 1997

012

BOUNDARY EXACT CONTROLLABILITY OF INTERFACE PROBLEMS WITH SINGULARITIES II: ADDITION OF INTERNAL CONTROLS*

SERGE NICAISE†

Abstract. We prove the exact controllability by boundary action of hyperbolic interface problems with singularities. The two proposed methods consist of acting by a classical boundary control whose support does not contain a neighborhood of the singular points and adding internal controls located near these singular points. (See [M.-T. Niane and O. Seck, *C. R. Acad. Sci. Paris Sér.* I, 318 (1994), pp. 945–948; A. Heibig and M. Moussaoui, *Disc. Cont. Dynam. Syst.*, 2 (1996), pp. 367–386; M.-T. Niane and O. Seck, *C. R. Acad. Sci. Paris Sér.* I, 316 (1993), pp. 695–700], where such a method was introduced.)

Key words. interface problems, singularities, control

AMS subject classifications. 93C20, 35B37, 35L67

PII. S0363012995292032

1. Introduction.

This paper is the second of two whose purpose is the boundary exact controllability of hyperbolic interface (or transmission) problems presenting singularities at the vertices. For the sake of convenience, we use the notation and definitions from part I [20] without comment.

As we explained in [20], our motivation is twofold: first, various models of multiple-link flexible structures, consisting of finitely many interconnected flexible elements, like strings, beams, plates, or shells, or combinations of them, are of particular interest for mechanical applications and were recently derived in [8, 1, 10, 11, 2]. Second, the problem of controllability, or even stabilizability, of such structures is considered very little. Let us quote the recent works of [22, 9, 10, 21, 6, 15, 16, 6, 7]. In all these works, either no singularity occurs or, if there are some singularities, they are cancelled by an appropriate choice of the multiplier. In both cases, this is possible under strong geometrical conditions.

In order to avoid strong geometrical conditions, Niane and Seck [13, 14] and Heibig and Moussaoui [5] proposed, for the wave equation on domains with slits or with mixed boundary conditions, using classical boundary controls whose support stays far from the singular points and adding internal controls located in a small neighborhood of the singular vertices. Here we show that the method proposed by Niane and Seck [13, 14] and Heibig and Moussaoui [5] also works for transmission problems. We even proposed two strategies: the first is similar to the above authors' [5, 14]; the second consists of adding internal controls with support concentrated on small circles centered at the singular vertices.

Let us recall that for the wave equation on two-dimensional (2-d) networks, we showed [20] how to manage the presence of singularities and the controllability problem using the Hilbert uniqueness method (HUM) of Lions [12]. The strategy consisted of replacing the boundary control by its regular part and adding the coefficients of

*Received by the editors September 20, 1995; accepted for publication (in revised form) February 9, 1996.

http://www.siam.org/journals/sicon/35-2/29203.html

†Université de Valenciennes et du Hainaut Cambrésis, LIMAV and URA D 751 CNRS "GAT," Institut des Sciences et Techniques de Valenciennes, B.P. 311, F-59304 Valenciennes Cedex, France (snicaise@univ-valenciennes.fr).

the singularities to the space of controls. This led to a classical boundary control but with an internal control, which is a distribution with a support equal to the singular vertices.

The order of part II is the following. In section 2, we formulate the main results of this paper. In sections 3 and 6, we establish different inequalities with multipliers, which will be useful in the application of HUM. These inequalities yield an estimate of the energy. The weak solution of the wave equation is considered in sections 4 and 7, as well as its interpretation in terms of partial differential equations. Sections 5 and 8 are devoted to the setting of the HUM, i.e., to the proof of the boundary exact controllability.

The results of this paper were presented in [19].

2. Main results. On the 2-d polygonal topological network Ω (see [20, section 2]), we consider the following boundary value problem: given $f \in L^2(\Omega)$, let u be a solution of

$$-\triangle u_i = f_i \text{ in } P_i \,\forall i \in \mathcal{I}, \tag{2.1}$$

$$\gamma_{ij} u_i = 0 \text{ on } \Gamma_{ij} \,\forall\, \Gamma_{ij} \in \mathcal{D}, \tag{2.2}$$

$$\gamma_{ij} u_i = \gamma_{kl} u_k \text{ when } \Gamma_{ij} = \Gamma_{kl}, \tag{2.3}$$

$$\sum_{i\in\mathcal{I}:\exists j:\Gamma_{ij}=A} \alpha_i \gamma_{ij} \frac{\partial u_i}{\partial \nu_{ij}} = 0 \text{ on } A \text{ when } A \in \mathcal{N}. \tag{2.4}$$

Recall that the associated self-adjoint operator A on $L^2(\Omega)$ has the property that any $u \in D(A)$ has the singular expansion (2.13) of [20] (see also [3]).

In this paper we are dealing with the boundary exact controllability of the hyperbolic transmission problem associated with (2.1)–(2.4): given $T > 0$ and $(y_0, y_1) \in L^2(\Omega) \times V'$, find controls $v_{i_A}, A \in \mathcal{D}$ and $w_{i_A}, A \in \mathcal{N}_{ext}$ such that the solution y of

$$\left\{\begin{array}{l} y'' + Ay = 0 \text{ in } Q, \\ y(0) = y_0, \qquad y'(0) = y_1, \\ y_{i_A} = v_{i_A} \text{ on } \Sigma_A \,\forall A \in \mathcal{D}, \\ \frac{\partial y_{i_A}}{\partial \nu_{i_A}} = w_{i_A} \text{ on } \Sigma_A \,\forall A \in \mathcal{N}_{ext} \end{array}\right. \tag{2.5}$$

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

For this problem, the HUM [12] is based on the estimate of the energy

$$E_0 = \frac{1}{2} \int_\Omega \{\alpha |\nabla \varphi_0|^2 + |\varphi_1|^2\}\, dx \tag{2.6}$$

for $(\varphi_0, \varphi_1) \in D(A) \times V$ with respect to the L^2-norms of $\frac{\partial \varphi}{\partial \nu}_{|\Sigma_A}$ for $A \in \mathcal{D}$, $\frac{\partial \varphi}{\partial \tau}_{|\Sigma_A}$, and $\varphi'_{|\Sigma_A}$ for $A \in \mathcal{N}_{ext}$, where φ is the unique solution of the homogeneous hyperbolic transmission problem

$$\left\{\begin{array}{l} \varphi'' + A\varphi = 0 \text{ in } Q, \\ \varphi(0) = \varphi_0, \qquad \varphi'(0) = \varphi_1, \\ \varphi_{i_A} = 0 \text{ on } \Sigma_A \,\forall A \in \mathcal{D}, \\ \frac{\partial \varphi_{i_A}}{\partial \nu_{i_A}} = 0 \text{ on } \Sigma_A \,\forall A \in \mathcal{N}_{ext}. \end{array}\right. \tag{2.7}$$

In our case, the implementation of HUM is not direct due to the presence of the singularities $S^{S,n}$ in the decomposition (2.13) in [20] of φ. Indeed these singularities imply that $\frac{\partial \varphi}{\partial \nu}$ and $\frac{\partial \varphi}{\partial \tau}$ are not square integrable in a neighborhood of the singular vertices (i.e. vertices for which there exists $\lambda_{S,n} \le 1/2$).

As we explained in the introduction, to overcome this difficulty we propose two methods. The first, called addition of surfacial internal control, uses classical boundary controls whose support stays far from the singular points and adds internal controls located in a small neighborhood of the singular vertices. The other consists of adding internal controls with support concentrated on small circles centered at the singular vertices; that is why we call it addition of circular internal controls.

In order to give our two main results, let us introduce the following notation: the set of singular vertices $\mathcal{S}_{sing}$ is the set of vertices S such that there exists at least one $\lambda_{S,n} \in (0, 1/2]$. As in [20, section 4], we also fix points $x_{0i} \in \Pi_i, i \in \mathcal{I}$, define the multiplier m on Ω by $m_i(x) = x - x_{0i}$, and suppose that the geometrical conditions (H1) to (H4) in [20] are satisfied (see Remark 2.3 below). For three positive real numbers δ, γ, T such that $0 < \delta < \gamma$, we set

$$\left\{\begin{array}{l} Q_i = P_i \times (0,T)\, \forall i \in \mathcal{I}, \\ A_\delta = A \setminus \cup_{S \in \mathcal{S}_{sing}} B(S,\delta)\, \forall A \in \mathcal{A}, \\ \Sigma_{A\delta} = A_\delta \times (0,T)\, \forall A \in \mathcal{A}, \\ V_S = (B(S,\gamma) \cap \Omega) \times (0,T)\, \forall S \in \mathcal{S}_{sing}, \\ C_S = (S_1(S,\delta) \cap \Omega) \times (0,T)\, \forall S \in \mathcal{S}_{sing}, \end{array}\right. \tag{2.8}$$

where, as usual, $B(S,\delta)$ (resp., $S_1(S,\delta)$) is the ball (resp., sphere) of center S and radius δ.

THEOREM 2.1. *Suppose that the geometrical conditions* (H1) *to* (H4) [20] *are satisfied. Then there exists $T_0 > 0$ such that for all $T > T_0$ and $(y_0, y_1) \in L^2(\Omega) \times V'$ there exists $(\varphi_0, \varphi_1) \in V \times L^2(\Omega)$ such that the solution y of*

$$\left\{\begin{array}{l} y'' + Ay = \sum_{S \in \mathcal{S}_{sing}} D_S \text{ in } Q, \\ y(0) = y_0, \qquad y'(0) = y_1, \\ y_{i_A} = \left\{\begin{array}{l} \alpha_{i_A}^{-1} \frac{\partial \varphi_{i_A}}{\partial \nu_{i_A}} \text{ on } \Sigma_{A\delta}\, \forall A \in \mathcal{D}^+, \\ 0 \text{ on the remainder of the Dirichlet boundary,} \end{array}\right. \\ \frac{\partial y_{i_A}}{\partial \nu_{i_A}} = \left\{\begin{array}{l} \alpha_{i_A}^{-1} \frac{d}{dt} \varphi'_{i_A} \text{ on } \Sigma_{A\delta}\, \forall A \in \mathcal{N}_{ext}^+, \\ -\alpha_{i_A}^{-1} \frac{\partial^2 \varphi}{\partial \tau_{i_A}^2} \text{ on } \Sigma_{A\delta}\, \forall A \in \mathcal{N}_{ext}^-, \\ 0 \text{ on the rest of the external Neumann boundary} \end{array}\right. \end{array}\right. \tag{2.9}$$

satisfies $y(T) = y'(T) = 0$, where φ is the solution of (2.7) *and $D_S = (\varphi - \frac{d}{dt}\varphi')\chi_{V_S}$.*

THEOREM 2.2. *Under the geometrical assumptions* (H1) *to* (H4) *in* [20]*, there exists $T_0 > 0$ such that for all $T > T_0$ and $(y_0, y_1) \in L^2(\Omega) \times V'$ there exists $(\varphi_0, \varphi_1) \in V \times L^2(\Omega)$ such that if φ is the solution of* (2.7)*, then the solution y of* (2.9) *satisfies $y(T) = y'(T) = 0$. Here D_S is a distribution defined by*

$$\langle D_S, \eta \rangle = - \sum_{j \in \mathcal{I}(S)} \int_{C_{Sj}} \left\{ \varphi'_j \eta'_j + \frac{\partial \varphi_j}{\partial \nu^j} \frac{\partial \eta_j}{\partial \nu^j} + \frac{\partial \varphi_j}{\partial \tau^j} \frac{\partial \eta_j}{\partial \tau^j} \right\} d\sigma dt, \tag{2.10}$$

where $C_S = \cup_{j \in \mathcal{I}(S)} C_{Sj}$, with $C_{Sj} = C_S \cap Q_j$, and on C_{Sj} we have set $\nu^j = (\cos\theta_j, \sin\theta_j)$, $\tau^j = (-\sin\theta_j, \cos\theta_j)$.

Remark 2.3. In Theorems 2.1 and 2.2, the boundary controls are quite classical. The influence of the singularities is balanced by the addition of the internal control

D_S for all singular vertices $S \in \mathcal{S}_{sing}$, which can be seen as a distributional internal control with a support concentrated on V_S in the first case and on C_S in the second one. Their introduction avoids the regularity hypothesis $D(A) \hookrightarrow \mathcal{H}^{3/2+\varepsilon}(\Omega)$ for some $\varepsilon > 0$, leading to strong geometrical conditions on the domains Ω [10, Chaps. 4, 7]. On the other hand, as already explained in part I, the conditions (H1) to (H4) [20] are not related to the singularities but are linked to the multiplier method, since they were introduced to avoid control on internal interfaces (see [20, Remark 6.2]).

3. Estimate of the energy I. The aim of this section is to prove the estimate of the energy with respect to an appropriate norm, leading to the addition of surfacial internal controls with the help of HUM. Usually [12, 4], the energy estimate is based on an identity with multiplier, but here we use two inequalities with multipliers, unlike [13, 5, 14], where an identity with a remainder involving the coefficients of singularities is used. The first multiplier has a support concentrated near the singular points S and vanishes at S in order to balance the singularities, while the second one vanishes in a neighborhood of the singular vertices. More precisely, we prove the following two lemmas.

LEMMA 3.1. *Let $\varphi \in C([0,T], D(A)) \cap C^1([0,T], V) \cap C^2([0,T], H)$ be the solution of (2.7). For any $S \in \mathcal{S}_{sing}$, set $m_i^S(x) = (r_i \cos\theta_i, r_i \sin\theta_i)$ for all $i \in \mathcal{I}(S)$ and $\tilde{m}^S = \eta_S m^S$, where $\eta_S \equiv \eta_S(r)$ is a cut-off function such that $\eta_S \equiv 1$ on $B(S,\delta) \cap \Omega$ with support in $B(S,\delta') \cap \Omega$ for $\delta < \delta' < \gamma$. Then the following identity holds:*

$$\begin{aligned}(3.1)\qquad &\frac{1}{2}\int_Q \operatorname{div}\,\tilde{m}^S\{(\varphi')^2 - \alpha|\nabla\varphi|^2\}\,dxdt\\ &+ \sum_{k,l=1,2}\int_Q \alpha D_k\tilde{m}_l^S D_k\varphi D_l\varphi\,dxdt = -\int_\Omega \varphi'\tilde{m}^S\cdot\nabla\varphi dx|_0^T.\end{aligned}$$

Proof. As $\tilde{m}^S \cong r$ near S and since $\tilde{m}^S$ is identically equal to 0 far from the other vertices, $\tilde{m}^S$ balances the singularities (see [20, Theorem 3.5]), and, therefore, one may apply formula (3.4) of [4]; i.e., the following identity holds on each P_i:

$$\begin{aligned}(3.2)\qquad &\int_{P_i}\triangle\varphi_i\tilde{m}_i^S\cdot\nabla\varphi_i\,dx = -\sum_{k,l=1,2}\int_{P_i} D_k\tilde{m}_{i,l}^S D_k\varphi_i D_l\varphi_i\,dx\\ &+\frac{1}{2}\int_{P_i}\operatorname{div}\,\tilde{m}_i^S|\nabla\varphi_i|^2\,dx + \sum_{j=1}^{N_i}\left\{-\frac{1}{2}\int_{\Gamma_{ij}}\tilde{m}_i^S\cdot\nu_{ij}|\nabla\varphi_i|^2\,d\sigma\right.\\ &\qquad\left.+\int_{\Gamma_{ij}}\frac{\partial\varphi_i}{\partial\nu_{ij}}\tilde{m}_i^S\cdot\nabla\varphi_i\,d\sigma\right\},\end{aligned}$$

where $\tilde{m}_{i,l}^S$ stands for the lth component of $\tilde{m}_i^S$. By the choice of the multiplier $\tilde{m}^S$, we always have

$$(3.3)\qquad \tilde{m}_i^S\cdot\nu_{ij} = 0 \text{ on } \Gamma_{ij}\ \forall j = 1,\dots,N_i.$$

Taking into account this property, multiplying the identity (3.2) by α_i, and summing on $i \in \mathcal{I}$, we get

$$\begin{aligned}(3.4)\qquad &\int_\Omega A\varphi\tilde{m}^S\cdot\nabla\varphi\,dx = \sum_{k,l=1,2}\int_\Omega \alpha D_k\tilde{m}_l^S D_k\varphi D_l\varphi\,dx\\ &-\frac{1}{2}\int_\Omega\alpha\operatorname{div}\,\tilde{m}^S|\nabla\varphi|^2\,dx - \sum_{i\in\mathcal{I}}\sum_{j=1}^{N_i}\int_{\Gamma_{ij}}\alpha_i\frac{\partial\varphi_i}{\partial\nu_{ij}}\tilde{m}_i^S\cdot\tau_{ij}\frac{\partial\varphi_i}{\partial\tau_{ij}}\,d\sigma.\end{aligned}$$

Since $\tilde{m}^S$ is continuous through the edges of Ω, the boundary term of the above identity may be transformed as follows:

$$\sum_{i\in\mathcal{I}}\sum_{j=1}^{N_i}\int_{\Gamma_{ij}}\alpha_i\frac{\partial\varphi_i}{\partial\nu_{ij}}\tilde{m}_i^S\cdot\tau_{ij}\frac{\partial\varphi_i}{\partial\tau_{ij}}\,d\sigma=\sum_{A\in\mathcal{A}}\int_A\left\{\sum_{i\in\mathcal{I}_A}\alpha_i\frac{\partial\varphi_i}{\partial\nu_i}\right\}\tilde{m}^S\cdot\tau\frac{\partial\varphi}{\partial\tau}\,d\sigma.$$

As φ satisfies the "boundary conditions" (2.2) and (2.4), this right-hand side cancels. Accordingly, the identity (3.4) is reduced to

$$\int_\Omega A\varphi\tilde{m}^S\cdot\nabla\varphi\,dx=\sum_{k,l=1,2}\int_\Omega\alpha D_k\tilde{m}_l^S D_k\varphi D_l\varphi\,dx - \frac{1}{2}\int_\Omega\alpha\ \mathrm{div}\ \tilde{m}^S|\nabla\varphi|^2\,dx. \tag{3.5}$$

For the term $\int_Q D_t^2\varphi\tilde{m}^S\cdot\nabla\varphi\ dxdt$, one integration by parts in t and Green's formula directly yield

$$\int_{Q_i}D_t^2\varphi_i\tilde{m}_i^S\cdot\nabla\varphi_i\,dxdt=\int_{P_i}D_t\varphi_i\tilde{m}_i^S\cdot\nabla\varphi_i\,dx|_0^T + \frac{1}{2}\int_{Q_i}\mathrm{div}\ \tilde{m}_i^S(D_t\varphi_i)^2\,dxdt \tag{3.6}$$

because of (3.3).

Integrating (3.5) with respect to $t\in(0,T)$ and summing the result with the sum of (3.6) on $i\in\mathcal{I}$, we arrive at (3.1), since $D_t^2\varphi+A\varphi=0$. □

By the preceding choice of $\tilde{m}^S$, we remark that

$$D_k\tilde{m}_{i,l}^S(x)=\eta_{S,i}(x)\delta_{k,l}+D_k\eta_{S,i}(x)(x-S)_l\ \forall x\in P_i, i\in\mathcal{I}(S).$$

This allows us to rewrite (3.1) in the following way:

$$\int_Q\eta_S(D_t\varphi)^2\,dxdt=R_S, \tag{3.7}$$

where we set

$$R_S=-\frac{1}{2}\int_Q\nabla\eta_S\cdot(x-S)\{(\varphi')^2-\alpha|\nabla\varphi|^2\}\,dxdt - \sum_{k,l=1,2}\int_Q\alpha D_k\eta_S(x-S)_l D_k\varphi D_l\varphi\,dxdt - \int_\Omega\varphi'\tilde{m}^S\cdot\nabla\varphi dx|_0^T. \tag{3.8}$$

The trick, as we shall see later on, is that this remainder is bounded by

$$|R_S|\le C\left[E_0+\int_{V_S}\{\varphi^2+(\varphi')^2\}\,dxdt\right] \tag{3.9}$$

for some positive constant C, depending upon η_S, T, and the geometry of Ω but independent of the initial data φ_0,φ_1.

Accordingly, summing (3.7) on all singular vertices S, we obtain

$$\int_Q \sum_{S\in\mathcal{S}_{sing}} \eta_S (D_t\varphi)^2\, dxdt = \sum_{S\in\mathcal{S}_{sing}} R_S. \tag{3.10}$$

The left-hand side of (3.10) differs from the left-hand side of the classical identity with multiplier (see, e.g., [4, eq. (3.6)] or [12, eq. (I.5.2)]) by the presence of the factor $\sum_{S\in\mathcal{S}_{sing}} \eta_S$ in front of $(D_t\varphi)^2$. Therefore it remains to cover the interior part of Ω. For this part, we take as multiplier $\tilde{m} = (1 - \sum_{S\in\mathcal{S}_{sing}} \eta_S)m$, where m is the multiplier introduced in section 4 of [20].

LEMMA 3.2. *The solution* $\varphi \in C([0,T], D(A)) \cap C^1([0,T], V) \cap C^2([0,T], H)$ *of* (2.7) *satisfies the inequality*

$$\begin{aligned}
&\frac{1}{2}\int_Q \operatorname{div}\tilde{m}\{(\varphi')^2 - \alpha|\nabla\varphi|^2\}\, dxdt + \sum_{k,l=1,2}\int_Q \alpha D_k\tilde{m}_l D_k\varphi D_l\varphi\, dxdt \\
&\le -\int_\Omega \varphi'\tilde{m}\cdot\nabla\varphi dx|_0^T + \frac{1}{2}\sum_{A\in\mathcal{D}}\int_{\Sigma_A} \alpha_{i_A}\tilde{m}_{i_A}\cdot\nu_{i_A}\left(\frac{\partial\varphi_{i_A}}{\partial\nu_{i_A}}\right)^2 d\sigma dt \\
&+ \frac{1}{2}\sum_{A\in\mathcal{N}_{ext}}\int_{\Sigma_A}\tilde{m}_{i_A}\cdot\nu_{i_A}\left\{(D_t\varphi_{i_A})^2 - \alpha_{i_A}(\frac{\partial\varphi_{i_A}}{\partial\tau_{i_A}})^2\right\} d\sigma dt.
\end{aligned} \tag{3.11}$$

Proof. The proof of (3.11) is analogous to the proof of Proposition 4.2 in [20] with $\varphi = \varphi_R$. Indeed, the integrations by parts are allowed since our multiplier $\tilde{m}$ vanishes in a neighborhood of the singular vertices; on the other hand, the continuity of η_S and the property $0 \le \eta_S \le 1$ imply that $\tilde{m}$ still satisfies the conditions (H1) to (H4) in [20]. □

As previously, the multiplier satisfies

$$D_k\tilde{m}_l = \left(1 - \sum_{S\in\mathcal{S}_{sing}}\eta_S\right)\delta_{k,l} - \sum_{S\in\mathcal{S}_{sing}} D_k\eta_S m_l.$$

This allows us to transform (3.11) into

$$\int_Q \left(1 - \sum_{S\in\mathcal{S}_{sing}}\eta_S\right)(D_t\varphi)^2\, dxdt \le R, \tag{3.12}$$

where we define

$$\begin{aligned}
R = &-\int_\Omega \varphi'\tilde{m}\cdot\nabla\varphi dx|_0^T \\
&+ \frac{1}{2}\sum_{A\in\mathcal{N}_{ext}}\int_{\Sigma_A}\tilde{m}_{i_A}\cdot\nu_{i_A}\left\{(D_t\varphi_{i_A})^2 - \alpha_{i_A}\left(\frac{\partial\varphi_{i_A}}{\partial\tau_{i_A}}\right)^2\right\} d\sigma dt \\
&+ \frac{1}{2}\sum_{A\in\mathcal{D}}\int_{\Sigma_A}\alpha_{i_A}\tilde{m}_{i_A}\cdot\nu_{i_A}\left(\frac{\partial\varphi_{i_A}}{\partial\nu_{i_A}}\right)^2 d\sigma dt \\
&+ \frac{1}{2}\int_Q \sum_{S\in\mathcal{S}_{sing}}(\nabla\eta_S\cdot m)\{(\varphi')^2 - \alpha|\nabla\varphi|^2\}\, dxdt \\
&+ \sum_{k,l=1,2}\int_Q \alpha\sum_{S\in\mathcal{S}_{sing}} D_k\eta_S m_l D_k\varphi D_l\varphi\, dxdt.
\end{aligned} \tag{3.13}$$

For any $\{\varphi_0, \varphi_1\} \in D(A) \times V$, let $\varphi \in C([0,T], D(A)) \cap C^1([0,T], V)$ be the solution of (2.7), and define

$$
\begin{aligned}
|||\{\varphi_0, \varphi_1\}|||^2 = & \sum_{A \in \mathcal{D}^+} \int_{\Sigma_{A\delta}} \left(\frac{\partial \varphi_{i_A}}{\partial \nu_{i_A}} \right)^2 d\sigma dt \\
& + \sum_{A \in \mathcal{N}_{ext}^-} \int_{\Sigma_{A\delta}} \left(\frac{\partial \varphi_{i_A}}{\partial \tau_{i_A}} \right)^2 d\sigma dt \\
& + \sum_{A \in \mathcal{N}_{ext}^+} \int_{\Sigma_{A\delta}} \left(D_t \varphi_{i_A} \right)^2 d\sigma dt \\
& + \sum_{S \in \mathcal{S}_{sing}} \int_{V_S} [\varphi^2 + (\varphi')^2] \, dx dt.
\end{aligned} \tag{3.14}
$$

We are now ready to establish the main result of this section.

PROPOSITION 3.3. *Let $\varphi \in C([0,T], D(A)) \cap C^1([0,T], V) \cap C^2([0,T], H)$ be a solution of* (2.7). *Then there exists a minimal time $T_0 > 0$ such that for all $T > T_0$, there exists a constant $C > 0$ (depending on T but not on φ_0, φ_1) such that*

$$
(T - T_0) E_0 \leq C |||\{\varphi_0, \varphi_1\}|||^2. \tag{3.15}
$$

Proof. We first start as in Proposition 4.3 of [20], i.e., by Theorem 4.3 and Remark 4.4 of [20] and the identity (4.24) of [15], we may write

$$
T E_0 = \int_Q |D_t \varphi|^2 dx dt - \frac{1}{2} \int_\Omega D_t \varphi \varphi \, dx \, |_0^T \, . \tag{3.16}
$$

The estimation of the term $\int_\Omega D_t \varphi \varphi \, dx \, |_0^T$ with respect to E_0 is classical (see [12, 4]); therefore, it remains to estimate $\int_Q |D_t \varphi|^2 dx dt$.

The sum of (3.10) and (3.12) yields a positive constant C depending only upon the functions η_S and the geometry of Ω such that

$$
\begin{aligned}
\int_Q |D_t \varphi|^2 dx dt \leq & \, C |||\{\varphi_0, \varphi_1\}|||^2 + C E_0 \\
& + C \sum_{S \in \mathcal{S}_{sing}} \int_0^T \int_{B(S,\delta')} [(\varphi')^2 + \alpha |\nabla \varphi|^2] \, dx dt
\end{aligned} \tag{3.17}
$$

because $\nabla \eta_S \equiv 0$ outside the ball $B(S, \delta')$. We therefore have only to estimate

$$
\int_0^T \int_{B(S,\delta')} \alpha |\nabla \varphi|^2 \, dx dt
$$

for all $S \in \mathcal{S}_{sing}$. This is done using Heibig and Moussaoui's trick [5]. We fix another cut-off function $\tilde{\eta}_S$ similar to η_S but satisfying $\tilde{\eta}_S \equiv 1$ on $B(S, \delta')$ and with a support on $B(S, \gamma)$. Then we multiply the equation

$$
D_t^2 \varphi_i - \alpha_i \triangle \varphi_i = 0
$$

by $\tilde{\eta}_{S,i} \varphi_i$ for all $i \in \mathcal{I}$ and integrate on Q_i, leading to

$$
\int_Q (D_t^2 \varphi - \alpha \triangle \varphi) \tilde{\eta}_S \varphi \, dx dt = 0.
$$

One integration by parts with respect to t and Green's formula on each P_i yields

$$\begin{aligned} 0 = &- \int_Q (D_t\varphi)^2 \tilde{\eta}_S \, dxdt \\ &+ \int_\Omega (D_t\varphi)\varphi\tilde{\eta}_S \, dx|_0^T \\ &+ \int_Q \alpha\nabla\varphi\cdot\nabla(\tilde{\eta}_S\varphi) \, dxdt. \end{aligned}$$

The boundary term on ∂P_i is cancelled due to the continuity of $\tilde{\eta}_S\varphi$ through the edges of Ω and since φ satisfies (2.2)–(2.4). By Leibniz's rule, the above identity may be written as

$$\begin{aligned} \int_Q \alpha|\nabla\varphi|^2\tilde{\eta}_S \, dxdt = &\int_Q (D_t\varphi)^2\tilde{\eta}_S \, dxdt \\ &- \int_\Omega (D_t\varphi)\varphi\tilde{\eta}_S \, dx|_0^T \\ &- \int_Q \alpha\varphi\nabla\varphi\cdot\nabla\tilde{\eta}_S \, dxdt. \end{aligned} \tag{3.18}$$

As $\nabla\tilde{\eta}_S$ is identically equal to 0 near the vertices, Green's formula again leads to

$$\begin{aligned} \int_Q \alpha\varphi\nabla\varphi\cdot\nabla\tilde{\eta}_S \, dxdt = &- \int_Q \alpha\varphi \text{ div } (\varphi\nabla\tilde{\eta}_S) \, dxdt \\ &+ \sum_{i\in\mathcal{I}}\sum_{j=1}^{N_i} \alpha_i \int_0^T \int_{\Gamma_{ij}} (\varphi_i)^2 \frac{\partial\tilde{\eta}_{S,i}}{\partial\nu_{ij}} \, d\sigma dt. \end{aligned}$$

Since $\frac{\partial\tilde{\eta}_{S,i}}{\partial\nu_{ij}} \equiv 0$ on all Γ_{ij}, the previous identity becomes

$$\int_Q \alpha\varphi\nabla\varphi\cdot\nabla\tilde{\eta}_S \, dxdt = -\frac{1}{2}\int_Q \alpha\varphi^2\triangle\tilde{\eta}_S \, dxdt.$$

Inserting this last identity into (3.18), we arrive at

$$\begin{aligned} \int_Q \alpha|\nabla\varphi|^2\tilde{\eta}_S \, dxdt = &\int_Q (D_t\varphi)^2\tilde{\eta}_S \, dxdt \\ &- \int_\Omega (D_t\varphi)\varphi\tilde{\eta}_S \, dx|_0^T \\ &+ \frac{1}{2}\int_Q \alpha\varphi^2\triangle\tilde{\eta}_S \, dxdt. \end{aligned} \tag{3.19}$$

Since $\tilde{\eta}_S \equiv 1$ on $B(S,\delta')$ and has its support on $B(S,\gamma)$, the identity (3.19) yields a positive constant C such that

$$\int_0^T \int_{B(S,\delta')} \alpha|\nabla\varphi|^2 \, dxdt \le C\left[E_0 + \int_{V_S} \{\varphi^2 + (\varphi')^2\} \, dxdt\right]. \tag{3.20}$$

The estimates (3.17) and (3.20) prove (3.15). □

The inverse estimate is only possible under some geometrical conditions. Therefore, we simply note that

$$|||\{\varphi_0,\varphi_1\}|||^2 \le C\{\|\varphi_0\|_{D(A)} + \|\varphi_1\|_V\}, \tag{3.21}$$

which is a consequence of Theorem 3.5 of [20] and the Poincaré estimate

$$\int_\Omega \varphi^2\, dx \le ca(\varphi,\varphi)$$

for some $c > 0$.

We now fix $T > T_0$ such that the inequality (3.15) holds. Then the application

$$D(A)\times V \to \mathbf{R}^+ : \{\varphi_0,\varphi_1\} \to |||\{\varphi_0,\varphi_1\}|||$$

is a norm stronger than the norm induced by $V \times H$ due to Proposition 3.3. As in [12, 4, 15], we define F as the closure of $D(A)\times V$ for this new norm. Due to (3.15) and (3.21) we have the algebraic and topological inclusions:

$$D(A)\times V \hookrightarrow F \hookrightarrow V \times H. \tag{3.22}$$

For the inhomogeneous wave equation (3.1) of [20], we can now state the following result.

PROPOSITION 3.4. *Let $\{\varphi_0,\varphi_1\} \in F$ and $f \in L^1(0,T;V)$. Then the unique solution $\varphi \in C([0,T],V)\cap C^1([0,T],H)$ of* (3.1) *of* [20] *satisfies*

$$\frac{\partial \varphi_{i_A}}{\partial \nu_{i_A}} \in L^2(\Sigma_{A\delta})\,\forall A \in \mathcal{D}^+, \tag{3.23}$$

$$\frac{\partial \varphi_{i_A}}{\partial \tau_{i_A}} \in L^2(\Sigma_{A\delta})\,\forall A \in \mathcal{N}^-_{ext}, \tag{3.24}$$

$$D_t\varphi_{i_A} \in L^2(\Sigma_{A\delta})\,\forall A \in \mathcal{N}^+_{ext}. \tag{3.25}$$

Moreover, there exists a constant $C > 0$ (independent of $\{\varphi_0,\varphi_1\}$ and f) such that

$$\begin{aligned}&\left\{\sum_{A\in\mathcal{D}^+}\int_{\Sigma_{A\delta}}\left(\frac{\partial\varphi_{i_A}}{\partial\nu_{i_A}}\right)^2 d\sigma dt + \sum_{A\in\mathcal{N}^-_{ext}}\int_{\Sigma_{A\delta}}\left(\frac{\partial\varphi_{i_A}}{\partial\tau_{i_A}}\right)^2 d\sigma dt\right.\\ &\left.+\sum_{A\in\mathcal{N}^+_{ext}}\int_{\Sigma_{A\delta}}(D_t\varphi_{i_A})^2\, d\sigma dt + \sum_{S\in\mathcal{S}_{sing}}\int_{V_S}[\varphi^2+(\varphi')^2]\,dxdt\right\}^{1/2}\\ &\le C\left\{|||\{\varphi_0,\varphi_1\}||| + \|f\|_{L^1(0,T;V)}\right\}.\end{aligned} \tag{3.26}$$

Proof. The proof is standard (see [4, Theorem 5.6] or [20, Proposition 4.4]). It is based on the definition of F, Theorem 2.2 of [20] and the usual trace theorems. Indeed the solution $\varphi^{(2)}$ of the wave equation (3.1) of [20] with data $\varphi_0 = \varphi_1 = 0$ and $f \in L^1(0,T;V)$ has the regularity $\varphi^{(2)} \in C([0,T],D(A))\cap C^1([0,T],V)$, as a consequence of Theorem 3.1 of [20]. By Theorem 2.2 and the estimate (3.2) of [20] and the usual trace theorems, it clearly satisfies the estimate (3.26). ∎

4. Weak solutions of the wave equation I. We transpose Proposition 3.4 to get Theorem 4.1.

THEOREM 4.1. *For all $u_0 \in H$, $u_1 \in V'$, $w_A \in L^2(\Sigma_{A\delta})$, where $A \in \mathcal{D}^+ \cup \mathcal{N}^+_{ext} \cup \mathcal{N}^-_{ext}$, and all $w_S \in H^1((0,T);L^2(B(S,\gamma)\cap\Omega)), S \in \mathcal{S}_{sing}$, there exist unique $u \in L^\infty(0,T;V')$, $\{\psi_1,\psi_0\} \in F'$, which satisfy*

$$\begin{aligned}(4.1)\qquad &\int_0^T \langle u(t), f(t)\rangle_{V'-V} dt + \langle\{\psi_1,\psi_0\},\{\varphi_0,-\varphi_1\}\rangle_{F'-F}\\ &= \langle u_1,\varphi(0)\rangle_{V'-V} - \langle u_0,\varphi'(0)\rangle_{H'-H} - \sum_{A\in\mathcal{D}^+}\int_{\Sigma_{A\delta}} w_A \frac{\partial\varphi_{i_A}}{\partial\nu_{i_A}}\, d\sigma dt\\ &- \sum_{A\in\mathcal{N}^-_{ext}}\int_{\Sigma_{A\delta}} w_A \frac{\partial\varphi_{i_A}}{\partial\tau_{i_A}}\, d\sigma dt - \sum_{A\in\mathcal{N}^+_{ext}}\int_{\Sigma_{A\delta}} w_A D_t\varphi_{i_A}\, d\sigma dt\\ &- \sum_{S\in\mathcal{S}_{sing}}\int_{V_S}\{w_S\varphi + w'_S\varphi'\}\, dxdt\end{aligned}$$

for all $f \in L^1(0,T;V)$, $\{\varphi_0,-\varphi_1\} \in F$, where $\varphi \in C([0,T],V)\cap C^1([0,T],H)$ is the unique solution of

$$(4.2)\qquad \begin{cases} \varphi''(t) + A\varphi(t) = f(t), & t\in[0,T],\\ \varphi(T)=\varphi_0, \varphi'(T)=\varphi_1. \end{cases}$$

For the interpretation of (4.1) in terms of partial differential equations, we need the next density result.

LEMMA 4.2. *Let X be a separable Hilbert space and denote by*

$$K = \{w \in H^2((0,T);X) : w'(0) = w'(T) = 0\}.$$

Then K is dense in $H^1((0,T);X)$.

Proof. Let $u \in K^\perp$; then it fulfills

$$(4.3)\qquad \int_0^T \{(u(t),w(t))_X + (u'(t),w'(t))_X\}\, dt = 0\,\forall w \in K,$$

where $(\cdot,\cdot)_X$ denote the inner product of X. Fix an arbitrary element w of X and introduce the function

$$U : t \to (u(t),w)_X.$$

Clearly, it belongs to $H^1((0,T))$ and satisfies

$$U'' - U = 0 \text{ in } \mathcal{D}'((0,T))$$

by taking as test function $w(t)$ in (4.3) the function $w(t) = \varphi(t)w$, where $\varphi \in \mathcal{D}((0,T))$. This implies that U is a linear combination of e^t and e^{-t}. But going back to (4.3) with $w(t) = w\varphi(t)$, where $\varphi \in C^\infty([0,T])$ such that $\varphi'(T) = \varphi'(0) = 0$, we deduce that $U \equiv 0$, because $\varphi(0), \varphi(T)$ are free.

Since w was arbitrary in X, we conclude that $u \equiv 0$. □

We apply Lemma 4.2 with $X = L^2(B(S,\gamma)\cap\Omega)$, for any $S \in \mathcal{S}_{sing}$. Let us then denote by K_S the corresponding space K.

THEOREM 4.3. *Let $u \in L^\infty(0,T;V')$, $\{\psi_1,\psi_0\} \in F'$ be the unique solution of (4.1) with data $u_0 \in V$, $u_1 \in H$, $w_A \in \mathcal{D}(\Sigma_{A\delta})$, where $A \in \mathcal{D}^+ \cup \mathcal{N}^+_{ext} \cup \mathcal{N}^-_{ext}$, and with $w_S \in K_S, S \in \mathcal{S}_{sing}$. Then $u \in C^1([0,T],H)$ satisfies the boundary conditions*

$$u_{i_A} = \begin{cases} \alpha_{i_A}^{-1} w_A \text{ on } \Sigma_{A\delta}\, \forall A \in \mathcal{D}^+, \\ 0 \text{ on } \Sigma_A \setminus \Sigma_{A\delta}\, \forall A \in \mathcal{D}^+, \\ 0 \text{ on } \Sigma_A\, \forall A \in \mathcal{D} \setminus \mathcal{D}^+ \end{cases} \tag{4.4}$$

and (4.5) *to* (4.7) *below:*

$$u_i'' - \alpha_i \triangle u_i = \sum_{S \in \mathcal{S}_{sing}} \{w_{S,i} - w_{S,i}''\}\chi_{V_S,i} \text{ in } \mathcal{D}'(Q_i)\, \forall i \in \mathcal{I}, \tag{4.5}$$

$$u(0) = u_0, \qquad u'(0) = u_1, \tag{4.6}$$

$$u(T) = \psi_0, \qquad u'(T) = \psi_1. \tag{4.7}$$

Proof. We proceed as in Theorem 5.3 of [15] or in Theorem 5.4 of [20] with the necessary adaptations. Let us fix $v \in \mathcal{D}(0,T,\prod_{i\in\mathcal{I}} C^\infty(\overline{P}_i))$ fulfilling (5.3) and (5.4) of [20] (on $\Sigma_A \setminus \Sigma_{A\delta}$, we simply take $w_A = 0$) and obtained in [20, Lemma 5.3]. (v can be chosen equal to 0 in a neighborhood of $S \times [0,T]$ for all $S \in \mathcal{S}_{sing}$; see Lemma 7.2). Define

$$f = v'' - \alpha \triangle v - \sum_{S \in \mathcal{S}_{sing}} \{w_S - w_S''\}\chi_{V_S}.$$

Since $f \in L^2(0,T;H)$, Lemma I.3.4 of [12] guarantees the existence of a unique solution $\psi \in C([0,T],V) \cap C^1([0,T],H) \cap H^2(0,T;V')$ of

$$\begin{cases} \langle \psi''(t), w\rangle + a(\psi(t), w) \\ = -\int_\Omega f(t) w dx, \text{ a.e. } t \in [0,T]\, \forall w \in V, \\ \psi(0) = u_0, \qquad \psi'(0) = u_1. \end{cases} \tag{4.8}$$

Let us now show that

$$u = \psi + v \tag{4.9}$$

is the unique solution of (4.1) when $\psi_0 = u(T), \psi_1 = u'(T)$. We remark that the fact that $v \equiv 0$ near $t = 0$ and $t = T$ leads to the initial conditions (4.6).

From Theorem 4.2 of [15], it suffices to check (4.1) for $\varphi \in C([0,T],D(A)) \cap C^1([0,T],V) \cap C^2([0,T],H)$. Since $u \in H^2(0,T;V')$, the integrations by parts over $(0,T)$ are allowed. Taking into account the initial conditions satisfied by φ and u, we get

$$\begin{aligned} &\int_0^T \langle u(t), \varphi''(t) + A\varphi(t)\rangle dt - \langle u(T), \varphi_1\rangle + \langle u'(T), \varphi_0\rangle \\ &= \langle u_1, \varphi(0)\rangle - \langle u_0, \varphi'(0)\rangle \\ &+ \int_0^T \{\langle \psi''(t), \varphi(t)\rangle + a(\psi(t), \varphi(t)) \\ &\qquad + \langle v''(t), \varphi(t)\rangle + (v(t), A\varphi(t))_H\} dt. \end{aligned} \tag{4.10}$$

As $v(t)$ is identically equal to 0 in a neighborhood of the singular vertices, Green's formula on each P_i is allowed and leads to

$$\begin{aligned}(v(t),A\varphi(t))_H &= (Av(t),\varphi(t))_H \\ &\quad+\sum_{A\in\mathcal{N}_{ext}}\int_A \alpha_{i_A}\frac{\partial v_{i_A}}{\partial\nu_{i_A}}\varphi_{i_A}\,d\sigma \\ &\quad-\sum_{A\in\mathcal{D}}\int_A \alpha_{i_A}v_{i_A}\frac{\partial\varphi_{i_A}}{\partial\nu_{i_A}}\,d\sigma.\end{aligned}$$

Integrating this identity on $(0,T)$ and taking into account (5.3) and (5.4) of [20], we obtain, after integrations by parts on each $\Sigma_{A\delta}$ for all $A\in\mathcal{N}_{ext}^+\cup\mathcal{N}_{ext}^-$,

$$\begin{aligned}\int_0^T (v(t),A\varphi(t))_H\,dt &= \int_0^T (Av(t),\varphi(t))_H\,dt \\ &\quad-\sum_{A\in\mathcal{N}_{ext}^+}\int_{\Sigma_{A\delta}} w_A D_t\varphi_{i_A}\,d\sigma dt \\ &\quad-\sum_{A\in\mathcal{N}_{ext}^-}\int_{\Sigma_{A\delta}} w_A\frac{\partial\varphi_{i_A}}{\partial\tau_{i_A}}\,d\sigma dt \\ &\quad-\sum_{A\in\mathcal{D}^+}\int_{\Sigma_{A\delta}} w_A\frac{\partial\varphi_{i_A}}{\partial\nu_{i_A}}\,d\sigma dt.\end{aligned}$$

Inserting this identity into (4.10), we arrive at the conclusion, because the choice of $w_S\in K_S$ yields

$$\int_{V_S}(w_S-w_S'')\varphi\,dxdt=\int_{V_S}(w_S\varphi+w_S'\varphi')\,dxdt. \qquad \square$$

Roughly speaking, ψ introduced in the above proof satisfies (2.4), and therefore, as v satisfies (5.4), one can say that u satisfies

$$\frac{\partial u_{i_A}}{\partial\nu_{i_A}}=\begin{cases}\alpha_{i_A}^{-1}D_t w_A \text{ on } \Sigma_{A\delta}\,\forall A\in\mathcal{N}_{ext}^+,\\ \alpha_{i_A}^{-1}\frac{\partial w_A}{\partial\tau_{i_A}} \text{ on } \Sigma_{A\delta}\,\forall A\in\mathcal{N}_{ext}^-,\\ 0 \text{ on } \Sigma_A\setminus\Sigma_{A\delta}\,\forall A\in(\mathcal{N}_{ext}^+\cup\mathcal{N}_{ext}^-),\\ 0 \text{ on } \Sigma_A\,\forall A\in\mathcal{N}_{ext}\setminus(\mathcal{N}_{ext}^+\cup\mathcal{N}_{ext}^-)\end{cases} \tag{4.11}$$

in a weak sense.

All these considerations lead us to call the solution u of (4.1) the weak solution of (4.5), (4.6), (4.4), and (4.11). Moreover, with the help of Lemma 4.2, Theorem 5.7 of [20] still holds for u, which allows us to give a meaning to the final conditions (4.7). We even have the following result.

THEOREM 4.4. *Under the assumption of Theorem* 4.1*, let* u*,* $\{\psi_1,\psi_0\}$ *be the solutions of* (4.1)*. Then* $u\in C([0,T],V')\cap C^1([0,T],D(A)')$ *and* u *satisfies the final conditions* (4.7)*.*

5. Addition of surfacial internal controls. The application of the Hilbert uniqueness method of Lions [12] is now standard. First, by Proposition 3.4 for $\{\varphi_0,\varphi_1\}\in F$ there exists a unique solution $\varphi\in C([0,T],V)\cap C^1([0,T],H)$ of (2.7) (or (3.1) of [20] with $f=0$), satisfying (3.26). Second, consider $\psi\in L^\infty(0,T;V')$, $\{\chi_1,-\chi_0\}\in F'$, the unique solution of

$$
\begin{aligned}
(5.1)\qquad & \int_0^T \langle \psi(t), g(t)\rangle dt - \langle \{\chi_1, -\chi_0\}, \{\eta_0, \eta_1\}\rangle \\
&= -\sum_{A\in\mathcal{D}^+} \int_{\Sigma_{A\delta}} \frac{\partial \varphi_{i_A}}{\partial \nu_{i_A}} \frac{\partial \eta_{i_A}}{\partial \nu_{i_A}}\, d\sigma dt - \sum_{A\in\mathcal{N}^-_{ext}} \int_{\Sigma_{A\delta}} \frac{\partial \varphi_{i_A}}{\partial \tau_{i_A}} \frac{\partial \eta_{i_A}}{\partial \tau_{i_A}}\, d\sigma dt \\
&- \sum_{A\in\mathcal{N}^+_{ext}} \int_{\Sigma_{A\delta}} D_t\varphi_{i_A} D_t\eta_{i_A}\, d\sigma dt \\
&- \sum_{S\in\mathcal{S}_{sing}} \int_{V_S} \{\varphi\eta + \varphi'\eta'\}\, dxdt
\end{aligned}
$$

for all $g \in L^1(0,T;V)$, $\{\eta_0, \eta_1\} \in F$, where $\eta \in C([0,T],V) \cap C^1([0,T],H)$ is the unique solution of

$$
(5.2)\qquad \begin{cases} \eta''(t) + A\eta(t) = g(t), & t \in [0,T], \\ \eta(0) = \eta_0, \qquad \eta'(0) = \eta_1. \end{cases}
$$

Its existence comes from Theorem 4.1 by inverting the order of time; moreover, Theorem 4.4 gives a meaning to the initial conditions

$$\psi(0) = \chi_0, \qquad \psi'(0) = \chi_1.$$

Accordingly, the operator

$$\Lambda : F \to F' : \{\varphi_0, \varphi_1\} \to \{\chi_1, -\chi_0\}$$

is well defined and is an isomorphism, because the identity (5.1) with $\eta = \varphi$ yields

$$\langle \Lambda\{\varphi_0, \varphi_1\}, \{\varphi_0, \varphi_1\}\rangle = |||\{\varphi_0, \varphi_1\}|||^2 \ \forall \{\varphi_0, \varphi_1\} \in F.$$

This leads to Theorem 2.1, which we reformulate as follows.

THEOREM 5.1. *For all $u_0 \in H$, $u_1 \in V'$ there exist $w_A \in L^2(\Sigma_{A\delta})$, $A \in \mathcal{D}^+ \cup \mathcal{N}^+_{ext} \cup \mathcal{N}^-_{ext}$, and $v_S \in H^1((0,T); L^2(B(S,\gamma)\cap\Omega)$, $S \in \mathcal{S}_{sing}$, such that the weak solution $u \in C([0,T],V') \cap C^1([0,T],D(A)')$ of the wave equation* (5.3) *below* (*in the sense of* (4.1)) *satisfies $u(T) = u'(T) = 0$:*

$$
(5.3)\qquad \begin{cases} u''(t) + Au(t) = \sum_{S\in\mathcal{S}_{sing}} \{v_S - \frac{d}{dt} v'_S\}\chi_{V_S}, & t \in [0,T], \\ u(0) = u_0, \qquad u'(0) = u_1, \\ u \text{ satisfies (4.4) and (4.11).} \end{cases}
$$

Proof. Since $\{u_1, -u_0\} \in V' \times H \subset F'$, there exists a unique solution $\{\varphi_0, \varphi_1\} \in F$ of

$$\Lambda\{\varphi_0, \varphi_1\} = \{u_1, -u_0\}.$$

We take the solution φ of (2.7) and the solution ψ of (5.1) and set $u = \psi$, $w_A = \frac{\partial \varphi_{i_A}}{\partial \nu_{i_A}}$ for all $A \in \mathcal{D}^+$, $w_A = D_t\varphi_{i_A}$ for all $A \in \mathcal{N}^+_{ext}$, $w_A = \frac{\partial \varphi_{i_A}}{\partial \tau_{i_A}}$ for all $A \in \mathcal{N}^-_{ext}$, $v_S = \varphi$ for all $S \in \mathcal{S}_{sing}$. Because of the time reversibility of the wave equation, the conclusion follows from Proposition 3.4. ∎

6. Estimate of the energy II. The main idea of the second method consists of isolating the singular vertices S by the introduction of an artificial interface C_S. We then use the classical multiplier far from the singular vertices and the multiplier m^S on $B(S,\delta)\cap\Omega$. Namely, we establish the following result.

LEMMA 6.1. *Introduce the multiplier q defined on Ω by $q(x) = m^S(x)$ if $x \in B(S,\delta)\cap\Omega$ for all $S \in \mathcal{S}_{sing}$ and $q(x) = m(x)$ if $x \in \Omega \setminus \cup_{S\in\mathcal{S}_{sing}} B(S,\delta)$. Then the solution $\varphi \in C([0,T], D(A)) \cap C^1([0,T], V)$ of (2.7) satisfies the inequality*

$$
\begin{aligned}
(6.1)\quad \frac{1}{2}\int_Q (\varphi')^2\, dxdt \le & -\int_\Omega \varphi' q\cdot\nabla\varphi dx|_0^T \\
& + \frac{1}{2}\sum_{A\in\mathcal{D}}\int_{\Sigma_{A\delta}} \alpha_{i_A} m_{i_A}\cdot\nu_{i_A}\left(\frac{\partial\varphi_{i_A}}{\partial\nu_{i_A}}\right)^2 d\sigma dt \\
& + \frac{1}{2}\sum_{A\in\mathcal{N}_{ext}}\int_{\Sigma_{A\delta}} m_{i_A}\cdot\nu_{i_A}\left\{(D_t\varphi_{i_A})^2 - \alpha_{i_A}\left(\frac{\partial\varphi_{i_A}}{\partial\tau_{i_A}}\right)^2\right\} d\sigma dt \\
& + \sum_{S\in\mathcal{S}_{sing}}\sum_{j\in\mathcal{I}(S)}\int_{C_{Sj}}\left[\frac{1}{2}(m_j^S - m_j)\cdot\nu^j\left\{(\varphi_j')^2 + \alpha_j\left(\frac{\partial\varphi_j}{\partial\nu_j}\right)^2 - \alpha_j\left(\frac{\partial\varphi_j}{\partial\tau_j}\right)^2\right\}\right. \\
& \qquad \left. + \alpha_j(m_j^S - m_j)\cdot\tau^j\frac{\partial\varphi_j}{\partial\nu_j}\frac{\partial\varphi_j}{\partial\tau_j}\right] d\sigma dt.
\end{aligned}
$$

Proof. We follow the proof of Proposition 4.2 of [20], except that the integration by parts on Ω are here made on $B(S,\delta)\cap\Omega$ for all $S\in\mathcal{S}_{sing}$ and on the remainder, i.e., on $\Omega\setminus\cup_{S\in\mathcal{S}_{sing}} B(S,\delta)$. That explains the presence of the boundary terms on C_S. More precisely, as $m^S \cong r$ near the singular vertices S, m^S balances the singularities (see [20, Theorem 3.5]) and, therefore, one may apply formula (3.4) of [4] in $P_i\cap B(S,\delta)$ and in $P_i\setminus\cup_{S\in\mathcal{S}_{sing}} B(S,\delta)$. Summing the results, we arrive at the following identity on each P_i:

$$
\begin{aligned}
(6.2)\quad \int_{P_i}\triangle\varphi_i q_i\cdot\nabla\varphi_i\, dx = & \sum_{j=1}^{N_i}\left\{-\frac{1}{2}\int_{\Gamma_{ij}} q_i\cdot\nu_{ij}|\nabla\varphi_i|^2\, d\sigma\right. \\
& \qquad \left. + \int_{\Gamma_{ij}}\frac{\partial\varphi_i}{\partial\nu_{ij}} q_i\cdot\nabla\varphi_i\, d\sigma\right\} \\
& + \sum_{S\in\mathcal{S}_{sing}\cap\bar{P}_i}\int_{S_1(S,\delta)\cap P_i}\left[\frac{1}{2}(m_i^S - m_i)\cdot\nu^i\left\{\left(\frac{\partial\varphi_i}{\partial\nu_i}\right)^2 - \left(\frac{\partial\varphi_i}{\partial\tau_i}\right)^2\right\}\right. \\
& \qquad \left. + (m_i^S - m_i)\cdot\tau^i\frac{\partial\varphi_i}{\partial\nu_i}\frac{\partial\varphi_i}{\partial\tau_i}\right] d\sigma dt.
\end{aligned}
$$

As usual, we multiply this identity by α_i and sum on $i\in\mathcal{I}$. Since q is continuous on $B(S,\delta)\cap\Omega$ for $S\in\mathcal{S}_{sing}$ and satisfies

$$
q_i\cdot\nu_{ij} = 0 \text{ on } \Gamma_{ij}\cap B(S,\delta)\ \forall j = 1,\dots,N_i,
$$

the boundary terms are equal to 0 on $\Gamma_{ij}\cap B(S,\delta)$ for all $S\in\mathcal{S}_{sing}$. Moreover, as in Lemma 4.1 of [20], taking into account the conditions (H1) to (H4) [20] satisfied by m, the identity (6.2) implies the estimate

$$
\begin{aligned}
(6.3)\quad \int_\Omega A\varphi q\cdot\nabla\varphi dx \ge & -\frac{1}{2}\sum_{A\in\mathcal{D}}\int_{A_\delta}\alpha_{i_A} m_{i_A}\cdot\nu_{i_A}\left(\frac{\partial\varphi_{i_A}}{\partial\nu_{i_A}}\right)^2 d\sigma \\
& + \frac{1}{2}\sum_{A\in\mathcal{N}_{ext}}\int_{A_\delta}\alpha_{i_A} m_{i_A}\cdot\nu_{i_A}\left(\frac{\partial\varphi_{i_A}}{\partial\tau_{i_A}}\right)^2 d\sigma
\end{aligned}
$$

$$+ \sum_{S\in\mathcal{S}_{sing}} \sum_{i\in\mathcal{I}(S)} \alpha_i \int_{S_1(S,\delta)\cap P_i} \left[\frac{1}{2}(m_i^S - m_i)\cdot\nu^i \left\{ \left(\frac{\partial\varphi_i}{\partial\nu_i}\right)^2 - \left(\frac{\partial\varphi_i}{\partial\tau_i}\right)^2 \right\} + (m_i^S - m_i)\cdot\tau^i \frac{\partial\varphi_i}{\partial\nu_i}\frac{\partial\varphi_i}{\partial\tau_i} \right] d\sigma dt.$$

For the term $\int_Q D_t^2\varphi q\cdot\nabla\varphi\, dxdt$ one integration by parts in t leads to

$$\int_Q D_t^2\varphi q\cdot\nabla\varphi\, dxdt = \int_\Omega D_t\varphi q\cdot\nabla\varphi\, dx\,|_0^T - \int_Q D_t\varphi q\cdot\nabla D_t\varphi\, dxdt. \tag{6.4}$$

The second term of the right-hand side is transformed using Green's formula in $P_i\cap B(S,\delta)$ for any $S\in\mathcal{S}_{sing}$ and in the remainder for all $i\in\mathcal{I}$. This leads to

$$\begin{aligned}\int_Q D_t\varphi q\cdot\nabla D_t\varphi\, dxdt = &-\int_Q (D_t\varphi)^2 dxdt \\ &+ \frac{1}{2}\sum_{A\in\mathcal{A}}\int_{\Sigma_A}(D_t\varphi)^2\left(\sum_{i\in\mathcal{I}_A} q_i\cdot\nu_i\right)d\sigma dt \\ &+ \frac{1}{2}\sum_{S\in\mathcal{S}_{sing}}\sum_{j\in\mathcal{I}(S)}\int_{C_{Sj}}(m_j^S - m_j)\cdot\nu^j(\varphi')^2\, d\sigma dt.\end{aligned}$$

Taking into account the above properties of q and the condition (H1) [20], we arrive at

$$\begin{aligned}\int_Q D_t\varphi q\cdot\nabla D_t\varphi\, dxdt = &-\int_Q (D_t\varphi)^2 dxdt \\ &+ \frac{1}{2}\sum_{A\in\mathcal{N}_{ext}}\int_{\Sigma_{A\delta}}(D_t\varphi)^2 q_{i_A}\cdot\nu_{i_A} d\sigma dt \\ &+ \frac{1}{2}\sum_{S\in\mathcal{S}_{sing}}\sum_{j\in\mathcal{I}(S)}\int_{C_{Sj}}(m_j^S - m_j)\cdot\nu^j(\varphi')^2\, d\sigma dt.\end{aligned} \tag{6.5}$$

Integrating (6.3) on $(0,T)$ and summing the result with (6.5), we obtain the inequality (6.1). □

The right-hand side of (6.1) legitimates the next definition. For any $\{\varphi_0,\varphi_1\}\in D(A)\times V$, let $\varphi\in C([0,T],D(A))\cap C^1([0,T],V)$ be the solution of (2.7) and define

$$\begin{aligned}|||\{\varphi_0,\varphi_1\}|||^2 = &\sum_{A\in\mathcal{D}^+}\int_{\Sigma_{A\delta}}\left(\frac{\partial\varphi_{i_A}}{\partial\nu_{i_A}}\right)^2 d\sigma dt \\ &+ \sum_{A\in\mathcal{N}_{ext}^-}\int_{\Sigma_{A\delta}}\left(\frac{\partial\varphi_{i_A}}{\partial\tau_{i_A}}\right)^2 d\sigma dt \\ &+ \sum_{A\in\mathcal{N}_{ext}^+}\int_{\Sigma_{A\delta}}(D_t\varphi_{i_A})^2\, d\sigma dt \\ &+ \sum_{S\in\mathcal{S}_{sing}}\sum_{j\in\mathcal{I}(S)}\int_{C_{Sj}}\left\{(\varphi_j')^2 + \left(\frac{\partial\varphi_j}{\partial\nu^j}\right)^2 + \left(\frac{\partial\varphi_j}{\partial\tau^j}\right)^2\right\} d\sigma dt.\end{aligned} \tag{6.6}$$

The identity (3.16) and the estimate (6.1) directly lead to the next estimate of the energy.

PROPOSITION 6.2. *Let $\varphi \in C([0,T],D(A)) \cap C^1([0,T],V) \cap C^2([0,T],H)$ be a solution of* (2.7). *Then there exists a minimal time $T_0 > 0$ such that for all $T > T_0$ there exists a constant $C > 0$ (independent upon φ_0, φ_1) such that*

$$(T - T_0)E_0 \leq C|||\{\varphi_0, \varphi_1\}|||^2. \tag{6.7}$$

As in section 3, the inverse estimate (3.21) also holds owing to Theorem 3.5 of [20] and the usual trace theorems. We then define F as the closure of $D(A) \times V$ for the norm (6.6) (F clearly satisfies (3.22)). The analogue of Proposition 3.4 here is Proposition 6.3.

PROPOSITION 6.3. *Let $\{\varphi_0, \varphi_1\} \in F$ and $f \in L^1(0,T;V)$. Then the unique solution $\varphi \in C([0,T],V) \cap C^1([0,T],H)$ of* [20, *eq.* (3.1)] *satisfies* (3.23) *to* (3.25) *and*

$$\varphi_j', \frac{\partial \varphi_j}{\partial \nu^j}, \frac{\partial \varphi_j}{\partial \tau^j} \in L^2(C_{Sj})\, \forall j \in \mathcal{I}(S), S \in \mathcal{S}_{sing}.$$

Moreover, there exists a constant $C > 0$ (independent of $\{\varphi_0, \varphi_1\}$ and f) such that

$$\begin{aligned} &\left\{ \sum_{A \in \mathcal{D}^+} \int_{\Sigma_{A\delta}} \left(\frac{\partial \varphi_{i_A}}{\partial \nu_{i_A}} \right)^2 d\sigma dt + \sum_{A \in \mathcal{N}_{ext}^-} \int_{\Sigma_{A\delta}} \left(\frac{\partial \varphi_{i_A}}{\partial \tau_{i_A}} \right)^2 d\sigma dt \right. \\ &\quad + \sum_{A \in \mathcal{N}_{ext}^+} \int_{\Sigma_{A\delta}} (D_t \varphi_{i_A})^2 \, d\sigma dt \\ &\quad \left. + \sum_{S \in \mathcal{S}_{sing}} \sum_{j \in \mathcal{I}(S)} \int_{C_{Sj}} \left\{ (\varphi_j')^2 + \left(\frac{\partial \varphi_j}{\partial \nu^j} \right)^2 + \left(\frac{\partial \varphi_j}{\partial \tau^j} \right)^2 \right\} d\sigma dt \right\}^{1/2} \\ &\quad \leq C \left\{ |||\{\varphi_0, \varphi_1\}||| + \|f\|_{L^1(0,T;V)} \right\}. \end{aligned} \tag{6.8}$$

Proof. The proof is similar to that of Proposition 3.4, since in the left-hand side of (6.8) derivatives of $\varphi^{(2)}$ appear only far from the singular vertices, and by Theorem 2.2 of [20], they are square integrable. ∎

7. Weak solutions of the wave equation II. The transposition of Proposition 6.3 yields the following theorem.

THEOREM 7.1. *For all $u_0 \in H$, $u_1 \in V'$, $w_A \in L^2(\Sigma_{A\delta})$, where $A \in \mathcal{D}^+ \cup \mathcal{N}_{ext}^+ \cup \mathcal{N}_{ext}^-$ and all $w_{Sj}^k \in L^2(C_{Sj}), j \in \mathcal{I}(S), S \in \mathcal{S}_{sing}, k = 1,2,3$, there exist unique $u \in L^\infty(0,T;V')$, $\{\psi_1, \psi_0\} \in F'$, which satisfy*

$$\begin{aligned} &\int_0^T \langle u(t), f(t) \rangle_{V'-V} dt + \langle \{\psi_1, \psi_0\}, \{\varphi_0, -\varphi_1\} \rangle_{F'-F} \\ &= \langle u_1, \varphi(0) \rangle_{V'-V} - \langle u_0, \varphi'(0) \rangle_{H'-H} - \sum_{A \in \mathcal{D}^+} \int_{\Sigma_{A\delta}} w_A \frac{\partial \varphi_{i_A}}{\partial \nu_{i_A}} \, d\sigma dt \\ &\quad - \sum_{A \in \mathcal{N}_{ext}^-} \int_{\Sigma_{A\delta}} w_A \frac{\partial \varphi_{i_A}}{\partial \tau_{i_A}} \, d\sigma dt - \sum_{A \in \mathcal{N}_{ext}^+} \int_{\Sigma_{A\delta}} w_A D_t \varphi_{i_A} \, d\sigma dt \\ &\quad - \sum_{S \in \mathcal{S}_{sing}} \sum_{j \in \mathcal{I}(S)} \int_{C_{Sj}} \left\{ w_{Sj}^1 \varphi_j' + w_{Sj}^2 \frac{\partial \varphi_j}{\partial \nu^j} + w_{Sj}^3 \frac{\partial \varphi_j}{\partial \tau^j} \right\} d\sigma dt \end{aligned} \tag{7.1}$$

for all $f \in L^1(0,T;V)$, $\{\varphi_0,-\varphi_1\} \in F$, *where* $\varphi \in C([0,T],V) \cap C^1([0,T],H)$ *is the unique solution of* (4.2).

The interpretation of (7.1) in terms of partial differential equations is based on the following trace lifting result, which is easily proved (compare with [20, Lemma 5.3]).

LEMMA 7.2. *Let* $w_A \in \mathcal{D}(\Sigma_{A\delta}), A \in \mathcal{D}^+ \cup \mathcal{N}^+_{ext} \cup \mathcal{N}^-_{ext}$, *and* $w^k_{Sj} \in \mathcal{D}(C_{Sj}), j \in \mathcal{I}(S), S \in \mathcal{S}_{sing}, k = 1,2,3$. *Then there exists a function* v *defined on* Q *and equal to zero in a neighborhood of* $S \times [0,T]$ *for all* $S \in \mathcal{S}_{sing}$ *having the regularity*

$$\begin{cases} v_i \in \mathcal{D}\left(0,T,C^\infty\left(\overline{P_i \setminus \cup_{S\in\mathcal{S}_{sing}} B(S,\delta)}\right)\right) \ \forall i \in \mathcal{I}, \\ v_i \in \mathcal{D}\left(0,T,C^\infty\left(\overline{P_i \cap B(S,\delta)}\right)\right) \ \forall S \in \mathcal{S}_{sing},\ i \in \mathcal{I}(S) \end{cases}$$

and fulfilling (2.3), (2.4) *for all* $A \in \mathcal{N}_{int}$, *the boundary conditions* (4.4) *and* (4.11), *as well as the following interface conditions through* C_S *for all* $S \in \mathcal{S}_{sing}$:

$$\text{(7.2)} \qquad \begin{cases} v^S_j - v^{far}_j = \alpha_j^{-1} w^2_{Sj} \ \text{on } C_{Sj}\ \forall j \in \mathcal{I}(S), \\ \frac{\partial v^S_j}{\partial \nu^j} - \frac{\partial v^{far}_j}{\partial \nu^j} = \alpha_j^{-1}\left(D_t w^1_{Sj} + \frac{\partial w^3_{Sj}}{\partial \tau^j}\right) \ \text{on } C_{Sj}\ \forall j \in \mathcal{I}(S), \end{cases}$$

where v^S_j *(resp.,* v^{far}_j*) stands for the trace on* C_{Sj} *of the restriction of* v_j *to* $\{B(S,\delta) \cap P_j\} \times (0,T)$ *(resp.,* $\{P_j \setminus B(S,\delta)\} \times (0,T)$).

THEOREM 7.3. *Let* $u \in L^\infty(0,T;V')$, $\{\psi_1,\psi_0\} \in F'$ *be the unique solution of* (4.1) *with data* $u_0 \in V$, $u_1 \in H$, $w_A \in \mathcal{D}(\Sigma_{A\delta}), A \in \mathcal{D}^+ \cup \mathcal{N}^+_{ext} \cup \mathcal{N}^-_{ext}$, *and* $w^k_{Sj} \in \mathcal{D}(C_{Sj}), j \in \mathcal{I}(S), S \in \mathcal{S}_{sing}, k = 1,2,3$. *Then* $u \in C^1([0,T],H)$ *satisfies the boundary conditions* (4.4), *the initial conditions* (4.6), *the final conditions* (4.7), *and the hyperbolic equation*

$$\text{(7.3)} \qquad u_i'' - \alpha_i \triangle u_i = \sum_{S\in\mathcal{S}_{sing}} D_{Si} \ \text{in } \mathcal{D}'(Q_i)\ \forall i \in \mathcal{I},$$

where the distribution $D_{Si} \in \mathcal{D}'(Q_i)$ *is defined by*

$$\text{(7.4)} \qquad \langle D_{Si}, \eta\rangle = -\int_{C_{Si}} \left\{ w^1_{Si}\eta' + w^2_{Si}\frac{\partial \eta}{\partial \nu^i} + w^3_{Si}\frac{\partial \eta}{\partial \tau^i}\right\} d\sigma dt$$

if $i \in \mathcal{I}(S)$ *and* $D_{Si} = 0$ *else.*

Proof. Since it is similar to that of Theorem 4.3, we only explain the differences. From the function v obtained in Lemma 7.2, we define the function f on Q by

$$f_i = v_i'' - \alpha_i \triangle v_i \text{ on } Q_i \setminus \cup_{S\in\mathcal{S}_{sing}} C_{Si}\ \forall i \in \mathcal{I}.$$

Note that the Laplacian of v_i is computed outside $\cup_{S\in\mathcal{S}_{sing}} C_{Si}$ because v_i is not necessarily continuous through C_{Si} for $S \in \mathcal{S}_{sing}$ due to (7.2).

As $f \in L^2(0,T;H)$, there exists a unique solution $\psi \in C([0,T],V) \cap C^1([0,T],H) \cap H^2(0,T;V')$ of (4.8). The conclusion follows with $u = \psi + v$ by applying Green's formula on $P_i \setminus \cup_{S\in\mathcal{S}_{sing}} B(S,\delta)$ and on $P_i \cap B(S,\delta)$ for all $S \in \mathcal{S}_{sing}$, as usual.

Let us finally remark that we readily deduce (7.3) from (4.8) and the conditions (7.2) satisfied by v. ∎

As before, we then call the solution u of (7.1) the weak solution of (7.3), (4.6), (4.4), and (4.11). Moreover, from the density of $\mathcal{D}(V)$ in $L^2(V)$ for any open set V, the conclusion of Theorem 4.4 holds for our solution u.

8. Addition of circular internal controls. Theorem 2.2 is now a direct consequence of the HUM of Lions [12]. As Proposition 3.4 is replaced by Proposition 6.3, we then solve (5.1), the right-hand side being unchanged except for the last term, which is replaced by

$$-\sum_{S\in\mathcal{S}_{sing}}\sum_{j\in\mathcal{I}(S)}\int_{C_{Sj}}\left\{\varphi_j'\eta_j'+\frac{\partial\varphi_j}{\partial\nu^j}\frac{\partial\eta_j}{\partial\nu^j}+\frac{\partial\varphi_j}{\partial\tau^j}\frac{\partial\eta_j}{\partial\tau^j}\right\}d\sigma dt.$$

THEOREM 8.1. *For all $u_0\in H$, $u_1\in V'$, there exist $w_A\in L^2(\Sigma_{A\delta})$, $A\in\mathcal{D}^+\cup\mathcal{N}_{ext}^+\cup\mathcal{N}_{ext}^-$, and $w_{Sj}^k\in L^2(C_{Sj}), j\in\mathcal{I}(S), S\in\mathcal{S}_{sing}, k=1,2,3$, such that the weak solution $u\in C([0,T],V')\cap C^1([0,T],D(A)')$ of the wave equation* (8.1) *below* (*in the sense of* (7.1)) *satisfies $u(T)=u'(T)=0$.*

$$\left\{\begin{array}{l} u_i''-\alpha_i\triangle u_i=\sum_{S\in\mathcal{S}_{sing}}D_{Si}\ \text{in}\ \mathcal{D}'(Q_i)\,\forall i\in\mathcal{I},\\ u(0)=u_0,\qquad u'(0)=u_1,\\ u\ \text{satisfies}\ (4.4)\ \text{and}\ (4.11),\end{array}\right.\tag{8.1}$$

where D_{Si} is defined by (7.4).

Proof. Since $\{u_1,-u_0\}\in V'\times H\subset F'$, there exists a unique solution $\{\varphi_0,\varphi_1\}\in F$ of

$$\Lambda\{\varphi_0,\varphi_1\}=\{u_1,-u_0\}.$$

We take the solution φ of (2.7) and then the solution ψ of (5.1) (with the new right-hand side). Because of the time reversibility of the wave equation and Proposition 6.3, the conclusion follows with $u=\psi$, $w_A=\frac{\partial\varphi_{i_A}}{\partial\nu_{i_A}}$ for all $A\in\mathcal{D}^+$, $w_A=D_t\varphi_{i_A}$ for all $A\in\mathcal{N}_{ext}^+$, $w_A=\frac{\partial\varphi_{i_A}}{\partial\tau_{i_A}}$ for all $A\in\mathcal{N}_{ext}^-$, $w_{Si}^1=\varphi_i', w_{Si}^2=\frac{\partial\varphi_i}{\partial\nu^i}, w_{Si}^3=\frac{\partial\varphi_i}{\partial\tau^i}$ for all $S\in\mathcal{S}_{sing}$. □

Let us finally remark that with the above choice of the w_{Si}^k and (7.4), we arrive at the form (2.10) of D_{Si} given in Theorem 2.2. The proof of Theorem 2.2 is then complete.

Acknowledgments. We would like to thank Professor J. E. Lagnese for his interest in this work, as well as his valuable remarks.

REFERENCES

[1] P. G. CIARLET, H. LE DRET, AND R. NZENGWA, *Junctions between three-dimension and two-dimensional linearly elastic structures*, J. Math. Pures Appl., 68 (1989), pp. 261–295.

[2] C. CONCA AND E. ZUAZUA, *Asymptotic analysis of a multidimensional vibrating structure*, SIAM J. Math. Anal., 25 (1994), pp. 836–858.

[3] P. GRISVARD, *Elliptic Problems in Nonsmooth Domains*, Monographs and Studies in Mathematics 21, Pitman, Boston, MA, 1985.

[4] P. GRISVARD, *Contrôlabilité exacte des solutions de l'équation des ondes en présence de singularités,* J. Math. Pures Appl., 68 (1989), pp. 215–259.

[5] A. HEIBIG AND M. MOUSSAOUI, *Exact controllability of the wave equation for domains with slits and for mixed boundary conditions*, Discrete Cont. Dynam. Systems, 2 (1996), pp. 367–386.

[6] J. E. LAGNESE, *Controllability of systems of interconnected membranes*, Disc. Cont. Dynam. Syst., 1 (1995), pp. 17–33.

[7] J. E. LAGNESE, *Modelling and controllability of plate-beam systems*, J. Math. Syst., Estimation Control, 5 (1995), pp. 141–188.

[8] J. E. LAGNESE, G. LEUGERING, AND E. J. P. G. SCHMIDT, *Modeling of dynamic networks of thin thermoelastic beams*, Math. Methods Appl. Sci., 16 (1993), pp. 327–358.

[9] J. E. Lagnese, G. Leugering, and E. J. P. G. Schmidt, *Control of planar networks of Timoshenko beams*, SIAM J. Control Optim., 31 (1993), pp. 780–811.
[10] J. E. Lagnese, G. Leugering, and E. J. P. G. Schmidt, *Modeling, analysis and control of dynamic elastic multi-link structures*, Birkhäuser Boston, Cambridge, MA, 1994.
[11] H. Le Dret, *Problèmes variationnels dans les multi-domaines. Modélisation des jonctions et applications*, Recherches en Mathématiques Appliques 19, Masson, Paris, 1991.
[12] J.-L. Lions, *Contrôlabilité exacte, perturbations et stabilisation de systèmes distribués,* tome 1, Recherches en Mathématiques Appliques 8, Masson, Paris, 1988.
[13] M.-T. Niane and O. Seck, *Contrôlabilité exacte de l'équation des ondes avec conditions mêlées,* C. R. Acad. Sci. Paris, Sér. I, 318 (1994), pp. 945–948.
[14] M.-T. Niane and O. Seck, *Contrôlabilité exacte frontière de l'équation des ondes en présence de fissures,* C. R. Acad. Sci. Paris, Sér. I, 316 (1993), pp. 695–700.
[15] S. Nicaise, *Exact controllability of a pluridimensional coupled problem*, Rev. Mat. Univ. Complut. Madrid, 5 (1992), pp. 91–135.
[16] S. Nicaise, *About the Lamé system in a polygonal or a polyhedral domain and a coupled problem between the Lamé system and the plate equation* II*: Exact controllability*, Ann. Scuola Norm. Sup. Pisa 20 (1993), pp. 163–191.
[17] S. Nicaise, *Polygonal interface problems*, Methoden und Verfahren der Mathematischen Physik 39, Verlag Peter D. Lang, Frankfurt an Main, 1993.
[18] S. Nicaise, *Contrôlabilité exacte frontière des problèmes de transmission avec singularités*, C. R. Acad. Sci. Paris, Sér. I, 320 (1995), pp. 663–668.
[19] S. Nicaise, *Contrôlabilité exacte frontière de problèmes de transmission par adjonction de contrôles internes*, C. R. Acad. Sci. Paris, Sér. I, 321 (1995), pp. 969–974.
[20] S. Nicaise, *Boundary exact controllability of interface problems with singularities* I*: Addition of the coefficients of singularities*, SIAM J. Control Optim., 34 (1996), pp. 1512–1533.
[21] J. P. Puel and E. Zuazua, *Exact controllability for a model of multidimensional flexible structure*, Proc. Roy. Soc. Edinburgh, 123 A, (1993), pp. 323–344.
[22] E. J. P. G. Schmidt, *On the modelling and exact controllability of networks of vibrating strings*, SIAM J. Control Optim., 30 (1992), pp. 229–245.

SIAM J. CONTROL OPTIM.
Vol. 35, No. 2, pp. 604–625, March 1997

013

INFINITE-DIMENSIONAL LINEAR PROGRAMMING APPROACH TO SINGULAR STOCHASTIC CONTROL*

MICHAEL I. TAKSAR†

Abstract. We consider a multidimensional singular stochastic control problem with state-dependent diffusion matrix and drift vector and control cost depending on the position and direction of displacement of the controlled process. The objective is to minimize the total expected discounted cost. We write an equivalent infinite-dimensional linear programming problem on a subspace of the space conjugate to $\mathcal{C}(\mathbb{R}^n) \times \mathcal{C}(\mathbb{R}^n \times B)$, where B is the unit sphere in $\mathbb{R}^n$. We write a dual linear program and prove absence of duality gap. The dual program characterizes the optimal cost function as a maximal solution to the variational inequality with gradient constraints.

Key words. stochastic control, stochastic differential equations, controlled diffusion processes, primary and dual linear programs, variational inequalities

AMS subject classifications. Primary, 93E20; Secondary, 90C48

PII. S036301299528685X

1. Introduction.

One of the methods to study the optimal control problems is via mathematical programming on suitable spaces. Most probably the earliest result obtained by this approach was Pontryagin's maximum principle (see, e.g., [6]). Recently there was a renewed interest in applying linear and convex programming techniques to deterministic and stochastic optimal control problems (see [31], [17], [30], [8], [9], [16], [14], [15]). A certain duality was established between the problem of minimizing a cost functional over a set of all admissible controls and minimizing a linear or convex functional in the space of measures. The measures in question correspond to occupational measures of the controlled process and are described by linear or convex constraints. The original cost functional can then be represented as an integral with respect to the occupational measure. This gives a possibility to interpret the original control problem as a linear or convex programming problem in an infinite-dimensional space regardless of the structure of the cost function. In [9] convex programming methods were used to study finite horizon and infinite horizon control problems. Infinite-dimensional linear programming methods were employed in studying deterministic continuous-time control problems and discrete-time Markov decision processes with discounted cost (see [14], [15]). The main challenge in the linear programming approach is to show an absence of duality gap between the primal and the dual programs (*strong duality*) and establishing equality of values of the original optimal control problem and its linear programming counterpart.

In this paper we apply linear programming techniques to singular diffusion control. In singular control models the control is described by a functional ν of bounded variation rather than by a classical "input-output" process. We will study a general singular control model with dimension of the control functional being equal to that of the state process.

*Received by the editors May 30, 1995; accepted for publication (in revised form) February 15, 1996. This research was supported by NSF grant DMS 9301200 and NATO International Science Exchange Program grant CRG-900147.

http://www.siam.org/journals/sicon/35-2/28685.html

†Department of Applied Mathematics, State University of New York at Stony Brook, Stony Brook, NY 11794-3600 (taksar@ams.sunysb.edu).

In classical stochastic control models there is a set $\mathcal{U}$ of controls available at each state $x \in \mathbb{R}^n$. To relate the original control problem to linear or convex programming, one considers measures on $\mathbb{R}^n \times \mathcal{U}$ generated by the controlled process and identifies constraints which these measures must satisfy. If we had chosen to follow the same route for singular control problems then we would have had to make $\mathcal{U}$ the set of values of generalized functions. Defining measures on such a space is technically cumbersome, and thus it would be rather difficult to implement such an approach. Instead we consider a linear space X consisting of pairs (M, N), where M is a linear functional on $\mathcal{C}(\mathbb{R}^n)$ and N is a linear functional on $\mathcal{C}(\mathbb{R}^n \times B)$, where B is a unit sphere in $\mathbb{R}^n$. ($\mathcal{C}(\mathcal{X})$ stands for the set of bounded continuous functions on $\mathcal{X}$.) Any measure on $\mathbb{R}^n$ or $\mathbb{R}^n \times B$ is identified with one such functional. On the other hand, with each admissible control ν we associate two measures M^ν and N^ν on $\mathbb{R}^n$ and $\mathbb{R}^n \times B$, respectively. The measure M^ν is the occupational measure of the controlled process in the state space. To calculate N^ν, we replace real time t by $|\nu|(\cdot)$, where $|\nu|$ is the total variation of the process ν. The measure N^ν is the joint occupational measure of the controlled process and $\frac{d\nu}{d|\nu|}$. The generalized Ito formula yields linear constraints which $(M^\nu, N^\nu) \in X$ must satisfy.

We consider the linear program on X and its dual. We prove strong duality and we also show that the dual program is equivalent to finding the maximal solution to a variational inequality.

The structure of the paper is the following. In the next section we describe notations and outline the main assumptions. In section 3 we formulate the singular control problem, and in section 4 we formulate its linear programming counterpart. Section 5 is devoted to the proof of the main results of the paper: the absence of a duality gap and the equivalence of the values of the linear programming problem and the optimal control problem. In section 6 we summarize the principle results from the infinite-dimensional linear programming used in this paper. Most of the technical issues such as the proof of the generalized Ito formula, the existence of a smooth classical solution to the partial differential equations employed in the paper, and the like are resolved in section 7.

2. Notations and main assumptions. In this section we describe the parameters of the problem and introduce notations used in the paper. We denote by $\mathbb{R}^n$ the n-dimensional Euclidean space and by B a unit sphere in it. The exogenous parameters of the singular control problem are the following:

• a *diffusion matrix* $\sigma(x) = \|\sigma_{ij}(x)\|$ and a drift vector $b(x) = (b_1(x), \dots, b_n(x))$ associated with each $x \in \mathbb{R}^n$.

• a *discount factor* $\alpha > 0$.

• a *holding cost* function $h(x), x \in \mathbb{R}^n$. It corresponds to the rate of increase of the cost when the controlled process is at the point x.

• a *control cost function* $c(x, y), x \in \mathbb{R}^n, y \in B$, which corresponds to the cost of a unit displacement in the direction y when the controlled process is at the point x.

Put

$$a(x) \equiv \|a_{ij}(x)\| = \sigma(x)\sigma(x)^T.$$

The conditions we impose upon the parameters of the problem are the following. (Below K stands for a generic constant which can be different in different formulas. The norm of the matrix $\sigma(x)$ is $(tr[\sigma(x)\sigma(x)^T])^{1/2}$.)

(H1) There exists $K > 0$ such that for all $x, y \in \mathbb{R}^n$,

$$|b(x)| \le K. \tag{2.1}$$

(H2) There exist constants $C_1, C_2 > 0$ such that for each $\xi = (\xi_1, \dots, \xi_n) \in \mathbb{R}^n$ and each x,

$$C_1|\xi|^2 \le \sum_{ij} a_{ij}(x)\xi_i\xi_j \le C_2|\xi|^2. \tag{2.2}$$

(H3) There exist $\beta_1, \beta_2 \ge 0$ such that

$$\|\sigma(x) - \sigma(y)\| \le \beta_1|x - y|, \tag{2.3}$$

$$|b(x) - b(y)| \le \beta_2|x - y|, \tag{2.4}$$

and

$$\beta_1^2/2 + \beta_2 < \alpha. \tag{2.5}$$

(H4) $h(x) \ge 0$ for any $x \in \mathbb{R}^n$. There exist $K > 0$ and $\delta > 0$ such that for all $|x - y| < \delta$,

$$|h(x) - h(y)| \le K|x - y|h(x). \tag{2.6}$$

(H5) There exist constants $\lambda_1, \lambda_2 > 0$, and $K > 0$ such that

$$\lambda_1 \le c(\cdot, \cdot) \le \lambda_2, \tag{2.7}$$

$$|c(x_1, y_1) - c(x_2, y_2)| \le K(|x_1 - x_2| + |y_1 - y_2|). \tag{2.8}$$

The following notations will be used throughout the paper.

$$L = \frac{1}{2}\sum_{ij} a_{ij}(x)\frac{\partial^2}{\partial x_i \partial x_j} + \sum_i b_i(x)\frac{\partial}{\partial x_i} - \alpha. \tag{2.9}$$

If $\mathcal{X}$ is a metric space, then $\mathcal{C}(\mathcal{X})$ stands for the Banach space of bounded continuous functions f on $\mathcal{X}$ with

$$\|f\| = \sup_{x \in \mathcal{X}} |f(x)|.$$

The set of k times continuously differentiable functions on $\mathcal{X}$ with bounded derivatives is denoted by $\mathcal{C}^k(\mathcal{X})$. If g is a continuous strictly positive function on $\mathcal{X}$, then we set

$$\|f\|_g = \sup_{x \in \mathcal{X}} |f(x)|/g(x).$$

The linear space of continuous functions with norm $\|\cdot\|_g$ is denoted by $\mathcal{C}_g(\mathcal{X})$. In the sequel $\mathcal{X}$ will be either $\mathbb{R}^n$ or $\mathbb{R}^n \times B$. Define

$$\mathcal{C}^1(\mathbb{R}^n) = \left\{ f : \frac{\partial f}{\partial x_i} \in \mathcal{C}(\mathbb{R}^n) \forall i \le n \right\},$$

$$\mathcal{C}_g^1(\mathbb{R}^n) = \mathcal{C}_g(\mathbb{R}^n) \cap \mathcal{C}^1(\mathbb{R}^n),$$

$$\mathcal{C}_g^2(\mathbb{R}^n) = \left\{ f \in \mathcal{C}_g^1(\mathbb{R}^n) : \frac{\partial^2 f}{\partial x_i \partial x_j} \in \mathcal{C}_g(\mathbb{R}^n) \forall i, j \le n \right\}.$$

Note that in $\mathcal{C}_g^1(\mathbb{R}^n)$ and $\mathcal{C}_g^2(\mathbb{R}^n)$ we require $\frac{\partial f}{\partial x_i}$ to be bounded while f and its second derivatives are bounded only with respect to the function g. The space $\mathcal{C}_g^2(\mathbb{R}^n)$ is a Banach space with the norm

$$\|f\|_{2,g} = \|f\|_g + \sum_i \left\| \frac{\partial f}{\partial x_i} \right\| + \sum_{i,j} \left\| \frac{\partial^2 f}{\partial x_i \partial x_j} \right\|_g,$$

as are $\mathcal{C}^1(\mathbb{R}^n)$ and $\mathcal{C}_g^1(\mathbb{R}^n)$.

If X is a generic Banach space then we denote by X^* its conjugate, i.e., the set of all bounded continuous functionals on X. Put

$$\mathbb{L}(\mathcal{X}) = \mathcal{C}(\mathcal{X})^*, \mathbb{L}_g(\mathcal{X}) = \mathcal{C}_g(\mathcal{X})^*.$$

Define $\mathcal{M}(\mathcal{X})$ to be the set of all signed measures on $\mathcal{X}$ with finite variation. We use the notation $\langle M, f \rangle$ for the integral of f with respect to M. Let

$$\mathcal{M}_g(\mathcal{X}) = \{M \in \mathcal{M} : |\langle M, g \rangle| < \infty\}.$$

We will identify measures with the functionals they generate on the space of bounded continuous functions. Thus $\mathcal{M}(\mathcal{X}) \subset \mathbb{L}(\mathcal{X})$ and $\mathcal{M}_g(\mathcal{X}) \subset \mathbb{L}_g(\mathcal{X})$.

For a linear space X we denote by X^+ the closed cone of its positive elements. For example, $\mathcal{C}_g(X)^+$ is the set of all nonnegative continuous functions on $\mathcal{X}$ with finite $\|\cdot\|_g$ norm. The set $\mathcal{M}(\mathcal{X})^+$ would be the set of all bounded (nonnegative) measures on $\mathcal{X}$, etc.

If $f \in \mathcal{C}^1(\mathbb{R}^n)$ then $(\nabla f, \cdot)$ is a function on $\mathcal{C}(\mathbb{R}^n, B)$ defined as

$$(\nabla f, \cdot)(x, y) = \frac{\partial f(x)}{\partial y} \equiv \sum_{i=1}^n \frac{\partial f(x)}{\partial x_i} y_i, \qquad x \in \mathbb{R}^n,\ y = (y_1, \dots, y_n) \in B. \tag{2.10}$$

The rest of the section is devoted to the description and properties of vector-valued functions of bounded variation.

A deterministic vector-valued function $l(t) \in \mathbb{R}^n, 0 \le t < \infty$ is called *cadlag* if it is right continuous and has left limits. For any cadlag function l we put $l(0-) = 0$. Let $\|\cdot\|$ be any Minkowski norm in $\mathbb{R}^n$. Set

$$\|l\|(t) = \lim \sum_{i=1}^k \|l(t_i) - l(t_{i_1} -)\|, \tag{2.11}$$

where the limit in (2.11) is taken over divisions $0 = t_0 < t_1 < \cdots < t_k = t$ such that $\max_i |t_i - t_{i-1}| \to 0$. When $\|\cdot\|$ coincides with the ordinary Euclidean metric, (2.11) becomes the definition of the total variation $|l|(t)$ of the function l on $[0, t]$. If $|l|(t) < \infty$ for all $t \le \infty$, then we say that l is a function of *bounded variation*. Since all norms in a finite-dimensional space are equivalent, $\|l\|(t) < \infty$ for any Minkowski norm $\|\cdot\|$ whenever l is a function of bounded variation.

If $f(s)$ is a measurable function and l is a (real or vector-valued) function of bounded variation then it is possible to define a Lebesgue–Stieltjes integral

$$\int_0^t f(s) dl(s).$$

Following the usual convention, we interpret this integral as an integral over $[0-, t]$ with respect to the (vector-valued) measure with distribution function l. Thus when $l(0) \neq 0$ a point mass equal to $l(0)$ is concentrated at $t = 0$.

Set

$$\begin{aligned} \Delta l(s) &= l(s) - l(s-), \\ \Lambda(l) &= \{s : \Delta l(s) \neq 0\}, \\ l^d(t) &= \sum_{s\in\Lambda(l), s\leq t} \Delta l(s), \\ l^c(t) &= l(t) - l^d(t). \end{aligned} \tag{2.12}$$

Obviously, $l^c(t)$ is a continuous process with $l(0) = 0$ and $|l^c|(t) \leq |l|(t)$.

For any function l of bounded variation there exists a vector function $\chi_l(s)$ such that $|\chi_l(s)| = 1$ for all s and

$$l(t) = \int_0^t \chi_l(s) d|l|(s). \tag{2.13}$$

We will use notations $dl(s)/d|l|(s)$ and $\chi_l(s)$ interchangeably. If $s \in \Lambda(l)$ then one can verify that

$$\frac{dl}{d|l|}(s) = \frac{\Delta l(s)}{|\Delta l(s)|}.$$

The vector function $\chi_l(\cdot)$ is unique up to a measure $d|l|(\cdot)$ on the real line. It is also easy to see that if l is a discontinuous functional then it has at most countable number of discontinuities. Thus $\chi_l(\cdot) = \chi_{l^c}(\cdot)$ almost everywhere (a.e.) $d|l^c|(\cdot)$ and

$$l^c(t) = \int_0^t \chi_l(s) d|l^c|(s) = \int_0^t \chi_{l_c}(s) d|l^c|(s). \tag{2.14}$$

3. Formulation of the singular control problem. We start with a probability space $(\Omega, \mathcal{F}, \mathcal{F}_t, P)$ and a standard n-dimensional Brownian motion $w(t)$ on it. An n-dimensional cadlag process $\nu(t)$ is called a *control* if $\nu(t)$ is $\mathcal{F}_t$-measurable for each $t \geq 0$. Given control ν and an initial position $x \in \mathbb{R}^n$ we define the state process $x(\cdot)$ as the solution of the following stochastic differential equation:

$$x(t) = x + \int_0^t \sigma(x(s)) dw(s) + \int_0^t b(x(s)) ds + \nu(t). \tag{3.1}$$

Note that according to our convention $x(0-) = x$ and $x(0) = x + \nu(0)$.

The control ν is called *admissible at point* x if (3.1) has a unique solution and

$$E\left\{\int_0^\infty e^{-\alpha t} d|\nu|(t)\right\} < \infty. \tag{3.2}$$

Obviously (3.2) implies that ν is a process of bounded variation. We denote by $\mathcal{A}(x)$ the set of all controls admissible at x. Given the initial position x and an admissible control ν, we define the cost functional

$$\begin{aligned} J_x(\nu) = E\Bigg\{&\int_0^\infty e^{-\alpha t} h(x(t)) dt + \int_0^\infty e^{-\alpha t} c(x(t), \chi_\nu(t)) d|\nu^c|(t) \\ &+ \sum_{s\in\Lambda(\nu)} e^{-\alpha s} \int_0^{|\Delta\nu(s)|} c(x(s-) + z\chi_\nu(s), \chi_\nu(s)) dz \Bigg\}. \end{aligned} \tag{3.3}$$

The first integral on the right-hand side of (3.3) is called the *holding cost*. The remaining two terms are called the *control cost*. The objective is to find

$$v(x) = \inf_{\nu \in \mathcal{A}(x)} J_x(\nu). \tag{3.4}$$

The usual framework leading to singular control problems appears in the models where the "input" (the control functional) is additive and the control cost is linear. If we consider a particle whose position is represented by the state process of our system then in absence of control the particle's dynamics are governed by a diffusion process. The control (input) corresponds to the displacement of the particle from the original trajectory with the cost proportional to the distance the particle is forced to move due to the exerted control. The coefficient of proportionality, however, may depend on the direction of the displacement as well as the position of the particle. On the right-hand side of (3.3) the "quantity" $d|\nu^c|(s)$ in the second integral corresponds to the length of displacement of the particle while it moves continuously and $\chi_\nu(s) = \frac{d\nu}{d|\nu|}(s)$ is the direction of displacement. The quantity $c(x(s), \chi_\nu(s))$ is the unit cost of moving in the direction $\chi_\nu(s)$ from the point $x(s)$. To understand the last term in (3.3), consider any $s \in \Lambda(\nu)$. This is the moment when the functional $\nu(\cdot)$ as well as the process $x(\cdot)$ are discontinuous. It corresponds to an instantaneous jump of the particle from $x(s-)$ in the direction $\Delta\nu(s)/|\Delta\nu(s)| \equiv \frac{d\nu}{d|\nu|}(s)$ to the position $x(s)$. At this moment of time we assume that the real time stops and the "internal clock" is turned on. During the "internal" time the particle travels the distance $|\nu(s) - \nu(s-)|$ with a constant unit velocity, thus moving from $x(s-)$ to $x(s)$. After z units of "internal" time it reaches the position $x(s-) + z\chi_\nu(s)$. The cost incurred during the time interval $[z, z + dz]$ is then equal to $c(x(s-) + z\chi_\nu(s), \chi_\nu(s))dz$. Integrating this cost yields the last term in (3.3).

In one of the classical situations of an "additive input–linear control cost" problem, the dynamics of the state process are described by

$$dx(t) = \sigma(x(t))dw(t) + (b(x(t)) + u(t))dt,$$

where $u(t)$ is a control functional with values restricted to a compact set. The control cost is given by

$$\int_0^\infty e^{-\alpha t}|u(t)|c(x(t), u(t)/|u(t)|)dt.$$

Suppose that in (3.1) we have a sequence of admissible controls $\nu^n(t)$ which are absolutely continuous with derivatives $d\nu^n(t)/dt = u^n(t)$. Suppose that $x^n(\cdot)$ and $\nu^n(\cdot)$ converge to a stochastic process $x(\cdot)$ and a bounded variation process $\nu(\cdot)$, respectively, in M_1-Skorohod topology. (In this case, M_1-convergence of ν^n to ν corresponds to the convergence of $\nu^n(t)$ to $\nu(t)$ at each point t at which $\nu(t)$ is continuous.) Then one can easily verify that

$$\begin{aligned}\int_0^\infty e^{-\alpha t}|u^n(t)|c(x^n(t), u^n(t)/|u^n(t)|)dt \to \int_0^\infty e^{-\alpha t}c(x(t), \chi_\nu(t))d|\nu^c|(t)\\ + \sum_{s\in\Lambda(\nu)} e^{-\lambda s}\int_0^{|\Delta\nu(s)|} c(x(s-) + z\chi_\nu(s), \chi_\nu(s))dz.\end{aligned} \tag{3.5}$$

This gives another justification for the definition of the control cost in (3.3).

When $c(x,y)$ does not depend on $x \in \mathbb{R}^n (= c(y))$, the expression for the control cost in (3.3) reduces to

$$\int_0^\infty e^{-\alpha t} c(\chi_\nu(t)) d|\nu|(t). \tag{3.6}$$

The most frequently considered case is the one in which $c(\cdot) \equiv \text{const}$ and the holding cost function is convex (e.g., [13], [18], [19], [25], [24], [27], [28], [29]). If $c(y) = \|y\|$, where $\|\cdot\|$ is any Minkowski norm in $\mathbb{R}^n$, then (3.5) becomes

$$\int_0^\infty e^{-\alpha t} d\|\nu\|(t), \tag{3.7}$$

where $\|\nu\|(t)$ is defined by (2.11). The case of $\|(y_1,\dots,y_n)\| = \sum_i (a_i y_i^+ + b_i y_i^-), a_i, b_i > 0$, was studied in [25]. It also appears in diffusion approximation of controlled queues (see [32], [33]).

4. Formulation of the linear programming problem. In this section we formulate linear program (P) and its dual (P*) related to the original singular control problem.

We start by associating two measures $M^\nu \in \mathcal{M}(\mathbb{R}^n)^+$ and $N^\nu \in \mathcal{M}(\mathbb{R}^n \times B)^+$ with each admissible control ν.

$$M^\nu(\Gamma) = E\left\{\int_0^\infty e^{-\alpha t} 1_\Gamma(x(t)) dt\right\}, \quad \Gamma \subset \mathbb{R}^n.$$

$$N^\nu(\Gamma) = E\left\{\int_0^\infty e^{-\alpha t} 1_\Gamma(x(t), \chi_\nu(t)) d|\nu^c|(t) \right. \tag{4.1}$$

$$\left. + \sum_{s\in\Lambda(\nu)} e^{-\alpha s} \int_0^{|\Delta\nu(s)|} 1_\Gamma(x(s-) + z\chi_\nu(s), \chi_\nu(s)) dz \right\}, \qquad \Gamma \subset \mathbb{R}^n \times B.$$

Let $g(x), x \in \mathbb{R}^n$ be a smooth convex function such that

$$\begin{aligned} 1 + |x| \le g(x) \le 2 + |x|, &\qquad x \in \mathbb{R}^n, \\ g(x) = 1 + |x|, &\qquad |x| \ge 1. \end{aligned} \tag{4.2}$$

PROPOSITION 4.1. *For each admissible control ν,*

$$M^\nu \in \mathcal{M}_g(\mathbb{R}^n)^+.$$

The proof of this proposition is given in section 7.

It follows from (3.3) that

$$J_x(\nu) = \langle M^\nu, h\rangle + \langle N^\nu, c\rangle.$$

Expression (4.1) shows that for any admissible control ν, with $J_x(\nu) < \infty$, we have $M^\nu \in \mathcal{M}_h(\mathbb{R}^n)$. Together with Proposition 4.1 this yields

$$M^\nu \in \mathcal{M}_{\tilde g}(\mathbb{R}^n), \tag{4.3}$$

where $\tilde g = g + h$.

To understand the constraints the measures M^ν and N^ν must satisfy, consider $f \in \mathcal{C}_g^2(\mathbb{R}^n)$ and apply Proposition 7.1 to it (below (x, y) stands for the inner product of vectors $x, y \in \mathbb{R}^n$).

$$-f(x) = E\left\{\int_0^\infty e^{-\alpha t} Lf(x(t))dt + \int_0^\infty e^{-\alpha t}\nabla f(x(t))d\nu^c(t) + \sum_{s\in\Lambda(\nu)} e^{-\alpha s}[f(x(s)) - f(x(s-))]\right\} \equiv$$

$$E\left\{\int_0^\infty e^{-\alpha t} Lf(x(t))dt + \int_0^\infty e^{-\alpha t}(\nabla f(x(t)), \chi_\nu(t))d|\nu^c|(t) + \sum_{s\in\Lambda(\nu)} \int_0^{|\Delta\nu(s)|} (\nabla f(x(s-) + z\chi_\nu(s)), \chi_\nu(s))dz\right\}$$

$$= M^\nu(Lf) + N^\nu((\nabla f, \cdot)). \tag{4.4}$$

We introduce a dual pair (X, Y),

$$X = \{(M, N) : \mathcal{M} \in \mathbb{L}_{\tilde{g}}(\mathbb{R}^n), N \in \mathbb{L}(\mathbb{R}^n \times B)\},$$

$$Y = \{(f, j) : f \in \mathcal{C}_{\tilde{g}}(\mathbb{R}^n), j \in \mathcal{C}(\mathbb{R}^n \times B)\}.$$

Define

$$\langle (M, N), (f, j)\rangle = \langle M, f\rangle + \langle N, j\rangle.$$

To write the linear program (P), we need another dual pair (Z, W):

$$W = \mathcal{C}_g^2(\mathbb{R}^n),$$

$$Z = \mathcal{C}_g^2(\mathbb{R}^n)^*.$$

Note that Z contains all the functionals of the form δ_x such that

$$\langle \delta_x, f\rangle = f(x).$$

Let $\mathcal{L} : W \to Y$ be defined as follows

$$\mathcal{L}(f) = (-Lf, -(\nabla f, \cdot)).$$

In a more detailed way $\mathcal{L}f(x) = (F(x), G(x, y))$, where

$$F(x) = -\frac{1}{2}\sum_{i,j=1}^n a_{ij}(x)\frac{\partial^2 f(x)}{\partial x_i \partial x_j} - \sum_{i=1}^n b_i(x)\frac{\partial f(x)}{\partial x_i} + \alpha f(x),$$

$$G(x, y) = -\sum_{i=1}^n \frac{\partial f(x)}{\partial x_i} y_i, \qquad x \in \mathbb{R}^n,\ y = (y_1, \ldots, y_n) \in B.$$

Since $\tilde{g} \geq g$ we have $\mathcal{C}_g(\mathbb{R}^n) \subseteq \mathcal{C}_{\tilde{g}}(\mathbb{R}^n)$; thus the operator $\mathcal{L} : W \to Y$ is well defined for all $f \in W$. Having a dual pair (X, Y), we will always consider the weak topology $\sigma(X, Y)$ on X (see [1] or [26]). The same applies for (Z, W).

For each $\tilde{M} = (M, N) \in X$ consider the following linear functional on W:

$$T_{\tilde{M}}(f) = \langle \tilde{M}, \mathcal{L}f \rangle \equiv -\langle M, Lf \rangle - \langle N, (\nabla f, \cdot) \rangle. \tag{4.5}$$

It follows from (2.1), (2.2), and the definition of $\mathcal{C}_g^2(\mathbb{R}^n)$ that $\mathcal{L}$ is a bounded linear operator from $\mathcal{C}_g^2(\mathbb{R}^n)$ into $\mathcal{C}_{\tilde{g}}(\mathbb{R}^n)$. Since $\tilde{M}$ is a bounded functional on $\mathcal{C}_{\tilde{g}}(\mathbb{R}^n)$ and $\tilde{g} \geq g$, we conclude that $\tilde{M}$ is a bounded functional on $\mathcal{C}_g(\mathbb{R}^n)$ as well. Thus (4.5) implies that $T_{\tilde{M}}$ is a bounded functional on $\mathcal{C}_g^2(\mathbb{R}^n)$ and there exists $\tilde{z} \in Z$ such that $T_{\tilde{M}}(f) = \langle \tilde{z}, f \rangle$. We denote $\tilde{z} = \mathcal{L}^* \tilde{M}$. Standard arguments show that $\mathcal{L}^*$ is a continuous linear map. Let δ_x be a unit measure concentrated at the point x. Formula (4.4) and the definition of the operator $\mathcal{L}^*$ imply that for any admissible control ν and any $f \in \mathcal{C}_g^2(\mathbb{R}^n)$

$$\langle \delta_x, f \rangle = -\langle M^\nu, Lf \rangle - \langle N^\nu, (\nabla f, \cdot) \rangle \equiv \langle \tilde{M}^\nu, \mathcal{L}f \rangle \equiv \langle \mathcal{L}^* \tilde{M}^\nu, f \rangle. \tag{4.6}$$

Relation (4.6) provides linear constraints which all (M^ν, N^ν) must satisfy. This enables us to formulate the linear program. Put $\tilde{h} = (h, c)$ below.

(P) *Minimize* $\langle \tilde{M}, \tilde{h} \rangle \equiv \langle M, h \rangle + \langle N, c \rangle$ *subject to*

$$\tilde{M} \in X^+,$$

$$\mathcal{L}^* \tilde{M} = \delta_x.$$

The dual program (see section 6) is

(P*) *Maximize* $\delta_x(\phi) \equiv \phi(x)$ *subject to*

$$\phi \in W,$$

$$\mathcal{L}\phi \leq \tilde{h}.$$

The dual program can be rewritten in a more conventional way:
Maximize $\phi(x)$ *subject to* $\phi(x) \in \mathcal{C}_g^2(\mathbb{R}^n)$ *and*

$$\frac{1}{2} \sum_{i,j} a_{ij}(x) \frac{\partial^2 \phi(x)}{\partial x_i \partial x_j} + \sum_i b_i(x) \frac{\partial \phi(x)}{\partial x_i} - \alpha \phi(x) + h(x) \geq 0 \quad \forall x \in \mathbb{R}^n, \tag{4.7}$$

$$(\nabla \phi(x), y) + c(x, y) \geq 0 \quad \forall x \in \mathbb{R}^n \forall y \in B. \tag{4.8}$$

If $c(x, y) = \|y\|$, then (4.8) is equivalent to

$$\|\nabla \phi(x)\|^* \leq 1, \tag{4.9}$$

where $\|\cdot\|^*$ is the norm in $\mathbb{R}^n$ dual to the Minkowski norm $\|\cdot\|$. When $c(x, y) \equiv c$, inequality (4.9) becomes a standard gradient constraint variational inequality

$$|\nabla \phi(x)| \leq c \tag{4.10}$$

(see [24], [29]).

5. Consistency and absence of duality gap. The program (P*) is consistent (see the definitions in section 6) since we can take $\phi \equiv 0$ in (4.7). According to

Proposition 7.2, for each x there exists a feasible control ν such that $J_x(\nu) < \infty$. The corresponding pair of measures $\tilde{M}^\nu = (M^\nu, N^\nu)$ is consistent with finite value. Since for any admissible ν, the measure $\tilde{M}^\nu$ satisfies (4.6), we deduce that the program (P) is consistent with finite value and

$$\inf(\mathrm{P}) \le v(x),$$

where $v(x)$ is given by (3.4). Therefore, by virtue of Proposition 6.1,

$$\sup(\mathrm{P})^* \le \inf(\mathrm{P}) \le v(x). \tag{5.1}$$

Our main results are given by the following theorems. They show that the inequalities in (5.1) are tight.

THEOREM 5.1. *There is no duality gap:*

$$\inf(\mathrm{P}) = \sup(\mathrm{P})^*.$$

THEOREM 5.2. *The linear program is consistent with the optimal control problem:*

$$v(x) = \inf(\mathrm{P}).$$

To prove Theorem 5.1, we need to show that the subset D of $Z \times \mathbb{R}$ defined below is closed (see Theorem 6.1).

$$D = \{(\mathcal{L}^*\tilde{M}, \langle \tilde{M}, \tilde{h}\rangle), \tilde{M} \ge 0\}.$$

Let $\gamma \in \mathcal{T}$ be a net and $\tilde{M}_\gamma = (M_\gamma, N_\gamma) \in X^+$ be a family of functionals such that there exist

$$\lim_{\gamma\in\mathcal{T}} \mathcal{L}^*\tilde{M}_\gamma = V, \tag{5.2}$$

$$\lim_{\gamma\in\mathcal{T}} \langle \tilde{M}_\gamma, \tilde{h}\rangle = d. \tag{5.3}$$

We need to show that there exists $\tilde{M} = (M, N)$ such that

$$V = \mathcal{L}^*\tilde{M}, \tag{5.4}$$

$$d = \langle \tilde{M}, \tilde{h}\rangle. \tag{5.5}$$

(We need to use nets in (5.2), (5.3) because the space Z is not separable and to prove closure of D one cannot consider only sequences. See [20] or [3, section I.7] for more details on nets and convergence.) Since M_γ, N_γ are nonnegative functionals and $h, c \ge 0$, relation (5.3) implies that $\langle M_\gamma, h\rangle$ and $\langle N_\gamma, c\rangle$ are bounded. Since $c(\cdot,\cdot) \ge \lambda_1 > 0$ (see (2.5)) we see that $\langle N_\gamma, F\rangle$ are bounded for any $F \in \mathcal{C}(\mathbb{R}^n \times B)^+$. Since the functionals N_γ are positive the latter implies that $\langle N_\gamma, F\rangle \equiv \langle N_\gamma, F^+\rangle - \langle N_\gamma, F^-\rangle$ is bounded for any $F \in \mathcal{C}(\mathbb{R}^n \times B)$. Therefore there exists $N \in \mathbb{L}(\mathbb{R}^n \times B)$ and a subnet $\gamma_1 \in \mathcal{T}_1$ such that

$$\lim_{\gamma_1\in\mathcal{T}_1} N_{\gamma_1} = N. \tag{5.6}$$

Using the definition of the topology in Z and the definition of the operator $\mathcal{L}^*$, we can rewrite (5.2) as

$$\lim_{\gamma\in\mathcal{T}} (-\langle M_\gamma, Lf\rangle - \langle N_\gamma, (\nabla f, \cdot)\rangle) = V(f) \tag{5.7}$$

for all $f \in \mathcal{C}_g^2(\mathbb{R}^n)$. By virtue of Proposition 7.3, there exists a function $f \in \mathcal{C}_g^2(\mathbb{R}^n)$ such that $-Lf = g$. Applying (5.7) to this f, we have

$$\lim_{\gamma\in\mathcal{T}}(\langle M_\gamma, g\rangle - \langle N_\gamma, (\nabla f, \cdot)\rangle) = V(f). \tag{5.8}$$

Since $\frac{\partial f}{\partial x_i} \in \mathcal{C}(\mathbb{R}^n)$, we have $(\nabla f, \cdot) \in \mathcal{C}(\mathbb{R}^n \times B)$. We have already established boundedness of $\langle N_\gamma, F\rangle$ for all $F \in \mathcal{C}(\mathbb{R}^n \times B)$; thus (5.8) implies boundedness of $\langle M_\gamma, g\rangle$. This and the boundedness of $\langle M_\gamma, h\rangle$ show that $\langle M_\gamma, h + g\rangle \equiv \langle M_\gamma, \tilde{g}\rangle$ is bounded as well. Using positivity of the functional M_γ, we conclude that $\langle M_\gamma, G\rangle$ is bounded for all $G \in \mathcal{C}_{\tilde{g}}(\mathbb{R}^n)$. Therefore there exists M and a subnet $\gamma_2 \in \mathcal{T}_2$ of the net $\gamma_1 \in \mathcal{T}_1$ such that

$$\lim_{\gamma_2\in\mathcal{T}_2} M_{\gamma_2} = M. \tag{5.9}$$

Combining (5.9) and (5.6), we conclude

$$\lim_{\gamma_2\in\mathcal{T}_2} \langle (M_{\gamma_2}, N_{\gamma_2}), (h, c)\rangle = \langle (M, N), (h, c)\rangle \tag{5.10}$$

and for any $f \in \mathcal{C}_g^2(\mathbb{R}^n)$

$$\lim_{\gamma_2\in\mathcal{T}_2} \langle (M_{\gamma_2}, N_{\gamma_2}), (-Lf, -(\nabla f, \cdot))\rangle = \langle (M, N), (-Lf, -(\nabla f, \cdot))\rangle. \tag{5.11}$$

Taking $\tilde{M} = (M, N)$, we see that (5.11) implies (5.4) and (5.10) implies (5.5). This proves that D is closed.

Proof of Theorem 5.2. Suppose $\inf(\mathrm{P}) < v(x)$. Then there exists $\tilde{M} = (M, N) \in X^+$ such that $\mathcal{L}^*\tilde{M} = \delta_x$ and

$$\langle M, h\rangle + \langle N, c\rangle < v(x). \tag{5.12}$$

By virtue of Proposition 7.4, there exists a family of functions $h^\epsilon(\cdot) \in \mathcal{C}(\mathbb{R}^n), v^\epsilon(\cdot) \in \mathcal{C}_g^2(\mathbb{R}^n), c^\epsilon(\cdot, \cdot) \in \mathcal{C}(\mathbb{R}^n \times B), \epsilon > 0$ subject to (7.31)–(7.35).

Applying $\mathcal{L}^*\tilde{M}$ to v^ϵ, we get

$$\begin{aligned} v^\epsilon(x) \equiv \delta_x(v^\epsilon) &= \langle \mathcal{L}^*\tilde{M}, v^\epsilon\rangle = \langle \tilde{M}, \mathcal{L}v^\epsilon\rangle \\ &= \langle (M, N), (-Lv^\epsilon, (-\nabla v^\epsilon, \cdot))\rangle = \langle M, -Lv^\epsilon\rangle + \langle N, (\nabla v^\epsilon, \cdot)\rangle \\ &\le \langle M, h^\epsilon\rangle + \langle N, c^\epsilon\rangle. \end{aligned} \tag{5.13}$$

The last inequality in (5.13) is due to (7.31) and (7.32). Letting $\epsilon \to 0$ in (5.13) and applying (7.33), (7.34), and (7.35), we obtain

$$v(x) \le \langle M, h\rangle + \langle N, c\rangle,$$

which contradicts (5.12).

6. Appendix I. Infinite-dimensional linear programming. In this section we present some well-known results from the theory of linear programming in infinite-dimensional spaces. Further details can be found in [1, Chapter 3].

DEFINITION. *Two linear vector spaces X and Y are called a* dual pair *with respect to a bilinear form $\langle\cdot,\cdot\rangle$ on $X \times Y$ if*

(i) *for each $x \in X, x \neq 0$ there exists $y \in Y$ such that $\langle x, y\rangle \neq 0$,*

(ii) *for each $y \in Y, y \neq 0$ there exists $x \in X$ such that $\langle x, y\rangle \neq 0$.*

The space Y is called dual *to X, and X is called dual to Y. If (X,Y) is a dual pair, then by $\sigma(X,Y)$ we denote the weakest topology on X in which $\langle x,y\rangle$ are continuous for all $y\in Y$. Similarly, the topology $\sigma(Y,X)$ is the weakest topology on Y in which $\langle x,y\rangle$ are continuous for all $x\in X$.*

If X^+ is a closed cone of positive elements in X, then the dual cone Y^+ is defined as the set of all $y\in Y$ such that $\langle x,y\rangle\geq 0$.

Let (X,Y) and (W,Z) be two dual pairs and $\mathcal{L}:X\to Z$ be a continuous linear map of X into Z. We define the adjoint map $\mathcal{L}^*:W\to Y$ via the relation

$$\langle \mathcal{L}x,w\rangle=\langle x,\mathcal{L}^*w\rangle \quad \textit{for all } x\in X, w\in W. \tag{6.1}$$

The mapping $\mathcal{L}^*$ is continuous (*see* [26, *section* II.6, *Proposition* 12]).

Consider the linear program

(P) *Minimize $\langle x,c\rangle$ subject to*

$$\mathcal{L}x=b,$$

$$x\in X^+,$$

where $b\in Z$ and $c\in Y$ are given vectors. The *dual* of (P) is

(P*) *Maximize $\langle b,w\rangle$ subject to*

$$w\in W$$

$$-\mathcal{L}^*w+c\in Y^+.$$

DEFINITION. *A linear program is* consistent with finite value (*or just* consistent) *if it has a feasible* (*i.e., satisfying the constraints*) *solution x. If* (P) (*respectively,* (P*)) *is consistent, then its* value *is defined as the infimum of $\langle x,c\rangle$* (*respectively, supremum of $\langle b,w\rangle$*) *over all feasible x* (*respectively, w*) *and is denoted by* inf(P) (*respectively,* sup(P*)).

PROPOSITION 6.1. *If* (P) *and* (P*) *are both consistent, then*

$$\sup(\mathrm{P}^*)\leq\inf(\mathrm{P}).$$

The proof of this proposition can be found in [1].

DEFINITION. *If both* (P) *and* (P*) *are consistent with finite values and*

$$\sup(\mathrm{P})=\inf(\mathrm{P}^*),$$

then it is said that there is no duality gap.

Conditions ensuring absence of a duality gap are given by the following theorem (see [1, Theorems 3.10 and 3.22]).

THEOREM 6.1. *Let D be the subset of $Z\times\mathbb{R}$ defined as*

$$D=\{(\mathcal{L}x,\langle x,c\rangle),x\in X^+\}.$$

If (P) *is consistent with finite value and the set D is closed, then there is no duality gap.*

7. Appendix II. Proofs of auxiliary results. This appendix is devoted to the proof of the technical results used in the previous sections. Below, K stands for a generic constant whose value might differ in different formulas.

Proof of Proposition 4.1. Let ν be an admissible control and $x(\cdot)$ be the solution of (3.1). To show that $\langle M^\nu, g\rangle < \infty$, we need to prove

$$E\left\{\int_0^\infty e^{-\alpha t}(1+|x(t)|)dt\right\} < \infty. \tag{7.1}$$

Obviously the expectation of the first term in the integrand in (7.1) is equal to $1/\alpha$. Let

$$\eta_t = \sup_{u\le t}\left|\int_0^u \sigma(x(s))dw(s)\right|,$$

$$\zeta_t = \sup_{u\le t}\left|\int_0^u b(x(s))ds\right|.$$

The dynamics equation (3.1) implies

$$|x(t)| \le \eta_t + \zeta_t + |\nu|(t).$$

Thus

$$E\left\{\int_0^\infty e^{-\alpha t}|x(t)|dt\right\} \le \int_0^\infty e^{-\alpha t}E\{\eta_t\}dt + \int_0^\infty e^{-\alpha t}E\{\zeta_t\}dt + E\left\{\int_0^\infty e^{-\alpha t}|\nu|(t)dt\right\}. \tag{7.2}$$

We can use (2.2) and Theorem I.9.2 of [23] to get

$$E\{\eta_t\} \le (E\{\eta_t^2\})^{1/2} \le E\left\{\sum_{i=1}^n \sup_{u\le t}\left(\int_0^u \sigma(x(s))dw(s)\right)_i^2\right\}^{1/2}$$

$$\le 2E\left\{\sum_{i=1}^n\left(\int_0^t \sigma(x(s))dw(s)\right)_i^2\right\}^{1/2} \le 2E\left\{\int_0^t \|\sigma(x(s))\|^2 ds\right\}^{1/2} \le Kt^{1/2}. \tag{7.3}$$

Inequality (2.1) implies

$$E\{|\zeta_t|\} \le E\left\{\int_0^t |b(x(s))|ds\right\} \le Kt. \tag{7.4}$$

Integrating by parts, we get

$$E\left\{\int_0^s e^{-\alpha t}d|\nu|(t)\right\} = E\{e^{-\alpha s}|\nu|(s)\} + E\left\{\int_0^s \alpha e^{-\alpha t}|\nu|(t)dt\right\}.$$

Letting $s\to\infty$,

$$E\left\{\int_0^\infty e^{-\alpha t}|\nu|(t)dt\right\} \le \alpha^{-1}E\left\{\int_0^\infty e^{-\alpha t}d|\nu|(t)\right\}. \tag{7.5}$$

The right-hand side of (7.5) is finite for any admissible control ν. Inequalities (7.2)–(7.5) yield (7.1).

PROPOSITION 7.1. *Let ν be an admissible control and $x(\cdot)$ be given by* (3.1) *and $f \in \mathcal{C}_g^2(\mathbb{R}^n)$. Then*

$$-f(x) = E\left\{\int_0^\infty e^{-\alpha t} Lf(x(t))dt + \int_0^\infty e^{-\alpha t}\nabla f(x(t))d\nu^c(t) + \sum_{s\in\Lambda(\nu)} e^{-\alpha s}[f(x(s)) - f(x(s-))]\right\}. \tag{7.6}$$

Proof. Using the generalized Ito's formula (see Theorem VIII.27 in [2]), we can write

$$\begin{aligned} e^{-\alpha T}f(x(T)) - f(x(0-)) &= \int_0^T e^{-\alpha t}\nabla f(x(t))dw(t) + \int_0^T e^{-\alpha t} Lf(x(t-))dt \\ &+ \int_0^T e^{-\alpha t}\nabla f(x(t-))d\nu(t) + \sum_{t\le T, t\in\Lambda(\nu)} e^{-\alpha t}[f(x(t)) - f(x(t-)) - \nabla f(x(t))]. \end{aligned} \tag{7.7}$$

Since $\Delta x(s) = \Delta\nu(s)$, we have

$$\begin{aligned} &\int_0^T e^{-\alpha t}\nabla f(x(t-))d\nu(t) - \sum_{t\le T} e^{-\alpha t}\nabla f(x(t-))\Delta x(t) \\ &= \int_0^T e^{-\alpha t}\nabla f(x(t-))d(\nu^c(t) + \nu^d(t)) - \sum_{t\le T} e^{-\alpha t}\nabla f(x(t-))\Delta\nu(t) \\ &= \int_0^T e^{-\alpha t}\nabla f(x(t-))d\nu^c(t) = \int_0^T e^{-\alpha t}\nabla f(x(t))d\nu^c(t). \end{aligned} \tag{7.8}$$

Since $\nabla f(\cdot)$ is bounded, the first term on the right-hand side of (7.7) is a square integrable martingale and its expectation vanishes. Thus, taking into account (7.8) and recalling that $x(0-) = x$,

$$e^{-\alpha T}E\{f(x(T)\} - f(x) = E\left\{\int_0^T e^{-\alpha t} Lf(x(t))dt + \int_0^T e^{-\alpha t}\nabla f(x(t))d\nu^c(t) + \sum_{s\in\Lambda(\nu), s\le T} e^{-\alpha s}[f(x(s)) - f(x(s-))]\right\}. \tag{7.9}$$

To complete the proof we need only to justify passing to a limit in (7.9) as $T \to \infty$. Since $f \in \mathcal{C}_g^2(\mathbb{R}^n)$, we have

$$f(x), |Lf(x(t))| \le K(1 + |x(t)|) \le K(1 + \eta_t + \zeta_t + |\nu|(t)) \tag{7.10}$$

(see the proof of Proposition 4.1). By virtue of (7.3) and (7.4), the expectation of the absolute value of the first term on the right-hand side of (7.9) does not exceed $\int_0^T e^{-\alpha t}(1 + t + t^{1/2})dt$. Therefore the expectation of the first term on the right-hand side of (7.9) converges to that of the first term on the right-hand side of (7.6) thanks

to the dominated convergence theorem. Due to the boundedness of ∇f, we can write

$$\begin{aligned}\left|\int_0^T e^{-\alpha t}\nabla f(x(t))d\nu^c(t)\right| + \left|\sum_{s\in\Lambda(\nu),s\le T} e^{-\alpha s}[f(x(s)) - f(x(s-))]\right| \\ \le \left|\int_0^T e^{-\alpha t}Kd\nu^c(t)\right| + \sum_{s\in\Lambda(\nu),s\le T} e^{-\alpha s}K|x(s) - x(s-)| \\ \le \int_0^T e^{-\alpha t}Kd|\nu^c|(t) + \sum_{s\in\Lambda(\nu),s\le T} e^{-\alpha s}K|\nu(x(s)) - \nu(x(s-))| \\ = \int_0^T e^{-\alpha t}Kd|\nu|(t). \qquad (7.11)\end{aligned}$$

Inequality (7.11) implies that the expectation of the last two terms on the right-hand side of (7.9) converges to that of the last two terms on the right-hand side of (7.6) thanks to (3.2) and the dominated convergence theorem.

In view of (7.10), (7.3), and (7.4)

$$E\{f(x(T))\} \le K(1 + T + T^{1/2}).$$

Therefore the second term on the left-hand side of (7.9) converges to 0 as $T \to \infty$ and we get (7.6).

PROPOSITION 7.2. *For each $x \in \mathbb{R}^n$ there exists a control ν such that*

$$J_x(\nu) < \infty.$$

Proof. Let D be a closed unit ball in $\mathbb{R}^n$ with the center at x. Consider the solution to the Skorohod problem in D for the stochastic differential equation with normal reflection at the boundary of D. This solution consists of a pair of continuous $\mathcal{F}_t$-adapted processes $(X(t), \nu(t))$ such that $\nu(t)$ is a process of bounded variation and

$$X(t) = x + \int_0^t \sigma(X(s))dw(s) + \int_0^t b(X(s))ds + \nu(t) \in D \text{ for all } t > 0, \qquad (7.12)$$

$$\int_0^\infty 1_{X(s)\ne\partial D} d|\nu|(s) = 0, \qquad (7.13)$$

$$\nu(t) = \int_0^t n(X(s))d|\nu|(s), \qquad (7.14)$$

where $n(y), y \in \partial D$ is a unit inward normal, i.e., $n(y) = x - y$. Existence of such a solution was proved in [22].

Let $f \in \mathcal{C}^2(D)$ be the solution to the following boundary value problem (see [21, section III], where the existence of such a solution is shown)

$$Lf(y) = 0, \qquad y \in D, \qquad (7.15)$$

$$\frac{\partial f}{\partial n}(y) = -1, \qquad y \in \partial D. \qquad (7.16)$$

Applying Ito's formula to $f(X(t))$ and using (7.15), (7.16) together with (7.13) and (7.14), we get

$$f(x) - e^{-\alpha T}E\{f(X(t))\} = E\left\{-\int_0^T Lf(X(t))dt - \int_0^T \nabla f(X(t))e^{-\alpha t}d\nu(t)\right\}$$

$$= E\left\{\int_0^T e^{-\alpha t}\nabla f(X(t))n(X(t))d|\nu|(t)\right\} = E\left\{-\int_0^T e^{-\alpha t}\frac{\partial f}{\partial n}(X(t))d|\nu|(t)\right\}$$

$$= E\left\{\int_0^T e^{-\alpha t}d|\nu|(t)\right\}. \tag{7.17}$$

In view of (7.12), the process $X(\cdot)$ belongs to a compact domain; hence $f(X(T))$ is uniformly bounded in $\mathbb{R}\times\Omega$. Taking limit as $T\to\infty$, we obtain

$$E\left\{\int_0^\infty e^{-\alpha t}d|\nu|(t)\right\} = f(x) < \infty. \tag{7.18}$$

Consider

$$J_x(\nu) = E\left\{\int_0^\infty e^{-\alpha t}h(X(t))dt\right\} + E\left\{\int_0^\infty e^{-\alpha t}c(X(t),\chi_\nu(t))d|\nu|(t)\right\}. \tag{7.19}$$

By virtue of (7.12), the process $h(X(t))$ is uniformly bounded. Thus the first term on the right-hand side of (7.19) is bounded. In view of (2.7) the second term on the right-hand side of (7.19) does not exceed $\lambda_2 E\{\int_0^\infty e^{-\alpha t}d|\nu|(t)\}$, which is finite thanks to (7.18).

PROPOSITION 7.3. *Let g be given by* (4.2). *Then there exists $f\in\mathcal{C}_g^2(\mathbb{R}^n)$ such that*

$$Lf = g. \tag{7.20}$$

Proof. (1) Let $(X(t),P_x)$ be the Markov diffusion process with the infinitesimal generator $L+\alpha$ (see [4], [5], [10]); i.e., for each $x\in\mathbb{R}^n$, the measure P_x is the distribution of the solution to the following stochastic differential equation:

$$X(t) = x + \int_0^t \sigma(X(s))dw(s) + \int_0^t b(X(s))ds. \tag{7.21}$$

Put

$$f(x) = E_x\left\{\int_0^\infty e^{-\alpha t}g(X(t))dt\right\}. \tag{7.22}$$

Since $|X(t)|\le x+\eta_t+\zeta_t$, estimates (7.3) and (7.4) show that $E_x\{g(X(t))\}\le x+\text{const}$, whereas

$$f(x)\le K(1+|x|). \tag{7.23}$$

Therefore $f\in\mathcal{C}_g(\mathbb{R}^n)$. Let D be the unit ball with the center at x and τ be the first hitting time of ∂D by $X(t)$. The strong Markov property of the process $X(t)$ implies

$$f(x) = E\left\{\int_0^\tau e^{-\alpha\tau}g(X(t))dt + e^{-\alpha\tau}f(X(\tau))\right\}.$$

In view of Theorem 5.1 of [11, section 6], the function f given by the above formula is the unique solution in D to the Dirichlet problem (7.20) with the boundary conditions on ∂D given by the same function f. The latter shows that f given by (7.22) satisfies (7.20) in the whole space $\mathbb{R}^n$.

(2) Let D_1, D_2 be two bounded open sets in $\mathbb{R}^n$ such that $\bar{D}_1 \subset D_2$. Then a priori Schauder's inner estimates (see [12, Chapter III]) yield

$$\sup_{x\in D_1}\left(|f(x)|+\sum_i\left|\frac{\partial f(x)}{\partial x_i}\right|+\sum_{i,j}\left|\frac{\partial^2 f(x)}{\partial x_i\partial x_j}\right|\right)\le c\sup_{x\in D_2}\left(|f(x)|+|g(x)|+\sum_i\left|\frac{\partial g(x)}{\partial x_i}\right|\right), \tag{7.24}$$

where c is a constant which depends only on $C_1, C_2, \beta_1, \beta_2$ in (2.2)–(2.4), on the upper bound for $|b|(\cdot)$ in (2.1), on the *diameter* of the set D_2, and on the *distance* between D_1 and D_2. Let D_1 be a ball of radius 1 with the center x and D_2 be a ball of radius 2 with the center x. Then, using (7.23) and (7.24), we deduce the existence of a constant K such that

$$\left(|f(x)|+\sum_i\left|\frac{\partial f(x)}{\partial x_i}\right|+\sum_{i,j}\left|\frac{\partial^2 f(x)}{\partial x_i\partial x_j}\right|\right)\le K(1+|x|). \tag{7.25}$$

(3) To complete the proof we need only to show that $\frac{\partial f}{\partial x_i}$ is bounded in $\mathbb{R}^n$. Assume that $\sigma(x)$ and $b(x)$ are differentiable with

$$\|D_x\sigma(x)\|^2/2+\|D_xb(x)\|\le\beta_1^2/2+\beta_2\equiv\beta<\alpha, \tag{7.26}$$

where $D_xb(x)$ is an $n\times n$ matrix with $(D_xb(x))_{ij}=\frac{\partial b_i(x)}{\partial x_j}$ and $\|D_xb(x)\|^2=\ \mathrm{tr}(D_xb(x)\cdot D_xb(x)^T)$ and with $D_x\sigma(x)$ being an $n\times n\times n$ tensor with

$$D_x\sigma(x)_{ijk}=\frac{\partial\sigma(x)_{ij}}{\partial x_k}$$

and

$$\|D_x\sigma(x)\|^2=\sum_{ijk}D_x\sigma(x)^2_{ijk}.$$

Using the results of section 5.5 of [11], we can show that

$$\frac{\partial f(x)}{\partial x_i}=E\left\{\int_0^\infty e^{-\alpha t}\nabla g(X(t))X^{(i)}(t)dt\right\}, \tag{7.27}$$

where $X^{(i)}(t)$ is the solution to the following stochastic differential equation

$$X^{(i)}(t)=e_i+\int_0^t D_x\sigma(X(s))X^{(i)}(s)dw(s)+\int_0^t D_xb(X(s))X^{(i)}(s)ds, \tag{7.28}$$

where $e_i\in\mathbb{R}^n$ is a vector with $(e_i)_j=\delta_{ij}$ and $X(\cdot)$ is the solution to (7.21). Taking $Z(t)=|X^{(i)}(t)|^2$ and using Ito's formula, we get

$$\begin{aligned}E\{Z(t)\}&=1+\int_0^t E\{2Z(s)D_xb(X(s))+D_x\sigma(X(s))X^{(i)}(s)(D_x\sigma(X(s))X^{(i)}(s))^T\}dt\\&\le 1+\sup_{x\in\mathbb{R}}(2\|D_xb(x)\|+\|D_x\sigma(x)\|^2)\int_0^\infty e^{-\alpha t}E\{Z(t)\}dt\\&\le 1+2\beta\int_0^\infty e^{-\alpha t}E\{Z(t)\}dt.\end{aligned} \tag{7.29}$$

Application of Gronwall's inequality to (7.29) yields

$$E\{|X^{(i)}(t)|^2\} \equiv E\{Z(t)\} \le e^{2\beta t}.$$

Thus $E\{|X^{(i)}(t)|\} \le e^{\beta t}$. Substituting the latter inequality into (7.27), we get

$$\left|\frac{\partial f(x)}{\partial x_i}\right| \le \sup_{y\in\mathbb{R}} |\nabla g(y)| \int_0^\infty e^{-\alpha t} E\{|X^{(i)}(t)|\}dt \le \sup_{y\in\mathbb{R}} ||\nabla g(y)||/(\alpha-\beta) \le K. \tag{7.30}$$

In the case of nondifferentiable $\sigma(\cdot)$ and $b(\cdot)$, we can approximate both functions uniformly by differentiable functions $\sigma_n(\cdot)$ and $b_n(\cdot)$ subject to (7.26). The corresponding cost functions $f_n(\cdot)$ given by (7.22) converge uniformly to $f(\cdot)$ and are subject to (7.30). The latter implies that the limiting function f satisfies (7.30) as well.

PROPOSITION 7.4. *Let v be given by* (3.4). *Then there exists a family of functions $h^\epsilon(\cdot) \in \mathcal{C}(\mathbb{R}^n), v^\epsilon(\cdot) \in \mathcal{C}_g^2(\mathbb{R}^n), c^\epsilon(\cdot,\cdot) \in \mathcal{C}(\mathbb{R}^n \times B), \epsilon > 0$, such that*

$$-Lv^\epsilon \le h^\epsilon, \tag{7.31}$$

$$-(\nabla v^\epsilon, \cdot) \le c^\epsilon, \tag{7.32}$$

$$h^\epsilon \le (1+\delta(\epsilon))h, \tag{7.33}$$

where $\delta(\epsilon) \to 0$ as $\epsilon \to 0$, and

$$\lim_{\epsilon\to 0} \|c - c^\epsilon\| = 0, \tag{7.34}$$

$$\lim_{\epsilon\to 0} \|v^\epsilon(x) - v(x)\| = 0. \tag{7.35}$$

Proof. (1) Assume that σ and b are continuously differentiable with bounded derivatives, satisfying (7.26). Let $(X(t), P_x)$ be the same Markov process as in part 1 of the proof to Proposition 7.3. Let $p(t,x,y), t>0, x, y \in \mathbb{R}^n$ be the transition density of this Markov process. Put

$$v^\epsilon(x) = E_x\{v(X(\epsilon)\} \equiv \int_{\mathbb{R}^n} p(\epsilon,x,y)v(y)dy, \tag{7.36}$$

$$h^\epsilon(x) = E_x\{h(X(\epsilon)\} \equiv \int_{\mathbb{R}^n} p(\epsilon,x,y)h(y)dy. \tag{7.37}$$

The weak dynamic programming principle (see [7]) implies that for any $s > 0$

$$\begin{aligned} v(y) &\le E_y\left\{\int_0^s e^{-\alpha t} h(X(t))dt\right\} + E_y\{e^{-\alpha s} v(X(s))\} \\ &\equiv \int_0^s e^{-\alpha t} E_y\{h(X(t))\}dt \;+\; e^{-\alpha s} E_y\{v(X(s))\}. \end{aligned}$$

Multiplying both sides of the above inequality by $p(\epsilon,x,y)$ and integrating we get

$$\int_{\mathbb{R}^n} p(\epsilon,x,y)v(y)dy$$

$$\le \int_0^s e^{-\alpha t}\left(\int_{\mathbb{R}^n} p(\epsilon,x,y)E_y\{h(X(t))\}dy\right)dt + e^{-\alpha s}\int_{\mathbb{R}^n} p(\epsilon,x,y)E_y\{v(X(t))\}dy$$

$$\equiv \int_0^s e^{-\alpha t} E_x\{h(X(t+\epsilon))\}dt + e^{-\alpha s} E_x\{v(X(t+\epsilon))\}$$

$$\equiv \int_0^s e^{-\alpha t} E_x\{h^\epsilon(X(t))\}dt + e^{-\alpha s} E_x\{v^\epsilon(X(s))\}. \tag{7.38}$$

The last two equalities in (7.38) are due to (7.36), (7.37) and the Chapman–Kolmogorov equation for the transition density p. It is known that the transition density p is the fundamental solution to the parabolic equation with the operator $-\partial/\partial t+(L+\alpha)$ and is twice continuously differentiable in x (see [11], [10], [4]). Therefore v^ϵ is also twice continuously differentiable. Subtracting $v^\epsilon(x)$ from both sides of inequality (7.38), we can apply Ito's formula for $e^{-\alpha s}v^\epsilon(X(s))$ and use (7.21) to get

$$E_x\left\{\int_0^s e^{-\alpha t}[Lv^\epsilon(X(t))+h^\epsilon(X(t))]dt\right\}\geq 0. \tag{7.39}$$

Dividing both parts of (7.39) by s and letting $s\to 0$, we use standard limiting arguments to conclude (7.31).

(2) Extend function $c(x,y)$ to $\mathbb{R}^n\times\mathbb{R}^n$ in a homogeneous way, putting

$$c(x,y)=c(x,y/|y|)|y|, x,y\in\mathbb{R}^n.$$

One can see that the thus extended function $c(\cdot,\cdot)$ remains Lipschitz.

Starting with the initial position $x\in\mathbb{R}^n$ and considering controls ν such that $\nu(0)=\epsilon y, y\in\mathbb{R}^n$, we get

$$v(x)\leq v(x+\epsilon y)+\int_0^\epsilon c(x+zy,y)dz.$$

Subtracting $v(x)$ from both sides and dividing by ϵ, we get

$$\limsup_{\epsilon\to 0}-(v(x+\epsilon y)-v(x))/\epsilon\leq c(x,y). \tag{7.40}$$

Inequalities (7.40) and (2.7) show that v is Lipschitz. Therefore $\nabla v(\cdot)$ exists almost everywhere in $\mathbb{R}^n$ and for almost all $x\in\mathbb{R}^n$ for all $y\in\mathbb{R}^n$:

$$-(\nabla v(x),y)\leq c(x,y). \tag{7.41}$$

Employing the same arguments as in section 5.5 of [11], we can show

$$(\nabla v^\epsilon(x),y)=E_x\{(\nabla v(X(\epsilon)),Y^y(\epsilon))\}, \tag{7.42}$$

where $Y^y(\epsilon)=(Y_1^y(\epsilon),\dots,Y_n^y(\epsilon)), Y_i^y(\epsilon)=(X^{(i)}(\epsilon),y)$, with the process $X^{(i)}(\cdot)$ satisfying the following stochastic differential equation (below δ_{ij} denotes the Kronecker's delta)

$$X_j^{(i)}(t)=\delta_{ij}+\int_0^t\sum_{k=1}^n\sum_{l=1}^n\frac{\partial\sigma_{jk}(X(s))}{\partial x_l}X_l^{(i)}dw_k(s)+\int_0^t\sum_{l=1}^n\frac{\partial b(X(s))}{\partial x_l}X_l^{(i)}(s)ds. \tag{7.43}$$

Put

$$c^\epsilon(x,y)=E_x\{c(X(\epsilon),Y^y(\epsilon))\}.$$

Inequality (7.41) yields

$$(\nabla v^\epsilon(x),y)+c^\epsilon(x,y)=E_x\{(\nabla v(X(\epsilon)),Y^y(\epsilon))+c(X(\epsilon),Y^y(\epsilon))\}\geq 0,$$

whereas (7.32) follows.

(3) Inequality (7.41) implies $v(x) \le K(1+|x|)$. Repeating the arguments of part 1 of the proof to Proposition 7.3, we get $v^\epsilon \le K(1+|x|)$. Since $v^\epsilon(x) = f(x,\epsilon)$, where $f(x,t)$ is the solution to the parabolic equation

$$-\frac{\partial f(x,t)}{\partial t} + (L+\alpha)f(x,t) = 0, \qquad x \in \mathbb{R}^n,\ 0 \le t \le \epsilon,$$

$$f(x,0) = v(x),$$

(see [10], [11]), we can derive that v^ϵ is twice continuously differentiable (see [12]). Applying Schauder-type a priori estimates for the solution of the parabolic equations (see [12, section III]), we conclude that v^ϵ, its first and second derivatives, belong to $\mathcal{C}_g(\mathbb{R}^n)$. To prove $v^\epsilon \in \mathcal{C}_g^2(\mathbb{R}^n)$, we need to show boundedness of ∇v^ϵ in $\mathbb{R}^n$.

Similar to the last section of the previous proposition, we obtain

$$E\{|Y^y(\epsilon)|\} \le e^{\beta\epsilon}, \tag{7.44}$$

where β is the same as in (7.26). Therefore,

$$|\nabla v^\epsilon(x)| \le \lambda_2 e^{\beta\epsilon},$$

thanks to (2.7), (7.41), (7.42), and (7.44).

(4) It is shown in section I.6 of [12] that there exist constants $k > 0$ and $k_1 > 0$ such that

$$p(t,x,y) \le k_1 t^{-n/2} \exp(-k(x-y)^2/t). \tag{7.45}$$

In [14] it was shown that every function h subject to (2.6) satisfies

$$h(y)/h(x) \le \exp(K|x-y|). \tag{7.46}$$

Combining (7.45) and (7.46), we see

$$h^\epsilon(x)/h(x) = \int_{\mathbb{R}^n} p(\epsilon,x,y)h(y)/h(x)dy$$

$$\le \int_{|z|\le\epsilon} p(\epsilon,x,x+z)\exp(Kz)dz + \int_{|z|>\epsilon} k_1\epsilon^{-n/2}\exp(-kz^2/\epsilon)\exp(z)dz$$

$$\le \exp(K\epsilon) + \int_{|z|>\epsilon} k_1\epsilon^{-n/2}\exp(-kz^2/\epsilon)\exp(z)dz \equiv (1+\delta(\epsilon)). \tag{7.47}$$

By inspection, $\delta(\epsilon)$ given in (7.47) converges to 0 as $\epsilon \to 0$. Therefore, (7.33) follows.

Using Lipschitz continuity of v with a constant λ_2 (see (2.7) and (7.40)) and (7.45), we can write

$$v^\epsilon(x) - v(x) = E_x\{v(X(\epsilon)) - v(x)\} \le \lambda_2 E_x\{|X(\epsilon) - x|\}.$$

On the other hand

$$E_x\{|X(\epsilon)-x|\} = \int_{\mathbb{R}^n} p(\epsilon,x,y)|y-x|dy = \left(\int_{|x-y|\le\epsilon} \dots + \int_{|x-y|>\epsilon} \dots\right)$$

$$\le \epsilon + \int_{|z|>\epsilon} k_1\epsilon^{-n/2}\exp(-kz^2/\epsilon)|z|dz = \delta_1(\epsilon) \to 0 \tag{7.48}$$

as $\epsilon \to 0$, whereas (7.35) follows.

From (7.43) it follows that $Y^y(0) = y$. Applying the Lipschitz property of the function c, we get

$$
\begin{aligned}
|c^\epsilon(x,y) - c(x,y)| &= E_x\{|c(X(\epsilon), Y^y(\epsilon)) - c(x, Y^y(0))|\} \\
&\le K E_x\{|X(\epsilon) - x||Y^y(\epsilon) - Y^y(0)|\}.
\end{aligned} \tag{7.49}
$$

From (7.43), (7.44), and (7.48), we imply that the right-hand side of (7.49) converges to 0 as $\epsilon \to 0$. The latter implies (7.34).

Acknowledgment. This research was done while the author was visiting Departamento de Matematicas of the CINVESTAV of the Mexican Institute of Technology, whose hospitality is greatly appreciated. The author would like to thank Dr. O. Hernández-Lerma for the introduction to this field and for a series of helpful and stimulating discussions.

REFERENCES

[1] E.J. Anderson and P. Nash, *Linear Programming in Infinite-Dimensional Spaces*, John Wiley, Chichester, 1989.
[2] C. Dellacherie and P.-A. Meyer, *Probabilites et Potentiel. Theorie des Martingales*, Hermann, Paris, 1980.
[3] N. Dunford and J.T. Schwartz, *Linear Operators. Part* I: *General Theory*, John Wiley, New York, 1988.
[4] E.B. Dynkin, *Markov Processes* I, II, Springer-Verlag, Berlin, 1965.
[5] S.N. Ethier and T.G. Kurtz, *Markov Processes: Characterization and Convergence*, John Wiley, New York, 1986.
[6] W.H. Fleming and R.W. Rishel, *Deterministic and Stochastic Optimal Control*, Springer-Verlag, Berlin, 1975.
[7] W.H. Fleming and H.M. Soner, *Controlled Markov Processes and Viscosity Solutions*, Springer-Verlag, Berlin, 1992.
[8] W.H. Fleming and D. Vermes, *Convex duality approach to the optimal control of diffusions*, SIAM J. Control Optim., 27 (1989), pp. 1136–1155.
[9] W.H. Fleming and D. Vermes, *Generalized Solutions in the Optimal Control of Diffusions*, IMA Vol. Math. Appl. 10, W.H. Fleming and P.L. Lions, eds., Springer-Verlag, New York, 1985, pp. 119–127.
[10] M.I. Freidlin, *Functional Integration and Partial Differential Equations*, Princeton University Press, Princeton, NJ, 1985.
[11] A. Friedman, *Stochastic Differential Equations and Applications*, Vol. 1, Academic Press, New York, 1975.
[12] A. Friedman, *Partial Differential Equations of Parabolic Type*, Prentice-Hall, Englewood Cliffs, NJ, 1964.
[13] J.M. Harrison and M.I. Taksar, *Instantaneous control of Brownian motion*, Math. Oper. Res., 8 (1983), pp. 439–453.
[14] D. Hernandez-Hernandez, O. Hernandez-Lerma, and M.I. Taksar, *The linear programming approach to deterministic control problems*, Applicationes Mathematicae, 14 (1996), pp. 17–33.
[15] D. Hernandez-Hernandez and O. Hernandez-Lerma, *Discounted cost Markov decision processes on Borel spaces: The linear programming formulation*, J. Math. Anal. Appl., 183 (1994), pp. 335–351.
[16] O. Hernandez-Lerma and J.B. Lasserre, *Linear programming and average optimality of Markov control processes on Borel spaces—unbounded cost*, SIAM J. Control Optim., 32 (1994), pp. 480–500.
[17] R.M. Lewis and R.B. Vinter, *Relaxation of optimal control problems to equivalent convex programs*, J. Math. Anal. Appl., 74 (1980), pp. 475–493.
[18] I. Karatzas, *A class of singular stochastic control problems*, Adv. in Appl. Probab., 22 (1983), pp. 225–254.
[19] I. Karatzas and S.E. Shreve, *Connection between optimal stopping and singular stochastic control* I, II, SIAM J. Control Optim., 22 (1984), pp. 856–877; 23 (1985), pp. 433–541.
[20] J.L. Kelley, *General Topology*, Van Nostrand, New York, 1955.

[21] O.A. Ladyzhenskaia and N.N. Uraltseva, *Linear and Quasilinear Elliptic Equations*, Academic Press, New York, 1968.
[22] P.L. Lions and A.S. Sznitman, *Stochastic differential equations with reflecting boundary conditions*, Comm. Pure Appl. Math., 3 (1984), pp. 511–537.
[23] R.S. Liptser and A.N. Shiryayev, *Theory of Martingales*, Kluwer Academic Publ., Dordrecht, the Netherlands, 1989.
[24] J.L. Menaldi, M. Robin, and M.I. Taksar, *Singular ergodic control of multidimensional Gaussian process*, Math. Control Signals Systems, 5 (1992), pp. 93–114.
[25] J.L. Menaldi and M.I. Taksar, *Optimal correction problem of a multidimensional stochastic system*, Automatica, 25 (1989), pp. 223–232.
[26] A.P. Robertson and W. Robertson, *Topological Vector Spaces*, Cambridge University Press, Cambridge, UK, 1964.
[27] M.I. Taksar, *Average optimal singular control and a related stopping problem*, Math. Oper. Res., 10 (1985), pp. 63–81.
[28] M.I. Taksar, *Singular control in a multidimensional space with control costs proportional to displacement*, in Proc. International Conference on Optimization, Singapore, 1987, pp. 314–324.
[29] M.I. Taksar, *Convex solutions to variational inequalities and multidimensional singular control*, in The Dynkin Festschrift. Markov Processes and Their Applications, M. Freidlin, ed., Birkhäuser, Boston, Baden, Berlin, 1994, pp. 371–386.
[30] D. Vermes, *Optimal control of piecewise deterministic Markov processes*, Stochastics, 14 (1985), pp. 165–207.
[31] R.B. Vinter and R.M. Lewis, *The equivalence of strong and weak formulations for certain problems in optimal control*, SIAM J. Control Optim., 16 (1978), pp. 546–570.
[32] L. Wein, *Optimal control of a two-station Brownian network*, Math. Oper. Res., 15 (1990), pp. 215–242.
[33] L. Wein, *Scheduling networks of queues: Heavy traffic analysis of a multistation network with controllable inputs*, Oper. Res., 40 (1992), pp. 312–334.

SIAM J. CONTROL OPTIM.
Vol. 35, No. 2, pp. 626–640, March 1997

014

THE DIFFERENTIABILITY OF THE DRAG WITH RESPECT TO THE VARIATIONS OF A LIPSCHITZ DOMAIN IN A NAVIER–STOKES FLOW*

JUAN ANTONIO BELLO†, ENRIQUE FERNÁNDEZ-CARA†, JÉRÔME LEMOINE‡, AND JACQUES SIMON‡

Abstract. This paper is concerned with the computation of the drag T associated with a body traveling at uniform velocity in a fluid governed by the stationary Navier–Stokes equations. It is assumed that the fluid fills a domain of the form $\Omega+u$, where $\Omega \subset \mathbb{R}^3$ is a reference domain and u is a displacement field. We assume only that Ω is a Lipschitz domain and that u is Lipschitz-continuous. We prove that, at least when the velocity of the body is sufficiently small, $u \mapsto T(\Omega + u)$ is a C^∞ mapping (in a ball centered at 0). We also compute the derivative at 0.

Key words. domain optimization, hydrodynamic drag, Navier–Stokes equations, Lipschitz domains, optimal control

AMS subject classification. 49J20

PII. S0363012994278213

1. Introduction.

Formulation of the problem. In this paper, we study the behavior of the drag T associated with a body traveling at uniform velocity γ in a viscous incompressible fluid. It is assumed that the flow of this fluid is governed by the stationary Navier–Stokes equations. We are interested in viewing T as a function of the shape of the body.

More precisely, let B be a reference shape for the body and Ω be the corresponding fluid domain. The body variations are described by a field u, and we search for a formula of the kind

$$T(\Omega + u) = T(\Omega) + T'(\Omega; u) + o(u),$$

where the modified fluid domain is

$$\Omega + u = \{x \in \mathbb{R}^d;\ x = (I + u)(\xi),\ \xi \in \Omega\}.$$

We are thus led to an analysis of the differentiability of the function $u \mapsto T(\Omega + u)$.

The main results. We prove that when Ω is a Lipschitz domain, u is Lipschitz-continuous, and the velocity γ is sufficiently small, the function $u \mapsto T(\Omega + u)$ is differentiable. More precisely (see Theorem 4), we show that it is a C^∞ mapping in a small ball $\mathcal{W}$ whose elements are Lipschitz vector fields. We also compute explicitly $T'(\Omega; u)$, i.e., the derivative at 0 in the direction u.

In the similar but more simple case of an elliptic equation, differentiability results have been established by F. Murat and J. Simon in [9], [10] without any regularity hypothesis on Ω. The proof relies on the change of variables $x = (I+u)(\xi)$, by means of which one is led to a fixed domain. This method has been used for many equations by several authors.

*Received by the editors December 5, 1994; accepted for publication (in revised form) February 20, 1996.

http://www.siam.org/journals/sicon/35-2/27821.html

†Departamento de Ecuaciones Diferenciales y Análisis Numérico, Universidad de Sevilla, Aptdo. 1160, 41 080 Sevilla, Spain. The research of these authors was supported in part by Proyecto Dirección General de Investigación Cientifica y Tecnologica.

‡Laboratoire de Mathématiques Appliquées, Université Blaise Pascal (Clermont-Ferrand 2), 63 177 Aubière Cedex, France (simon@ucfma.univ-bpclermont.fr).

Some difficulties related to incompressibility. The general method in [9], [10] cannot be directly applied to the Stokes and Navier–Stokes cases. This is due to the incompressibility condition

$$\nabla \cdot y(u) = 0 \qquad \text{in } \Omega + u,$$

which has to be satisfied by the velocity field $y(u)$. This difficulty was surmounted when Ω is a $W^{2,\infty}$ domain by J. Simon [17] for Stokes flows and by J. A. Bello, E. Fernández-Cara, and J. Simon [1], [2] for Navier–Stokes flows. In [17], the author uses a variant of the implicit function theorem; in [1], [2], one introduces a family of isomorphisms which allow us to rewrite the equation $\nabla \cdot y(u) = 0$ appropriately. In this paper, the incompressibility equation is rewritten explicitly.

We will assume that Ω is a Lipschitz domain and that u is Lipschitz-continuous. This includes many interesting situations in which $\partial\Omega$ and/or $\partial(\Omega + u)$ possess "corner" points.

Recall that formal computations of the derivative were previously carried out by O. Pironneau [12] (see also [13]) using "normal" variations.

Some difficulties related to weak regularity. The "natural" expression of the derivative $T'(\Omega; u)$ (that is, the right-hand side of (15)) is not defined a priori since y is only $H^1(\Omega)^d$. Nevertheless, we will give a meaning for this expression using the technical result (17).

2. The definition of the drag. Let D and B be bounded open connected sets in $\mathbb{R}^d$, $d = 2$ or 3, with $B \subset\subset D$. Let us set $\Omega = D \backslash \overline{B}$. In the following discussion, it will be assumed that

$$\Omega \text{ is a Lipschitz domain;} \tag{1}$$

that is to say, its boundary $\partial\Omega$ is locally the graph of a Lipschitz-continuous function and Ω is the corresponding epigraph. (This is explained more in detail in the appendix.)

Let $\gamma \in \mathbb{R}^d$ be a given vector. We consider the stationary Navier–Stokes problem [4]

$$\begin{cases} y - g \in H_0^1(\Omega)^d, \\ p \in L^2(\Omega), \qquad \int_\Omega p = 0, \\ -\nu\,\Delta y + (y \cdot \nabla)\, y + \nabla p = 0, \\ \nabla \cdot y = 0. \end{cases} \tag{2}$$

Here, $g \in H^1(\mathbb{R}^d)^d$ and satisfies

$$\nabla \cdot g = 0,\ g = \gamma \text{ in a neighborhood of } \partial D,\ g = 0 \text{ in a neighborhood of } B. \tag{3}$$

When B is small with respect to D, any solution (y, p) to (2) provides good approximations to the velocity field and the pressure distribution of a viscous incompressible fluid in Ω having constant velocity far from B. It can be imagined that we have chosen spatial coordinates fixed with respect to B, D is an approximation to $\mathbb{R}^d$, the fluid is at rest at infinity, and B is the shape of a body traveling at constant velocity $-\gamma$.

The requirement $\int_\Omega p = 0$ provides uniqueness for the pressure p that, otherwise, would be defined up to an additive constant.

If γ is sufficiently small, problem (2) possesses exactly one solution, which is "small" and does not depend on the choice of g. More precisely, Theorem 2.1 in [9] gives the following lemma.

LEMMA 1. *There exists a constant $\alpha > 0$ such that, if $|\gamma| < \alpha\nu$, then* (2) *possesses exactly one solution, $(y, p) \in H^1(\Omega)^d \times L^2(\Omega)$. This solution does not depend on the choice of the function g satisfying* (3). *Furthermore, for each $\epsilon > 0$, the constant α can be chosen in such a way that*

$$\|y\|_{H^1(\Omega)^d} \leq \epsilon\nu.$$

If $\mathcal{O} \subset\subset D$ is given, one can also choose $\alpha = \alpha(\epsilon, \mathcal{O}, D)$ not depending on B, provided $B \subset \mathcal{O}$. Finally, if Ω is a $W^{2,\infty}$ domain, then $(y, p) \in H^2(\Omega)^d \times H^1(\Omega)$.

Thus, at least when γ is small, one can associate with Ω a drag

$$T(\Omega) = \frac{\nu}{2} \int_\Omega \sigma(y)^2, \tag{4}$$

where $\sigma(y)^2 = \sigma(y) \cdot \sigma(y) \equiv \sum_{ij} (\sigma_{ij}(y))^2$.

Remark. If Ω is regular enough, $T(\Omega)$ coincides with the usual hydrodynamical drag, which is given as follows (cf. [14]):

$$\mathcal{T}(\Omega) = -\gamma \cdot \int_{\partial B} (-p\,Id + \nu\,\sigma(y)) \cdot n\,ds.$$

Indeed, using the boundary condition, we obtain

$$\mathcal{T}(\Omega) = -\int_{\partial\Omega} (p\,(y-\gamma) - \nu\,\sigma(y) \cdot (y-\gamma)) \cdot n\,ds.$$

From Gauss formula and incompressibility, this gives

$$\begin{aligned}\mathcal{T}(\Omega) &= -\int_\Omega \nabla \cdot (p\,(y-\gamma) - \nu\,\sigma(y) \cdot (y-\gamma)) \\ &= \int_\Omega ((\nu\,\Delta y - \nabla p) \cdot (y-\gamma) + \nu\,\sigma(y) \cdot \nabla y).\end{aligned}$$

Note that, again using incompressibility,

$$(\nu\,\Delta y - \nabla p) \cdot (y-\gamma) = ((y \cdot \nabla)\,y) \cdot (y-\gamma) = \nabla \cdot (|y-\gamma|^2\,y).$$

Therefore,

$$\int_\Omega (\nu\,\Delta y - \nabla p) \cdot (y-\gamma) = \int_{\partial\Omega} |y-\gamma|^2\,y \cdot n\,ds = 0,$$

and, finally, since $\sigma(y) \cdot \nabla y = \frac{1}{2}\,\sigma(y)^2$, we have $\mathcal{T}(\Omega) = T(\Omega)$. □

3. The domain variations. We will choose fields $u \in W^{1,\infty}(\mathbb{R}^d, \mathbb{R}^d)$ such that $u = 0$ on ∂D. This condition expresses the fact that the outer boundary limiting the fluid is fixed.

We will also assume $\|u\|_{\text{Lip}} < c(\Omega)$, with $c(\Omega)$ being small enough to ensure that $\Omega + u$ is Lipschitzian and also that $B + u$ is included in a fixed open set $\mathcal{O}$ satisfying

$$B \subset\subset \mathcal{O} \subset\subset D.$$

Here, we have denoted by $\|u\|_{\mathrm{Lip}}$ the best Lipschitz constant for u. More precisely, we have the following obvious result (see [8] for a proof).

LEMMA 2. *Assume that $\mathcal{O}$ is as before. There exists $c(\Omega)$, $0 < c(\Omega) < 1$, such that*

$$B + u \subset \mathcal{O} \tag{5}$$

for all $u \in W^{1,\infty}(\mathbb{R}^d, \mathbb{R}^d)$ satisfying $u = 0$ on ∂D and $\|u\|_{\mathrm{Lip}} \leq c(\Omega)$.

We will also use the following result, whose proof is given in the appendix.

LEMMA 3. *There exists $c(\Omega)$, $0 < c(\Omega) < 1$, such that*

$$\Omega + u \quad \textit{is a bounded Lipschitz domain in} \quad \mathbb{R}^d \tag{6}$$

for all $u \in W^{1,\infty}(\mathbb{R}^d, \mathbb{R}^d)$ satisfying $\|u\|_{W^{1,\infty}(\mathbb{R}^d,\mathbb{R}^d)} \leq c(\Omega)$.

Remark. This lemma holds for each bounded Lipschitz domain $\Omega \subset \mathbb{R}^d$. □

For the subsequent discussion, we introduce

$$\mathcal{W} = \{u \in W^{1,\infty}(\mathbb{R}^d, \mathbb{R}^d);\ \|u\|_{W^{1,\infty}(\mathbb{R}^d,\mathbb{R}^d)} < c(\Omega),\ u = 0 \ \text{ on } \ \partial D\},$$

with $c(\Omega)$ being as in Lemmas 2 and 3. Observing that

$$\|u\|_{\mathrm{Lip}} \leq \|u\|_{W^{1,\infty}(\mathbb{R}^d,\mathbb{R}^d)}$$

we see that (5) and (6) are satisfied for all $u \in \mathcal{W}$.

It will also be assumed in the sequel that

$$|\gamma| < \alpha(\epsilon, \mathcal{O}, D)\, \nu, \tag{7}$$

where α is furnished by Lemma 1. The precise value of ϵ will be fixed below. Now, we choose g satisfying (3) and

$$g \equiv 0 \ \text{ in a neighborhood of } \ \mathcal{O}.$$

(Such a choice is always possible; for instance, one can take $g = a \wedge \nabla\psi$, where $a \in \mathbb{R}^3$, $a \,.\, \gamma = 0$, $|a| = 1$, $\psi \in C^\infty(\mathbb{R}^3)$, $\psi = 0$ in $\mathcal{O}$, $\psi(x) = (g \wedge a) \,.\, x$ in a neighborhood of ∂D.) If $u \in \mathcal{W}$, one has $g = 0$ in a neighborhood of $\partial B + u$. The Navier–Stokes problem in $\Omega + u$ can be written as follows:

$$\begin{cases} y(u) - g \in H_0^1(\Omega + u)^d, \\ p(u) \in L^2(\Omega + u), \qquad \displaystyle\int_\Omega p(u) \circ (I + u) = 0, \\ -\nu\, \Delta y(u) + (y(u) \,.\, \nabla)\, y(u) + \nabla p(u) = 0, \\ \nabla \,.\, y(u) = 0. \end{cases} \tag{8}$$

From Lemma 1, we know that (8) possesses exactly one solution $(y(u), p(u))$. Accordingly, the drag associated with $B + u$ can be defined and is given by

$$T(\Omega + u) = \frac{\nu}{2} \int_{\Omega+u} \sigma(y(u))^2. \tag{9}$$

Remark. In principle, it seems more natural to normalize $p(u)$ by imposing that $\int_{\Omega+u} p(u) = 0$. However, it will be seen below that the choice that we have made is more useful when one considers different fields $u \in \mathcal{W}$. (Indeed, it yields $\int_\Omega P(u) = 0$ for the transported pressure $P(u) = p(u) \circ (I + u)$; see (23).) □

4. A differentiability result for the drag. Our main interest in this section to describe the variations of $T(\Omega+u)$ with respect to u. As already mentioned in the introduction, we search for a formula

$$T(\Omega+u) = T(\Omega) + T'(\Omega;u) + o(u), \tag{10}$$

which must hold for all $u \in \mathcal{W}$, with $T'(\Omega;\ .\ )$ being a linear mapping and

$$o(u)/\|u\|_{W^{1,\infty}(\mathbb{R}^d,\mathbb{R}^d)} \to 0 \ \text{ as } \ \|u\|_{W^{1,\infty}(\mathbb{R}^d,\mathbb{R}^d)} \to 0.$$

That such a formula can be obtained stems from the next result, which is the most important in this article.

THEOREM 4. *There exists $\alpha > 0$ such that if $|\gamma| < \alpha\nu$, then $u \mapsto T(\Omega+u)$ is a C^∞ mapping in the set $\mathcal{W}$.*

In addition, the first derivative at 0 can be obtained from any of the expressions (11), (15), or (18).

THEOREM 5. *Assume $|\gamma| < \alpha\nu$.*

(i) *For all $u \in W^{1,\infty}(\mathbb{R}^d,\mathbb{R}^d)$ such that $u|_{\partial D} = 0$, one has*

$$T'(\Omega;u) = \nu \int_\Omega \sum_{ij}\sigma_{ij}(y)\left(\sigma_{ij}(\dot y(u)) - \sum_k (\partial_i u_k\,\partial_k y_j + \partial_j u_k\,\partial_k y_i) + \frac{1}{2}\sigma_{ij}(y)\,\nabla\cdot u\right) \tag{11}$$

with $(\dot y(u), \dot p(u))$ being the unique solution to the linear problem

$$\begin{cases} \dot y(u) \in H_0^1(\Omega)^d, \\ \dot p(u) \in L^2(\Omega), \qquad \displaystyle\int_\Omega \dot p(u) = 0, \\ -\nu\,\Delta \dot y(u) + (\dot y(u)\cdot\nabla)\,y + (y\cdot\nabla)\,\dot y(u) + \nabla \dot p(u) = G(u,y,p), \\ \nabla\cdot \dot y(u) = \sum_{ij}\partial_i u_j\,\partial_j y_i. \end{cases} \tag{12}$$

Here, $y = y(0)$, $p = p(0)$, and $G_k(u,y,p) \in H^{-1}(\Omega)$ is given as follows for $1 \le k \le d$:

$$\begin{aligned} G_k(u,y,p) = &-\nu \sum_{ij}(\partial_j(\partial_i u_j\,\partial_i y_k) + \partial_j(\partial_j u_i\,\partial_i y_k)) + \nu\sum_j \partial_j((\nabla\cdot u)\,\partial_j y_k) \\ &+ \sum_{ij} y_i\,\partial_i u_j\,\partial_j y_k - (y\cdot\nabla)\,y_k\,\nabla\cdot u \\ &+ \sum_j \partial_j(\partial_k u_j\,p) - \partial_k((\nabla\cdot u)\,p). \end{aligned} \tag{13}$$

Moreover, $y \in C^\infty(\Omega)^d$, $p \in C^\infty(\Omega)$, and, consequently,

$$G(u,y,p) = -\nu\,\Delta((u\cdot\nabla)\,y) + (((u\cdot\nabla)\,y)\cdot\nabla)\,y + (y\cdot\nabla)((u\cdot\nabla)\,y) + \nabla(u\cdot\nabla p). \tag{14}$$

(ii) *One also has*

$$T'(\Omega;u) = \nu\int_\Omega \sum_{ij}\left(\sigma_{ij}(y)\,\sigma_{ij}(y'(u)) + \frac{1}{2}\,\nabla\cdot\left(\sigma_{ij}(y)^2 u\right)\right), \tag{15}$$

with $(y'(u), p'(u))$ being the unique solution to

$$\begin{cases} y'(u) + (u\cdot\nabla)\,y \in H_0^1(\Omega)^d, \\ (p'(u) + u\cdot\nabla p) \in L^2(\Omega), \qquad \displaystyle\int_\Omega (p'(u) + u\cdot\nabla p) = 0, \\ -\nu\,\Delta y'(u) + (y'(u)\cdot\nabla)\,y + (y\cdot\nabla)\,y'(u) + \nabla p'(u) = 0, \\ \nabla\cdot y'(u) = 0. \end{cases} \tag{16}$$

Furthermore, $y'(u) \in H^1_{\rm loc}(\Omega)^d$ *and the sum in* (15) *satisfies*

$$\sum_{ij} \left(\sigma_{ij}(y)\, \sigma_{ij}(y'(u)) + \frac{1}{2} \nabla \cdot \left(\sigma_{ij}(y)^2 u \right) \right) \in L^1(\Omega). \tag{17}$$

(iii) *If* B *and* D *are* $W^{2,\infty}$ *domains and* $u \in W^{2,\infty}(\mathbb{R}^d, \mathbb{R}^d)$, *then* $(y,p) \in H^2(\Omega)^d \times H^1(\Omega)$ *and*

$$T'(\Omega; u) = \int_{\partial B} u \cdot n \left(\frac{\partial w}{\partial n} - \frac{\partial y}{\partial n} \right) \cdot \frac{\partial y}{\partial n}\, ds, \tag{18}$$

with (w,q) *being the unique solution to the "adjoint" problem*

$$\begin{cases} w \in H^1_0(\Omega)^d \cap H^2(\Omega)^d, \\ q \in H^1(\Omega), \qquad \displaystyle\int_\Omega q = 0, \\ -\nu\, \Delta w_i + \sum_j \partial_i y_j\, w_j - \sum_j y_j\, \partial_j w_i + \partial_i q = -2\nu\, \Delta y_i\,, \quad 1 \le i \le d, \\ \nabla \cdot w = 0. \end{cases} \tag{19}$$

Remark. In order to compute the derivative of the drag in several directions using (15), one has to solve, for each direction u, the corresponding partial differential problem (16). It is much more interesting to use the identity (18) because it suffices to solve (2) and (19) only once; then, for each u, one has only to compute an integral on ∂B. □

Remark. One can also obtain expressions for the derivatives of higher orders. This must be made with caution; indeed, $T''(\Omega;\, \cdot\,,\, \cdot\,)$ (i.e., the second derivative at 0 of $u \mapsto T(\Omega + u)$) does not coincide with $(T'(\Omega;\, \cdot\,)';\, \cdot\,)$ (i.e., the derivative at 0 of the mapping $u \mapsto T'(\Omega + u;\, \cdot\,)$). In fact, these two quantities are related by the following formula (see [16]):

$$T''(\Omega; u, v) = (T'(\Omega; u)'; v) - T'(\Omega; (u \cdot \nabla)\, v). \qquad \square$$

5. Differentiability results for the velocity and the pressure. In order to prove Theorem 4, we will first show that $u \mapsto y(u)$ is, in a certain sense, a "differentiable" mapping. An important difficulty arises here, because $y(u)$ is a function defined only for $x \in \Omega + u$, a domain which depends on u. This is why we introduce a suitable change of variables and we rewrite the equations satisfied by $y(u)$ and $p(u)$ in the fixed domain Ω. Then, we will have to differentiate the transported variable $Y(u) = y(u) \circ (I + u)$, which is defined in Ω.

In what follows, y and p stand for $y(0)$ and $p(0)$, respectively. We will check the following:

$$\dot{y}(u) = Y'(0) \cdot u \equiv \lim_{t \to 0} \frac{y(tu) \circ (I + tu) - y}{t}\,.$$

This is the "total derivative" of $y(u)$ at 0, used in (11) to give an expression of $T'(\Omega; u)$. We will also have to use the "local derivative." In fact, we will check that

$$y'(u) = \frac{d}{dv} y(v)|_\omega(0) \cdot u \equiv \lim_{t \to 0} \frac{y(tu)|_\omega - y|_\omega}{t} \qquad \text{in } \ \omega.$$

This defines $y'(u)$ in each open set $\omega \subset\subset \Omega$ and, consequently, in the whole domain Ω. The previous local derivative was used in (15) to give an expression of $T'(\Omega; u)$. More precisely, the following result holds.

THEOREM 6. *There exists $\alpha > 0$ such that if $|\gamma| < \alpha\nu$, then*

(i) *The mapping $u \mapsto (y(u), p(u)) \circ (I+u)$ is C^∞ in $\mathcal{W}$, with values in the product space $H^1(\Omega)^d \times L^2(\Omega)$. Its derivative at 0 in the direction u is the unique solution $(\dot{y}(u), \dot{p}(u))$ to (12).*

(ii) *For all $\omega \subset\subset \Omega$, the mapping $u \mapsto y(u)|_\omega$ is differentiable in $\mathcal{W}$, with values in $L^2(\omega)^d$. Its derivative at 0 in the direction u is $y'(u)|_\omega$, where $y'(u)$ is uniquely defined by (16). One also has*

$$y'(u) = \dot{y}(u) - (u \cdot \nabla)\, y. \tag{20}$$

Remark. From general results on local differentiability (see Lemma 2.1 in [15]), (ii) is implied by (i). ∎

Theorems 4, 5, and 6 will be demonstrated in several steps:

— differentiability at 0 of the velocity, the pressure (section 5), and the drag (section 6);

— differentiability at any point in $\mathcal{W}$ (section 7); higher-order differentiability (section 8).

6. Proof of differentiability at 0 of the velocity and the pressure. The goal of this section is to prove the following result.

LEMMA 7. *There exists $\alpha > 0$ such that, if $|\gamma| < \alpha\nu$, then the mapping $u \mapsto (y(u), p(u)) \circ (I+u)$, which is defined in $\mathcal{W}$ and takes values in $H^1(\Omega)^d \times L^2(\Omega)$, is differentiable at 0. Its derivative, denoted by $(\dot{y}(u), \dot{p}(u))$, is uniquely determined by (12).*

The proof is based on the implicit function theorem. We will show that this lemma holds with α being of the form $\alpha(\epsilon, \mathcal{O}, D)$ (as in Lemma 1) for an appropriate constant ϵ. First, we will have to rewrite the equations (8) in the fixed domain Ω. For this, we have to "transport" all the terms, some of which belong to $H^{-1}(\Omega+u)$. But it is not clear for a distribution $f \in H^{-1}(\Omega+u)$ how $f \circ (I+u)$ can be defined. Contrarily, following [10, Definition 4.1], one can give a definition of $(f \circ (I+u))\,\mathrm{Jac}(I+u)$.

DEFINITION 8. *Assume $u \in \mathcal{W}$ and $f \in H^{-1}(\Omega+u)$. Then*

$$(f \circ (I+u))\,\mathrm{Jac}(I+u) \in H^{-1}(\Omega)$$

is defined as follows: for any $\varphi \in H_0^1(\Omega)$, one has

$$\langle (f \circ (I+u))\,\mathrm{Jac}(I+u), \varphi \rangle_{H^{-1}(\Omega)\times H_0^1(\Omega)} = \langle f, \varphi \circ (I+u)^{-1} \rangle_{H^{-1}(\Omega+u)\times H_0^1(\Omega+u)}. \tag{21}$$

Remark. Rigorously speaking, $(f \circ (I+u))\,\mathrm{Jac}(I+u)$ is not a good notation, because $f \circ (I+u)$ is not defined. However, it will be used in subsequent discussion for convenience. ∎

Note that (21) makes sense; indeed, $\varphi \circ (I+u)^{-1} \in H_0^1(\Omega+u)$ (see [10, Lemma 4.1]). It does not change the usual definition of $(f \circ (I+u))\,\mathrm{Jac}(I+u)$ when $f \in L^1_{\mathrm{loc}}(\Omega+u)$.

In order to rewrite (8), we denote by $D(u)$ the operator whose components $D_i(u)$ are given as follows:

$$D_i(u) = \textstyle\sum_j M_{ij}(u)\, \partial_j, \quad M(u) = {}^t\,[\partial_j (I+u)_i]^{-1}. \tag{22}$$

Here, ${}^t\,[\partial_j (I+u)_i]^{-1}$ is the transpose of the inverse of the matrix of components $\partial_j (I+u)_i$. We will use the following three lemmas (see [9] and [10]).

LEMMA 9. *Assume $u \in \mathcal{W}$ and $f \in H^1(\Omega + u)$. Then*

$$(\partial_i f) \circ (I + u) = \textstyle\sum_j M_{ij}(u)\, \partial_j (f \circ (I + u)) = D_i(u)(f \circ (I + u)).$$

LEMMA 10. *If $u \in \mathcal{W}$ and $f \in L^2(\Omega + u)$, then*

$$((\partial_i f) \circ (I + u))\, \mathrm{Jac}(I + u) = \textstyle\sum_j \partial_j (M_{ij}(u)\, (f \circ (I + u))\, \mathrm{Jac}(I + u)).$$

LEMMA 11. *Assume $u \in \mathcal{W}$ and $f \in H^1(\Omega + u)$. Then*

$$((\Delta f) \circ (I + u))\, \mathrm{Jac}(I + u) = \textstyle\sum_{ij} \partial_j (M_{ij}(u)\, \mathrm{Jac}(I + u)\, D_i(u)(f \circ (I + u))).$$

The Navier–Stokes problem (8) can now be written as follows:

$$\text{(23)} \qquad \begin{cases} Y(u) - g \in H^1_0(\Omega)^d, \\ P(u) \in L^2(\Omega), \qquad \displaystyle\int_\Omega P(u) = 0, \\ -\nu \textstyle\sum_{ij} \partial_j (M_{ij}(u)\, \mathrm{Jac}(I + u)\, D_i(u) Y_k(u)) \\ \qquad\qquad + (Y(u) \cdot D(u)) Y_k(u)\, \mathrm{Jac}(I + u) \\ \qquad\qquad + \textstyle\sum_j \partial_j (M_{kj}(u)\, P(u)\, \mathrm{Jac}(I + u)) = 0, \qquad 1 \le k \le d, \\ D(u) \cdot Y(u)\, \mathrm{Jac}(I + u) = 0. \end{cases}$$

Here, we have set $Y(u) = y(u) \circ (I + u)$ and $P(u) = p(u) \circ (I + u)$.

We will also introduce in (23) the new variable $X(u) = Y(u) - g$. This leads to the following system, equivalent to (23) (which is, in turn, equivalent to (8)):

$$\text{(24)} \qquad \begin{cases} X(u) \in H^1_0(\Omega)^d, \\ P(u) \in L^2(\Omega), \qquad \displaystyle\int_\Omega P(u) = 0, \\ -\nu \textstyle\sum_{ij} \partial_j (M_{ij}(u)\, \mathrm{Jac}(I + u)\, D_i(u)(X(u) + g)_k) \\ \qquad\qquad + ((X(u) + g) \cdot D(u))\, (X(u) + g)_k\, \mathrm{Jac}(I + u) \\ \qquad\qquad + \textstyle\sum_j \partial_j (M_{kj}(u)\, P(u)\, \mathrm{Jac}(I + u)) = 0, \qquad 1 \le k \le d, \\ D(u) \cdot (X(u) + g)\, \mathrm{Jac}(I + u) = 0. \end{cases}$$

This equation can be written

$$\text{(25)} \qquad H(u; X(u), P(u)) = 0,$$

where the function H is defined, from $\mathcal{W} \times H^1_0(\Omega)^d \times L^2_0(\Omega)$ into $H^{-1}(\Omega)^d \times L^2_0(\Omega)$, by

$$\text{(26)} \qquad \begin{cases} H(u; \chi, \pi) = (F(u; \chi, \pi), R(u; \chi, \pi)), \quad F = (F_1, \dots, F_d), \\ F_k(u; \chi, \pi) = -\nu \textstyle\sum_{ij} \partial_j (M_{ij}(u)\, \mathrm{Jac}(I + u)\, D_i(u)(\chi + g)_k) \\ \qquad\qquad + ((\chi + g) \cdot D(u))(\chi + g)_k\, \mathrm{Jac}(I + u) \\ \qquad\qquad + \textstyle\sum_j \partial_j (M_{kj}(u)\, \pi\, \mathrm{Jac}(I + u)), \qquad 1 \le k \le d, \\ R(u; \chi, \pi) = D(u) \cdot (\chi + g)\, \mathrm{Jac}(I + u). \end{cases}$$

The fact that $R(u;\chi,\pi) \in L^2_0(\Omega)$ is crucial. This is true because

$$\begin{aligned}\int_\Omega (D(u)\,.\,Y(u))\,\mathrm{Jac}(I+u) &= \int_{\Omega+u} (D(u)\,.\,Y(u))\circ(I+u)^{-1}\\ &= \int_{\Omega+u} \nabla\,.\,(Y(u)\circ(I+u)^{-1})\\ &= 0.\end{aligned}$$

Now, we check that the assumptions of the implicit function theorem are satisfied. First, H is C^1 in a neighborhood of $(0;X,P)$, where we have set $X = X(0) = y - g$, $P = P(0) = p$. Indeed, the coefficients in $D(u)$ and $M(u)$ are C^1 since, according to the results in [10], the mapping $u \mapsto M_{ij}(u)$ is C^1 in a neighborhood of 0 in $W^{1,\infty}(\mathbb{R}^d,\mathbb{R}^d)$, with values in $L^\infty(\mathbb{R}^d,\mathbb{R}^{d^2})$.

On the other hand, let us see that the differential operator $L = D_{(\chi,\pi)}H(0;X,P)$ is an isomorphism from $H^1_0(\Omega)^d \times L^2_0(\Omega)$ onto $H^{-1}(\Omega)^d \times L^2_0(\Omega)$. For each $(\chi,\pi) \in H^1_0(\Omega)^d \times L^2_0(\Omega)$, one has

$$L(\chi,\pi) = (\,-\nu\,\Delta\chi + (\chi\,.\,\nabla)\,y + (y\,.\,\nabla)\,\chi + \nabla\pi\,,\,\nabla\,.\,\chi). \tag{27}$$

The operator L is linear and bounded from $H^1_0(\Omega)^d \times L^2_0(\Omega)$ into $H^{-1}(\Omega)^d \times L^2_0(\Omega)$. Hence, we have to check that, for each $f \in H^{-1}(\Omega)^d$ and $\phi \in L^2_0(\Omega)$, there exists a unique solution $(\chi,\pi) \in H^1_0(\Omega)^d \times L^2_0(\Omega)$ to the system

$$\begin{cases} -\nu\,\Delta\chi + (\chi\,.\,\nabla)\,y + (y\,.\,\nabla)\,\chi + \nabla\pi = f,\\ \nabla\,.\,\chi = \phi \end{cases} \tag{28}$$

and, also, that this solution depends continuously on the data. Since Ω is a Lipschitz domain, Corollary 2.4 in [6] asserts

$$\forall\phi \in L^2(\Omega) \ \text{ such that } \int_\Omega \phi = 0, \ \text{ there exists } \ \psi \in H^1_0(\Omega)^d \ \text{ such that } \ \nabla\,.\,\psi = \phi. \tag{29}$$

Setting $\Phi = \chi - \psi$, system (28) reduces to

$$\begin{cases} \Phi \in V, \quad \pi \in L^2_0(\Omega),\\ -\nu\,\Delta\Phi + (\Phi\,.\,\nabla)\,y + (y\,.\,\nabla)\,\Phi + \nabla\pi = F, \end{cases}$$

where $V = \{v \in H^1_0(\Omega)^d;\ \nabla\,.\,v = 0\}$ and $F = f + \nu\Delta\psi - (\psi\,.\,\nabla)y - (y\,.\,\nabla)\psi$. This equation is elliptic with respect to Φ and possesses a unique solution depending continuously on the data if, for some appropriate $r = r(\mathcal{O},D) > 0$, one has

$$\|y\|_{H^1_0(\Omega)^d} < r\,\nu. \tag{30}$$

Hence, if we choose $\epsilon < r$, $\alpha = \alpha(\epsilon,\mathcal{O},D)$ as in Lemma 1 and

$$|\gamma| < \alpha\nu,$$

this condition holds and L is an isomorphism.

This allows us to apply the implicit function theorem to (25). We deduce that the mapping $u \mapsto (X(u),P(u))$, which takes values in the space $H^1_0(\Omega)^d \times L^2_0(\Omega)$, is

differentiable at 0. Since $y(u)\circ(I+u)=X(u)+g$ and $p(u)\circ(I+u)=P(u)$, the first part of Lemma 7 is proven.

Finally, let us deduce the equations satisfied by $(\dot y(u),\dot p(u))$. In accordance with the implicit function theorem,

$$L(\dot y(u),\dot p(u)) = -D_vH(0;X,P)\,.\,u$$

for all admissible u. Taking into account (26) and also the identities

$$M'_{ik}(0)\,.\,u=-\partial_i u_k \quad\text{and}\quad \frac{d}{dv}\mathrm{Jac}(I+v)(0)\,.\,u=\nabla\,.\,u \tag{31}$$

(see [10]), we find that $(\dot y(u),\dot p(u))$ is a solution to (12). But this problem possesses exactly one solution, since L is an isomorphism. Consequently, Lemma 7 is proven.

Remark. In order to solve (28), we have had to assume that Ω is a Lipschitz domain. The same requirement is found when one writes (28) as a mixed problem and one tries to apply general results concerning mixed variational formulations. □

7. Proof of differentiability at 0 of the drag. The goal of this section is to prove Theorem 5.

Proof of part (i). By definition, one has

$$\begin{aligned} T(\Omega+u) &= \frac{\nu}{2}\int_{\Omega+u}\textstyle\sum_{ij}(\partial_i y_j(u)+\partial_j y_i(u))^2\\ &= \frac{\nu}{2}\int_{\Omega}\textstyle\sum_{ij}(\sum_k(M_{ik}(u)\,\partial_k Y_j(u)+M_{jk}(u)\,\partial_k Y_i(u)))^2\,\mathrm{Jac}(I+u).\end{aligned}$$

We will deduce the differentiability of the mapping $u\mapsto T(\Omega+u)$ from the following result (Theorem 4.1 in [10]).

LEMMA 12. *Assume that $z(u)$ is well defined for all $u\in\mathcal{W}$ and, also, that*

$$u\mapsto z(u)\circ(I+u) \quad \textit{is differentiable at } 0,\ \textit{with values in}\ \ L^1(\Omega). \tag{32}$$

Then the mapping $u\mapsto S(\Omega+u)=\int_\Omega (z(u)\circ(I+u))\,\mathrm{Jac}(I+u)$ is also differentiable at 0. Its derivative at 0 in the direction u is given by

$$S'(\Omega;u)=\int_\Omega(\dot z(u)+z(0)\,\nabla\,.\,u).$$

We will apply this lemma with

$$z(u)\circ(I+u)=\textstyle\sum_{ij}(\sum_k(M_{ik}(u)\,\partial_k Y_j(u)+M_{jk}(u)\,\partial_k Y_i(u)))^2.$$

Obviously, $S(\Omega+u)\equiv T(\Omega+u)$ in this case; also, that (32) holds is deduced from the differentiability at 0 of the $H^1_0(\Omega)^d$-valued mapping $u\mapsto Y(u)$.

Let us compute $T'(\Omega;u)$. From (31) and the fact that $M(0)=Id$, one has

$$\begin{aligned}\dot z(u) &= 2\textstyle\sum_{ij}(\partial_i y_j+\partial_j y_i)(\partial_i\dot y_j(u)+\partial_j\dot y_i(u)-\sum_k\partial_i u_k\,\partial_k y_j-\sum_k\partial_j u_k\,\partial_k y_i)\\ &= 2\textstyle\sum_{ij}\sigma_{ij}(y)(\sigma_{ij}(\dot y(u))-\sum_k\partial_i u_k\,\partial_k y_j-\sum_k\partial_j u_k\,\partial_k y_i).\end{aligned}$$

Since $z(0)=\sum_{ij}\sigma_{ij}(y)^2$, we have

$$T'(\Omega;u)=\nu\int_\Omega\textstyle\sum_{ij}\sigma_{ij}(y)\left(\sigma_{ij}(\dot y(u))-\sum_k(\partial_i u_k\,\partial_k y_j+\partial_j u_k\,\partial_k y_i)+\frac12\,\sigma_{ij}(y)\,\nabla\,.\,u\right).$$

This proves (11). The regularity results are $y \in C^\infty(\Omega)^d$ and $p \in C^\infty(\Omega)$. (This is well known; for instance, see [7].) The identity (14) is then an easy consequence of (13).

Proof of part (ii). Let us set

$$y'(u) = \dot{y}(u) - (u \cdot \nabla)\, y, \quad p'(u) = \dot{p}(u) - u \cdot \nabla p.$$

Using (14) we see that (12) and (16) are equivalent. On the other hand, these definitions provide the following identity:

$$\begin{aligned}&\textstyle\sum_{ij} \left(\sigma_{ij}(y)\, \sigma_{ij}(y'(u)) + \frac{1}{2}\, \nabla \cdot (\sigma_{ij}(y)^2 u) \right) \\ &\quad = \textstyle\sum_{ij} \sigma_{ij}(y) \left(\sigma_{ij}(\dot{y}(u)) - \sum_k (\partial_i u_k\, \partial_k y_j + \partial_j u_k\, \partial_k y_i) + \frac{1}{2}\, \sigma_{ij}(y)\, \nabla \cdot u \right).\end{aligned}$$

Hence, (11) implies (17) and (15).

Proof of part (iii). Let us now suppose that Ω is a $W^{2,\infty}$ domain and $u \in W^{2,\infty}(\mathbb{R}^d, \mathbb{R}^d)$. According to Lemma 1, one has $y \in H^2(\Omega)^d$ and $p \in H^1(\Omega)$. Consequently, one obtains from (15)

$$T'(\Omega; u) = \nu \int_\Omega \textstyle\sum_{ij} \sigma_{ij}(y)\, \sigma_{ij}(y'(u)) + \frac{\nu}{2} \int_{\partial\Omega} \sum_{ij} \sigma_{ij}(y)^2\, u \cdot n\, ds. \tag{33}$$

Since $\dot{y}(u) = 0$ and $y \equiv$ const. on $\partial\Omega$, $y'(u) = -u \cdot n\, \frac{\partial y}{\partial n}$ on $\partial\Omega$. Therefore,

$$\begin{aligned}&\nu \int_\Omega \textstyle\sum_{ij} \sigma_{ij}(y)\, \sigma_{ij}(y'(u)) \\ &\qquad = -2\nu \int_\Omega \Delta y \cdot y'(u) - 2\nu \textstyle\sum_{ij} \int_{\partial\Omega} u \cdot n\, (\partial_i y_j + \partial_j y_i)\, \frac{\partial y_i}{\partial n}\, n_j\, ds.\end{aligned}$$

In addition, $\sum_i \partial_i y_i = 0$ imply $\sum_{ij} (\partial_i y_j + \partial_j y_i)\, \frac{\partial y_i}{\partial n} n_j = |\frac{\partial y}{\partial n}|^2$, whence

$$T'(\Omega; u) = -2\nu \int_\Omega \Delta y \cdot y'(u) - \nu \int_{\partial\Omega} \left| \frac{\partial y}{\partial n} \right|^2 u \cdot n\, ds.$$

If w and q are given by (19), after some manipulation, one obtains

$$\begin{aligned}T'(\Omega; u) = &\int_\Omega \textstyle\sum_i (-\nu\, \Delta w_i\, y_i'(u) + \sum_j (\partial_i y_j\, w_j - y_j \partial_j w_i) y_i'(u) + \partial_i q\, y_i'(u)) \\ &\qquad - \nu \int_{\partial\Omega} \left| \frac{\partial y}{\partial n} \right|^2 u \cdot n\, ds \\ = &\langle -\nu\, \Delta y'(u) + (y'(u) \cdot \nabla)\, y + (y \cdot \nabla)\, y'(u) + \nabla p'(u), w \rangle_{H^{-1}(\Omega)^d \times H^1_0(\Omega)^d} \\ &\qquad + \nu \int_{\partial\Omega} u \cdot n \left(\frac{\partial w}{\partial n} - \frac{\partial y}{\partial n} \right) \cdot \frac{\partial y}{\partial n}\, ds.\end{aligned}$$

Using (16) satisfied by $(y'(u), p'(u))$, one sees that the duality product on the right-hand side cancels. This proves (18), since $u = 0$ on ∂D. ∎

8. Proof of differentiability at any point in $\mathcal{W}$ of the velocity, the pressure, and the drag. In this section, we prove the following result.

LEMMA 13. *The mapping $u \mapsto (y(u), p(u)) \circ (I + u)$, which takes values in $H^1(\Omega)^d \times L^2(\Omega)$, is differentiable at any point $u_0 \in \mathcal{W}$. The mapping $u \mapsto T(\Omega + u)$ is also differentiable at any $u_0 \in \mathcal{W}$.*

Proof. Let $u_0 \in \mathcal{W}$ be given. We have

$$\Omega + (u_0 + v) = (\Omega + u_0) + v \circ (I + u_0)^{-1} \tag{34}$$

for $v \in \mathcal{W}$ small enough in order to have $u_0 + v \in \mathcal{W}$. According to the results in section 6, the mapping $w \mapsto T((\Omega + u_0) + w)$ is differentiable at 0. The mapping $v \mapsto v \circ (I + u_0)^{-1}$ is linear and bounded (therefore differentiable) from $W^{1,\infty}(\mathbb{R}^d, \mathbb{R}^d)$ into itself. Consequently,

$$v \mapsto T((\Omega + u_0) + v \circ (I + u_0)^{-1}) \quad \text{is differentiable at 0;}$$

i.e., $u \mapsto T(\Omega + u)$ is differentiable at u_0.

Now we will apply the previous results to some new reference domains different from Ω. So we introduce the more explicit notation $(y(\Omega; v), p(\Omega; v))$ for the solution to the Navier–Stokes problem in $\Omega + v$. We see from (34) that, for small v,

$$y(\Omega; u_0 + v) \circ (I + (u_0 + v)) = y(\Omega + u_0; v \circ (I + u_0)^{-1}) \circ (I + u_0 + v). \tag{35}$$

On the other hand, from Lemma 7, we know that the $H^1(\Omega + u_0)^d$-valued mapping $w \mapsto y(\Omega + u_0; w) \circ (I + w)$ is differentiable at 0. Thus, $v \mapsto y(\Omega; u_0 + v) \circ (I + u_0 + v)$ is differentiable at 0; i.e., $u \mapsto y(\Omega; u) \circ (I + u)$ is differentiable at u_0. A similar argument holds for the function $u \mapsto p(\Omega; u) \circ (I + u)$. ∎

Remark. Theorem 4.1 in [1] asserts that, when Ω is a $W^{2,\infty}$ domain, the mapping $u \mapsto (y(u), p(u)) \circ (I + u)$ is well defined for $u \in W^{2,\infty}(\mathbb{R}^d, \mathbb{R}^d) \cap \mathcal{W}$ and differentiable at 0, with values in $H^2(\Omega)^d \times H^1(\Omega)$. Adapting the previous argument, we can deduce differentiability at each point in a $W^{2,\infty}$-open ball centered at 0. ∎

9. Higher-order differentiability. In this section, we will prove Theorems 6 and 4.

Proof of part (i) *of Theorem* 6. It remains to prove that $u \mapsto (Y(u), P(u))$ is a C^∞ mapping. (The remainder of part (i) has already been proven in section 6, Lemma 7.)

Observe that the mapping H, introduced in section 5 and defined from $\mathcal{W} \times H_0^1(\Omega)^d \times L_0^2(\Omega)$ into $H^{-1}(\Omega)^d \times L_0^2(\Omega)$, is C^∞. This is a consequence of the fact that $u \mapsto M_{ij}(u)$ and $u \mapsto \mathrm{Jac}(I + u)$ are C^∞ mappings. In turn, this stems from the following:

(a) The mapping $u \mapsto \mathrm{Jac}(I + u)$ is multilinear and, consequently, is of class C^∞.

(b) The mapping $u \mapsto M(u) = {}^t[\partial_i (I + u)_j]^{-1}$ is C^∞ on $\mathcal{W}$, because the inversion operator is indefinitely differentiable in the set of the nonsingular matrices.

From the implicit function theorem, we deduce that $u \mapsto (Y(u), P(u))$ possesses derivatives of all orders at 0. Again using (35), which can be written in the form

$$Y(\Omega; u_0 + u) = Y(\Omega + u_0; u \circ (I + u_0)^{-1}) \circ (I + u_0),$$

one also sees that $u \mapsto Y(\Omega; u)$ is C^∞ at each point $u_0 \in \mathcal{W}$. The same is true for $u \mapsto P(\Omega; u)$. ∎

Proof of part (ii). The differentiability of the mapping $u \mapsto y(u)|_\omega$ at 0 in $L^2(\omega)^d$ and the identity (20) are consequences of the differentiability of $u \mapsto y(u) \circ (I + u)$

given in Lemma 7. This is a consequence of general results on differentiation with respect to domains (see Lemma 2.1 in [15]). On the other hand, (12) and (20) together imply (16). ∎

Proof of Theorem 4. We have to check that $u \mapsto T(\Omega+u)$ is a C^∞ mapping. This is deduced from the above results and the following equality, which has already been used in section 6:

$$T(\Omega+u) = \frac{\nu}{2}\int_\Omega \sum_{ij}\Big(\sum_k (M_{ik}(u)\,\partial_k Y_j(\Omega;u) + M_{jk}(u)\,\partial_k Y_i(\Omega;u))\Big)^2 \mathrm{Jac}(I+u). \qquad \blacksquare$$

10. Miscellaneous remarks. *The case of a non-Lipschitz domain.* Until now, we have assumed that Ω is a Lipschitz domain in order to ensure, among other things, that (29) is true. Actually, this assumption on Ω can be replaced by (29) itself:

$$\forall \phi \in L^2(\Omega) \ \text{ such that } \ \int_\Omega \phi = 0, \ \text{ there exists } \ \psi \in H_0^1(\Omega)^d \ \text{ such that } \ \nabla \cdot \psi = \phi;$$

i.e., the divergence operator maps $H_0^1(\Omega)^d$ onto $L_0^2(\Omega)$.

Under this weaker hypothesis, the results in the previous sections hold again with minor changes. Instead of $p \in C^\infty(\Omega) \cap L^2(\Omega)$, we now have only

$$p \in C^\infty(\Omega), \quad \nabla p \in H^{-1}(\Omega)^d. \tag{36}$$

On the other hand, we cannot normalize p and $\dot{p}(u)$ as before. Instead, a possibility is to fix a nonempty open set $\omega \subset\subset \Omega$ and to impose

$$\int_\omega p = 0, \quad \int_\omega \dot{p}(u) = 0.$$

Remark. The condition (29) requires some regularity on Ω, which is probably not far from being Lipschitz. ∎

Remark. It is important to note that, here, the difficulty is not related to nonlinearity. Even if we were concerned with Stokes flows (the term $(y \cdot \nabla)\, y$ disappears), (36) could not be improved unless a regularity assumption is required for Ω. This difficulty is connected with the fact that the equations are coupled by the incompressibility condition $\nabla \cdot y = 0$. ∎

Remark. For more simple (scalar) problems, we can obtain a result similar to Theorem 4, without any regularity hypothesis for Ω. For example, let y be the unique solution to

$$-\Delta y = f \ \text{ in } \ \Omega, \qquad y - g \in H_0^1(\Omega)^d, \tag{37}$$

and let us set

$$S(\Omega) = \int_\Omega |\nabla(y-z)|^2,$$

where $f \in L^2(\Omega)^d$, $g \in H^2(\mathbb{R}^d)$, and $z \in H^1(\mathbb{R}^d)$ are given and Ω is an arbitrary bounded open set in $\mathbb{R}^d$. Then, $u \mapsto S(\Omega+u)$ is well defined and differentiable in a neighborhood of 0 in $W^{1,\infty}(\mathbb{R}^d,\mathbb{R}^d)$ [10, Theorem 5.2, p. V.10]. ∎

The particular case of a polygonal two-dimensional body. Assume that B is a two-dimensional polygonal domain with vertices $s_1, s_2, \ldots, s_n$. Let us set $s = (s_1, \ldots, s_n)$,

and let us assume that the corresponding polygonal line, ∂B, does not cross itself. Thus, using the notation $s_{n+1} = s_1$, one has

$$[s_i, s_{i+1}[\bigcap [s_j, s_{j+1}[= \emptyset \quad \text{if} \quad 1 \leq i < j \leq n. \tag{38}$$

Also, assume that

$$B \subset\subset \mathcal{O} \subset\subset D. \tag{39}$$

It is then obvious that $\Omega_s = D \backslash \overline{B}$ satisfies (1). In this situation, the following is not difficult to prove:

The mapping $s \mapsto T(\Omega_s)$ is C^∞ at each point $s \in \mathbb{R}^{2n}$ satisfying (38) *and* (39).

Other examples. Above, the polygonal domain can be replaced by a spline depending on a finite number of parameters. In such a way, we obtain similar results for "NACA profiles" or other piecewise C^1 boundaries. Similar results hold for three-dimensional domains.

11. Appendix. In order to prove Lemma 3, we need some previous definitions and results.

DEFINITION 14. *Let Ω be a bounded open set in $\mathbb{R}^d$.*

(i) *We say that* Ω is a Lipschitz domain (*also that Ω is Lipschitzian; see* [11], [5]) *if there exist constants $a > 0$ and $b > 0$ such that, for each $z \in \partial\Omega$, one can find*

— *coordinates $(x_1, \ldots, x_d)$,*

— *a Lipschitz-continuous real-valued function ψ in Θ_* with best Lipschitz constant smaller than b, where $\Theta_* = \{x_*;\ |x_* - z_*| < a\}$, $x_* = (x_1, \ldots, x_{d-1})$, and $z_* = (z_1, \ldots, z_{d-1})$,*

such that, for each $x \in \Theta = \{x \in \mathbb{R}^d;\ |x_* - z_*| < a,\ |x_d - \psi(x_*)| < a\}$, one has

$$x \in \Omega \Longleftrightarrow x_d > \psi(x_*).$$

(ii) *We say that* Ω satisfies the cone property uniformly *if there exist constants $\alpha > 0$ and $b > 0$ such that, for each $z \in \partial\Omega$, one can find coordinates such that*

$$x \in \Omega \cap B(z; \alpha) \Longrightarrow x + \mathcal{C}_{b,\alpha} \subset \Omega.$$

Here, we have set $B(z; \alpha) = \{x \in \mathbb{R}^d;\ |x - z| < \alpha\}$ and

$$\mathcal{C}_{b,\alpha} = \{x \in \mathbb{R}^d;\ x_d > b\,|x_*|,\ |x| < \alpha\}.$$

The properties (i) and (ii) are equivalent. More precisely, we have the following result (see [3]).

LEMMA 15. *A bounded open set in $\mathbb{R}^d$ is Lipschitzian if and only if it satisfies the cone property uniformly.*

The following result was also used in the proof of Lemma 3.

LEMMA 16. *Assume that $\alpha > 0$ and $b > 0$ are given. There exist $\alpha' > 0$, $b' > 0$, and $l \in (0, 1)$ such that, whenever $v \in W^{1,\infty}(\mathbb{R}^d, \mathbb{R}^d)$, $\|v\|_{W^{1,\infty}(\mathbb{R}^d,\mathbb{R}^d)} \leq l$, and $v(0) = 0$, one has*

$$\mathcal{C}_{b',\alpha'} \subset (I + v)\,\mathcal{C}_{b,\alpha}.$$

Proof of Lemma 3. From Lemma 15, there exist $\alpha > 0$ and $b > 0$ such that, for each $z \in \partial\Omega$, one has

$$x \in \Omega \cap B(z; \alpha) \Longrightarrow x + \mathcal{C}_{b,\alpha} \subset \Omega. \tag{40}$$

Again from Lemma 15, it is enough to find α' and b' such that, for each $z' \in \partial(\Omega+u)$,

$$x' \in (\Omega+u) \cap B(z';\alpha') \Longrightarrow x' + \mathcal{C}_{b',\alpha'} \subset \Omega + u. \tag{41}$$

Given such an x', let $\xi' \in \mathcal{C}_{b',\alpha'}$, and define x and z by $x' = x + u(x)$, $z' = z + u(z)$. Lemma 16 with $v(\xi) = u(\xi + x) - u(x)$ gives the existence of $\xi \in \mathcal{C}_{b,\alpha}$ such that $\xi' = \xi + u(\xi + x) - u(x)$. Then

$$x' + \xi' = x + \xi + u(x + \xi).$$

This gives (41), provided that $x + \xi \in \Omega$. By (40), it is enough to check that $x \in \Omega$ (which is obvious) and $|x - z| \leq \alpha$, which is satisfied for $\alpha' \leq \alpha(1 - c)$ (indeed, $x' - z' = x - z + u(x) - u(z)$ implies $|x' - z'| \geq |x - z|(1 - c)$). □

REFERENCES

[1] J.A. Bello, E. Fernández-Cara, and J. Simon, *Optimal shape design for Navier-Stokes flow*, in System Modelling and Optimization, Lecture Notes in Control and Inform. Sci. 180, P. Kall, ed., Springer-Verlag, Berlin, 1992, pp. 481–489.

[2] J.A. Bello, E. Fernández-Cara, and J. Simon, *Variation par rapport au domaine de l'énergie visqueuse dissipée dans un fluide de Navier-Stokes*, C. R. Acad. Sci. Paris Sér. I Math., 313 (1991), pp. 447–450.

[3] D. Chenais, *Un résultat de compacité d'un ensemble de parties de* $\mathbb{R}^n$, C. R. Acad. Sci. Paris Sér. A, 277 (1973), pp. 905–907.

[4] R. Dautray and J.L. Lions, *Analyse mathématique et calcul numérique pour les sciences et les techniques, tome* 1, Masson, Paris, 1984.

[5] D. Gilbarg and N.S. Trudinger, *Elliptic Partial Differential Equations of Second Order*, Springer-Verlag, Berlin, 1983.

[6] V. Girault and P.A. Raviart *Finite Element Methods for Navier–Stokes Equations*, Springer-Verlag, Berlin, 1983.

[7] O.A. Ladyzhenskaya, *The Mathematical Theory of Viscous Incompressible Flow*, Gordon and Breach, New York, 1969.

[8] J. Lemoine, *Couplage et contrôle pour les équations de Navier–Stokes: Espaces fonctionnels*, Thesis, Université Blaise Pascal (Clermont–Ferrand 2), 1995.

[9] F. Murat and J. Simon, *Quelques résultats sur le contrôle par un domaine géométrique*, Report of L.A. 189 74003, Université Paris VI, 1974.

[10] F. Murat and J. Simon, *Sur le contrôle par un domaine géométrique*, Report of L.A. 189 76015, Université Paris VI, 1976.

[11] J. Nečas, *Les méthodes directes en théorie des équations elliptiques*, Masson, Paris, 1967.

[12] O. Pironneau, *On optimum design in fluid mechanics*, J. Fluid. Mech., 64 (1974), pp. 97–110.

[13] O. Pironneau, *Optimal Shape Design for Elliptic Systems*, Springer-Verlag, New York, 1984.

[14] H. Schlichting, *Boundary Layer Theory*, Academic Press, New York, 1970.

[15] J. Simon, *Differentiation with respect to the domain in boundary value problems*, Numer. Funct. Anal. Optim., 2 (1980), pp. 649–687.

[16] J. Simon, *Second variation in domain optimization problems*, in Control and Estimation of Distributed Parameter Systems, Internat. Ser. Numer. Math. 91, F. Kappel, K. Kunisch, and W. Schappacher, eds., Birkhäuser, Basel, 1989, pp. 361–378.

[17] J. Simon, *Domain variation for drag in Stokes flow*, in Control Theory of Distributed Parameter Systems and Applications, Lecture Notes in Control and Inform. Sci. 159, X. Li and J. Yong, eds., Springer-Verlag, Berlin, 1991, pp. 28–42.

SIAM J. CONTROL OPTIM.
Vol. 35, No. 2, pp. 641–653, March 1997

015

PARTIAL EXACT CONTROLLABILITY FOR SPHERICAL MEMBRANES*

PAOLA LORETI† AND VANDA VALENTE†

Abstract. In this paper the partial exact controllability of an elastic spherical membrane is proved. The reachability problem for the integrodifferential equation, introduced for the vibrations of the meridional displacement, is solved. The main result is a generalization of a Ingham's theorem on nonharmonic Fourier series.

Key words. partial exact controllability, reachability problem, almost periodic functions

AMS subject classifications. 35Q99, 49E15

PII. S0363012994269624

Introduction. Following Love's and Koiter's linear shell theory [17], [10], in previous works [3], [4], [5], [6] we studied the problem of exact controllability for a thin elastic shell. The mathematical model is a system of partial differential equations where the unknown is the displacement vector $\mathbf{v}$ of the middle shell surface. We denoted by $\mathbf{A}^m$ and $\mathbf{A}^f$ the operators associated with the membrane energy and flexion energy, respectively, and by ε a small parameter depending on h (shell thickness) with $\lim_{h\to 0}\varepsilon = 0$. The spectrum behavior of $\mathbf{A} = \mathbf{A}^m + \varepsilon\mathbf{A}^f$ can be utilized to prove some results of exact controllability for the vibrations of thin shells. In the particular case of an hemispherical shell, the existence of an asymptotic gap for the eigenvalues λ_j of $\mathbf{A}$ allowed us to give a controllability time (depending on ε) and to prove exact controllability in suitable initial data spaces. In the general case we pointed out that when the thickness of the shell goes to zero, the number of eigenvalues of $\mathbf{A}$ less than a fixed $\lambda \geq \lambda^0$ goes to infinity; that is,

$$N_\lambda(\mathbf{A}) = \sum_{\lambda_j<\lambda} 1 \to \infty \quad \text{as} \quad h \to 0.$$

It suggests therefore that an accumulation point for the eigenvalues of the limit problem $\varepsilon = 0$ (the so-called membrane approximation) may occur; moreover, we proved that exact controllability of the limit problem generally fails, and an example of nonexact controllability is constructed in the case of hemispherical membrane approximation (see [6]).

In this paper we give a result of partial controllability for a spherical membrane; i.e., we want to control only one of the displacement components without conditions for the other components. The axially symmetric vibrations of a spherical membrane are described in section 1 by a pair of partial differential equations in the meridional and radial displacements u and w. The partial exact controllability problem is given in terms of the reachability problem for the integrodifferential equation for the vibrations of the meridional displacement. We propose the reverse or reachability Hilbert uniqueness method (RHUM) [11], [14], [15], [16] to construct our control function. In section 2, the well posedness of the corresponding homogeneous problem is proved. In section 4 we give a result of partial exact controllability (PEC) taking into account

*Received by the editors October 15, 1995; accepted for publication (in revised form) February 22, 1996.

http://www.siam.org/journals/sicon/35-2/26962.html

†Istituto per le Applicazioni del Calcolo, Consiglio Nationale delle Ricerche, viale del Policlinico, 137, 00161 Rome, Italy.

a generalization of the Ingham theorem [8], which we prove in section 3. Results on the reachability problem for plate equation with memory can be found in [9], [11], [12], [13].

1. Statement of the problem. We consider the axially symmetric vibrations of an elastic spherical membrane with middle ray R and opening angle θ_0. The meridional and radial components of the displacement vector $\mathbf{v} = (u(\theta,t),\ w(\theta,t))$ satisfy in $Q = (0,T)\times(0,\theta_0)$ the system

$$\begin{cases} du_{tt} - \mathcal{L}(u) - (1-\nu)u + (1+\nu)w' = 0, \\ dw_{tt} - \frac{(1+\nu)}{\sin\theta}(u\sin\theta)' + 2(1+\nu)w = 0 \end{cases} \tag{1.1}$$

with

$$u(0,t) = 0, \quad u(\theta_0,t) = g(t)$$

and

$$\mathbf{v}\,(\theta,0) = \mathbf{v}^{\,0}, \quad \mathbf{v}_t(\theta,0) = \mathbf{v}^{\,1},$$

where

$$\mathcal{L}(u) = \left(\frac{(u\sin\theta)'}{\sin\theta}\right)'. \tag{1.2}$$

The prime stands for the first derivative with respect to θ, $\nu \in (-1, 1/2)$, and $d = d_0\ R^2(1-\nu^2)/E$, where E is an elastic positive constant and d_0 is the density of the material. In what follows we change $t = \sqrt{d}\ t$. We introduce the following spaces:

$$L^2 = L^2(0,\theta_0;\sin\theta\,\mathrm{d}\theta) = \left\{ f : \int_0^{\theta_0} |f|^2 \sin\theta\,\mathrm{d}\theta < +\infty \right\},$$

$$\mathcal{U} = \left\{ u :\ \frac{\partial u}{\partial\theta}\,,\ u\cot\theta \in L^2(0,\theta_0;\sin\theta\,\mathrm{d}\theta)\ ,\quad u(0) = u(\theta_0) = 0 \right\}.$$

$\|f\|_0$ is the norm induced by the scalar product $(f,g)_0 = \int_0^{\theta_0} f\cdot g\sin\theta\,\mathrm{d}\theta$, and $\|u\|^2_{\mathcal{U}} = \|u'\|^2_0 + \|u\cot\theta\|^2_0$. The exact controllability problem requires that we find a control function $g(t)$ that drives the system to the rest in a finite time T. In [6] we proved that the membrane approximation is not exactly controllable for any $\{\mathbf{v}^{\,1}, \mathbf{v}^{\,0}\} \in (\mathcal{U}'\times L^2)\times\ (L^2\times L^2)$. We observed that for the hemispherical membrane there exists a subsequence of eigenfunctions $\mathbf{v}^{\,*}_n(\theta) = (u^*_n(\theta), w^*_n(\theta))$ such that

$$\lim_{n\to\infty} u^*_n(\pi/2) - (1+\nu)w^*_n(\pi/2) = 0.$$

Then the sequence $\{\mathbf{v}^{\,*}_n\ ,0\}$ with $\|\mathbf{v}^{\,*}_n\|_{\mathcal{U}\times L^2} = 1$ (initial data for the homogeneous problem associated with (1.1)) does not satisfy the necessary (and sufficient) condition of exact controllability.

Since the exact controllability for membrane approximation generally fails, we look for a partial result; i.e., we look for a PEC result.

The problem (1.1) is equivalently written in the form

$$u_{tt} - \mathcal{L}(u) - (1-\nu)u + (1+\nu)(w^0)' \cos t\sqrt{2(1+\nu)} + \frac{(1+\nu)}{\sqrt{2(1+\nu)}}(w^1)'$$
$$\cdot \sin t\sqrt{2(1+\nu)} + \frac{(1+\nu)^{3/2}}{\sqrt{2}} \int_0^t \sin((t-s)\sqrt{2(1+\nu)})\mathcal{L}(u(\theta,s))\,\mathrm{d}s = 0.$$

We consider the following partial controllability (PC) problem.

(PC) Given $T > 0$ and two functions $u^0(\theta)$, $u^1(\theta)$, find $g(t)$ such that the system

$$\begin{cases} u_{tt} - \mathcal{L}(u) - (1-\nu)u + (1+\nu)w' = 0, \\ w_{tt} - \frac{(1+\nu)}{\sin\theta}(u\sin\theta)' + 2(1+\nu)w = 0, \end{cases} \tag{1.3}$$

with boundary conditions

$$u(0,t) = 0, \quad u(\theta_0,t) = g(t) \tag{1.4}$$

and null initial data, i.e.,

$$\mathbf{v}\,(\theta,0) = 0, \quad \mathbf{v}_t(\theta,0) = 0, \tag{1.5}$$

verifies the final conditions

$$u(\theta,T) = u^0, \quad u_t(\theta,T) = u^1 \tag{1.6}$$

or, equivalently,

(PC)′ Given $T > 0$ and two functions $u^0(\theta)$, $u^1(\theta)$, find $g(t)$ such that the system

$$u_{tt} - \mathcal{L}(u) - (1-\nu)u + \frac{(1+\nu)^{3/2}}{\sqrt{2}} \int_0^t \sin((t-s)\sqrt{2(1+\nu)})\mathcal{L}(u(\theta,s))\,\mathrm{d}s = 0,$$

with

$$u(0,t) = 0, \quad u(\theta_0,t) = g(t)$$

and

$$\mathbf{v}\,(\theta,0) = 0, \quad \mathbf{v}_t(\theta,0) = 0,$$

verifies the conditions

$$u(\theta,T) = u^0, \quad u_t(\theta,T) = u^1.$$

To solve the problem (PC)′ we apply the RHUM method [11], [16].

We consider the adjoint system

$$z_{tt} - \mathcal{L}(z) - (1-\nu)z + \frac{(1+\nu)^{3/2}}{\sqrt{2}} \int_t^T \sin((s-t)\sqrt{2(1+\nu)})\mathcal{L}(z(\theta,s))\,\mathrm{d}s = 0 \tag{1.7}$$

with homogeneous boundary conditions

$$z(0,t) = 0, \quad z(\theta_0,t) = 0 \tag{1.8}$$

and final data

$$z(\theta,T)=z^0,\quad z_t(\theta,T)=z^1. \tag{1.9}$$

Then we consider the problem

$$\ddot{\phi}_{tt}-\mathcal{L}(\phi)-(1-\nu)\phi+\frac{(1+\nu)^{3/2}}{\sqrt{2}}\int_0^t \sin((t-s)\sqrt{2(1+\nu)})\mathcal{L}(\phi(\theta,s))\,\mathrm{d}s=0$$

with

$$\phi(0,t)=0,\quad \phi(\theta_0,t)=-\frac{(1+\nu)^{3/2}}{\sqrt{2}}\int_t^T \sin((s-t)\sqrt{2(1+\nu)})z'(\theta_0,s)\,\mathrm{d}s+z'(\theta_0,t)$$

and

$$\phi(\theta,0)=0,\quad \phi_t(\theta,0)=0.$$

We define $\mu\{z^0,z^1\}=\{\phi_t^0(T),-\phi_0(T)\}$ and put

$$G_z(\theta,t)=\int_0^t \sin((t-s)\sqrt{2(1+\nu)})z(\theta,s)\,\mathrm{d}s$$

and the adjoint operator of G by G^*, where

$$G_z^*(\theta,t)=\int_t^T \sin((s-t)\sqrt{2(1+\nu)})z(\theta,s)\,\mathrm{d}s. \tag{1.10}$$

With some simple computations we can prove

$$\langle\mu\{z^0,z^1\},\ \{z^0,z^1\}\rangle=\int_0^T \sin\theta_0\phi(\theta_0,t)z'(\theta_0,t)\,\mathrm{d}t$$
$$-\frac{(1+\nu)^{3/2}}{\sqrt{2}}\int_0^T G_{z'}^*(\theta_0,t)\phi(\theta_0,t)\sin\theta_0\,\mathrm{d}t.$$

Our aim is to prove that $\nu\{z^0,z^1\}$ is invertible in a suitable function space.

2. Analysis of the homogeneous problem. In order to prove existence and uniqueness results for (1.7), (1.8), and (1.9), we introduce the energy

$$E(t)=\frac{1}{2}\int_0^{\vartheta_0}(z_t^2+(z')^2+z^2\cot^2\vartheta+\nu z^2)\sin\vartheta\,\mathrm{d}\vartheta$$

and

$$E_T=\frac{1}{2}\int_0^{\vartheta_0}(z^{1^2}+(z^{0'})^2+(z^0)^2\cot^2\vartheta+\nu(z^0)^2)\sin\vartheta\,\mathrm{d}\vartheta.$$

We have the following proposition.

PROPOSITION 2.1. *There exist two constants $C_1(T,\nu)$ and $C_2(T,\nu)$ such that $E(t)\le C_1(T,\nu)E_T\cdot e^{C_2(T,\nu)(T-t)}$.*

Proof. To prove Proposition 2.1 we argue in similar way to [16, vol. 2, p. 242]. To this end we introduce

$$r_0(t) = \int_0^{\vartheta_0} \int_t^T \sin((s-t)\sqrt{2(1+\nu)})(z(\vartheta,s)\sin\vartheta)' \cdot \frac{1}{\sqrt{\sin\vartheta}}\,\mathrm{d}s[z(\vartheta,t)\sin\vartheta]'\frac{1}{\sqrt{\sin\vartheta}}\,\mathrm{d}\vartheta$$

and

$$r_1(t) = \int_0^{\vartheta_0} \int_t^T \cos((s-t)\sqrt{2(1+\nu)})(z(\vartheta,s)\sin\vartheta)' \cdot \frac{1}{\sqrt{\sin\vartheta}}\,\mathrm{d}s[z(\vartheta,t)\sin\vartheta]'\frac{1}{\sqrt{\sin\vartheta}}\,\mathrm{d}\vartheta.$$

Applying the Schwarz inequality after simple computations we obtain

$$|r_0(t)| \le 4\frac{E(t)}{\sqrt{27}} + \frac{\sqrt{27}}{16}\left\{\frac{T-t}{2} - \frac{1}{4}\frac{\sin 2(T-t)\sqrt{2(1+\nu)}}{\sqrt{2(1+\nu)}}\right\}\int_t^T E(s)\,\mathrm{d}s,$$

and similarly we obtain

$$|r_1(t)| \le E(t) + \frac{1}{4}\left(\frac{T-t}{2} + \frac{1}{4}\frac{\sin 2(T-t)\sqrt{2(1+\nu)}}{\sqrt{2(1+\nu)}}\right)\int_t^T E(s)\,\mathrm{d}s.$$

On the other hand, it is easy to show that

$$\frac{d}{dt}\left(E(t) + \frac{\sqrt{2}}{2}\sqrt{(1+\nu)^3}r_0(t)\right) = -(1+\nu)^2 r_1(t).$$

Hence,

$$E(t) \le E_T + \frac{\sqrt{2}}{2}\sqrt{(1+\nu)^3}|r_0(t)| + (1+\nu)^2\int_t^T |r_1(s)|\,\mathrm{d}s.$$

We compute

$$\int_t^T |r_1(s)|\,\mathrm{d}s \le \left(1 + \frac{1}{4}\int_t^T M_1^2(s)\,ds\right)\int_t^T E(s)\,\mathrm{d}s,$$

where

$$M_1^2 = \frac{T-t}{2} + \frac{1}{4}\frac{\sin 2(T-t)\sqrt{2(1+\nu)}}{\sqrt{2(1+\nu)}},$$

$$E(t) \le E_T + 2\sqrt{2(1+\nu)^3}\frac{E(t)}{\sqrt{27}} + \sqrt{27}\frac{1}{64}\left(T + \frac{1}{2\sqrt{2(1+\nu)}}\right)\sqrt{2(1+\nu)^3}\int_t^T E(s)\,\mathrm{d}s + (1+\nu)^2\left(1 + \frac{1}{4}\int_t^T M_1^2(s)\,\mathrm{d}s\right)\int_t^T E(s)\,\mathrm{d}s.$$

It is straightforward to find

$$\int_t^T M_1^2(t)\,\mathrm{d}t \le \frac{T^2}{4} + \frac{1}{8(1+\nu)}.$$

From the above estimates we conclude choosing

$$c_1(T,\nu) = \frac{E_T}{\left(1 - \frac{\sqrt{2}}{2}\sqrt{(1+\nu)^3}\frac{2}{\sqrt{27}}\right)},$$

$$c_2(T,\nu) = \left((1+\nu)^2\frac{T^2}{16} + \sqrt{27}\frac{1}{64}\sqrt{2(1+\nu)^3}T + \left(\sqrt{27}\frac{1}{128} + (1+\nu) + \frac{1}{8}\right)\right)(1+\nu). \qquad \square$$

From the above result we obtain the following.

PROPOSITION 2.2. *The problem* (1.2) *has a unique solution such that*

$$\{z, z'\} \in C([0,T];\ \mathcal{U} \times L^2).$$

Now we shall find the explicit solution of the equation (1.2) with boundary homogeneous conditions and the final state z^0, z^1.

So we consider the eigenvalues problem

$$-\mathcal{L}(z_k) = \lambda_k z_k, \tag{2.1}$$

where $\mathcal{L}$ is the operator given in (1.2) and $z_k(0) = z_k(\vartheta_0) = 0$.

We assume that the solution of our problem can be written as

$$z(\vartheta,t) = \sum_{k=1}^{+\infty} f_k(t) z_k(\vartheta), \tag{2.2}$$

where z_k is an orthonormal base in $(0,\theta_0)$.

Substituting z, given by (2.2), using the eigenvalue problem (2.1), and multiplying the equation by z_k we find the following integrodifferential equation in the unknown f (here the dot denotes the derivative with respect to t):

$$\ddot{f}_k(t) + (\lambda_k - 1 + \nu) f_k(t) - \lambda_k \frac{(1+\nu)^2}{\sqrt{2(1+\nu)}} \int_0^t \sin((t-s)\sqrt{2(1+\nu)}) f_k(s)\,\mathrm{d}s = 0.$$

Using [18, p. 149] and the method of Evans [18, p. 67], we find the solution

$$\begin{aligned} f_k(t) = {} & (f_k^T + \dot{f}_k^T t)\left(1 + \frac{b_k^1}{a_k^+} + \frac{b_k^2}{a_k^-}\right) - f_k^T\left(\frac{b_k^1}{a_k^+}\cos(a_k^+ t) + \frac{b_k^2}{a_k^-}\cos(a_k^- t)\right) \\ & - \dot{f}_k^T\left(\frac{b_k^1}{(a_k^+)^2}\sin(a_k^+ t) + \frac{b_k^2}{(a_k^-)^2}\sin(a_k^- t)\right) \end{aligned}$$

and b_k^1, b_k^2, a_k^+, a_k^- given, respectively, by

$$b_k^1 = -\frac{(\lambda_k - 1 + \nu)[(a_k^-)^2 - (\lambda_k - 1 + \nu)] - \lambda_k(1+\nu)^2}{a_k^+[(a_k^-)^2 - (a_k^+)^2]}, \tag{2.3}$$

$$b_k^2 = \frac{(\lambda_k - 1 + \nu)[(a_k^+)^2 - (\lambda_k - 1 + \nu)] - \lambda_k(1+\nu)^2}{a_k^-[(a_k^-)^2 - (a_k^+)^2]}, \tag{2.4}$$

$$a_k^{\pm} = \sqrt{\frac{(\lambda_k + 1 + 3\nu) \pm \sqrt{(\lambda_k + 1 + 3\nu)^2 - 4(1-\nu^2)(\lambda_k - 2)}}{2}}. \tag{2.5}$$

Doing the substitution $t = T - t$ in (2.2) we find the solution of (1.7)–(1.9): $z = \sum_{k=1}^{+\infty} f_k(t) z_k(\theta)$ with

$$\begin{aligned} f_k(t) = {} & (f_k^T - \dot{f}_k^T(T-t)) \left(1 + \frac{b_k^1}{a_k^+} + \frac{b_k^2}{a_k^-}\right) \\ & - f_k^T \left(\frac{b_k^1}{a_k^+}\cos(a_k^+(T-t)) + \frac{b_k^2}{a_k^-}\cos(a_k^-(T-t))\right) \\ & + \dot{f}_k^T \left(\frac{b_k^1}{(a_k^+)^2}\sin(a_k^+(T-t)) + \frac{b_k^2}{(a_k^-)^2}\sin(a_k^-(T-t))\right). \end{aligned}$$

Taking into account that

$$\frac{b_k^1}{a_k^+} + \frac{b_k^2}{a_k^-} = -1 \quad \forall\, \nu \quad \forall\, k,$$

we have

$$\begin{aligned} f_k(t) = {} & -f_k^T \left\{\frac{b_k^1}{a_k^+}\cos(a_k^+(T-t)) + \frac{b_k^2}{a_k^-}\cos(a_k^-(T-t))\right\} \\ & + \dot{f}_k^T \left\{\frac{b_k^1}{a_k^{+2}}\sin(a_k^+(T-t)) + \frac{b_k^2}{(a_k^-)^2}\sin(a_k^-(T-t))\right\}. \end{aligned}$$

PROPOSITION 2.3. *The following properties of the coefficients a_k^+ and a_k^- hold:*

$$(a_k^+)^2 = \lambda_k - 1 + \nu + c_k^+(1+\nu)^2,$$

$$(a_k^-)^2 = 2(1+\nu) + c_k^-(1+\nu)^2,$$

with

$$\lim_{k\to\infty} c_k^+ = +1, \qquad \lim_{k\to\infty} c_k^- = -1,$$

$$\lim_{k\to\infty} a_n^+ = +\infty, \qquad \lim_{k\to\infty} a_k^- = \sqrt{1-\nu^2}.$$

Remark 2.1. The solution of this problem can be easily given at least in the case $\vartheta_0 = \frac{\pi}{2}$. More precisely we have $z_k = a_k P_k'$ and $\lambda_k = k(k+1)$ with $k = 2, 4, 6, \ldots$, and P_k is the k–Legendre polynomial and a_k is the normalization factor; i.e., $(a_k P_k', a_j P_j') = \delta_j^k$.

3. Trigonometrical inequalities for almost periodic functions. The following theorem, on almost periodic functions [2], is a generalization of a result due to Ingham (see [8, Theorem 1]).

THEOREM 3.1. *Let*

$$f(t) = \sum_{n=-\infty}^{+\infty} A_n^+ e^{-ia_n^+ t} + A_n^- e^{-ia_n^- t} = \sum_n A_n^+ e^{-ia_n^+ t} + A_n^- e^{-ia_n^- t}, \tag{3.1}$$

where we assume $\sum_n A_n^+ e^{-ia_n^+ t} + A_n^- e^{-ia_n^- t}$ *is uniformly convergent in* $[-T, T]$ *and*

$$\exists \gamma_1 > 0 : |a_n^+ - a_{n-1}^+| \geq \gamma_1 \quad \forall\, n, \tag{3.2}$$

$$\exists \gamma_2 > 0 : |a_n^+ - a_j^-| \geq \gamma_2 \quad \forall\, n\ \forall\, j, \tag{3.3}$$

$$\exists \alpha \geq 1\,, C_2 > 0 :\ \forall\, k \qquad |A_k^-| \leq \frac{C_2}{k^\alpha}\,|A_k^+|. \tag{3.4}$$

Then $\exists\, T_0 : \forall\, T > T_0 : \exists C_3(T)$, $C_4(T) > 0$ *such that*

$$C_3(T) \sum_n |A_n^+|^2 \leq \int_{-T}^{T} |f|^2 \,\mathrm{d}t \leq C_4(T) \sum_n |A_n^+|^2.$$

Proof. We prove the first inequality. Let $h(t)$ be a nonnegative integrable function over $(-\infty, +\infty)$. We consider

$$\begin{aligned}
&\int_{-\infty}^{+\infty} h(t) |f(t) - \sum_k A_k^+ e^{-ia_k^+ t}|^2 \,\mathrm{d}t \\
&\quad = \int_{-\infty}^{+\infty} h(t) \left(f(t) - \sum_k A_k^+ e^{-ia_k^+ t} \right) \cdot \left(\overline{f}(t) - \sum_j \overline{A}_j^+ e^{ia_j^+ t} \right) \mathrm{d}t \\
&\quad = \int_{-\infty}^{+\infty} h(t) |f(t)|^2 \,\mathrm{d}t + \int_{-\infty}^{+\infty} \sum_k \sum_j A_k^+ \overline{A}_j^+ e^{i(a_j^+ - a_k^+)t} h(t) \,\mathrm{d}t \\
&\quad - \int_{-\infty}^{+\infty} h(t) \overline{f}(t) \cdot \sum_k A_k^+ e^{-ia_k^+ t} \,\mathrm{d}t - \int_{-\infty}^{+\infty} h(t) f(t) \sum_j \overline{A}_j^+ e^{+ia_j^+ t} \,\mathrm{d}t \\
&\quad = \int_{-\infty}^{+\infty} h(t) |f(t)|^2 \,\mathrm{d}t + \sum_k \sum_j A_k^+ \overline{A}_j^+ K(a_k^+ - a_j^+) \\
&\quad - \int_{-\infty}^{+\infty} h(t) \left(\sum_j \overline{A}_j^+ e^{ia_j^+ t} + \overline{A}_j^- e^{ia_j^- t} \right) \cdot \sum_k A_k^+ e^{-ia_k^+ t} \,\mathrm{d}t \\
&\quad - \int_{-\infty}^{+\infty} h(t) \left(\sum_k A_k^+ e^{-ia_k^+ t} + A_k^- e^{-ia_k^- t} \right) \cdot \sum_j \overline{A}_j^+ e^{ia_j^+ t} \,\mathrm{d}t,
\end{aligned}$$

where

$$K(u) = \int_{-\infty}^{+\infty} e^{-itu} \cdot h(t) \,\mathrm{d}t.$$

Hence, using $\sum_k \sum_j'$ to denote the sum for $k \neq j$,

$$\begin{aligned} 0 \leq & -\sum_k |A_k^+|^2 K(0) - \sum_k \sum_j{}' A_k^+ \overline{A}_j^+ K(a_k^+ - a_j^+) \\ & - \sum_j \sum_k \overline{A}_j^- A_k^+ K(a_k^+ - a_j^-) \\ & - \sum_k \sum_j A_k^- \overline{A}_j^+ K(a_k^- - a_j^+) + \int_{-\infty}^{+\infty} h(t)|f|^2 \, \mathrm{d}t, \end{aligned}$$

we take

$$h(t) = \begin{cases} \cos \frac{\pi t}{2T}, & |t| \leq T, \\ 0, & |t| > T, \end{cases}$$

with $T > \frac{\pi}{2\overline{\gamma}}$ and $\overline{\gamma} = \min\{\gamma_1, \gamma_2\}$

$$\int_{-\infty}^{+\infty} h(t) e^{i(a_k^+ - a_j^+)t} \, \mathrm{d}t = \begin{cases} \frac{4T\pi \cos(a_k^+ - a_j^+)T}{\pi^2 - 4T^2(a_k^+ - a_j^+)^2}, & k \neq j, \\ \frac{4T}{\pi}, & k = j. \end{cases}$$

By assumption (3.2)

$$\sum_k \left| \frac{4T\pi \cos(a_k^+ - a_j^+)}{\pi^2 - 4T^2(a_k^+ - a_j^+)^2} \right| \leq \frac{4\pi}{T\gamma_1^2} \leq \frac{4\pi}{T\overline{\gamma}^2}$$

and by assumption (3.3)

$$\sum_k \left| \frac{4T\pi \cos(a_k^+ - a_j^-)}{\pi^2 - 4T^2(a_k^+ - a_j^-)^2} \right| \leq \frac{4\pi}{T\gamma_2^2} \leq \frac{4\pi}{T\overline{\gamma}^2}$$

so that

$$\begin{aligned} \int_{-T}^{T} |f(t)|^2 \, \mathrm{d}t &\geq \frac{4T}{\pi} \sum_k |A_k^+|^2 + \sum_k \sum_j{}' A_k^+ \overline{A}_j^+ K(a_k^+ - a_j^+) \\ &\quad + \sum_k \sum_j A_k^- \overline{A}_j^+ K(a_k^- - a_j^+) + \sum_j \sum_k \overline{A}_j^- A_k^+ K(a_k^+ - a_j^-) \\ &\geq \frac{4T}{\pi} \sum_k |A_k^+|^2 - \frac{4\pi}{T\gamma_1^2} \sum_k |A_k^+|^2 \\ &\quad - C_2 \sum_k \sum_j |A_k^+||A_j^+| \frac{|K(a_k^- - a_j^+)|}{k^\alpha} - C_2 \sum_k \sum_j \frac{|A_j^+|}{j^\alpha} |A_k^+||K(a_k^+ - a_j^-)| \\ &\geq \sum_k |A_k^+|^2 \left(\frac{4T}{\pi} - \frac{4\pi}{T\gamma_1^2} \right) - \sum_k \sum_j |A_k^+|^2 |K(a_j^+ - a_k^-)| \\ &\quad - C_2^2 \sum_k \sum_j |A_j^+|^2 \frac{|K(a_j^+ - a_k^-)|}{k^{2\alpha}}. \end{aligned}$$

If we put $S = C_2^2 \sum_J \frac{1}{J^{2\alpha}}, \alpha \geq 1$, we have

$$\int_{-T}^{T} |f(t)|^2 \, \mathrm{d}t \geq \left(\frac{4T}{\pi} - \frac{4\pi}{T\gamma_1^2} - \frac{4\pi}{T\gamma_2^2} - \frac{S4\pi}{T\gamma_2^2} \right) \cdot \sum_k |A_k^+|^2.$$

This inequality is verified for any $T > T_0$ with

$$T_0 = \left\{ \inf T : \frac{4T}{\pi} - \frac{4\pi}{T\gamma_1^2} - \frac{4\pi}{T\gamma_2^2} - \frac{S4\pi}{T\gamma_2^2} > 0 \right\} = \pi \sqrt{\frac{1+S}{\gamma_2^2} + \frac{1}{\gamma_1^2}}. \tag{3.5}$$

Next, we prove the second inequality.

We compute

$$\begin{aligned}
&\int_{-T}^{T} |f(t)|^2 h(t) \, \mathrm{d}t \\
&\quad = \int_{-T}^{T} \left(\sum_k A_k^+ e^{-ia_k^+ t} + A_k^- e^{-ia_k^- t} \right) \cdot \left(\sum_j \overline{A}_j^+ e^{+ia_j^+ t} + \overline{A}_j^- e^{ia_j^- t} \right) h(t) \, \mathrm{d}t \\
&= \int_{-T}^{T} \sum_k \sum_j A_k^+ \overline{A}_j^+ e^{i(a_j^+ - a_k^+)t} h(t) \, \mathrm{d}t + \int_{-T}^{T} \sum_k \sum_j A_k^- \overline{A}_j^- e^{i(a_j^- - a_k^-)t} \cdot h(t) \, \mathrm{d}t \\
&+ \int_{-T}^{T} \sum_k \sum_j A_k^+ \overline{A}_j^- e^{i(a_j^- - a_k^+)t} \cdot h(t) \, \mathrm{d}t + \int_{-T}^{T} \sum_k \sum_j A_k^- \overline{A}_j^+ e^{i(a_j^+ - a_k^-)t} \cdot h(t) \, \mathrm{d}t,
\end{aligned}$$

$$\begin{aligned}
\int_{-T}^{T} |f|^2 h(t) \, \mathrm{d}t &= \sum_k \sum_j A_k^+ \overline{A}_j^+ K(a_k^+ - a_j^+) \\
&\quad + \sum_k \sum_j A_k^- \overline{A}_j^- K(a_k^- - a_j^-) + \sum_k \sum_j A_k^+ \overline{A}_j^- K(a_k^+ - a_j^-) \\
&\quad + \sum_k \sum_j A_k^- \overline{A}_j^+ K(a_k^- - a_j^+) \leq \sum_k \sum_j |A_k^+| |A_j^+| |K(a_k^+ - a_j^+)| \\
&\quad + \sum_k \sum_j \frac{|A_k^+|}{k^\alpha} \frac{|A_j^+|}{j^\alpha} |K(a_k^- - a_j^-)| + 2 \sum_k \sum_j |A_k^+| \frac{|A_j^+|}{j^\alpha} |K(a_k^+ - a_j^-)|
\end{aligned}$$

$$\begin{aligned}
&\leq \left(\frac{4T}{\pi} + \frac{4\pi}{T\overline{\gamma}^2} \right) \sum_k |A_k^+|^2 + \frac{4T}{\pi} C_2^2 \left\{ \sum_k |A_k^+|^2 \sum_j \frac{1}{j^{2\alpha}} + \sum_j |A_j^+|^2 \sum_k \frac{1}{k^{2\alpha}} \right\} \\
&\quad + C_2^2 \sum_k |A_k^+|^2 \sum_j \frac{|K(a_j^- - a_k^+)|}{j^{2\alpha}} + \sum_j |A_j^+|^2 \sum_k |K(a_j^- - a_k^+)| \\
&\leq \left(\frac{4T}{\pi} + \frac{4\pi}{T\overline{\gamma}^2} + \frac{8T\ S}{\pi} \right) \sum_k |A_k^+|^2 + \left(\frac{4\pi}{T\overline{\gamma}^2} S + \frac{4\pi}{T\overline{\gamma}^2} \right) \sum_k |A_k^+|^2 = \overline{C}_4(T) \sum_k |A_k^+|^2.
\end{aligned}$$

We conclude the proof choosing $C_4(T) = \frac{2}{\sqrt{2}} \overline{C}_4(2T)$. □

Remark 3.1. In an analogous way to [1] (see also [7]), the assumptions (3.2) and (3.3) can be relaxed with

$$\exists \gamma_1 > 0 : \quad |a_{k+1}^+ - a_k^+| \geq \gamma_1 \quad \forall |k| > K, \tag{3.2'}$$

$$\exists \gamma_2 > 0 : \quad |a_k^+ - a_j^-| \geq \gamma_2 \qquad \forall |k| > K \quad \text{and} \quad \forall j. \tag{3.3'}$$

If $\gamma_2 \to +\infty$ as $K \to +\infty$ (as in the case of spherical membranes, see Proposition 2.3), we find from (3.5) that the proof of Theorem 3.1 holds with

$$T_0 = \frac{\pi}{\overline{\gamma}}. \tag{3.6}$$

4. Controllability. In what follows we put $\alpha = \frac{(1+\nu)^{3/2}}{\sqrt{2}}$.

LEMMA 4.1. *For every fixed $T > 0$, there exist two positive constants $c_0 = c_0(T, \alpha)$ and $c_1 = c_1(T, \alpha)$ such that*

$$c_0 \int_0^T (z' - \alpha G_{z'}^*(\theta_0, t))^2 \, \mathrm{d}t \leq \int_0^T (z'(\theta_0, t))^2 \, \mathrm{d}t \leq c_1 \int_0^T (z' - \alpha G_{z'}^*(\theta_0, t))^2 \, \mathrm{d}t.$$

Proof. The left inequality follows by simple computations. To prove the right inequality we assume that there exists a sequence $\tilde{z}_n'$ such that $\|\tilde{z}_n'\|_{L^2} = 1$ and

$$\tilde{z}_n'(t) - \alpha G^* \tilde{z}_n'(t) = \tilde{f}_n(t), \tag{4.1}$$

with

$$\lim_{n \to \infty} \|\tilde{f}_n\|_{L^2} = 0.$$

By simple computation [18, p. 45] we have by (4.1),

$$\tilde{z}_n'(t) = \tilde{f}_n(t) - \int_t^T Q(s,t) \tilde{f}_n(s) \, \mathrm{d}s, \tag{4.2}$$

where Q is a reciprocal kernel of (4.1), (1.10) given by

$$Q(s,t) = \sum_{h=1}^{\infty} P^{(h)}(s,t),$$

where

$$|P^{(h)}(s,t)| \leq \frac{|\alpha|^h |s-t|^{h-1}}{(h-1)!}.$$

It is easy to prove that

$$|Q(s,t)| \leq |\alpha| e^{|\alpha| \, |s-t|}.$$

On the other hand,

$$\int_t^T Q^2(s,t) \, \mathrm{d}s \leq \alpha^2 \int_t^T e^{2|\alpha| \, |s-t|} \, \mathrm{d}s = \frac{\alpha^2}{2|\alpha|} (e^{(T-t) \cdot 2|\alpha|} - 1)$$

and

$$\int_0^T \int_t^T Q^2(s,t)\,\mathrm{d}s \le \frac{|\alpha|}{2}\int_0^T (e^{(T-t)2|\alpha|}-1)\,\mathrm{d}t = \frac{1}{4}e^{2T|\alpha|} - \frac{1}{4} - \frac{|\alpha|}{2}T,$$

so that

$$\int_0^T (\tilde{z}_n'(\theta_0,t)^2\,\mathrm{d}t \le 2\int_0^T (\tilde{f}_n(s))^2\,\mathrm{d}s\left\{\frac{3}{4}+\frac{1}{4}e^{2T|\alpha|}-\frac{|\alpha|}{2}T\right\}.$$

Hence we find a contradiction. □

THEOREM 4.1. *If the generalized assumptions of Theorem* 3.1 *are verified, then we can solve the partial controllability problem in a suitable space.*

Proof. We take

$$A_k^+ = \frac{b_k^1}{(a_k^+)^2}\, z_k'(\theta_0)\left[a_k^+ f_k^T + \frac{1}{i}\,\dot{f}_k^T\right],$$

$$A_k^- = \frac{b_k^2}{(a_k^-)^2}\, z_k'(\theta_0)\left[a_k^- f_k^T + \frac{1}{i}\,\dot{f}_k^T\right],$$

where a_k^+, a_k^-, b_k^1, and b_k^2 are given by (2.3), (2.4), and (2.5). By Lemma 4.1 and Theorem 3.1 we have that there exist two positive constants $C_1(T)$ and $C_2(T)$ such that

$$C_1(T)\sum_k |A_k^+|^2 \le \langle \mu(z_0,z_1),(z_0,z_1)\rangle \le C_2(T)\sum_k |A_k^+|^2.$$

This ends the proof. □

Example 4.1 (PEC for the hemispherical shell).

For $\theta_0 = \pi/2$, we have

$$\lambda_k = 2k(2k+1), \qquad k = 1,2,\ldots,$$

$$a_k^+ \sim 2k+\frac{1}{2} \qquad \text{as} \quad k\to\infty,$$

$$|a_{k+1}^+ - a_k^+| = 2\ , \qquad \lim_{k\to\infty} a_k^- = \sqrt{1-\nu^2} < a_1^+.$$

Moreover,

$$\frac{b_k^1}{(a_k^+)^2} \sim \frac{\text{const}}{k} \quad \text{as} \qquad k\to\infty$$

and

$$\frac{b_k^2}{(a_k^-)^2} \sim \frac{\text{const}}{k^2} \quad \text{as} \quad k\to\infty.$$

Moreover, we assume the data $\{f^T, \dot{f}^T\}$ are given in a suitable space in order for $\sum_k |A_k^+| < +\infty$. Hence the hypotheses of Theorem 3.1 are verified, and we can apply Theorem 4.1.

REFERENCES

[1] J. M. Ball and M. Slemrod, *Nonharmonic Fourier series and the stabilization of distributed semi–linear control systems*, Comm. Pure Appl. Math., 37 (1979), pp. 555–587.

[2] A. S. Besicovitch, *Almost Periodic Functions*, Cambridge University Press, Cambridge, UK, 1932.

[3] G. Geymonat, P. Loreti, and V. Valente, *Controllabilité exacte d'une modele de coque mince*, C. R. Acad. Sci. Paris Sér. I Math., 313 (1991), pp. 81–86.

[4] G. Geymonat, P. Loreti, and V. Valente, *Exact controllability of a shallow shell model*, in Optimization, Optimal Control and Partial Differential Equations (Iaşi, 1992), International Series of Numerical Mathematics 107, Birkhäuser Verlag, Basel, 1992, pp. 85–97.

[5] G. Geymonat, P. Loreti, and V. Valente, *Exact controllability of a thin elastic hemispherical shell via harmonic analysis*, in Boundary Value Problems for Partial Differential Equations and Applications, Masson, Paris, 1993.

[6] G. Geymonat, P. Loreti, and V. Valente, *Spectral problem for thin shells and exact controllability*, in Spectral Analysis of Complex Structures, 49, Hermann, Paris, 1995, pp. 35–59.

[7] A. Haraux, *Series lacunaires et controle semi-interne des vibrations d'une plaque rectangulaire*, J. Math. Pures Appl., 68 (1989), pp. 457–465.

[8] A. E. Ingham, *Some trigonometrical inequalities with applications to the theory of series*, Math. Z., 41 (1936), pp. 367–379.

[9] J. U. Kim, *Control of a plate equation with large memory*, Differential Integral Equations, 5 (1992), pp. 261–279.

[10] W. T. Koiter, *On the foundations of the linear theory of thin elastic shells*, Proc. Kon. Nederl. Akad. Wetensch., B73 (1970), pp. 169–195.

[11] J. E. Lagnese and J. L. Lions, *Modeling Analysis and Control of Thin Plates*, Masson, Paris, 1988.

[12] I. Lasiecka, *Controllability of a viscoelastic Kirchhoff plate*, in Control and Estimation of Distributed Parameter Systems (Vorau, 1988), International Series of Numerical Mathematics 91, Birkhäuser Verlag, Basel, 1989, pp. 237–247.

[13] G. Leugering, *Exact boundary controllability of an integro–differential equation*, Appl. Math. Optim., 15 (1987), pp. 223–250.

[14] J. L. Lions, *Controllabilité exacte des systemes distribues*, C. R. Acad. Sci. Paris Sér. I Math., 302 (1986), pp. 471–475.

[15] J. L. Lions, *Exact controllability, stabilization and perturbations for distributed systems*, SIAM Rev., 30 (1988), pp. 1–68.

[16] J. L. Lions, *Controllabilité exacte perturbation et stabilization des systemes distribues*, 1 and 2, Masson, Paris, 1988.

[17] E. H. A. Love, *A Treatise on the Mathematical Theory of Elasticity*, Cambridge University Press, Cambridge, UK, 1927.

[18] V. Volterra, *Theory of Functionals and of Integral and Integro–differential Equations*, Dover Publications, New York, 1959.

SIAM J. CONTROL OPTIM.
Vol. 35, No. 2, pp. 654–677, March 1997

016

DYNAMICS FOR CONTROLLED NAVIER–STOKES SYSTEMS WITH DISTRIBUTED CONTROLS*

L. S. HOU† AND Y. YAN‡

Abstract. The long-time behavior of solutions for an optimal distributed control problem associated with the Navier–Stokes equations is studied. First, a linear feedback solution for the Navier–Stokes equations is constructed; this feedback solution possesses decay (in time) properties. Then, some preliminary estimates for the long-time behavior of all solutions of the Navier–Stokes equations are derived. Next, the existence of a solution for the optimal control problem is proved. Finally, the long-time decay properties for the optimal solutions are established.

Key words. dynamics, optimal control, feedback control, Navier–Stokes equations

AMS subject classifications. 35B40, 49J20, 49K20, 76D05, 93B52

PII. S0363012994274926

1. Introduction.

In this article we study the long-time behavior of the solutions for optimal control problems associated with the Navier–Stokes equations on the infinite time interval. In this paper, we concentrate on the following prototype problem: minimize the functional

$$\mathcal{J}(\mathbf{u},\mathbf{f}) = \frac{\alpha}{2}\int_0^\infty\int_\Omega |\mathbf{u}-\mathbf{U}|^2\,d\mathbf{x}\,dt + \frac{\beta}{2}\int_0^\infty\int_\Omega |\mathbf{f}-\mathbf{F}|^2\,d\mathbf{x}\,dt \tag{1.1}$$

subject to the Navier–Stokes equations

$$\partial_t\mathbf{u} - \nu\Delta\mathbf{u} + (\mathbf{u}\cdot\nabla)\mathbf{u} + \nabla p = \mathbf{f} \qquad \text{in } \Omega\times(0,\infty)\,, \tag{1.2}$$

$$\operatorname{div}\mathbf{u} = 0 \qquad \text{in } \Omega\times(0,\infty)\,, \tag{1.3}$$

$$\mathbf{u} = \mathbf{0} \qquad \text{on } \partial\Omega\times(0,\infty), \tag{1.4}$$

and

$$\mathbf{u}(\cdot,0) = \mathbf{u}_0 \qquad \text{in } \Omega\,. \tag{1.5}$$

Here, $\alpha,\beta>0$ are given constants, Ω is a bounded domain in $\mathbb{R}^2$ which is of class C^2 or convex, $\partial\Omega$ denotes the boundary of Ω, $\mathbf{U}$ is a given desired flow field, and $\mathbf{F}$ is a given body force. Also, $\mathbf{f}$ is the distributed control (body force), and $(\mathbf{u},p)$ denote the velocity field and the pressure field. The first term in the functional (1.1) measures the $L^2(0,\infty;\mathbf{L}^2(\Omega))$-distance between the candidate flow and the desired flow. Thus, the physical objective of this minimization problem is to match a desired flow field (in the L^2 sense) by adjusting (controlling) the body force $\mathbf{f}$. The second term in the functional measures the size of the control with respect to some fixed force

*Received by the editors September 30, 1994; accepted for publication (in revised form) February 23, 1996.

http://www.siam.org/journals/sicon/35-2/27492.html

†Department of Mathematics and Statistics, York University, North York, ON M3J 1P3, Canada (hou@mathstat.yorku.ca). The research of this author was supported in part by Natural Science and Engineering Research Council of Canada grant OGP-0169786.

‡Department of Mathematics, Virginia Polytechnic Institute and State University, Blacksburg, VA 24061 (yan@math.vt.edu). The research of this author was supported in part by National Science Foundation grant DMS-9626154.

$\mathbf{F}$. The inclusion of this term in the functional will keep the body force $\mathbf{f}$ (i.e., the control) within a reasonable distance from $\mathbf{F}$ so that the optimal $\mathbf{f}$ we find can still be physically realized. This functional reflects a trade-off between achieving a physical objective and minimizing the work. It is worth mentioning that $(\mathbf{U}, \mathbf{F})$ in general is not an optimal solution, for $\mathbf{U}(\cdot, 0) \not\equiv \mathbf{u}_0$ in general.

Through the change of variables $\mathbf{v} = \mathbf{u} - \mathbf{U}$ and $\mathbf{g} = \mathbf{f} - \mathbf{F}$, we may interpret the optimal control problem from another physical point of view; i.e., one seeks a candidate flow $\mathbf{v}$ and a candidate body force $\mathbf{g}$ such that the $L^2(\Omega \times (0, \infty))$-energy and the total work done by the body force in the flow's entire life span is minimized. Many other physical cost functionals such as the drag functional can be formulated in a similar optimal control setting wherein one minimizes the total cost such as the total drag and the total work done by the body force in the flow's entire life span.

If our goal is only to match the desired flow $\mathbf{U}$ by the designed flow $\mathbf{u}$ (without considering minimizing the energy and work), then a linear body force feedback control can be used instead of the optimal control. In this paper we will study both optimal control and feedback control. The feedback control solution we will construct has exponential decay properties, but, unlike the optimal solution, the feedback solution can be too costly to realize physically.

The study of optimal flow control problems in the infinite time interval is of great importance in many physical applications, such as in drag and turbulence minimization in the entire life span of a flow. (Of course, the functional will be chosen differently in these cases from the functional introduced above.) Much effort has been made by mathematicians and scientists in the study of the asymptotic behaviors and dynamics of solutions for the Navier–Stokes equations. Naturally, we are motivated to study the asymptotic behaviors and dynamics of solutions for the controlled Navier–Stokes equations. There is also an interest in controlling the dynamics of flows which will be studied elsewhere. Although the methods and techniques used in this paper are applicable to the study of optimal control problems for many other cost functionals, we will only deal with the functional (1.1) throughout this paper. As was explained above, the physical objective behind (1.1) is to match the candidate flow field with a desired one. Ideally, one wishes to match the desired flow at each time instance; the functional that describes this ideal objective involves L^∞-norm in t. Such an ideal objective is in general too costly to achieve physically, and numerical solutions of such an ideal control problem can be computationally expensive (numerical methods will be discussed elsewhere). It is natural to introduce the time-averaged functional (1.1) for the matching objective. But how good is the optimizer for (1.1) as an optimizer for pointwise matching in t? This will be the main question to be answered in this paper. Our main conclusion is that for large t, the time-averaged optimizer will indeed give us pointwise matching in t.

We comment on some works extant in the literature that are related to this paper. In [2], [8], [9], [10], and [14] optimal distributed control problems for the time-dependent Navier–Stokes equations on finite time intervals were studied. In [8], [9], and [10], the existence of optimal distributed controls was shown, an optimality system of equations was derived, and the question of the uniqueness of optimal solutions was resolved. The optimality system of equations established in these references on finite time intervals will be very useful in our derivation of an analogous system for the infinite time interval.

Controllability is somewhat related to our interest in controlling the pointwise-in-t behavior for the solutions of the Navier–Stokes equations. In [7], [11], and [12]

the distributed approximate controllability problem was studied. It was shown in [7] that in the special case of control acting everywhere in the domain, the Navier–Stokes system is approximately controllable; i.e., one may match a given flow field with arbitrary accuracy at time instance $t = T$. In [11] it was proved that one can exactly match the zero vector field at $t = T$ by distributed control. However, the techniques used in [7] and [12] cannot be applied to more general controllability or approximate controllability problems for the Navier–Stokes system. The results of [12] on the noncontrollability of the Burgers equation seem to indicate that in general one cannot obtain approximate controllability for the Navier–Stokes system. In any case, the controllability approach does not give us information on the matching of flow fields over a time period, nor could it give us any information beyond $t = T$.

We also point out that some results on the asymptotic behaviors/dynamics for the (uncontrolled) Navier–Stokes equations can be found in the literature, such as [3], [15], [16], [18], and [19]. Also, in [4] and [5], the dynamics for the controlled Burgers equation was studied. Although the properties for the controlled Burgers equations can be substantially different from that for the controlled Navier–Stokes equations, results on Burgers equations often give us some insight into the results for the Navier–Stokes equations.

We summarize the major components of this paper as follows.

• A feedback control solution is constructed which can be treated as a quasi optimizer for the optimal control problem.

• The results of [8], [9], and [10] for finite time intervals are generalized to the infinite time interval $[0, \infty)$; i.e., we prove the existence of a solution for the distributed optimal control problem of minimizing (1.1) subject to (1.2)–(1.5) and derive an optimality system of equations from which optimal solutions may be deduced.

• The long-time behavior (dynamics) of the optimal solution is derived and the main result is that $\|\widehat{\mathbf{u}}(t) - \mathbf{U}(t)\|$ and $\|\nabla\widehat{\mathbf{u}}(t) - \nabla\mathbf{U}(t)\|$; i.e., the $\mathbf{L}^2(\Omega)$-distance and the $\mathbf{H}^1(\Omega)$-distance between the optimal solution $\widehat{\mathbf{u}}(t)$ and the desired flow field $\mathbf{U}(t)$ both decay to zero as time $t \to \infty$. Note that the distances are measured pointwise in t despite the fact that the (time-averaged) functional (1.1) seems to provide only L^2-information in t. Moreover, although the functional (1.1) provides only $\mathbf{L}^2$-information in $\mathbf{x}$, the decay property can be upgraded to the $\mathbf{H}^1(\Omega)$-distance in $\mathbf{x}$.

• We also obtain as by-products some estimates for the solutions of the optimality system on both finite and infinite intervals.

Our plan of the paper is as follows. In section 2, we construct a feedback control solution and obtain some preliminary estimates for all solution of the Navier–Stokes equations and, in particular, for optimal solutions. In section 3, we first recall the results of [8], [9], and [10] on finite time intervals and establish some estimates for the Lagrange multiplier (the adjoint state variable). We then prove the existence of an optimal solution on the infinite time interval. Finally, in section 4, we prove the decay of the controlled dynamics to the desired dynamics.

2. Statement of the problem, feedback control, and preliminary estimates.

2.1. Functional spaces and notations, statement of the problem. Throughout this paper, C denotes a generic constant depending only on the physical domain Ω. We will use the standard notations for the function space $L^r(\Omega)$ and the Sobolev spaces $H^m(\Omega)$ with its norm denoted by $\|\cdot\|_m$. $H^0(\Omega) = L^2(\Omega)$. Also, $H_0^r(\Omega)$ is the closure of $C_0^\infty(\Omega)$ under the $\|\cdot\|_r$-norm. The dual space of $H_0^r(\Omega)$ is denoted by $H^{-r}(\Omega)$, $r > 0$. The vector-valued ($\mathbb{R}^2$-valued) counterparts of these spaces are

denoted by $\mathbf{L}^r(\Omega)$, $\mathbf{H}^m(\Omega)$, $\mathbf{H}_0^r(\Omega)$, and $\mathbf{H}^{-r}(\Omega)$. For details, see [1] and [13]. We introduce the solenoidal spaces

$$\mathbf{V}^r(\Omega) = \{\mathbf{u} \in \mathbf{H}_0^r(\Omega) \ : \ \nabla \cdot \mathbf{u} = 0\}$$

equipped with the norm $\|\cdot\|_r$ for $r \geq 0$,

$$\mathbf{V} = \mathbf{V}^1(\Omega) = \{\mathbf{u} \in \mathbf{H}_0^1(\Omega) \ : \ \nabla \cdot \mathbf{u} = 0\}$$

equipped with the norm $\|\cdot\|_1$, and

$$\mathbf{W} = \{\mathbf{u} \in \mathbf{L}^2(\Omega) \ : \ \nabla \cdot \mathbf{u} = 0, \ (\mathbf{u} \cdot \mathbf{n})|_\Gamma = 0\}$$

(i.e., the closure of div-free $\mathbf{C}_0^\infty(\Omega)$-functions under the $\|\cdot\|_0$-norm) equipped with the norm $\|\cdot\|_0$. The dual space of $\mathbf{V}^r(\Omega)$, $r > 0$, is denoted by $\mathbf{V}^{-r}(\Omega)$. The dual space of $\mathbf{V}$ is denoted by $\mathbf{V}^*$. We identify the dual space of $\mathbf{W}$ with $\mathbf{W}$ itself under the $\mathbf{L}^2(\Omega)$-inner product. We next introduce the temporal–spatial function spaces defined on $Q_T = \Omega \times (0,T)$ for $T \in (0,\infty]$ (note that Q_∞ is also simply denoted by Q):

$$L^m(0,T;\mathbf{H}^r(\Omega))$$

equipped with the norm

$$\|\mathbf{u}\|_{L^m(0,T;\mathbf{H}^r(\Omega))} = \left(\int_0^T \|\mathbf{u}(t)\|_r^m \, dt\right)^{1/m},$$

and the solenoidal temporal–spatial function space

$$\mathcal{V}^{(s)}(Q_T) = \{\mathbf{v} \in L^2(0,T;\mathbf{V}^s(\Omega)) : \partial_t \mathbf{v} \in L^2(0,T;\mathbf{V}^{s-2}(\Omega))\}$$

with the norm

$$\|\mathbf{v}\|_{\mathcal{V}^{(s)}(Q_T)}^2 = \|\mathbf{v}\|_{L^2(0,T;\mathbf{V}^s(\Omega))}^2 + \|\partial_t \mathbf{v}\|_{L^2(0,T;\mathbf{V}^{s-2}(\Omega))}^2 .$$

For a function $\mathbf{u}$ in the temporal–spatial space, we often use the abbreviated notation

$$\mathbf{u}(t) \stackrel{\text{def}}{=\!=} \mathbf{u}(\cdot,t),$$

which is defined over the spatial domain Ω.

We introduce the simplified norm notation

$$\|\cdot\| \stackrel{\text{def}}{=\!=} \|\cdot\|_{L^2(\Omega)} . \tag{2.1}$$

Let $\lambda_1 > 0$ be the greatest real number satisfying the Poincaré inequality

$$\|\nabla \mathbf{w}\|^2 \geq \lambda_1 \|\mathbf{w}\|^2 \qquad \forall\, \mathbf{w} \in \mathbf{H}_0^1(\Omega) . \tag{2.2}$$

We now define the solution for the Navier–Stokes equations in a weak sense (see [6] or [18]). To this end, we introduce two continuous linear forms:

$$a(\mathbf{u},\mathbf{v}) = \nu \int_\Omega \nabla \mathbf{u} : \nabla \mathbf{v} \, d\mathbf{x} \qquad \forall\, \mathbf{u},\mathbf{v} \in \mathbf{H}^1(\Omega)$$

and

$$c(\mathbf{u},\mathbf{v},\mathbf{w}) = \int_\Omega (\mathbf{u}\cdot\nabla)\mathbf{v}\cdot\mathbf{w}\,d\mathbf{x} \qquad \forall\,\mathbf{u},\mathbf{v},\mathbf{w}\in\mathbf{H}^1(\Omega)\,,$$

where the colon notation : denotes the inner product on $\mathbb{R}^{2\times 2}$. Also, $(\cdot,\cdot)$ denotes the $\mathbf{L}^2(\Omega)$-inner product and $\langle\cdot,\cdot\rangle$ the duality pairing between a Banach space and its dual.

DEFINITION 2.1. *Given $T\in(0,\infty)$, $\mathbf{u}_0\in\mathbf{W}$, and $\mathbf{f}\in L^2(0,T;\mathbf{V}^*)$, $\mathbf{u}$ is said to be a solution of the Navier–Stokes equations iff $\mathbf{u}\in\mathcal{V}^{(1)}(Q_T)$ and $\mathbf{u}$ satisfies*

$$\begin{aligned}&\langle\partial_t\mathbf{u}(t),\mathbf{w}\rangle + a(\mathbf{u}(t),\mathbf{w}) + c(\mathbf{u}(t),\mathbf{u}(t),\mathbf{w})\\ &\qquad = \langle\mathbf{f}(t),\mathbf{w}\rangle \qquad \forall\,\mathbf{w}\in\mathbf{V}\ \text{(almost everywhere)}\ t\in(0,T)\end{aligned}\tag{2.3}$$

and

$$\lim_{t\to 0^+}\mathbf{u}(t) = \mathbf{u}_0 \qquad \text{in } \mathbf{W}.\tag{2.4}$$

We point out that $\mathbf{u}\in\mathcal{V}^{(1)}(Q_T)$ implies $\mathbf{u}\in C([0,T];\mathbf{W})$, so that (2.4) makes sense. It is well known that if $T\in(0,\infty)$ and $\mathbf{f}\in L^2(0,T;\mathbf{L}^2(\Omega))$, then there is indeed a solution $\mathbf{u}$ for the Navier–Stokes equations. Furthermore, $\mathbf{u}$ satisfies the regularity result $\mathbf{u}\in L^2_{\rm loc}(0,T;\mathbf{H}^2(\Omega))$.

For $T=\infty$, we define a solution for the Navier–Stokes equations as follows.

DEFINITION 2.2. *Given $\mathbf{u}_0\in\mathbf{W}$ and $\mathbf{f}\in L^2_{\rm loc}(0,\infty;\mathbf{V}^*)$, $\mathbf{u}$ is said to be a solution of the Navier–Stokes equations on $(0,\infty)$ iff $\mathbf{u}\in L^2_{\rm loc}(0,\infty;\mathbf{V})\cap L^\infty(0,\infty;\mathbf{W})$, $\partial_t\mathbf{u}\in L^2_{\rm loc}(0,\infty;\mathbf{V}^*)$, and $\mathbf{u}$ satisfies* (2.3), (2.4) *with $T=\infty$.*

Intuitively, if a flow field $\mathbf{u}$ is close to the desired field $\mathbf{U}$, then the body forces corresponding to the two fields $\mathbf{u}$ and $\mathbf{U}$ should also be close. Hence, in order that the optimal control solution of the Navier–Stokes flow is close to the desired flow $\mathbf{U}$, we must place some restrictions on the desired body force $\mathbf{F}$ involved in the cost functional (1.1). In fact, throughout this paper we will simply choose

$$\mathbf{F} = \mathbf{N}(\mathbf{U}) \stackrel{\rm def}{=\!=} \partial_t\mathbf{U} - \nu\Delta\mathbf{U} + (\mathbf{U}\cdot\nabla)\mathbf{U},\tag{2.5}$$

which is the body force corresponding to the desired flow $\mathbf{U}$. Here note that, for convenience, the pressure term is not included in the definition of $\mathbf{F}=\mathbf{N}(\mathbf{U})$, since it is not involved in the cost functional. The omission of the pressure term will not affect any of the results in this paper. In fact, the pressure term does not even appear in our definition of the solutions (in the weak sense) for the Navier–Stokes equations. The restriction (2.5) implies that when the functional (1.1) is minimized, $\mathbf{u}$ is close to the desired flow $\mathbf{U}$ and $\mathbf{f}$ is close to the body force corresponding to the desired flow. Note that the pair $(\mathbf{U},\mathbf{F})$ in general is not an optimal solution, for $\mathbf{U}$ in general does not satisfy the initial condition (2.4).

Throughout this paper, in addition to (2.5), we make the following hypothesis for the desired flow $\mathbf{U}$ and the fixed body force $\mathbf{F}=\mathbf{N}(\mathbf{U})$:

$$\begin{cases}\mathbf{U}=\mathbf{U}(\mathbf{x},t)\in L^\infty(0,\infty;\mathbf{H}^2(\Omega)\cap\mathbf{V})\,,\\ \mathbf{F}=\mathbf{N}(\mathbf{U})\in L^\infty(0,\infty;\mathbf{L}^2(\Omega))\,.\end{cases}\tag{2.6}$$

Hypothesis (2.6) implies

$$\partial_t\mathbf{U}\in L^\infty(0,\infty;\mathbf{L}^2(\Omega))\cap L^2_{\rm loc}(0,\infty;\mathbf{L}^2(\Omega))\,.$$

Note that these hypotheses permit the special case of steady state $\mathbf{U}$. Thus one application of the optimal control problem is to match a steady state flow field through the control of external forces.

We also introduce the following simplified norm notations:

$$|||\cdot||| \stackrel{\text{def}}{=\!=} \|\cdot\|_{L^\infty(0,\infty;\mathbf{L}^2(\Omega))} \tag{2.7}$$

and

$$|||\cdot|||\!| \stackrel{\text{def}}{=\!=} \|\cdot\|_{\mathbf{L}^\infty(\Omega\times(0,\infty))}\,. \tag{2.8}$$

These norms will be applied solely to $\mathbf{U}$, $\nabla\mathbf{U}$, ..., etc.

We now turn to the precise statement of the optimal control problem.

For each $T \in (0,\infty]$ we define the functional $\mathcal{J}_T$ by

$$\mathcal{J}_T(\mathbf{u},\mathbf{f}) = \frac{\alpha}{2}\int_0^T\int_\Omega |\mathbf{u}-\mathbf{U}|^2\,d\mathbf{x}\,dt + \frac{\beta}{2}\int_0^T\int_\Omega |\mathbf{f}-\mathbf{F}|^2\,d\mathbf{x}\,dt \tag{2.9}$$

for all $\mathbf{u} \in \mathbf{U} + \mathbf{L}^2(\Omega\times(0,T))$ and $\mathbf{f} \in \mathbf{N}(\mathbf{U}) + \mathbf{L}^2(\Omega\times(0,T))$. Note that $\mathcal{J}_\infty$ is also simply denoted by $\mathcal{J}$.

We point out that in the case of $T=\infty$, which we will eventually consider, if we choose the control $\mathbf{f}$ in the space $\mathbf{L}^2(\Omega\times(0,\infty))$, it may happen (e.g., in the case of a steady state $\mathbf{U}$) that the value of the cost functional $\mathcal{J}_\infty(\mathbf{u},\mathbf{f})$ is always infinite for every pair $(\mathbf{u},\mathbf{f})$ under consideration. Therefore, the choice of the control set should also involve $\mathbf{U}$ and $\mathbf{F}$. We define the admissible elements as follows with $\mathbf{X}_T$ and $\mathbf{Y}_T$ denoting, respectively, the functional spaces

$$\mathbf{X}_T = \begin{cases} \mathcal{V}^{(1)}(Q_T) & \text{if } T \in (0,\infty)\,, \\ \{\mathbf{u} \in L^2_{\text{loc}}(0,\infty;\mathbf{V}) \cap L^\infty(0,\infty;\mathbf{W}) : \partial_t\mathbf{u} \in L^2_{\text{loc}}(0,\infty;\mathbf{V}^*)\} & \text{if } T=\infty \end{cases}$$

and

$$\mathbf{Y}_T = \begin{cases} L^2(0,T;\mathbf{V}^*) & \text{if } T \in (0,\infty)\,, \\ L^2_{\text{loc}}(0,\infty;\mathbf{V}^*) & \text{if } T=\infty\,. \end{cases}$$

DEFINITION 2.3. *For a given $T \in (0,\infty]$, a pair $(\mathbf{u},\mathbf{f}) \in \mathbf{X}_T \times \mathbf{Y}_T$ is called an* admissible element *if $\mathcal{J}_T(\mathbf{u},\mathbf{f}) < \infty$ and $(\mathbf{u},\mathbf{f})$ satisfies (2.3)–(2.4). The set of all admissible elements is denoted by $\mathcal{U}_{ad}(T)$.*

Now for each $T \in (0,\infty]$, we state the optimal control problem on $(0,T)$ as follows:

$$\begin{aligned} &\text{find a } (\mathbf{u},\mathbf{f}) \in \mathcal{U}_{ad}(T) \text{ such that} \\ &\mathcal{J}_T(\mathbf{u},\mathbf{f}) \le \mathcal{J}_T(\mathbf{w},\mathbf{h}) \qquad \forall\,(\mathbf{w},\mathbf{h}) \in \mathcal{U}_{ad}(T)\,. \end{aligned} \tag{2.10}$$

We point out that in general, the initial state $\mathbf{u}_0$ is a certain distance away from the desired flow, or, $\mathbf{u}_0 \neq \mathbf{U}(\cdot,t)$ for all t, the cost functional generally has a positive minimum. Therefore our optimal control problem has nontrivial solutions.

With the change of variables

$$\mathbf{v} = \mathbf{u} - \mathbf{U} \qquad \text{and} \qquad \mathbf{g} = \mathbf{f} - \mathbf{N}(\mathbf{U})\,, \tag{2.11}$$

(2.3), (2.4) are equivalent to $(\mathbf{v},\mathbf{g}) \in \mathbf{X}_T \times \mathbf{Y}_T$ satisfying

$$\begin{aligned} &\langle \partial_t\mathbf{v}(t),\mathbf{w}\rangle + a(\mathbf{v}(t),\mathbf{w}) + c(\mathbf{v}(t),\mathbf{v}(t),\mathbf{w}) + c(\mathbf{U}(t),\mathbf{v}(t),\mathbf{w}) \\ &\qquad + c(\mathbf{v}(t),\mathbf{U}(t),\mathbf{w}) = (\mathbf{g}(t),\mathbf{w}) \qquad \forall\,\mathbf{w} \in \mathbf{V} \text{ a.e. } t \in (0,T) \end{aligned} \tag{2.12}$$

and

$$\lim_{t\to 0^+} \mathbf{v}(t) = \mathbf{u}_0 - \mathbf{U}_0 \qquad \text{in } \mathbf{W}\,. \tag{2.13}$$

The cost functional can be rewritten as

$$\mathcal{K}_T(\mathbf{v},\mathbf{g}) \stackrel{\text{def}}{=\!=} \mathcal{J}_T(\mathbf{v}+\mathbf{U},\mathbf{g}+\mathbf{N}(\mathbf{U})) = \frac{\alpha}{2}\int_0^T\int_\Omega |\mathbf{v}|^2\,d\mathbf{x}dt + \frac{\beta}{2}\int_0^T\int_\Omega |\mathbf{g}|^2 d\mathbf{x}dt\,. \tag{2.14}$$

By defining

$$\mathcal{V}_{\text{ad}}(T) \stackrel{\text{def}}{=\!=} \{(\mathbf{v},\mathbf{g}) \in \mathbf{X}_T \times \mathbf{Y}_T : \mathcal{K}_T(\mathbf{v},\mathbf{g}) < \infty,\ (\mathbf{v},\mathbf{g}) \text{ satisfies (2.12), (2.13)}\}$$

for each $T \in (0,\infty]$, we can restate the optimization problem (2.10) in terms of the auxiliary variables $(\mathbf{v},\mathbf{g})$:

$$\begin{aligned} &\text{find a } (\mathbf{v},\mathbf{g}) \in \mathcal{V}_{\text{ad}}(T) \text{ such that}\\ &\mathcal{K}_T(\mathbf{v},\mathbf{g}) \le \mathcal{K}_T(\mathbf{w},\mathbf{h}) \qquad \forall\,(\mathbf{w},\mathbf{h}) \in \mathcal{V}_{\text{ad}}(T)\,. \end{aligned} \tag{2.15}$$

2.2. A linear feedback distributed control—a quasi optimizer. To estimate the dynamics of the optimal control solution, we need to find a sharp bound for the value of $\inf_{(\mathbf{u},\mathbf{f})\in\mathcal{U}_{ad}(T)} \mathcal{J}_T(\mathbf{u},\mathbf{f})$. It is important that this bound is uniform in T. We now construct a quasi optimizer $(\widetilde{\mathbf{u}},\widetilde{\mathbf{f}}) \in \mathcal{U}_{ad}(\infty)$ for $\mathcal{J}_\infty(\cdot,\cdot)$ by means of a linear feedback. We can in turn derive some preliminary estimates for the optimal solutions. By a quasi optimizer we mean an element $(\widetilde{\mathbf{u}},\widetilde{\mathbf{f}}) \in \mathcal{U}_{ad}(\infty)$ satisfying $\|\widetilde{\mathbf{u}}(t) - \mathbf{U}(t)\| \to 0$ as $t \to \infty$. The following theorem asserts the existence of such an element.

THEOREM 2.4. *There exists a pair* $(\widetilde{\mathbf{u}},\widetilde{\mathbf{f}}) \in \mathcal{U}_{ad}(\infty)$ *satisfying*

$$\|\widetilde{\mathbf{u}}(t) - \mathbf{U}(t)\|^2 \le \|\mathbf{u}_0 - \mathbf{U}_0\|^2 \exp\{-(\tilde{k} + \nu\lambda_1 - C_1\,|||\nabla\mathbf{U}|||^2/\nu)t\} \tag{2.16}$$

for some $\tilde{k} > M$, *where* $C_1 > 0$ *is a constant depending only on* Ω, $\lambda_1 > 0$ *is the Poincaré constant in* (2.2),

$$M = M(|||\nabla\mathbf{U}|||) \stackrel{\text{def}}{=\!=} \frac{C_1\,|||\nabla\mathbf{U}|||^2}{\nu} - \nu\lambda_1, \tag{2.17}$$

and

$$\mathcal{J}_T(\widetilde{\mathbf{u}},\widetilde{\mathbf{f}}) \le K\,\|\mathbf{u}_0 - \mathbf{U}_0\|^2 \left(1 - \exp\{-T\sqrt{\beta^2M^2 + 4\alpha\beta}/\beta\}\right) \quad \forall\, T \in (0,\infty], \tag{2.18}$$

where

$$K = K(\nabla\mathbf{U},\alpha,\beta) \stackrel{\text{def}}{=\!=} \frac{4\alpha\beta + \left(\sqrt{\beta^2M^2 + 4\alpha\beta} + \beta M\right)^2}{8\sqrt{\beta^2M^2 + 4\alpha\beta}}. \tag{2.19}$$

If, in addition, $\nabla\mathbf{U} \in L^\infty(\Omega\times(0,\infty))$, *then the constant* M *in* (2.18), (2.19) *can be replaced by*

$$M' = M'(|||\nabla\mathbf{U}|||) \stackrel{\text{def}}{=\!=} 4\,|||\nabla\mathbf{U}||| - 2\nu\lambda_1. \tag{2.20}$$

Proof. Let $k > 0$ be an arbitrary (fixed) constant, and we seek a solution $\mathbf{u} \in \mathbf{X}_\infty$ for the following Navier–Stokes equations with a linear feedback in the body force:

$$\begin{aligned} &\langle \partial_t \mathbf{u}(t), \mathbf{w}\rangle + a(\mathbf{u}(t), \mathbf{w}) + c(\mathbf{u}(t), \mathbf{u}(t), \mathbf{w}) \\ &\quad = -\frac{k}{2}(\mathbf{u}(t), \mathbf{w}) + (k\mathbf{U}(t)/2 + \mathbf{N}(\mathbf{U}(t)), \mathbf{w}) \qquad \forall\, \mathbf{w} \in \mathbf{V} \text{ a.e. } t \in (0, \infty) \end{aligned} \tag{2.21}$$

and

$$\lim_{t \to 0^+} \mathbf{u}(t) = \mathbf{u}_0 \qquad \text{in } \mathbf{W}\,. \tag{2.22}$$

As the term $-k(\mathbf{u}/2, \mathbf{w})$ has the right sign when moved to the left-hand side and by (2.6), $\mathbf{U} \in L^\infty(0, T; \mathbf{L}^2(\Omega))$, we can use the techniques for the Navier–Stokes equations (see [6] and [18]) to show that there exists a $\mathbf{u} \in \mathbf{X}_\infty$ that satisfies (2.21), (2.22) provided k is bounded from below by a certain finite number (the range of k will be determined later in the proof). By setting $\mathbf{v} = \mathbf{u} - \mathbf{U}$ we see that $\mathbf{v}$ satisfies

$$\begin{aligned} &\langle \partial_t \mathbf{v}(t), \mathbf{w}\rangle + a(\mathbf{v}(t), \mathbf{w}) + c(\mathbf{v}(t), \mathbf{v}(t), \mathbf{w}) + c(\mathbf{U}(t), \mathbf{v}(t), \mathbf{w}) \\ &\qquad + c(\mathbf{v}(t), \mathbf{U}(t), \mathbf{w}) = -\frac{k}{2}(\mathbf{v}(t), \mathbf{w}) \qquad \forall\, \mathbf{w} \in \mathbf{V} \text{ a.e. } t \in (0, \infty) \end{aligned} \tag{2.23}$$

and

$$\lim_{t \to 0^+} \mathbf{v}(t) = \mathbf{u}_0 - \mathbf{U}_0 \qquad \text{in } \mathbf{W}\,. \tag{2.24}$$

Setting $\mathbf{w} = \mathbf{v}(t)$ in (2.23), we obtain

$$\frac{1}{2}\frac{d}{dt}\|\mathbf{v}(t)\|^2 + \nu\|\nabla \mathbf{v}(t)\|^2 + \frac{k}{2}\|\mathbf{v}(t)\|^2 = -c(\mathbf{v}(t), \mathbf{U}(t), \mathbf{v}(t))\,. \tag{2.25}$$

We deduce from the Ladyzhenskaya inequality $\|\mathbf{v}\|^2_{\mathbf{L}^4(\Omega)} \le C\,\|\mathbf{v}\|\,\|\nabla \mathbf{v}\|$ for all $\mathbf{v} \in \mathbf{H}^1(\Omega)$ that

$$\begin{aligned} |c(\mathbf{v}(t), \mathbf{U}(t), \mathbf{v}(t))| &\le \|\mathbf{v}(t)\|^2_{\mathbf{L}^4(\Omega)}\, \|\nabla \mathbf{U}(t)\| \le C\,\|\mathbf{v}(t)\|\,\|\nabla \mathbf{v}(t)\|\,\, |||\nabla \mathbf{U}||| \\ &\le \frac{\nu}{2}\|\nabla \mathbf{v}(t)\|^2 + \frac{C}{\nu}\|\mathbf{v}(t)\|^2\,\, |||\nabla \mathbf{U}|||^2\,, \end{aligned} \tag{2.26}$$

so that from (2.25) and Poincaré inequality (2.2) we obtain

$$\frac{d}{dt}\|\mathbf{v}(t)\|^2 + \left(k + \nu\lambda_1 - \frac{C_1\,|||\nabla \mathbf{U}|||^2}{\nu}\right)\|\mathbf{v}(t)\|^2 \le 0\,, \tag{2.27}$$

where λ_1 is the Poincaré constant in (2.2) and $C_1 > 0$ is a constant depending only on Ω. Thus, if k satisfies

$$k > M = M\left(|||\nabla \mathbf{U}|||\right) \stackrel{\text{def}}{=\!=} \frac{C_1\,|||\nabla \mathbf{U}|||^2}{\nu} - \nu\lambda_1\,, \tag{2.28}$$

then we may apply the Gronwall inequality to (2.27) to obtain

$$\|\mathbf{v}(\cdot, t)\|^2 \le \|\mathbf{u}_0 - \mathbf{U}_0\|^2\, \exp\{-(k - M)t\}. \tag{2.29}$$

Setting $\mathbf{f} = -k\mathbf{v}/2 + \mathbf{N}(\mathbf{U})$ and $\mathbf{g} = -k\mathbf{v}/2$, we see that for each $T \in (0, \infty]$,

$$\begin{aligned} \mathcal{J}_T(\mathbf{u}, \mathbf{f}) = \mathcal{K}_T(\mathbf{v}, \mathbf{g}) &= \frac{\alpha}{2} \int_0^T \int_\Omega |\mathbf{v}|^2 \, d\mathbf{x} \, dt + \frac{\beta}{2} \int_0^T \int_\Omega |\mathbf{g}|^2 \, d\mathbf{x} \, dt \\ &\le \frac{\left(4\alpha + \beta k^2\right) \|\mathbf{u}_0 - \mathbf{U}_0\|^2}{8} \int_0^T \exp\{-(k-M)t\} \, dt \\ &= \frac{\left(4\alpha + \beta k^2\right) \|\mathbf{u}_0 - \mathbf{U}_0\|^2 \cdot (1 - \exp\{-(k-M)T\})}{8(k-M)} \, . \end{aligned} \tag{2.30}$$

We now choose a particular k and define $\widetilde{\mathbf{u}}$. We notice that the function

$$\theta(k) \stackrel{\text{def}}{=\!=} \frac{4\alpha + \beta k^2}{k - M}$$

defined for $k \in (M, \infty)$ attains its minimum at

$$\tilde{k} \stackrel{\text{def}}{=\!=} \frac{1}{\beta} \left(\sqrt{\beta^2 M^2 + 4\alpha\beta} + \beta M \right) > M \, . \tag{2.31}$$

We let $\widetilde{\mathbf{u}}$ denote the solution of (2.21), (2.22) with $k = \tilde{k}$. Upon setting $\widetilde{\mathbf{v}} = \widetilde{\mathbf{u}} - \mathbf{U}$, $\widetilde{\mathbf{f}} = -k\widetilde{\mathbf{v}}/2 + \mathbf{N}(\mathbf{U})$, and $\widetilde{\mathbf{g}} = -k\widetilde{\mathbf{v}}/2$, and using (2.30), we are led to

$$\begin{aligned} \mathcal{J}_T(\widetilde{\mathbf{u}}, \widetilde{\mathbf{f}}) &\le \frac{1}{8} \theta(\tilde{k}) \, \|\mathbf{u}_0 - \mathbf{U}_0\|^2 \, (1 - \exp\{-(\tilde{k} - M)T\}) \\ &= \frac{(4\alpha + \beta\tilde{k}^2) \|\mathbf{u}_0 - \mathbf{U}_0\|^2}{8(\tilde{k} - M)} \, (1 - \exp\{-(\tilde{k} - M)T\}) \end{aligned} \tag{2.32}$$

for all $T \in (0, \infty]$, so that (2.18) is proved with

$$K \stackrel{\text{def}}{=\!=} \frac{4\alpha + \beta\tilde{k}^2}{8(\tilde{k} - M)} = \frac{4\alpha\beta + \left(\sqrt{\beta^2 M^2 + 4\alpha\beta} + \beta M\right)^2}{8\sqrt{\beta^2 M^2 + 4\alpha\beta}} \, .$$

Also, (2.16) follows from (2.29) with $k = \tilde{k}$.

If, in addition, $\nabla \mathbf{U} \in \mathbf{L}^\infty(\Omega \times (0, \infty))$, then from (2.25) we have, instead of (2.26),

$$|c(\mathbf{v}(t), \mathbf{U}(t), \mathbf{v}(t))| \le 2 \, |||\nabla \mathbf{U}||| \, \|\mathbf{v}(t)\|^2$$

so that, instead of (2.27),

$$\frac{d}{dt} \|\widetilde{\mathbf{v}}(t)\|^2 + (k + 2\nu\lambda_1 - 4 \, |||\nabla \mathbf{U}|||) \, \|\widetilde{\mathbf{v}}(t)\|^2 \le 0 \, .$$

Evidently, upon replacing M by M' (defined by (2.20)) in the foregoing arguments after (2.27), we see that (2.16) and (2.18) now hold with $M = M'$. □

In what follows, the element $(\widetilde{\mathbf{u}}, \widetilde{\mathbf{f}})$ will always denote the one constructed in Theorem 2.4. As a consequence of (2.18) and the intermediate-value theorem we obtain the following bound for $\mathcal{J}_T(\widetilde{\mathbf{u}}, \widetilde{\mathbf{f}})$:

COROLLARY 2.5. *If the assumptions of Theorem* 2.4 *hold, then*

$$\mathcal{J}_T(\widetilde{\mathbf{u}}, \widetilde{\mathbf{f}}) \le \min \left\{ K, \frac{T}{8}(4\alpha + \beta\tilde{k}^2) \right\} \, \|\mathbf{u}_0 - \mathbf{U}_0\|^2 \tag{2.33}$$

for all $T \in (0, \infty]$, where $\tilde{k}$ is given by (2.31) *and K is given by* (2.19). □

It follows trivially from Corollary 2.5 that

$$\min_{(\mathbf{u},\mathbf{f})\in\mathcal{U}_{ad}(T)} \mathcal{J}_T(\mathbf{u},\mathbf{f}) = O(T) \quad \text{as } T \to 0^+. \tag{2.34}$$

Remark 2.6. In Theorem 2.4, we see that a quasi optimizer $(\widetilde{\mathbf{u}},\widetilde{\mathbf{f}})$ has been created in the sense that $\|\widetilde{\mathbf{u}}(t)-\mathbf{U}(t)\| \to 0$ and $\mathcal{J}_\infty(\widetilde{\mathbf{u}},\widetilde{\mathbf{f}})$ is finite. In fact, $\|\widetilde{\mathbf{u}}(t)-\mathbf{U}(t)\| \to 0$ exponentially as $t\to\infty$. Also, the computation of $(\widetilde{\mathbf{u}},\widetilde{\mathbf{f}}) \stackrel{\text{def}}{=\!=} (\mathbf{U}+\widetilde{\mathbf{v}}, \mathbf{N}(\mathbf{U})-\tilde{k}\widetilde{\mathbf{v}}/2)$ is straightforward—one only needs to integrate the initial value problem (2.21), (2.22) with $k=\tilde{k}$. However, we point out that $-\tilde{k}\widetilde{\mathbf{v}}/2+\mathbf{N}(\mathbf{U})$ acts as an external force to the flow, and both $\widetilde{\mathbf{v}}$ and $\tilde{k}$ can be large; therefore the work (external force) required to achieve the quasi optimizer may be prohibitively expensive. The true optimizer $\widehat{\mathbf{u}}$ is expected to have the property $\|\widehat{\mathbf{u}}(t)-\mathbf{U}(t)\| \to 0$ as $t\to\infty$ and, at the same time, minimize the work involved to realize and maintain the optimized flow.

2.3. Preliminary estimates for the dynamics of admissible elements. With the aid of the quasi optimizer constructed in Theorem 2.4, we are prepared to derive some estimates for the dynamics of all solutions of (2.3), (2.4). These estimates in turn will allow us to derive preliminary estimates for the dynamics of the optimal solutions. First, we consider the $L^\infty(0,T;\mathbf{L}^2(\Omega))$ estimates in terms of the initial data and the functional values.

THEOREM 2.7. *Let $T\in(0,\infty]$. Assume that $(\mathbf{u},\mathbf{f})\in\mathcal{U}_{ad}(T)$. Then*

$$\|\mathbf{u}(t)-\mathbf{U}(t)\|^2 \le \|\mathbf{u}_0-\mathbf{U}_0\|^2 + \frac{C}{\nu}\max\left\{\frac{|||\nabla\mathbf{U}|||^2}{\alpha},\frac{1}{\beta}\right\}\mathcal{J}_T(\mathbf{u},\mathbf{f}) \tag{2.35}$$

for all $t\in[0,T]$. If, in addition,

$$\mathcal{J}_T(\mathbf{u},\mathbf{f}) \le \mathcal{J}_T(\widetilde{\mathbf{u}},\widetilde{\mathbf{f}}), \tag{2.36}$$

then

$$\|\mathbf{u}(t)-\mathbf{U}(t)\|^2 \le K_0\cdot\|\mathbf{u}_0-\mathbf{U}_0\|^2, \tag{2.37}$$

where

$$K_0 = K_0(\nabla\mathbf{U},\alpha,\beta) \stackrel{\text{def}}{=\!=} 1+\frac{C}{\nu}\max\left\{\frac{|||\nabla\mathbf{U}|||^2}{\alpha},\frac{1}{\beta}\right\}\cdot K \tag{2.38}$$

with $K=K(\nabla\mathbf{U},\alpha,\beta)$ defined by (2.19).

Proof. Since $(\mathbf{v},\mathbf{g}) \stackrel{\text{def}}{=\!=} (\mathbf{u}-\mathbf{U},\mathbf{f}-\mathbf{N}(\mathbf{U}))$ satisfies (2.12), by setting $\mathbf{w}=\mathbf{v}(t)$ in (2.12) and using the inequality $\|\mathbf{v}\|^2_{\mathbf{L}^4(\Omega)} \le C\,\|\mathbf{v}\|\,\|\nabla\mathbf{v}\|$ for all $\mathbf{v}\in\mathbf{H}^1(\Omega)$, we obtain that for a.e. $t\in(0,T)$,

$$\begin{aligned}
\frac{1}{2}\frac{d}{dt}\|\mathbf{v}(t)\|^2+\nu\,\|\nabla\mathbf{v}(t)\|^2 &= c(\mathbf{v}(t),\mathbf{U}(t),\mathbf{v}(t))+(\mathbf{g}(t),\mathbf{v}(t))\\
&\le C\,|||\nabla\mathbf{U}|||\,\|\mathbf{v}(t)\|\,\|\nabla\mathbf{v}(t)\|+\|\mathbf{g}(t)\|\,\|\mathbf{v}(t)\|\\
&\le \frac{C}{\nu}\left(|||\nabla\mathbf{U}|||^2\,\|\mathbf{v}(t)\|^2+\|\mathbf{g}(t)\|^2\right)+\frac{\nu}{2}\|\nabla\mathbf{v}(t)\|^2
\end{aligned}$$

so that

$$\frac{d}{dt}\|\mathbf{v}(t)\|^2+\nu\|\nabla\mathbf{v}(t)\|^2 \le \frac{C}{\nu}\max\left\{\frac{|||\nabla\mathbf{U}|||^2}{\alpha},\frac{1}{\beta}\right\}\left(\alpha\|\mathbf{v}(t)\|^2+\beta\|\mathbf{g}(t)\|^2\right). \tag{2.39}$$

Multiplying the last inequality by $e^{\nu\lambda_1 t}$ and integrating over $(0,t)$, we arrive at

$$\|\mathbf{v}(t)\|^2 \le \|\mathbf{u}_0 - \mathbf{U}_0\|^2 e^{-\nu\lambda_1 t} + \frac{C}{\nu}\max\left\{\frac{|||\nabla\mathbf{U}|||^2}{\alpha}, \frac{1}{\beta}\right\}\int_0^t \left(\alpha\|\mathbf{v}(s)\|^2 + \beta\|\mathbf{g}(s)\|^2\right) e^{-\nu\lambda_1(t-s)}\, ds,$$

which easily yields (2.35). We see that (2.37) follows from (2.35) and (2.18). □

Now, we derive the $L^\infty(0,T;\mathbf{H}^1(\Omega))$ estimates. To this end, we recall two well-known results. The first result is the uniform Gronwall inequality (see [19, Lemma III.1.1]).

UNIFORM GRONWALL LEMMA. *Let g, h, and y be three nonnegative locally integrable functions on $t \in (t_0,\infty)$ such that y' is locally integrable on (t_0,∞) and that*

$$\frac{dy}{dt} \le gy + h \qquad \forall\, t \ge t_0\,,$$

$$\int_t^{t+r} g(s)ds \le a_1\,, \quad \int_t^{t+r} h(s)ds \le a_2\,, \quad \text{and} \quad \int_t^{t+r} y(s)ds \le a_3 \qquad \forall\, t \ge t_0,$$

where $r > 0$ and a_1, a_2, a_3 are constants. Then

$$y(t+r) \le \left(\frac{a_3}{r} + a_2\right) e^{a_1} \qquad \forall\, t \ge t_0. \tag{2.40}$$ □

The second result is that for the Leray operator

$$P : \mathbf{L}^2(\Omega) \to \mathbf{W} \tag{2.41}$$

(i.e., the orthogonal projection with respect to the $\mathbf{L}^2(\Omega)$-norm), it is well known (see [6]) that there is a constant $\gamma > 0$ depending only on Ω such that

$$\gamma\|\mathbf{w}\|_2 \le \|P\Delta\mathbf{w}\| \le \|\mathbf{w}\|_2 \qquad \forall\, \mathbf{w} \in \mathbf{H}^2(\Omega)\cap\mathbf{V}\,, \tag{2.42}$$

so that $\|P\Delta\cdot\|$ is equivalent to the $\mathbf{H}^2(\Omega)$-norm on $\mathbf{H}^2(\Omega)\cap\mathbf{V}$.

THEOREM 2.8. *Let $T \in (0,\infty]$. Assume that $(\mathbf{u},\mathbf{f}) \in \mathcal{U}_{ad}(T)$ and that (2.36) holds. Then for each $\varepsilon > 0$,*

$$\mathbf{u} - \mathbf{U} \in L^2(0,T;\mathbf{H}^1(\Omega)) \cap L^\infty(\varepsilon,T;\mathbf{H}^1(\Omega)) \cap C([\varepsilon,T];\mathbf{H}^1(\Omega)), \tag{2.43}$$

$$\int_0^T \|\nabla\mathbf{u}(s) - \nabla\mathbf{U}(s)\|^2 ds \le \overline{K}_1 \cdot \|\mathbf{u}_0 - \mathbf{U}_0\|^2, \tag{2.44}$$

and

$$\|\nabla\mathbf{u}(t) - \nabla\mathbf{U}(t)\|^2 \le K_1(\varepsilon)\cdot\|\mathbf{u}_0 - \mathbf{U}_0\|^2 \qquad \forall\, t \in [\varepsilon,T]\,, \tag{2.45}$$

where

$$\overline{K}_1 = \overline{K}_1(\nabla\mathbf{U},\alpha,\beta) \stackrel{\text{def}}{=\!=} \frac{1}{\nu} + \frac{C}{\nu^2}\max\left\{\frac{|||\nabla\mathbf{U}|||^2}{\alpha}, \frac{1}{\beta}\right\} K = K_0/\nu\,, \tag{2.46}$$

with K_0 and K given in Theorems 2.7 and 2.4, and

$$K_1(\varepsilon) = K_1(\varepsilon,\mathbf{U},\alpha,\beta) \stackrel{\text{def}}{=\!=} \exp\left\{\|\mathbf{u}_0 - \mathbf{U}_0\|^4 K_0\overline{K}_1/\nu^3\right\} \cdot \left\{\frac{1}{\varepsilon}\overline{K}_1 + \frac{C}{\nu^3}\left[(|||\mathbf{U}|||^2\ |||\nabla\mathbf{U}|||^2 + \nu^2\,|||\nabla\mathbf{U}|||\ |||\Delta\mathbf{U}|||)\overline{K}_1 + \frac{\nu^2 K}{\beta}\right]\right\}. \tag{2.47}$$

Proof. Let $T \in (0,\infty]$ be given and set $\mathbf{v} = \mathbf{u} - \mathbf{U}$. We first note that (2.45) follows from integrating (2.39). Thus $\mathbf{u} - \mathbf{U} \in L^2(0,T;\mathbf{H}^1(\Omega))$.

To show $\mathbf{u} - \mathbf{U} \in L^\infty(\varepsilon,T;\mathbf{H}^1(\Omega))$ for any $\varepsilon > 0$, it suffices to derive the estimate (2.45). We set $\mathbf{w} = -P\Delta\mathbf{v}(t)$ in (2.12) to obtain

$$\begin{aligned}
&\frac{1}{2}\frac{d}{dt}\|\nabla\mathbf{v}(t)\|^2 + \nu\|P\Delta\mathbf{v}(t)\|^2 \\
&\qquad = ([\mathbf{v}(t)\cdot\nabla]\mathbf{v}(t), -P\Delta\mathbf{v}(t)) - ([\mathbf{U}(t)\cdot\nabla]\mathbf{v}(t), -P\Delta\mathbf{v}(t)) \\
&\qquad\qquad - ([\mathbf{v}(t)\cdot\nabla]\mathbf{U}(t), -P\Delta\mathbf{v}(t)) + (\mathbf{g}(t), -P\Delta\mathbf{v}(t))\,.
\end{aligned}$$

Applying Sobolev imbedding and interpolation results, together with Young's inequality and (2.42), we have

$$\begin{aligned}
&\frac{1}{2}\frac{d}{dt}\|\nabla\mathbf{v}(t)\|^2 + \nu\,\|P\Delta\mathbf{v}(t)\|^2 \\
&\quad \le C(\|\mathbf{v}(t)\|_{\mathbf{L}^4(\Omega)}\,\|\nabla\mathbf{v}(t)\|_{\mathbf{L}^4(\Omega)} + \|\mathbf{U}(t)\|_{\mathbf{L}^4(\Omega)}\,\|\nabla\mathbf{v}(t)\|_{\mathbf{L}^4(\Omega)} \\
&\qquad\qquad + \|\mathbf{v}(t)\|_{\mathbf{L}^4(\Omega)}\,\|\nabla\mathbf{U}(t)\|_{\mathbf{L}^4(\Omega)} + \|\mathbf{g}(t)\|)\,\|P\Delta\mathbf{v}(t)\| \\
&\quad \le C(\|\mathbf{v}(t)\|^{1/2}\|\nabla\mathbf{v}(t)\|\,\|P\Delta\mathbf{v}(t)\|^{1/2} \\
&\qquad + \|\mathbf{U}(t)\|^{1/2}\,\|\nabla\mathbf{U}(t)\|^{1/2}\,\|\nabla\mathbf{v}(t)\|^{1/2}\,\|P\Delta\mathbf{v}(t)\|^{1/2} \\
&\qquad\quad + \|\nabla\mathbf{U}(t)\|^{1/2}\,\|P\Delta\mathbf{U}(t)\|^{1/2}\,\|\mathbf{v}(t)\|^{1/2}\,\|\nabla\mathbf{v}(t)\|^{1/2} + \|\mathbf{g}(t)\|)\,\|P\Delta\mathbf{v}(t)\| \\
&\quad \le \frac{\nu}{2}\|P\Delta\mathbf{v}(t)\|^2 + \frac{C}{\nu^3}(\|\mathbf{v}(t)\|^2\,\|\nabla\mathbf{v}(t)\|^4 + \|\mathbf{U}(t)\|^2\,\|\nabla\mathbf{U}(t)\|^2\,\|\nabla\mathbf{v}(t)\|^2 \\
&\qquad\qquad + \nu^2\|\nabla\mathbf{U}(t)\|\,\|\Delta\mathbf{U}(t)\|\,\|\nabla\mathbf{v}(t)\|^2 + \nu^2\|\mathbf{g}(t)\|^2).
\end{aligned}$$

Rearranging terms and applying Theorem 2.7 and (2.42), we obtain

$$\begin{aligned}
&\frac{d}{dt}\|\nabla\mathbf{v}(t)\|^2 + \nu\|P\Delta\mathbf{v}(t)\|^2 \le \frac{CK_0\|\mathbf{u}_0 - \mathbf{U}_0\|^2\,\|\nabla\mathbf{v}(t)\|^2}{\nu^3}\,\|\nabla\mathbf{v}(t)\|^2 \\
&+ \frac{C}{\nu^3}[(\|\mathbf{U}(t)\|^2\,\|\nabla\mathbf{U}(t)\|^2 + \nu^2\|\nabla\mathbf{U}(t)\|\,\|\Delta\mathbf{U}(t)\|)\|\nabla\mathbf{v}(t)\|^2 + \nu^2\|\mathbf{g}(t)\|^2]\,.
\end{aligned} \tag{2.48}$$

We introduce

$$\begin{aligned}
y(t) &\overset{\text{def}}{=\!=} \|\nabla\mathbf{v}(t)\|^2\,, \\
g(t) &\overset{\text{def}}{=\!=} \frac{CK_0\|\mathbf{u}_0 - \mathbf{U}_0\|^2\,\|\nabla\mathbf{v}(t)\|^2}{\nu^3},
\end{aligned}$$

and

$$h(t) \overset{\text{def}}{=\!=} \frac{C}{\nu^3}[(\|\mathbf{U}(t)\|^2\,\|\nabla\mathbf{U}(t)\|^2 + \nu^2\|\nabla\mathbf{U}(t)\|\,\|\Delta\mathbf{U}(t)\|)\|\nabla\mathbf{v}(t)\|^2 + \nu^2\|\mathbf{g}(t)\|^2]\,.$$

For each $\varepsilon > 0$, by Theorem 2.7 and (2.44) we have

$$\int_t^{t+\varepsilon} y(s)ds \le \|\mathbf{u}_0 - \mathbf{U}_0\|^2 \cdot \overline{K}_1\,,$$

$$\int_t^{t+\varepsilon} g(s)ds \le \|\mathbf{u}_0 - \mathbf{U}_0\|^4\,\frac{CK_0\overline{K}_1}{\nu^3},$$

and

$$\begin{aligned}
&\int_t^{t+\varepsilon} h(s)ds \\
&\le \|\mathbf{u}_0 - \mathbf{U}_0\|^2 \frac{C}{\nu^3} \left(|||\mathbf{U}|||^2 \, |||\nabla \mathbf{U}|||^2 + \nu^2 \, |||\nabla \mathbf{U}||| \, |||\Delta \mathbf{U}||| \right) \overline{K}_1 + \frac{C}{\nu} \int_t^{t+\varepsilon} \|\mathbf{g}(s)\|^2 ds \\
&\le \|\mathbf{u}_0 - \mathbf{U}_0\|^2 \frac{C}{\nu^3} \left(|||\mathbf{U}|||^2 \, |||\nabla \mathbf{U}|||^2 + \nu^2 \, |||\nabla \mathbf{U}||| \, |||\Delta \mathbf{U}||| \right) \overline{K}_1 + \frac{C}{\nu\beta} \mathcal{J}_T(\mathbf{u}, \mathbf{f})
\end{aligned}$$

so that by (2.36) and (2.18),

$$\begin{aligned}
&\int_t^{t+\varepsilon} h(s)ds \\
&\le \|\mathbf{u}_0 - \mathbf{U}_0\|^2 \frac{C}{\nu^3 \beta} \left\{ \beta \left[|||\mathbf{U}|||^2 \, |||\nabla \mathbf{U}|||^2 + \nu^2 \, |||\nabla \mathbf{U}||| \, |||\Delta \mathbf{U}||| \right] \overline{K}_1 + \nu^2 K \right\} .
\end{aligned}$$

Hence, (2.45) follows from the uniform Gronwall inequality (2.40) (applied to (2.48)) and the last three estimates for y, g, and h.

Finally, we prove $\mathbf{u} - \mathbf{U} \in C([\varepsilon, T]; \mathbf{H}^1(\Omega))$. Integrating (2.48) for $t \in [\varepsilon, T]$ and utilizing the bounds for $g(t)$, $h(t)$, and $y(t)$, we obtain

$$\begin{aligned}
\nu\lambda_1 \int_\varepsilon^T \|P\Delta v(t)\|^2 \, dt &\le y(\varepsilon) + \int_\varepsilon^T \left(g(t)y(t) + h(t) \right) \, dt \\
&\le \int_\varepsilon^T h(t) \, dt + \left(\sup_{t \in (\varepsilon, T)} y(t) \right) \cdot \left(1 + \int_\varepsilon^T g(t) \, dt \right) < \infty.
\end{aligned}$$

Hence $\mathbf{v} \in L^2(\varepsilon, T; \mathbf{H}^2(\Omega))$. From (2.39) we easily see that $\mathbf{v}_t = \nu P\Delta \mathbf{v} - P(\mathbf{v} \cdot \nabla)\mathbf{v} - P(\mathbf{U} \cdot \nabla)\mathbf{v} - P(\mathbf{v} \cdot \nabla)\mathbf{U} + P\mathbf{f} \in L^2(\varepsilon, T; \mathbf{L}^2(\Omega))$, where P is the Leray operator defined in (2.41). Therefore, $\nabla \mathbf{v} \in L^2(\varepsilon, T; (\mathbf{H}^1(\Omega))^2)$ and $(\nabla \mathbf{v})_t \in L^2(\varepsilon, T; (\mathbf{H}^{-1}(\Omega))^2)$. From [17, Lemma 5.5.1], we conclude that $\nabla \mathbf{v} \in C([\varepsilon, T]; \mathbf{L}^2(\Omega))$, which implies $\mathbf{v} \in C([\varepsilon, T]; \mathbf{H}^1(\Omega))$. ∎

An immediate consequence of Theorems 2.7 and 2.8 is the following preliminary estimates for the optimal solutions.

THEOREM 2.9. *Let $T \in (0, \infty]$. Assume $(\widehat{\mathbf{u}}, \widehat{\mathbf{f}}) \in \mathcal{U}_{ad}(T)$ is an optimal solution for* (2.10). *Then*

$$\|\widehat{\mathbf{u}}(t) - \mathbf{U}(t)\|^2 \le K_0 \cdot \|\mathbf{u}_0 - \mathbf{U}_0\|^2 \qquad \forall\, t \in [\varepsilon, T] \,, \tag{2.49}$$

$$\int_0^T \|\nabla \widehat{\mathbf{u}}(s) - \nabla \mathbf{U}(s)\|^2 ds \le \overline{K}_1 \cdot \|\mathbf{u}_0 - \mathbf{U}_0\|^2 \,, \tag{2.50}$$

and

$$\|\nabla \widehat{\mathbf{u}}(t) - \nabla \mathbf{U}(t)\|^2 \le K_1(\varepsilon) \cdot \|\mathbf{u}_0 - \mathbf{U}_0\|^2 \quad \forall\, t \in [\varepsilon, T] \,, \tag{2.51}$$

where all the constants are as defined in Theorems 2.7 *and* 2.8. ∎

Remark 2.10. The quantity $K_1(\varepsilon)$ is unbounded as $\varepsilon \to 0$:

$$K_1(\varepsilon) = O(1/\varepsilon) \qquad \text{as } \varepsilon \to 0^+. \tag{2.52}$$

Remark 2.11. If $T = \infty$, then the interval $[\varepsilon, T]$ in (2.43), (2.45), (2.49), and (2.51) (and elsewhere in this paper) should be understood as $[\varepsilon, \infty)$.

3. Existence of an optimal control.

3.1. The case of finite time intervals. In this subsection, we first quote the results of [8], [9], and [10] concerning the existence of an optimal solution for (2.10) with $T < \infty$ and concerning an optimality system. We then derive some estimates for the adjoint equations.

With the notion of admissible elements, admissible sets and optimal solutions were introduced in section 2.1 together with the functional spaces. We may state the results of [8], [9], and [10] as follows.

THEOREM 3.1. *Let $T \in (0,\infty)$. Then there exists a optimal solution $(\widehat{\mathbf{u}},\widehat{\mathbf{f}}) \in \mathcal{U}_{ad}(T)$ for the problem* (2.10)*. Furthermore, there exists $\widehat{\boldsymbol{\xi}} \in \mathcal{V}^{(1)}(Q_T)$ such that*

$$-\langle \partial_t \widehat{\boldsymbol{\xi}}(t), \mathbf{w}\rangle + a(\widehat{\boldsymbol{\xi}}(t),\mathbf{w}) + c(\widehat{\mathbf{u}}(t),\mathbf{w},\widehat{\boldsymbol{\xi}}(t)) + c(\mathbf{w},\widehat{\mathbf{u}}(t),\widehat{\boldsymbol{\xi}}(t)) = \alpha\,(\widehat{\mathbf{v}}(t),\mathbf{w}) \qquad \forall\,\mathbf{w}\in\mathbf{V} \text{ a.e. } t\in(0,T)\,, \tag{3.1}$$

$$\lim_{t\to T^-} \widehat{\boldsymbol{\xi}}(t) = \mathbf{0} \qquad \text{in } \mathbf{W}, \tag{3.2}$$

and

$$\widehat{\boldsymbol{\xi}} + \beta\,\widehat{\mathbf{g}} = \mathbf{0} \qquad \text{in } \Omega\times(0,T), \tag{3.3}$$

where $\widehat{\mathbf{v}} \overset{\text{def}}{=\!=} \widehat{\mathbf{u}} - \mathbf{U}$ and $\widehat{\mathbf{g}} \overset{\text{def}}{=\!=} \widehat{\mathbf{f}} - \mathbf{N}(\mathbf{U})$. □

From this theorem we see that the optimal solution $(\widehat{\mathbf{u}},\widehat{\mathbf{f}})$ together with the Lagrange multiplier $\widehat{\boldsymbol{\xi}}$ satisfies equations (2.12), (2.13) and (3.1), (3.3). Note that (3.3) allows us to eliminate the variable $\mathbf{f}$ in (2.12). We collect these equations here to form an optimality system of equations:

$$\langle \partial_t \widehat{\mathbf{u}}(t),\mathbf{w}\rangle + a(\widehat{\mathbf{u}}(t),\mathbf{w}) + c(\widehat{\mathbf{u}}(t),\widehat{\mathbf{u}}(t),\mathbf{w}) = (\widehat{\mathbf{f}}(t),\mathbf{w}) \qquad \forall\,\mathbf{w}\in\mathbf{V} \text{ a.e. } t\in(0,T), \tag{3.4}$$

$$\lim_{t\to 0^+} \widehat{\mathbf{u}}(t) = \mathbf{u}_0 \qquad \text{in } \mathbf{W}\,, \tag{3.5}$$

$$-\langle \partial_t \widehat{\boldsymbol{\xi}}(t),\mathbf{w}\rangle + a(\widehat{\boldsymbol{\xi}}(t),\mathbf{w}) + c(\widehat{\mathbf{u}}(t),\mathbf{w},\widehat{\boldsymbol{\xi}}(t)) + c(\mathbf{w},\widehat{\mathbf{u}}(t),\widehat{\boldsymbol{\xi}}(t)) = \alpha\,(\widehat{\mathbf{v}}(t),\mathbf{w}) \qquad \forall\,\mathbf{w}\in\mathbf{V} \text{ a.e. } t\in(0,T), \tag{3.6}$$

and

$$\lim_{t\to T^-} \widehat{\boldsymbol{\xi}}(t) = \mathbf{0} \qquad \text{in } \mathbf{W}\,. \tag{3.7}$$

Based on the optimality system and the preliminary estimates for the optimal solutions, we may obtain estimates for $\widehat{\boldsymbol{\xi}}$ on finite time intervals.

THEOREM 3.2. *Let $T \in (0,\infty)$. Assume that $(\widehat{\mathbf{u}},\widehat{\mathbf{f}}) \in \mathcal{U}_{ad}(T)$ is a solution for* (2.10) *and that $\widehat{\boldsymbol{\xi}} \in \mathcal{V}^{(1)}(Q_T)$ is a solution for* (3.1)–(3.3)*. Then, for each $\varepsilon > 0$,*

$$\|\widehat{\boldsymbol{\xi}}(t)\|^2 + \nu\int_t^T \|\nabla\widehat{\boldsymbol{\xi}}(s)\|^2 ds \le \frac{C}{\nu}\max\{\alpha,\beta\rho_1^2(\varepsilon)\}\mathcal{J}_T(\widehat{\mathbf{u}},\widehat{\mathbf{f}}) \qquad \forall\, t\in[\varepsilon,T]\,, \tag{3.8}$$

where

$$\rho_1(\varepsilon) \overset{\text{def}}{=\!=} K_1(\varepsilon)\,\|\mathbf{u}_0 - \mathbf{U}_0\| + |||\nabla\mathbf{U}||| < \infty\,. \tag{3.9}$$

Proof. For each $\varepsilon > 0$, we clearly have $\rho_1(\varepsilon) < \infty$. Equation (3.3) yields

$$\widehat{\boldsymbol{\xi}} = -\beta\, \widehat{\mathbf{g}}\,.$$

Setting $\mathbf{w} = \widehat{\boldsymbol{\xi}}(t)$ in (3.1), we have

$$\begin{aligned} -\frac{1}{2}\frac{d}{dt}\|\widehat{\boldsymbol{\xi}}(t)\|^2 + \nu\,\|\nabla\widehat{\boldsymbol{\xi}}(t)\|^2 &= \alpha\,(\widehat{\mathbf{v}}(t),\widehat{\boldsymbol{\xi}}(t)) - c(\widehat{\boldsymbol{\xi}}(t),\widehat{\mathbf{u}}(t),\widehat{\boldsymbol{\xi}}(t)) \\ &\le \alpha\,\|\widehat{\mathbf{v}}(t)\|\,\|\widehat{\boldsymbol{\xi}}(t)\| + C\,\|\nabla\widehat{\mathbf{u}}(t)\|\,\|\widehat{\boldsymbol{\xi}}(t)\|\,\|\nabla\widehat{\boldsymbol{\xi}}(t)\|\,. \end{aligned} \tag{3.10}$$

From the triangle inequality and estimate (2.51) we obtain

$$\|\nabla\widehat{\mathbf{u}}(t)\| \le \|\nabla\widehat{\mathbf{u}}(t) - \nabla\mathbf{U}(t)\| + \|\nabla\mathbf{U}(t)\| \le \rho_1(\varepsilon) \qquad \forall\, t \in [\varepsilon, T]\,.$$

Using Young's inequality in (3.10) and using the last relation, we are led to

$$\begin{aligned} -\frac{d}{dt}\|\widehat{\boldsymbol{\xi}}(t)\|^2 + \nu\,\|\nabla\widehat{\boldsymbol{\xi}}(t)\|^2 &\le \frac{C}{\nu}(\alpha^2\|\widehat{\mathbf{v}}(t)\|^2 + \|\nabla\widehat{\mathbf{u}}(t)\|^2\,\|\widehat{\boldsymbol{\xi}}(t)\|^2) \\ &\le \frac{C}{\nu}\max\left\{\alpha, \beta\rho_1^2(\varepsilon)\right\}\,\left(\alpha\|\widehat{\mathbf{v}}(t)\|^2 + \beta\|\widehat{\mathbf{g}}(t)\|^2\right)\,. \end{aligned}$$

Integrating both sides over the interval $(t, T') \subset (0, T)$ and using the fact that $\lim_{T'\to T^-}\widehat{\boldsymbol{\xi}}(T') = \mathbf{0}$, we obtain the desired estimate for $\boldsymbol{\xi}$. ∎

3.2. The case of the infinite time interval. We now prove the existence of an optimal solution for (2.10) on the infinite time interval $(0, \infty)$. We will make use of the existence results on finite time intervals.

THEOREM 3.3. *There exists a solution* $(\widehat{\mathbf{u}}, \widehat{\mathbf{f}}) \in \mathcal{U}_{ad}(\infty)$ *for* (2.10) *with* $T = \infty$. *Furthermore, there exists a* $\widehat{\boldsymbol{\xi}} \in \mathcal{V}^{(1)}_{\rm loc}(Q)$ *which satisfies* (3.1) *and* (3.3) *with* $T = \infty$.

Proof. For each $T \in (0, \infty)$, we may use Theorem 3.1 to choose a $(\mathbf{u}_T, \mathbf{f}_T) \in \mathcal{U}_{ad}(T)$ which solves (2.10). Thus, $(\mathbf{u}_T, \mathbf{f}_T)$ satisfies

$$\mathcal{J}_T(\mathbf{u}_T, \mathbf{f}_T) = \inf_{(\mathbf{w},\mathbf{h})\in\mathcal{U}_{ad}(T)} \mathcal{J}_T(\mathbf{w}, \mathbf{h})\,, \tag{3.11}$$

$$\begin{aligned} \langle \partial_t \mathbf{u}_T(t), \mathbf{w}\rangle + a(\mathbf{u}_T(t), \mathbf{w}) + c(\mathbf{u}_T(t), \mathbf{u}_T(t), \mathbf{w}) \\ = (\mathbf{f}_T(t), \mathbf{w}) \qquad \forall\, \mathbf{w} \in \mathbf{V} \text{ a.e. } t \in (0, T), \end{aligned} \tag{3.12}$$

and

$$\lim_{t\to 0^+} \mathbf{u}_T(t) = \mathbf{u}_0 \qquad \text{in } \mathbf{W}\,. \tag{3.13}$$

Furthermore, there exists a $\boldsymbol{\xi}_T \in \mathcal{V}^{(1)}(Q_T)$ which satisfies (3.1)–(3.3).

For each finite T, we obviously have $[\mathcal{U}_{ad}(\infty)]|_{(0,T)} \subset \mathcal{U}_{ad}(T)$. Thus, for each $(\mathbf{w}, \mathbf{h}) \in \mathcal{U}_{ad}(\infty)$,

$$\mathcal{J}_T(\mathbf{u}_T, \mathbf{f}_T) \le \mathcal{J}_T(\mathbf{w}, \mathbf{h}) \le \mathcal{J}_\infty(\mathbf{w}, \mathbf{h})\,. \tag{3.14}$$

Since $\mathcal{J}_\infty(\widetilde{\mathbf{u}}, \widetilde{\mathbf{f}}) < \infty$ so that $(\widetilde{\mathbf{u}}, \widetilde{\mathbf{f}}) \in \mathcal{U}_{ad}(\infty)$, where $(\widetilde{\mathbf{u}}, \widetilde{\mathbf{f}})$ is constructed in section 2.2, it follows that $\inf_{(\mathbf{w},\mathbf{h})\in\mathcal{U}_{ad}(\infty)} \mathcal{J}_\infty(\mathbf{w}, \mathbf{h}) < \infty$. For each $T \in (0, \infty)$, we obtain from (3.14) that

$$\mathcal{J}_T(\mathbf{u}_T, \mathbf{f}_T) \le \inf_{(\mathbf{w},\mathbf{h})\in\mathcal{U}_{ad}(\infty)} \mathcal{J}_\infty(\mathbf{w}, \mathbf{h}) < \infty\,. \tag{3.15}$$

For each integer $k > 0$, we denote by $(\mathbf{u}_k, \mathbf{f}_k)$ a solution of (3.11)–(3.13) for $T = k$ and denote by $\boldsymbol{\xi}_k$ the corresponding multiplier which satisfies (3.1)–(3.3). We set $(\mathbf{v}_k, \mathbf{g}_k) \stackrel{\text{def}}{=\!=} (\mathbf{u}_k - \mathbf{U}, \mathbf{f}_k - \mathbf{N}(\mathbf{U}))$. Then, $(\mathbf{v}_k, \mathbf{g}_k)$ is a solution of (2.12), (2.13) with $T = k$. Using (3.15) and standard estimates for the Navier–Stokes equations on finite time intervals, we obtain that $\|\mathbf{g}_k\|_{L^2(0,k;\mathbf{L}^2(\Omega))}$, $\|\mathbf{u}_k\|_{\mathcal{V}^{(1)}(Q_k)}$, and $\|\mathbf{u}_k\|_{L^\infty(0,k;\mathbf{W})}$ are uniformly bounded for all k. Using the estimate (3.8), we obtain that $\|\boldsymbol{\xi}_k\|_{\mathcal{V}^{(1)}(\Omega\times(\epsilon,k))}$ and $\|\boldsymbol{\xi}_k\|_{L^\infty(\epsilon,k;\mathbf{W})}$ are uniformly bounded for all $\varepsilon > 0$ and k. Hence, by induction we may choose successive subsequences of positive integers $\{k_n^{(m)}\}_{n=1}^\infty$ for $m = 1, 2, \dots$ such that $\{k_n^{(1)}\}_{n=1}^\infty \supset \{k_n^{(2)}\}_{n=1}^\infty \supset \{k_n^{(3)}\}_{n=1}^\infty \supset \cdots$ and

$$
\begin{aligned}
&\mathbf{v}_{k_n^{(m)}} \rightharpoonup \mathbf{v}^{(m)} \quad \text{in } \mathcal{V}^{(1)}(Q_m) \text{ as } n \to \infty\,,\\
&\mathbf{v}_{k_n^{(m)}} \stackrel{*}{\rightharpoonup} \mathbf{v}^{(m)} \quad \text{in } L^\infty(0, m; \mathbf{W}) \text{ as } n \to \infty\,,\\
&\mathbf{g}_{k_n^{(m)}} \rightharpoonup \mathbf{g}^{(m)} \quad \text{in } L^2(0, m; \mathbf{W}) \text{ as } n \to \infty\,,\\
&\boldsymbol{\xi}_{k_n^{(m)}} \rightharpoonup \boldsymbol{\xi}^{(m)} \quad \text{in } \mathcal{V}^{(1)}(\Omega \times (1/m, m)) \text{ as } n \to \infty,
\end{aligned}
$$

and

$$
\boldsymbol{\xi}_{k_n^{(m)}} \stackrel{*}{\rightharpoonup} \boldsymbol{\xi}^{(m)} \quad \text{in } L^\infty(1/m, m; \mathbf{W}) \text{ as } n \to \infty
$$

for some $\mathbf{v}^{(m)} \in \mathcal{V}^{(1)}(Q_m)$, $\mathbf{g}^{(m)} \in L^2(0, m; \mathbf{W})$, and $\boldsymbol{\xi}^{(m)} \in \mathcal{V}^{(1)}(\Omega \times (1/m, m))$. (We remark that $\mathcal{V}^{(1)}(Q_m) \rightharpoonup L^\infty(0, m; \mathbf{W})$ and $\mathcal{V}^{(1)}(\Omega \times (1/m, m)) \rightharpoonup L^\infty(\varepsilon, m; \mathbf{W})$.) Hence, by extracting the diagonal subsequence, we have that for each m',

$$
\mathbf{v}_{k_m^{(m)}} \rightharpoonup \mathbf{v}^{(m')} \quad \text{in } \mathcal{V}^{(1)}(Q_{m'}) \text{ as } m \to \infty\,, \tag{3.16}
$$

$$
\mathbf{v}_{k_m^{(m)}} \stackrel{*}{\rightharpoonup} \mathbf{v}^{(m')} \quad \text{in } L^\infty(0, m'; \mathbf{W}) \text{ as } m \to \infty\,, \tag{3.17}
$$

$$
\mathbf{g}_{k_m^{(m)}} \rightharpoonup \mathbf{g}^{(m')} \quad \text{in } L^2(0, m'; \mathbf{W}) \text{ as } m \to \infty\,, \tag{3.18}
$$

$$
\boldsymbol{\xi}_{k_m^{(m)}} \rightharpoonup \boldsymbol{\xi}^{(m')} \quad \text{in } \mathcal{V}^{(1)}(\Omega \times (1/m', m')) \text{ as } m \to \infty \tag{3.19}
$$

and

$$
\boldsymbol{\xi}_{k_m^{(m)}} \stackrel{*}{\rightharpoonup} \boldsymbol{\xi}^{(m')} \quad \text{in } L^\infty(1/m', m'; \mathbf{W}) \text{ as } m \to \infty\,. \tag{3.20}
$$

For each integer $m' > 0$, (3.16)–(3.18) and standard techniques for the Navier–Stokes equations (see, e.g., [18]) allow us to pass to the limit as $m \to \infty$ in the equation

$$
\begin{aligned}
&\int_0^{m'} \langle \partial_t \mathbf{v}_{k_m^{(m)}}(t), \mathbf{w}\rangle \psi(t)\, dt + \int_0^{m'} a(\mathbf{v}_{k_m^{(m)}}(t), \mathbf{w})\psi(t)\, dt\\
&\quad + \int_0^{m'} c(\mathbf{v}_{k_m^{(m)}}(t), \mathbf{v}_{k_m^{(m)}}(t), \mathbf{w})\psi(t)\, dt\\
&\quad + \int_0^{m'} c(\mathbf{U}(t), \mathbf{v}_{k_m^{(m)}}(t), \mathbf{w})\psi(t)\, dt + \int_0^{m'} c(\mathbf{v}_{k_m^{(m)}}(t), \mathbf{U}(t), \mathbf{w})\psi(t)\, dt\\
&\quad = \int_0^{m'} (\mathbf{g}_{k_m^{(m)}}(t), \mathbf{w})\psi(t)\, dt \qquad \forall\, \mathbf{w} \in \mathbf{V},\ \psi \in C[0, m'] \text{ with } \psi(m') = 0
\end{aligned} \tag{3.21}
$$

to obtain

$$
\begin{aligned}
(3.22)\quad &\int_0^{m'} \langle \partial_t \mathbf{v}^{(m')}(t), \mathbf{w}\rangle \psi(t)\,dt + \int_0^{m'} a(\mathbf{v}^{(m')}(t), \mathbf{w})\psi(t)\,dt \\
&\quad + \int_0^{m'} c(\mathbf{v}^{(m')}(t), \mathbf{v}^{(m')}(t), \mathbf{w})\psi(t)\,dt \\
&\quad + \int_0^{m'} c(\mathbf{U}(t), \mathbf{v}^{(m')}(t), \mathbf{w})\psi(t)\,dt + \int_0^{m'} c(\mathbf{v}^{(m')}(t), \mathbf{U}(t), \mathbf{w})\psi(t)\,dt \\
&\quad = \int_0^{m'} (\mathbf{g}^{(m')}(t), \mathbf{w})\psi(t)\,dt \qquad \forall\, \mathbf{w} \in \mathbf{V},\ \psi \in C[0, m'] \text{ with } \psi(m') = 0,
\end{aligned}
$$

which is equivalent to

$$
\begin{aligned}
(3.23)\quad &\langle \partial_t \mathbf{u}^{(m')}(t), \mathbf{w}\rangle + a(\mathbf{u}^{(m')}(t), \mathbf{w}) + c(\mathbf{u}^{(m')}(t), \mathbf{u}^{(m')}(t), \mathbf{w}) \\
&\qquad = (\mathbf{f}^{(m')}(t), \mathbf{w}) \qquad \forall\, \mathbf{w} \in \mathbf{V} \text{ a.e. } t \in (0, m'),
\end{aligned}
$$

where $\mathbf{u}^{(m')} = \mathbf{v}^{(m')} + \mathbf{U}$. Similarly, for each positive integer m', we may pass to the limit as $m \to \infty$ in the equations

$$
\begin{aligned}
(3.24)\quad &-\langle \partial_t \boldsymbol{\xi}_{k_m^{(m)}}(t), \mathbf{w}\rangle + a(\boldsymbol{\xi}_{k_m^{(m)}}(t), \mathbf{w}) + c(\mathbf{u}_{k_m^{(m)}}(t), \mathbf{w}, \boldsymbol{\xi}_{k_m^{(m)}}(t)) \\
&\quad + c(\mathbf{w}, \mathbf{u}_{k_m^{(m)}}(t), \boldsymbol{\xi}_{k_m^{(m)}}(t)) = \alpha\,(\mathbf{v}_{k_m^{(m)}}(t), \mathbf{w}) \qquad \forall\, \mathbf{w} \in \mathbf{V} \text{ a.e. } t \in (0, m')
\end{aligned}
$$

and

$$
(3.25)\qquad \boldsymbol{\xi}_{k_m^{(m)}} + \beta\, \mathbf{g}_{k_m^{(m)}} = \mathbf{0} \qquad \text{in } \Omega \times (1/m', m')
$$

to obtain

$$
\begin{aligned}
&-\langle \partial_t \boldsymbol{\xi}^{(m')}(t), \mathbf{w}\rangle + a(\boldsymbol{\xi}^{(m')}(t), \mathbf{w}) + c(\mathbf{u}^{(m')}(t), \mathbf{w}, \boldsymbol{\xi}^{(m')}(t)) \\
&\quad + c(\mathbf{w}, \mathbf{u}^{(m')}(t), \boldsymbol{\xi}^{(m')}(t)) = \alpha\,(\mathbf{v}^{(m')}(t), \mathbf{w}) \qquad \forall\, \mathbf{w} \in \mathbf{V} \text{ a.e. } t \in (0, m')
\end{aligned}
$$

and

$$
\boldsymbol{\xi}^{(m')} + \beta\, \mathbf{g}^{(m')} = \mathbf{0} \qquad \text{in } \Omega \times (1/m', m').
$$

By the uniqueness of weak limits, we have that $\mathbf{v}^{(m_1)}|_{(0,m_2)} = \mathbf{v}^{(m_2)}$, $\mathbf{g}^{(m_1)}|_{(0,m_2)} = \mathbf{g}^{(m_2)}$ and $\boldsymbol{\xi}^{(m_1)}|_{(0,m_2)} = \boldsymbol{\xi}^{(m_2)}$ for all m_1, m_2 with $m_1 < m_2$. Thus, the functions

$$
\widehat{\mathbf{v}}(t) \stackrel{\text{def}}{=\!=} \mathbf{v}^{(m)}(t) \qquad (\text{if } t \le m)\,,
$$

$$
\widehat{\mathbf{g}}(t) \stackrel{\text{def}}{=\!=} \mathbf{g}^{(m)}(t) \qquad (\text{if } t \le m),
$$

and

$$
\widehat{\boldsymbol{\xi}}(t) \stackrel{\text{def}}{=\!=} \boldsymbol{\xi}^{(m)}(t) \qquad (\text{if } (1/m) \le t \le m)
$$

are well defined on $(0, \infty)$; furthermore, $\widehat{\mathbf{v}} \in \mathcal{V}^{(1)}_{\text{loc}}(Q)$, $\widehat{\mathbf{g}} \in L^2(0, \infty; \mathbf{L}^2(\Omega))$, and $\widehat{\boldsymbol{\xi}} \in \mathcal{V}^{(1)}_{\text{loc}}(Q)$. Upon setting $\widehat{\mathbf{u}} = \widehat{\mathbf{v}} + \mathbf{U}$ and $\widehat{\mathbf{f}} = \mathbf{g} + \mathbf{F}$ and noting that m' is arbitrary in (3.23)–(3.25), we are easily led to

$$
\begin{aligned}
(3.26)\quad &\langle \partial_t \widehat{\mathbf{u}}(t), \mathbf{w}\rangle + a(\widehat{\mathbf{u}}(t), \mathbf{w}) + c(\widehat{\mathbf{u}}(t), \widehat{\mathbf{u}}(t), \mathbf{w}) \\
&\qquad = (\widehat{\mathbf{f}}(t), \mathbf{w}) \qquad \forall\, \mathbf{w} \in V \text{ a.e. } t \in (0, \infty)\,,
\end{aligned}
$$

$$-\langle \partial_t \widehat{\boldsymbol{\xi}}(t), \mathbf{w}\rangle + a(\widehat{\boldsymbol{\xi}}(t), \mathbf{w}) + c(\mathbf{u}(t), \mathbf{w}, \widehat{\boldsymbol{\xi}}(t)) + c(\mathbf{w}, \mathbf{u}(t), \widehat{\boldsymbol{\xi}}(t)) = \alpha\,(\widehat{\mathbf{v}}(t), \mathbf{w}) \qquad \forall\, \mathbf{w} \in \mathbf{V} \text{ a.e. } t \in (0,\infty)\,,$$

and

$$\widehat{\boldsymbol{\xi}} + \beta\,\widehat{\mathbf{g}} = \mathbf{0} \qquad \text{in } \Omega \times (0,\infty)\,.$$

We now examine the initial condition for $\widehat{\mathbf{u}}$. Let ψ be a continuously differentiable function in $[0,\infty)$ with a bounded support. Equation (3.21) can be rewritten as

$$\langle \partial_t \mathbf{u}_{k_m^{(m)}}(t), \mathbf{w}\rangle + a(\mathbf{u}_{k_m^{(m)}}(t), \mathbf{w}) + c(\mathbf{u}_{k_m^{(m)}}(t), \mathbf{u}_{k_m^{(m)}}(t), \mathbf{w}) = (\mathbf{f}_{k_m^{(m)}}(t), \mathbf{w}) \qquad \forall\, \mathbf{w} \in \mathbf{V} \text{ a.e. } t \in (0, m') \quad (m > m')\,.$$

Multiplying the last equation by $\psi(t)$, integrating by parts, and using the fact that $\mathbf{u}_{k_m^{(m)}}(0) = \mathbf{u}_0$, we are led to

$$-\int_0^\infty (\mathbf{u}_{k_m^{(m)}}(t), \mathbf{z})\,\psi'(t)\,dt + \nu\int_0^\infty\int_\Omega \psi(t)\nabla \mathbf{u}_{k_m^{(m)}}(t) : \nabla \mathbf{z}\,d\mathbf{x}\,dt + \int_0^\infty\int_\Omega \psi(t)\,(\mathbf{u}_{k_m^{(m)}}(t)\cdot\nabla)\mathbf{u}_{k_m^{(m)}}(t)\cdot\mathbf{z}\,d\mathbf{x}\,dt = \psi(0)\,(\mathbf{u}_0, \mathbf{z}) \qquad \forall\, \mathbf{z} \in \mathbf{V}\,.$$

Thus, by passing to the limit in the last equation we obtain

$$-\int_0^\infty (\widehat{\mathbf{u}}(t), \mathbf{z})\,\psi'(t)\,dt + \nu\int_0^\infty\int_\Omega \psi(t)\nabla \widehat{\mathbf{u}}(t) : \nabla \mathbf{z}\,d\mathbf{x}\,dt + \int_0^\infty\int_\Omega \psi(t)\,(\widehat{\mathbf{u}}(t)\cdot\nabla)\widehat{\mathbf{u}}(t)\cdot\mathbf{z}\,d\mathbf{x}\,dt = \psi(0)\,(\mathbf{u}_0, \mathbf{z}) \qquad \forall\, \mathbf{z} \in \mathbf{V}\,. \tag{3.27}$$

On the other hand, by multiplying (3.26) by $\psi(t)$ and integrating by parts we obtain

$$-\int_0^\infty (\widehat{\mathbf{u}}(t), \mathbf{z})\,\psi'(t)\,dt + \nu\int_0^\infty\int_\Omega \psi(t)\nabla \widehat{\mathbf{u}}(t) : \nabla \mathbf{z}\,d\mathbf{x}\,dt + \int_0^\infty\int_\Omega \psi(t)\,(\widehat{\mathbf{u}}(t)\cdot\nabla)\widehat{\mathbf{u}}(t)\cdot\mathbf{z}\,d\mathbf{x}\,dt = \psi(0)\,(\widehat{\mathbf{u}}(0), \mathbf{z}) \qquad \forall\, \mathbf{z} \in \mathbf{V}\,. \tag{3.28}$$

Here we have used the continuous imbedding result $\mathcal{V}^{(1)}(Q) \hookrightarrow \mathbf{C}([0,T];\mathbf{W})$ so that $\widehat{\mathbf{u}}(0)$ is well defined in $\mathbf{W}$. A comparison of (3.27) and (3.28) yields $\widehat{\mathbf{u}}(0) = \mathbf{u}_0$ in $\mathbf{W}$.

Finally, using the lower semicontinuity of the functional $\mathcal{J}_T(\cdot,\cdot)$ and the fact that $\widehat{\mathbf{v}} = \widehat{\mathbf{u}} - \mathbf{U} \in L^2(0,\infty;\mathbf{V})$ and $\widehat{\mathbf{g}} = \widehat{\mathbf{f}} - \mathbf{F} \in L^2(0,\infty;\mathbf{W})$, we obtain

$$\mathcal{J}_{k_m^{(m)}}(\widehat{\mathbf{u}}, \widehat{\mathbf{f}}) \le \liminf_{m\to\infty} \mathcal{J}_{k_m^{(m)}}(\mathbf{u}_{k_m^{(m)}}, \mathbf{f}_{k_m^{(m)}}) \le \mathcal{J}_\infty(\mathbf{w}, \mathbf{h}) \quad \forall\,(\mathbf{w}, \mathbf{h}) \in \mathcal{U}_{ad}(\infty)$$

so that by letting $m \to \infty$,

$$\mathcal{J}_\infty(\widehat{\mathbf{u}}, \widehat{\mathbf{f}}) \le \mathcal{J}_\infty(\mathbf{w}, \mathbf{h}) \quad \forall\,(\mathbf{w}, \mathbf{h}) \in \mathcal{U}_{ad}(\infty)\,.$$

Hence we have shown that $(\widehat{\mathbf{u}}, \widehat{\mathbf{f}})$ is the desired optimizer for (2.10) with $T = \infty$. □

4. Dynamics of optimal control solutions on the infinite time interval. As mentioned in the introduction, one of the motivations of considering infinite time optimal control is that one wishes the dynamics, or at least the eventual dynamics, of optimal control solutions to match well with the desired dynamics. Minimizing the cost functional (1.1) forces the controlled flow $\widehat{\mathbf{u}}$ to be close to the desired flow $\mathbf{U}$ only in the L^2 sense in t. For some t $\|\mathbf{u}(t) - \mathbf{U}(t)\|$ can still be very large. To reduce the error of the dynamics in t, e.g., to reduce the value of $\|\mathbf{u}(t) - \mathbf{U}(t)\|$ as $t \to \infty$, one needs to control the system for a long time. For many feedback control models, the controlled flow exponentially decays to the desired flow. For the optimal control system with a functional of the type (1.1), this is not expected to be true. But we may still obtain some results on their pointwise dynamics. Theorems 2.7 and 2.8 gave some preliminary results in this regard, i.e., $\|\mathbf{u}(t) - \mathbf{U}(t)\|$ and $\|\nabla\mathbf{u}(t) - \nabla\mathbf{U}(t)\|$ stay bounded. We will prove much stronger results in this section: $\|\mathbf{u}(t) - \mathbf{U}(t)\|$ and $\|\nabla\mathbf{u}(t) - \nabla\mathbf{U}(t)\|$ approach zero as t goes to ∞. We point out that these last results are not unique to the solutions of the optimal control system; these results can be proved under weaker conditions.

We first establish the "reverse" inequalities as opposed to the inequalities in Theorem 2.7.

LEMMA 4.1. *Let $T \in (0, \infty]$. Assume $(\mathbf{u}, \mathbf{f}) \in \mathcal{U}_{ad}(T)$. If $\|\mathbf{u}(t) - \mathbf{U}(t)\| > 0$ for all $t \in (t_1, t_2) \subset [0, T]$, then*

$$\begin{aligned} \|\mathbf{u}(t_1) - \mathbf{U}(t_1)\| &\geq \|\mathbf{u}(t_2) - \mathbf{U}(t_2)\| \\ &\quad - C\sqrt{t_2 - t_1}\left(\frac{|||\nabla\mathbf{U}|||^4}{\alpha\nu^2} + \frac{1}{\beta}\right)^{1/2} \left(\mathcal{J}_T(\mathbf{u}, \mathbf{f})\right)^{1/2}. \end{aligned} \tag{4.1}$$

Assume further that (2.36) *holds, i.e., $\mathcal{J}_T(\mathbf{u}, \mathbf{f}) \leq \mathcal{J}_T(\widetilde{\mathbf{u}}, \widetilde{\mathbf{f}})$, where $(\widetilde{\mathbf{u}}, \widetilde{\mathbf{f}})$ is as defined in* Theorem 2.4, *then*

$$\begin{aligned} \|\mathbf{u}(t_1) - \mathbf{U}(t_1)\| &\geq \|\mathbf{u}(t_2) - \mathbf{U}(t_2)\| \\ &\quad - C\sqrt{t_2 - t_1}\,\|\mathbf{u}_0 - \mathbf{U}_0\|\left(\frac{|||\nabla\mathbf{U}|||^4}{\alpha\nu^2} + \frac{1}{\beta}\right)^{1/2} \left(K(\nabla\mathbf{U}, \alpha, \beta)\right)^{1/2}. \end{aligned} \tag{4.2}$$

Proof. By setting $\mathbf{w} = \mathbf{v}(t)$ in (2.12) we obtain

$$\begin{aligned} \|\mathbf{v}\|\frac{d}{dt}\|\mathbf{v}(t)\| + \nu\|\nabla\mathbf{v}(t)\|^2 &\leq C\,|||\nabla\mathbf{U}|||\,\|\mathbf{v}(t)\|\,\|\nabla\mathbf{v}(t)\| + \|\mathbf{g}(t)\|\,\|\mathbf{v}(t)\| \\ &\leq \frac{\nu}{2}\|\nabla\mathbf{v}(t)\|^2 + \frac{C\,|||\nabla\mathbf{U}|||^2}{\nu}\|\mathbf{v}(t)\|^2 + \|\mathbf{g}(t)\|\,\|\mathbf{v}(t)\| \end{aligned}$$

for all $t \in (0, T)$. If $\|\mathbf{u}(t) - \mathbf{U}(t)\| > 0$ for all $t \in (t_1, t_2)$, then we may divide the last inequality by $\|\mathbf{v}(t)\|$ to obtain

$$\begin{aligned} \frac{d}{dt}\|\mathbf{v}(t)\| + \frac{\nu\lambda_1}{2}\|\mathbf{v}(t)\| &\leq C\left(\frac{|||\nabla\mathbf{U}|||^2}{\nu}\|\mathbf{v}(t)\| + \|\mathbf{g}(t)\|\right) \\ &\leq C\left(\frac{|||\nabla\mathbf{U}|||^4}{\alpha\nu^2} + \frac{1}{\beta}\right)^{1/2} \left(\alpha\|\mathbf{v}(t)\|^2 + \beta\|\mathbf{g}(t)\|^2\right)^{1/2} \end{aligned}$$

for all $t \in (t_1, t_2)$. Multiplying the last inequality by $e^{\nu\lambda_1 t/2}$ and integrating over (t_1, t_2), we are led to

$$\begin{aligned}
\|\mathbf{v}(t_2)\| &\le \|\mathbf{v}(t_1)\|\, e^{-\nu\lambda_1(t_2-t_1)/2} \\
&\quad + Ce^{-\nu\lambda_1 t_2/2}\left(\frac{|||\nabla\mathbf{U}|||^4}{\alpha\nu^2}+\frac{1}{\beta}\right)^{1/2}\int_{t_1}^{t_2}\left(\alpha\|\mathbf{v}(s)\|^2+\beta\|\mathbf{g}(s)\|^2\right)^{1/2} e^{\nu\lambda_1 s/2}\,ds \\
&\le \|\mathbf{v}(t_1)\| + C\left(\frac{|||\nabla\mathbf{U}|||^4}{\alpha\nu^2}+\frac{1}{\beta}\right)^{1/2} \\
&\quad \cdot\left(\int_{t_1}^{t_2}\left[\alpha\|\mathbf{v}(s)\|^2+\beta\|\mathbf{g}(s)\|^2\right]ds\right)^{1/2}\left(e^{-\nu\lambda_1 t_2}\int_{t_1}^{t_2} e^{\nu\lambda_1 s}\,ds\right)^{1/2} \\
&\le \|\mathbf{v}(t_1)\| + C\left(\frac{|||\nabla\mathbf{U}|||^4}{\alpha\nu^2}+\frac{1}{\beta}\right)^{1/2}\left(\mathcal{J}_T(\mathbf{u},\mathbf{f})\right)^{1/2}\left(\frac{1-\exp\{-\nu\lambda_1(t_2-t_1)\}}{\nu\lambda_1}\right)^{1/2}
\end{aligned}$$

so that by applying the mean value theorem to the last factor we have shown (4.1). (4.2) simply follows from the bound (2.18). □

We are now prepared to establish the asymptotic decay property of $\|\mathbf{u}(t)-\mathbf{U}(t)\|$ as $t\to\infty$ for any $(\mathbf{u},\mathbf{f})\in\mathcal{U}_{ad}(\infty)$.

THEOREM 4.2. *Assume that* $(\mathbf{u},\mathbf{f})\in\mathcal{U}_{ad}(\infty)$. *Then*

$$\lim_{t\to\infty}\|\mathbf{u}(t)-\mathbf{U}(t)\| = 0. \tag{4.3}$$

Proof. The theorem is trivial if $J_\infty(\mathbf{u},\mathbf{f})=0$. Thus we assume $J_\infty(\mathbf{u},\mathbf{f})>0$ and proceed to prove (4.3) by contradiction. Assume that (4.3) is false. Then we may choose an $\varepsilon_0>0$ and a sequence $\{t_n\}$ such that $t_n\to\infty$ and

$$\|\mathbf{u}(t_n)-\mathbf{U}(t_n)\|\ge\varepsilon_0>0.$$

Upon setting

$$\delta \stackrel{\text{def}}{=\!=} \frac{\varepsilon_0^2}{4C^2}\Big/\left[\left(\frac{|||\nabla\mathbf{U}|||^4}{\alpha\nu^2}+\frac{1}{\beta}\right)J_\infty(\mathbf{u},\mathbf{f})\right]>0,$$

we may assume, without loss of generality (by choosing a subsequence if necessary), that

$$|t_{n+1}-t_n|\ge\delta\,.$$

We claim that for each n

$$\|\mathbf{u}(t)-\mathbf{U}(t)\|>0 \qquad \forall\, t\in(t_n-\delta,t_n)\,.$$

If the claim were not true, then there would exist an $\bar n$ such that $\bar t_{\bar n}\stackrel{\text{def}}{=\!=}\sup\{t\in(-\infty,t_{\bar n}):\|\mathbf{u}(t)-\mathbf{U}(t)\|=0\}\in(t_{\bar n-1},t_{\bar n})$ satisfies $|\bar t_{\bar n}-t_{\bar n}|<\delta$ and $\|\mathbf{u}(t)-\mathbf{U}(t)\|>0$ on $(\bar t_{\bar n},t_{\bar n})$ so that by (4.1),

$$\begin{aligned}
\|\mathbf{u}(\bar t_{\bar n})-\mathbf{U}(\bar t_{\bar n})\| &\ge \|\mathbf{u}(t_{\bar n})-\mathbf{U}(t_{\bar n})\| - C\,\delta^{1/2}\left(\frac{|||\nabla\mathbf{U}|||^4}{\alpha\nu^2}+\frac{1}{\beta}\right)^{1/2}\left(J_\infty(\mathbf{u},\mathbf{f})\right)^{1/2} \\
&\ge \varepsilon_0-\varepsilon_0/2=\varepsilon_0/2>0\,,
\end{aligned}$$

which would contradict $\|\mathbf{u}(\bar{t}_{\bar{n}}) - \mathbf{U}(\bar{t}\bar{n})\| = 0$. This proves the claim. Now using (4.1) again, we have

$$\begin{aligned}\|\mathbf{u}(t) - \mathbf{U}(t)\| &\geq \|\mathbf{u}(t_n) - \mathbf{U}(t_n)\| - C\,\delta^{1/2}\left(\frac{|||\nabla \mathbf{U}|||^4}{\alpha\nu^2} + \frac{1}{\beta}\right)^{1/2} (J_\infty(\mathbf{u},\mathbf{f}))^{1/2} \\ &\geq \varepsilon_0 - \varepsilon_0/2 = \varepsilon_0/2 \qquad \forall\, t \in (t_n - \delta, t_n)\,.\end{aligned}$$

Thus

$$J_\infty(\mathbf{u},\mathbf{f}) \geq \frac{\alpha}{2}\sum_{n=2}^{\infty}\int_{t_n-\delta}^{t_n} \|\mathbf{u}(t) - \mathbf{U}(t)\|^2\, dt \geq \frac{\alpha\varepsilon_0^2}{8}\sum_{n=2}^{\infty}\int_{t_n-\delta}^{t_n} dt = \infty\,,$$

contradicting the assumption $J_\infty(\mathbf{u},\mathbf{f}) < \infty$. Hence, (4.3) is true. ∎

We now turn to the study of the asymptotic behavior of $\|\nabla\mathbf{u}(t) - \nabla\mathbf{U}(t)\|$. Note that Theorem 4.2 is true for arbitrary $(\mathbf{u},\mathbf{f}) \in \mathcal{U}_{ad}(\infty)$. Under some additional assumptions on $\mathbf{u}$, we can establish a similar asymptotic result for $\|\nabla\mathbf{u}(t) - \nabla\mathbf{U}(t)\|$. In particular, we can establish the asymptotic behavior for $\mathbf{u} = \widehat{\mathbf{u}}$, where $(\widehat{\mathbf{u}},\widehat{\mathbf{f}})$ is the optimizer for (2.10) with $T = \infty$. With $\widehat{\boldsymbol{\xi}} \in L^2(\varepsilon,\infty;\mathbf{V})$ for each $\varepsilon > 0$, we can prove an analog of (4.1) in the $H^1(\Omega)$ norm.

LEMMA 4.3. *Let $T \in (0,\infty]$. Assume that $(\widehat{\mathbf{u}},\widehat{\mathbf{f}}) \in \mathcal{U}_{ad}(T)$ is a solution of* (2.10). *Assume further that $\|\nabla(\widehat{\mathbf{u}} - \mathbf{U})(t)\| > 0$ for all $t \in (t_1,t_2) \subset [\varepsilon,T]$. Then*

$$\begin{aligned}&\|\nabla(\widehat{\mathbf{u}} - \mathbf{U})(t_1)\| \geq \|\nabla(\widehat{\mathbf{u}} - \mathbf{U})(t_2)\| - \frac{C}{\nu}(\sigma_0^2\sigma_1^3(\varepsilon) + \sigma_1(\varepsilon)\,|||\mathbf{U}|||^2\,|||\nabla\mathbf{U}|||^2 \\ &+ \nu^2\sigma_1(\varepsilon)\,|||\nabla\mathbf{U}|||\,|||\Delta\mathbf{U}|||)(t_2 - t_1) - \frac{C\,\max\{\sqrt{\alpha},\sqrt{\beta}\rho_1(\varepsilon)\}}{\beta\nu}[\mathcal{J}_T(\widehat{\mathbf{u}},\widehat{\mathbf{f}})]^{1/2}\sqrt{t_2 - t_1}\,,\end{aligned}$$

where $\rho_1(\varepsilon)$ is defined by (3.9) *and*

$$\begin{aligned}\sigma_0 &\stackrel{\text{def}}{=\!=} \sqrt{K_0}\|\mathbf{u}_0 - \mathbf{U}_0\|, \\ \sigma_1(\varepsilon) &\stackrel{\text{def}}{=\!=} \sqrt{K_1(\varepsilon)}\,\|\mathbf{u}_0 - \mathbf{U}_0\|\end{aligned}$$

with K_0 defined by (2.37) *and $K_1(\varepsilon)$ defined by* (2.45).

Proof. We first note that from (2.37) and (2.45),

$$\begin{aligned}\sup_{t\in[\varepsilon,T]}\|\widehat{\mathbf{u}}\| &\leq \rho_1(\varepsilon)\,, \\ \sup_{t\in[0,T]}\|\widehat{\mathbf{v}}(t)\| &\leq \sigma_0\,,\end{aligned}$$

and

$$\sup_{t\in[\varepsilon,T]}\|\nabla\widehat{\mathbf{v}}(t)\| \leq \sigma_1(\varepsilon)\,.$$

Setting $\mathbf{w} = -P\Delta\widehat{\mathbf{v}}(t)$ in (2.12) and by similar treatments as in the proof of Theorem 2.8, we obtain

$$\begin{aligned}&\frac{1}{2}\frac{d}{dt}\|\nabla\widehat{\mathbf{v}}(t)\|^2 + \frac{3\nu}{4}\|P\Delta\widehat{\mathbf{v}}(t)\|^2 \\ &\leq \frac{C}{\nu^3}\Big(\|\widehat{\mathbf{v}}(t)\|^2\|\nabla\widehat{\mathbf{v}}(t)\|^4 + \|\mathbf{U}(t)\|^2\|\nabla\mathbf{U}(t)\|^2\|\nabla\widehat{\mathbf{v}}(t)\|^2 \\ &\qquad +\nu^2\|\nabla\mathbf{U}(t)\|\cdot\|\Delta\mathbf{U}(t)\|\cdot\|\nabla\widehat{\mathbf{v}}(t)\|^2 + \frac{\nu^3}{\beta}|(\widehat{\boldsymbol{\xi}}(t), -P\Delta\widehat{\mathbf{v}}(t))|\Big)\,.\end{aligned} \tag{4.4}$$

Since $\operatorname{div}\widehat{\boldsymbol{\xi}}(t)=\mathbf{0}$, we have $\widehat{\boldsymbol{\xi}}(t)=P\widehat{\boldsymbol{\xi}}(t)$. Thus

$$\left|(\widehat{\boldsymbol{\xi}}(t),-P\Delta\widehat{\mathbf{v}}(t))\right|=(\nabla\widehat{\boldsymbol{\xi}}(t),\nabla\widehat{\mathbf{v}}(t))\le\|\nabla\widehat{\boldsymbol{\xi}}(t)\|\cdot\|\nabla\widehat{\mathbf{v}}(t)\|\ .$$

Using (2.42) and dividing both sides by $\|\nabla\widehat{\mathbf{v}}\|$, we have

$$\frac{1}{2}\frac{d}{dt}\|\nabla\widehat{\mathbf{v}}(t)\|+\frac{\gamma\nu}{2}\|\nabla\widehat{\mathbf{v}}(t)\|\le\frac{C}{\nu^3}\Big(\sigma_0^2\sigma_1^3(\varepsilon)+\sigma_1(\varepsilon)\,|||\mathbf{U}|||^2\,|||\nabla\mathbf{U}|||^2 +\nu^2\sigma_1(\varepsilon)\,|||\nabla\mathbf{U}|||\cdot|||\Delta\mathbf{U}|||+\frac{\nu^3}{\beta}\|\nabla\widehat{\boldsymbol{\xi}}(t)\|\Big)$$

for all $t\in(t_1,t_2)$. Multiplying both sides by $e^{\gamma\nu\lambda_1 t}$ and integrating over (t_1,t_2), we are led to

$$\begin{aligned}\|\nabla\widehat{\mathbf{v}}(t_2)\|\le{}&\|\nabla\widehat{\mathbf{v}}(t_1)\|\,\exp\{-\gamma\nu(t_2-t_1)\}\\&+\frac{C}{\nu^3}(\sigma_0^2\sigma_1^3(\varepsilon)+\sigma_1(\varepsilon)\,|||\mathbf{U}|||^2\,|||\nabla\mathbf{U}|||^2\\&\qquad+\sigma_1(\varepsilon)\nu^2\,\big|\big|\big|\nabla\mathbf{U}\big|\big|\big|\cdot|||\Delta\mathbf{U}|||)\,\frac{1-\exp\{-\gamma\nu(t_2-t_1)\}}{\gamma\nu}\\&+\frac{C\exp\{-\gamma\nu t_2\}}{\beta}\int_{t_1}^{t_2}\|\nabla\widehat{\boldsymbol{\xi}}(s)\|\,\exp\{\gamma\nu s\}\,ds.\end{aligned}$$

By the Schwarz inequality and (3.8),

$$\begin{aligned}&\int_{t_1}^{t_2}\|\nabla\widehat{\boldsymbol{\xi}}(s)\|\,\exp\{\gamma\nu s\}\,ds\\&\le\left(\int_{t_1}^{t_2}\|\nabla\widehat{\boldsymbol{\xi}}(s)\|^2\,ds\right)^{1/2}\left(\int_{t_1}^{t_2}\exp\{2\gamma\nu s\}\,ds\right)^{1/2}\\&\le\frac{C}{\nu}\,\max\{\sqrt{\alpha},\sqrt{\beta}\rho_1(\varepsilon)\}\,\left(\mathcal{J}_\tau(\widehat{\mathbf{u}},\widehat{\mathbf{f}})\right)^{1/2}\left(\frac{\exp\{2\gamma\nu t_2\}-\exp\{2\gamma\nu t_1\}}{2\gamma\nu}\right)^{1/2}.\end{aligned}$$

Hence, by the mean value theorem,

$$\begin{aligned}\|\nabla\widehat{\mathbf{v}}(t_2)\|\le{}&\|\nabla\widehat{\mathbf{v}}(t_1)\|+\frac{C}{\nu}\,(\sigma_0^2\sigma_1^3(\varepsilon)+\sigma_1(\varepsilon)\,|||\mathbf{U}|||^2\,|||\nabla\mathbf{U}|||^2\\&\qquad+\sigma_1(\varepsilon)\nu^2\,|||\nabla\mathbf{U}|||\cdot|||\Delta\mathbf{U}|||)\,(t_2-t_1)\\&+\frac{C}{\beta\nu}\,\max\{\sqrt{\alpha},\sqrt{\beta}\rho_1(\varepsilon)\}\,[\mathcal{J}_\tau(\widehat{\mathbf{u}},\widehat{\mathbf{f}})]^{1/2}\,\sqrt{t_2-t_1}.\end{aligned}$$

□

Based on Lemma 4.3, we may establish the long-time behavior for $\|\nabla\widehat{\mathbf{u}}(t)-\nabla\mathbf{U}(t)\|$.

THEOREM 4.4. *Let* $(\widehat{\mathbf{u}},\widehat{\mathbf{f}})$ *be an solution for* (2.10) *with* $T=\infty$. *Then*

$$\lim_{t\to\infty}\|\nabla\widehat{\mathbf{u}}(t)-\nabla\mathbf{U}(t)\|=0. \tag{4.5}$$

The proof is similar to that of Theorem 4.2 (now we use the bound (2.44) of Theorem 2.8 in place of $\mathcal{J}_\tau(\mathbf{u},\mathbf{f})<\infty$) and is omitted here.

Remark 4.5. From Theorem 4.4 and the proof of Lemma 4.3, we see that the condition $(\widehat{\mathbf{u}},\widehat{\mathbf{f}})$ being an optimizer for (2.10) is not essential. If a pair $(\mathbf{u},\mathbf{f})\in\mathcal{U}_{ad}(\infty)$ satisfies

$$\int_0^\infty\int_\Omega|\mathbf{u}-\mathbf{U}|^2d\mathbf{x}dt+\int_0^\infty\int_\Omega|\nabla(\mathbf{f}-\mathbf{F})|^2d\mathbf{x}dt<\infty, \tag{4.6}$$

$$(\mathbf{f}-\mathbf{F})\cdot\mathbf{n}=0 \qquad \text{in } \Omega, \tag{4.7}$$

and

$$\operatorname{div}(\mathbf{f}-\mathbf{F})=0 \qquad \text{on } \partial\Omega\,, \tag{4.8}$$

then (4.5) holds for such $\mathbf{u}$. The condition (4.6) requires higher regularity of $\mathbf{f}-\mathbf{F}$, (4.7) requires that the boundary conditions of $\mathbf{f}$ and $\mathbf{F}$ agree with each other, and (4.8) requires $\operatorname{div}\mathbf{f}=\operatorname{div}\mathbf{F}$, where we recall that $\mathbf{F}=\partial_t\mathbf{U}-\nu\Delta\mathbf{U}+(\mathbf{U}\cdot\nabla)\mathbf{U}$. ∎

For each finite T, $\widehat{\boldsymbol{\xi}}_T$ satisfies $\widehat{\boldsymbol{\xi}}_T(\cdot,T)=\mathbf{0}$. This is also true for $T=\infty$, namely $\lim_{t\to\infty}\widehat{\boldsymbol{\xi}}_\infty(\cdot,t)=\mathbf{0}$. To prove this, we first establish the following inequality.

LEMMA 4.6. *Let $T\in(0,\infty]$ and $\varepsilon>0$. Assume that $(\widehat{\mathbf{u}},\widehat{\mathbf{f}})$ is a solution for* (2.10) *and $\widehat{\boldsymbol{\xi}}$ is a solution of* (3.1)–(3.3). *Assume further that $\|\widehat{\boldsymbol{\xi}}(t)\|>0$ for all $t\in(t_1,t_2)\subset[\varepsilon,T]$. Then*

$$\|\widehat{\boldsymbol{\xi}}(t_2)\|\ge\|\widehat{\boldsymbol{\xi}}(t_1)\|-\sqrt{t_2-t_1}\left(\alpha+\frac{C\beta\sigma_1^4(\varepsilon)}{\nu^2}\right)\left(\mathcal{J}_T(\widehat{\mathbf{u}},\widehat{\mathbf{f}})\right)^{1/2},$$

where $\rho_1(\varepsilon)$ is defined as in (3.9).

Proof. (3.10) and Young's inequality yield

$$-\frac{1}{2}\frac{d}{dt}\|\widehat{\boldsymbol{\xi}}(t)\|^2+\nu\,\|\nabla\widehat{\boldsymbol{\xi}}(t)\|^2\le\alpha\|\widehat{\mathbf{v}}(t)\|\,\|\widehat{\boldsymbol{\xi}}(t)\|+\frac{C\|\nabla\widehat{\mathbf{u}}(t)\|^2}{\nu}\,\|\widehat{\boldsymbol{\xi}}(t)\|^2.$$

Then, by applying Poincaré inequality, dividing both sides by $\|\widehat{\boldsymbol{\xi}}(t)\|$, and then applying the Schwarz inequality, we obtain

$$-\frac{d}{dt}\,\|\widehat{\boldsymbol{\xi}}(t)\|+\nu\,\lambda_1\,\|\widehat{\boldsymbol{\xi}}(t)\|\le\left(\alpha+\frac{C\beta\,\|\nabla\widehat{\mathbf{u}}\|^4}{\nu^2}\right)^{1/2}\left(\alpha\|\widehat{\mathbf{v}}(t)\|^2+\frac{1}{\beta}\|\widehat{\boldsymbol{\xi}}(t)\|^2\right)^{1/2}.$$

By integrating both sides over $(t_1,t_2)\subset[\varepsilon,T]$ and applying the Schwarz inequality, the lemma is proved. ∎

Similar to the proof of Theorem 4.2 (we now use the bound (3.8) of Theorem 3.2), we have the following result on the long-time behavior of $\|\boldsymbol{\xi}(t)\|$ as $t\to\infty$.

THEOREM 4.7. *Assume that the hypotheses of Lemma* 4.6 *hold. Then*

$$\lim_{t\to\infty}\|\widehat{\boldsymbol{\xi}}(t)\|=\lim_{t\to\infty}\beta\,\|\mathbf{f}-\mathbf{F}\|=0\,. \qquad \blacksquare$$

Remark 4.8. We make some comparison between the optimizer $(\widehat{\mathbf{u}},\widehat{\mathbf{f}})$ and the quasi optimizer $(\widetilde{\mathbf{u}},\widetilde{\mathbf{f}})$. The optimizer $\widehat{\mathbf{u}}(t)$ matches well with $\mathbf{U}(t)$ for large t and $\nabla\widehat{\mathbf{u}}(t)$ matches well with $\nabla\mathbf{U}(t)$ for large t. We see from the functional (1.1) that in obtaining a good matching of $\widehat{\mathbf{u}}(t)$ with $\mathbf{U}(t)$ for large t, the work done by the external force $\mathbf{f}$ is also minimized. However, we are unable to prove the exponential decay of $\|\widehat{\mathbf{u}}(t)-\mathbf{U}(t)\|$. The optimizer is costly to compute numerically (one needs to solve the optimality system). The quasi optimizer $\widetilde{\mathbf{u}}$ gives us an acceptable matching of $\widetilde{\mathbf{u}}(t)$ with $\mathbf{U}(t)$ with an exponential decay rate for $\|\widetilde{\mathbf{u}}(t)-\mathbf{U}(t)\|$. $\widetilde{\mathbf{u}}$ is defined by a Navier–Stokes–like equation and is therefore easy to compute (compared with the optimizer). However, we have no control over the work involved (i.e., the corresponding external force) in maintaining the quasi optimizer. The work to maintain the quasi optimizer can be formidable.

Remark 4.9. In Theorems 2.4, 2.7, 2.8, 2.9, and 3.2 and Lemmas 4.1, 4.3, and 4.6, the quantities on the right-hand side of all the estimates were independent of T (such

as K_0, K_1, $|||\nabla \mathbf{U}|||$, etc). This facilitated the derivation of estimates on $(0,\infty)$ and the passage to the limit as $T \to \infty$. However, if one is only interested in corresponding estimates on the finite time interval, then one can replace the quantities on the right-hand side of all the estimates by their finite time counterparts (such as K_0 replaced by $K_0(T)$, K_1 by $K_1(T)$, $|||\nabla \mathbf{U}|||$ by $|||\nabla \mathbf{U}|||_T \stackrel{\text{def}}{=\!=} \|\mathbf{U}\|_{L^\infty(0,T;\mathbf{L}^2(\Omega))}$, etc., where the definitions of the quantities $K_0(T)$, $K_1(T)$, etc., are the obvious modifications of the definitions for the quantities K_0, K_1, etc.).

REFERENCES

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

[2] F. Abergel and R. Temam, *On some control problems in fluid mechanics*, Theoret. Comput. Fluid Dynamics, 1 (1990) pp. 303–325.

[3] V. Alekseev, V. Tikhomirov, and S. Fomin, *Optimal Control*, Consultants Bureau, New York, 1987.

[4] J. Burns and S. Kang, *A control problem for Burgers' equation with bounded input/output*, Nonlinear Dynamics, 2 (1991), pp. 235–262.

[5] J. Burns and S. Kang, *A stabilization problem for Burgers' equation with unbounded control and observation*, in Estimation and Control of Distributed Parameter Systems, Internat. Ser. Numer. Math., 100, Birkhäuser, Basel, 1991, pp. 51–72.

[6] P. Constantin and C. Foias, *Navier-Stokes Equations*, Univ. of Chicago Press, Chicago, 1988.

[7] J. Diaz and A. Fursikov, *A simple proof of the approximate controllability from the interior for nonlinear evolution problems*, Appl. Math. Lett., 7 (1994), pp. 85–87.

[8] A. Fursikov, *On some control problems and results concerning the unique solvability of a mixed boundary value problems for the three-dimensional Navier-Stokes and Euler systems*, Soviet Math. Dokl., 3 (1980), pp. 889–893.

[9] A. Fursikov, *Control problems and theorems concerning the unique solvability of a mixed boundary value problems for the three-dimensional Navier-Stokes and Euler equations*, Math USSR Sb., 43 (1982), pp. 281–307.

[10] A. Fursikov, *Properties of solutions of some extremal problems connected with the Navier-Stokes system*, Math USSR Sb., 46 (1983), pp. 323–351.

[11] A. Fursikov and O. Imanuvilov, *On exact boundary zero controllability of two-dimensional Navier-Stokes equations*, Acta Appl. Math., 37 (1994), pp. 67–76.

[12] A. Fursikov and O. Imanuvilov, *On controllability of certain systems simulating a fluid flow*, in Flow Control, IMA Vol. Math. Appl. 68, M. D. Gunzburger, ed., Springer-Verlag, New York, 1995, pp. 149–184.

[13] V. Girault and P. Raviart, *Finite Element Methods for Navier-Stokes Equations*, Springer-Verlag, Berlin, 1986.

[14] J.-L. Lions, *Control of Distributed Singular systems*, Bordas, Paris, 1985.

[15] G. Raugel and G.R. Sell, *Navier-Stokes equations on thin* 3*D domains* I*: Global regularity of solutions*, J. Amer. Math. Soc., 6 (1993), pp. 503–568.

[16] G. Raugel and G.R. Sell, *Navier-Stokes equations in thin* 3*D domains* II*: Global regularity of spatially periodic solutions*, in Proc. College de France, to appear.

[17] H. Tanabe, *Equations of Evolution*, Pitman, London, 1979.

[18] R. Temam, *Navier-Stokes Equations, Theory and Numerical Methods*, North-Holland, Amsterdam, 1979.

[19] R. Temam, *Infinite Dimensional Dynamical Systems in Mechanics and Physics*, Springer-Verlag, New York, 1988.

SIAM J. CONTROL OPTIM.
Vol. 35, No. 2, pp. 678–713, March 1997

017

ON-LINE PARAMETER ESTIMATION FOR INFINITE-DIMENSIONAL DYNAMICAL SYSTEMS*

J. BAUMEISTER†, W. SCONDO‡, M. A. DEMETRIOU§, AND I. G. ROSEN¶

Abstract. The on-line or adaptive identification of parameters in abstract linear and nonlinear infinite-dimensional dynamical systems is considered. An estimator in the form of an infinite-dimensional linear evolution system having the state and parameter estimates as its states is defined. Convergence of the state estimator is established via a Lyapunov estimate. The finite-dimensional notion of a plant being sufficiently rich or persistently excited is extended to infinite dimensions. Convergence of the parameter estimates is established under the additional assumption that the plant is persistently excited. A finite-dimensional approximation theory is developed, and convergence results are established. Numerical results for examples involving the estimation of both constant and functional parameters in one-dimensional linear and nonlinear heat or diffusion equations and the estimation of stiffness and damping parameters in a one-dimensional wave equation with Kelvin–Voigt viscoelastic damping are presented.

Key words. on-line estimation, adaptive identification, parameter convergence, persistence of excitation, distributed parameter systems, infinite-dimensional systems, finite-dimensional approximation

AMS subject classifications. 93B30, 93C25, 93C20, 65J10

PII. S0363012994270928

1. Introduction.

In this paper we develop an abstract framework for the on-line, or adaptive, identification of unknown parameters for a class of infinite-dimensional and, in general, nonlinear dynamical systems. The estimator we construct takes the form of an infinite-dimensional linear evolution system with time-varying *coefficients* whose states consist of an estimator for the state of the plant and an estimator for the unknown parameters. Our scheme can estimate both constant and functional (e.g., spatially varying) parameters including the nonlinearity itself. That is, both the state space of the plant and the parameter space may be infinite dimensional.

The results reported here were in fact obtained independently by two separate groups of researchers. The efforts of the first two authors (Baumeister and Scondo) culminated in the Ph.D. thesis of Dr. Scondo [30], while the investigation by the second two authors (Demetriou and Rosen) led to the Ph.D. thesis of Dr. Demetriou [6]. In this paper we have attempted to capture the essence of the problem and its solution as treated independently by both groups of researchers in a clear and coherent manner. However, it should be noted that, as would be expected, there are some variations between the two treatments. Thus, the interested reader may also wish to consult the two theses, [6] and [30], in addition to the study that we present here.

*Received by the editors July 8, 1994; accepted for publication (in revised form) February 26, 1996.

http://www.siam.org/journals/sicon/35-2/27092.html

†Fachbereich Mathematik, Johann Wolfgang Goethe-Universität, D-60054 Frankfurt am Main, Germany (baumeister@math.uni-frankfurt.de).

‡Balduinstraße 112, D-60599 Frankfurt am Main, Germany. The research of this author was supported in part by the Deutsche Forschungsgemeinschaft.

§Department of Mathematics and Computer Science, Boise State University, Boise, ID 83726 (mdemetri@math.idbsu.edu). The research of this author was supported in part by Air Force Office of Scientific Research grant AFOSR 91-0076.

¶Center for Applied Mathematical Sciences, Department of Mathematics, University of Southern California, Los Angeles, CA 90089-1113 (rosen@mathc.usc.edu). The research of this author was supported in part by Air Force Office of Scientific Research grant AFOSR 91-0076.

The approach we take here represents an infinite-dimensional analogue, or extension, of some portion of the finite-dimensional treatment in [22] (see also [23] and [24]). Convergence of the state estimator is established using a Lyapunov estimate–based argument and an argument in the spirit of the one used to verify Barbălat's lemma (see [27]). In order to establish the convergence of the parameter estimates, we require an additional assumption. This assumption, which is a *richness* condition on the plant data, is an infinite-dimensional analogue of the notion of *persistence of excitation* defined in [22] and [23]. In mathematical terms, we establish that the solution to the error equations derived from the plant dynamics and the estimator with arbitrary initial data, tends to the trivial solution as time tends to infinity. Our primary motivation for studying these on-line identification schemes is that we ultimately intend to use them as a part of an indirect adaptive control algorithm for distributed parameter systems.

Because the estimator is infinite dimensional, its implementation requires some form of finite-dimensional approximation. Consequently, we have also developed a rather complete approximation theory and established corresponding convergence results. In addition, while our treatment is in the context of abstract first-order systems, we have also shown how our theory can be applied to certain classes of abstract second-order systems. A number of examples along with numerical studies have been included to demonstrate the feasibility of our schemes.

There has been a great deal of research activity in the area of identification of distributed parameter systems over the past two decades. An extensive treatment of off-line schemes (e.g., output least squares, equation error, etc.) together with a rather comprehensive survey of the literature can be found in the monograph by Banks and Kunisch [2]. In the case of on-line, or adaptive, schemes, the available literature is less extensive and more recent. In [1] Alt, Hoffmann, and Sprekels developed an asymptotic embedding method for the identification of functional parameters in linear elliptic (stationary) partial differential equations. In their approach, the elliptic equation is embedded in a nonautonomous pseudoparabolic evolution equation in such a way that the elliptic equation's solution is an asymptotic steady state of the evolution equation. In [14] Hoffmann and Sprekels introduce a form of regularization into their embedding equations, and in [15] an abstract functional analytic framework for the earlier results summarized above is developed. They also extend their earlier results to certain classes of stationary elliptic and evolutionary parabolic nonlinear variational inequalities.

In [4] Baumeister and Scondo consider parameter estimation techniques for finite-dimensional evolution equations, while in [5] they treat linear elliptic partial differential equations. The elliptic equation is embedded in a pseudoparabolic evolution equation having the solution to the elliptic equation and the true parameters as an equilibrium point. Using a *richness*-like assumption and linear semigroup theory, they are able to establish uniform exponential convergence to this equilibrium as $t \to \infty$. The extension of the treatment in [5] to abstract evolution equations via infinite-dimensional analogues of the arguments in [22] is, to a large extent, the contribution of the effort that we are reporting on here.

Recently Hong and Bentsman (see [16] and [17]) have studied model reference adaptive control (MRAC) of linear n-dimensional parabolic partial differential equations. Although, strictly speaking, MRAC is not the same problem that we treat here, there are some connections (and a number of significant differences) between their efforts and ours. For example, the resulting error equations are formally the same, and both treatments are concerned with state and parameter convergence. On

the other hand, however, they deal with a specific system, while our approach is more abstract. Their analysis is more classical, while ours is more functional analytic in nature.

In a recent series of papers [9], [10], [11], [12], [13], [25] Duncan and Pasik-Duncan and their coworkers have developed and analyzed adaptive control algorithms for classes of linear stochastic distributed parameter systems. In particular, they have considered indirect adaptive control schemes in the form of consistent least squares and maximum likelihood estimators for the unknown parameters combined with linear quadratic (LQ) control design techniques. They consider a variety of classes of infinite-dimensional systems, including hereditary systems [10] and systems involving unbounded input, such as boundary and point control [9], [13]. The schemes that they propose and techniques that they use to argue convergence are, in general, very different from and largely unrelated to the theory that we develop here.

An outline of the remainder of the paper is as follows. In section 2 we define the plant and the estimator equations. In section 3 we establish convergence of the state estimator. We define the notion of persistence of excitation and establish parameter convergence. The notion of partial persistence of excitation also is defined, and a corresponding partial parameter convergence result is given. Our approximation results are presented in section 4, and the extension of our results to a class of abstract second order systems is discussed in section 5. Examples together with the results of our numerical studies are presented in section 6.

In general all notation is standard. For X and Y Banach spaces, $\mathcal{L}(X,Y)$ denotes the space of *bounded* linear operators from X into Y. All inner products, $\langle\cdot,\cdot\rangle$, are assumed to be linear in the first argument and conjugate linear in the second. Finally, for X a linear space and Y a space of linear or conjugate linear functionals on X, $\langle x,\varphi\rangle = \langle x,\varphi\rangle_{X,Y}$ denotes the action of the linear functional $\varphi \in Y$ on the element $x \in X$, and $\langle \varphi, x\rangle = \langle \varphi, x\rangle_{Y,X}$ denotes the action of the conjugate linear functional $\varphi \in Y$ on the element $x \in X$.

2. The plant and the estimator. Let H be a Hilbert space with inner product $\langle\cdot,\cdot\rangle$ and corresponding induced norm $|\cdot|$. Let V be a reflexive Banach space with norm denoted by $\|\cdot\|$, and assume that V is embedded densely and continuously in H. Let V^* denote the *conjugate* dual of V (i.e., the space of continuous conjugate linear functionals on V) and $\|\cdot\|_*$ denote the usual uniform operator norm on V^*. It follows that

$$V \hookrightarrow H \hookrightarrow V^*, \tag{2.1}$$

with both embeddings dense and continuous. In particular we assume that

$$|\varphi| \le K\|\varphi\|, \qquad \varphi \in V, \tag{2.2}$$

for some positive constant K. The notation $\langle\cdot,\cdot\rangle$ will also be used to denote the duality pairing between V^* and V induced by the continuous and dense embeddings given in (2.1). We note that while we have chosen to develop our theory in the generality of complex Hilbert and Banach spaces H and V, all that follows is easily modified (simplified) to allow for H and V to be chosen to be real.

Let Q be a Hilbert space with inner product $\langle\cdot,\cdot\rangle_Q$ and corresponding induced norm $|\cdot|_Q$. Let $Q^* = Q$ denote the conjugate dual of Q. The Hilbert space Q is known as the parameter space. The Hilbert space Q could be taken to be real as well.

For each $q \in Q$, let $A_0(q) : V \to V^*$ be a, in general, nonlinear operator satisfying the following assumptions:

(A1) (Q-linearity) The map $q \to A_0(q)\varphi$ is affine from Q into V^* for each $\varphi \in V$. That is, for $q \in Q$ and $\varphi \in V$ we have $A_0(q)\varphi = A_1(q)\varphi + A_2\varphi$, where $A_1(q) : V \to V^*$ and $A_2 : V \to V^*$ are, in general, nonlinear operators from V into V^* with the map $q \to A_1(q)\varphi$ from Q into V^* linear for each $\varphi \in V$.

(A2) ($V \mapsto V^*$-boundedness) There exist $\alpha_1, \alpha_2 > 0$ such that $|\langle A_1(q)\varphi, \psi\rangle| \leq \alpha_1 |q|_Q \|\varphi\| \|\psi\|$ for $\varphi, \psi \in V$ and $q \in Q$, and $|\langle A_2\varphi, \psi\rangle| \leq \alpha_2 \|\varphi\| \|\psi\|$ for $\varphi, \psi \in V$.

In order to simplify our treatment we have assumed that for $q \in Q$ the operator $A_0(q)$ is time invariant. However, it would be relatively straightforward to extend all of the results in this section and those in the subsequent sections to the case of a time-dependent operator, $A_0(t; q)$, $t \geq 0$. Of course for some of these results to remain valid, additional, but rather standard, assumptions on the regularity of the map $t \to A_0(t; q)$, $t \geq 0$ would be required (see, for example, [3], [19], [26], [32]).

For $\varphi \in V$, let $B(\varphi) : V \to Q$ be the linear operator defined by

$$\langle B(\varphi)\psi, q\rangle_Q = \overline{\langle A_1(q)\varphi, \psi\rangle}, \qquad \psi \in V, \quad q \in Q. \tag{2.3}$$

Note that assumption (A2) implies that for $\varphi \in V$, $B(\varphi) \in \mathcal{L}(V, Q)$ with

$$\|B(\varphi)\|_{\mathcal{L}(V,Q)} \leq \alpha_1 \|\varphi\|. \tag{2.4}$$

Let $u_0 \in H$, let $f \in L_2(0, T, V^*)$ for all $T > 0$, and for each $q \in Q$ consider the initial value problem

$$D_t u(t) + A_0(q)u(t) = f(t), \qquad \text{almost every (a.e.) } t > 0, \tag{2.5}$$

$$u(0) = u_0. \tag{2.6}$$

By a solution to the initial value problem (2.5), (2.6) we mean a weak or variational solution. That is, a function $u \in L_2(0, T; V)$ with $D_t u \in L_2(0, T; V^*)$ for all $T > 0$ which satisfies (2.5) and (2.6). Note that if u is a solution to (2.5), (2.6), then for all $T > 0$, u agrees almost everywhere with a function in $C([0, T]; H)$ (see [20]). Note also that $A_0(q)$ monotone (i.e., $\langle A_0(q)\varphi - A_0(q)\psi, \varphi - \psi\rangle \geq 0$), hemicontinuous (i.e., $\lim_{\lambda \to 0}\langle A_0(q)\{\varphi + \lambda\psi\}, \chi\rangle = \langle A_0(q)\varphi, \chi\rangle$, $\varphi, \psi, \chi \in V$), and coercive (i.e., $\mathrm{Re}\langle A_0(q)\varphi, \varphi\rangle \geq \beta_0(q)\|\varphi\|^2 + \lambda_0(q)$, $\varphi \in V$, for some $\beta_0(q), \lambda_0(q) \in \mathbf{R}$ with $\beta_0(q) > 0$), for example, is sufficient to guarantee the existence of a unique solution to (2.5), (2.6) (see, for example, [3], [19], [26], [32]).

DEFINITION 2.1. *A plant is a pair* $(\overline{q}, \overline{u})$ *for which* $\overline{q} \in Q$, $\overline{u}$ *is a solution to* (2.5), (2.6) *with* $q = \overline{q}$, *and there exists a constant* $\gamma = \gamma(\overline{u})$ *such that* $|\langle B(\overline{u}(t))\varphi, q\rangle_Q| \leq \gamma(\overline{u})|q|_Q\|\varphi\|$, $t > 0$, $q \in Q$, $\varphi \in V$.

Note that if $(\overline{q}, \overline{u})$ is a plant, then $B(\overline{u}(\cdot)) \in L_2(0, T; \mathcal{L}(V, Q))$ for all $T > 0$. Note also that (2.4) implies that if $(\overline{q}, \overline{u})$ is such that $\|\overline{u}(t)\| \leq \gamma$, a.e. $t > 0$, for some $\gamma > 0$, then $(\overline{q}, \overline{u})$ is a plant.

To demonstrate that it is in fact possible to provide sufficient conditions for a pair $(\overline{q}, \overline{u})$ to be a plant, let $\overline{q} \in Q$, let $\overline{u}$ be the solution to the initial value problem (2.5), (2.6) with $q = \overline{q}$, and assume that $A_0(\overline{q}) \in \mathcal{L}(V, V^*)$ is coercive,

$$\mathrm{Re}\,\langle A_0(\overline{q})\varphi, \varphi\rangle \geq \beta_0(\overline{q})\|\varphi\|^2, \qquad \varphi \in V, \tag{2.7}$$

for some $\beta_0(\overline{q}) > 0$. It follows (see [32, Theorem 3.6.1]) that $-A_0(\overline{q}) : V \subset V^* \to V^*$ is the infinitesimal generator of an analytic semigroup, $\{T_0(t; \overline{q}) : t \geq 0\}$, of bounded

linear operators on V^* and that

$$\overline{u}(t) = T_0(t;\overline{q})u_0 + \int_0^t T_0(t-s;\overline{q})f(s)ds, \qquad t \geq 0. \tag{2.8}$$

Suppose further that $u_0 \in V$ and that $f \in C([0,\infty];V^*)$ is uniformly V^*-Hölder continuous. That is, there exists $C > 0$ and $\rho \in (0,1)$ such that $\|f(t)-f(s)\|_* \leq C|t-s|^\rho$, $0 \leq t,s < \infty$. Assume also that there exists $f_\infty \in V^*$ such that $\lim_{t\to\infty}\|f(t) - f_\infty\|_* = 0$.

It can be shown (see [32, Theorem 5.6.1]) that $\overline{u}(t) \in V$, $t \geq 0$, there exists an element $\overline{u}_\infty \in V$ such that $A_0(\overline{q})\overline{u}_\infty = f_\infty$, and

$$\lim_{t\to\infty} \|A_0(\overline{q})\overline{u}(t) - A_0(\overline{q})\overline{u}_\infty\|_* = 0. \tag{2.9}$$

Now

$$\|\overline{u}(t)\| \leq \|\overline{u}(t) - \overline{u}_\infty\| + \|\overline{u}_\infty\|, \tag{2.10}$$

and for $t \geq 0$, coercivity (i.e., (2.7)) implies that

$$\begin{aligned}
\|\overline{u}(t) - \overline{u}_\infty\|^2 &\leq \frac{1}{\beta_0(\overline{q})}\mathrm{Re}\,\langle A_0(\overline{q})\{\overline{u}(t) - \overline{u}_\infty\}, \overline{u}(t) - \overline{u}_\infty\rangle \\
&\leq \frac{1}{\beta_0(\overline{q})}\,|\langle A_0(\overline{q})\{\overline{u}(t) - \overline{u}_\infty\}, \overline{u}(t) - \overline{u}_\infty\rangle| \\
&\leq \frac{1}{\beta_0(\overline{q})}\|A_0(\overline{q})\overline{u}(t) - A_0(\overline{q})\overline{u}_\infty\|_*\|\overline{u}(t) - \overline{u}_\infty\|.
\end{aligned}$$

Consequently, (2.9) yields

$$\lim_{t\to\infty} \|\overline{u}(t) - \overline{u}_\infty\| = 0. \tag{2.11}$$

Similarly, recalling that $u_0 \in V$, (2.7) and (2.8) imply that for all $t \geq 0$

$$\begin{aligned}
\|\overline{u}(t) - u_0\| &\leq \frac{1}{\beta_0(\overline{q})}\|A_0(\overline{q})\{\overline{u}(t) - u_0\}\|_* \\
&\leq \frac{1}{\beta_0(\overline{q})}\|A_0(\overline{q})\{T_0(t;\overline{q})u_0 - u_0\}\|_* + \frac{1}{\beta_0(\overline{q})}\|A_0(\overline{q})\overline{v}(t)\|_* \\
&= \frac{1}{\beta_0(\overline{q})}\|T_0(t;\overline{q})A_0(\overline{q})u_0 - A_0(\overline{q})u_0\|_* + \frac{1}{\beta_0(\overline{q})}\|A_0(\overline{q})\overline{v}(t)\|_*,
\end{aligned}$$

where $\overline{v}(t) = \int_0^t T_0(t-s;\overline{q})f(s)ds$, $t \geq 0$. Since $\{T_0(t;\overline{q}) : t \geq 0\}$ is an analytic semigroup on V^* and f was assumed to be uniformly Hölder continuous, Lemma IX.1.28 in [18] (see also the proof of Theorem 3.34 in [32]) implies that $\lim_{t\to 0}\|A_0(\overline{q})\overline{v}(t)\|_* = 0$. It follows from the elementary properties of strongly continuous semigroups that

$$\lim_{t\to 0} \|\overline{u}(t) - u_0\| = 0. \tag{2.12}$$

Since for $t \geq 0$, $\|\overline{u}(t)\| \leq \|\overline{u}(t) - u_0\| + \|u_0\|$, (2.10), (2.11), and (2.12) yield that $\|\overline{u}(t)\|$ is bounded on $[0,\infty)$. This, together with assumption (A2), implies that $(\overline{q},\overline{u})$ is a plant.

There is also another, alternative set of assumptions on $A_0(\overline{q})$, f, and u_0 that lead to the conclusion that $\|\overline{u}(t)\|$ is uniformly bounded for $t \geq 0$ and therefore that

$(\overline{q}, \overline{u})$ is a plant. Indeed, suppose once again that $A_0(\overline{q}) \in \mathcal{L}(V, V^*)$ is coercive; that is, (2.7) holds. Suppose further that $A_0(\overline{q})$ is symmetric in the sense that $\langle A_0(\overline{q})\varphi, \psi\rangle = \langle A_0(\overline{q})\psi, \varphi\rangle$, $\varphi, \psi \in V$, and that $f \in L_2(0, \infty; H)$ and $u_0 \in V$. Then, if we consider the operator $A_0(\overline{q})$ restricted to the subspace of H, $\mathrm{Dom}(A_0(\overline{q})) = \{\varphi \in V : A_0(\overline{q})\varphi \in H\}$, then $A_0(\overline{q}) : \mathrm{Dom}(A_0(\overline{q})) \subset H \to H$ is positive definite and self-adjoint. It follows that the square root of $A_0(\overline{q})$, $A_0(\overline{q})^{\frac{1}{2}}$, can be defined with $\mathrm{Dom}(A_0(\overline{q})^{\frac{1}{2}}) = V$ (see, for example, Theorem 2.2.3 on page 29 of [32]). Moreover, for $\varphi \in V$, $\|\varphi\|_0 = |A_0(\overline{q})^{\frac{1}{2}}\varphi|$ defines a norm on V and, by assumption (A2) and (2.7), we have that

$$\beta_0(\overline{q})\|\varphi\|^2 \le \langle A_0(\overline{q})\varphi, \varphi\rangle = \langle A_0(\overline{q})^{\frac{1}{2}}\varphi, A_0(\overline{q})^{\frac{1}{2}}\varphi\rangle = \|\varphi\|_0^2 = \langle A_0(\overline{q})\varphi, \varphi\rangle \le \alpha_0(\overline{q})\|\varphi\|^2$$

for all $\varphi \in V$, where $\alpha_0(\overline{q}) = \alpha_1|\overline{q}|_Q + \alpha_2$. Thus the two norms $\|\cdot\|$ and $\|\cdot\|_0$ on V are equivalent.

We require also that $\overline{u}(t) \in \mathrm{Dom}(A_0(\overline{q}))$ for almost all $t > 0$. Note that since $\{T_0(t; \overline{q}) : t \ge 0\}$, the semigroup of bounded linear operators on H generated by the operator $-A_0(\overline{q})$, is analytic, this can be guaranteed if, for example, we require that f be H-Hölder continuous for $t \ge 0$. That is, if for $t \ge 0$, $|f(t) - f(s)| \le C|t - s|^\rho$, $0 \le t, s < \infty$, where $C > 0$ and $\rho \in (0, 1]$ (see, for example, [18] and [26]). Then, from (2.5) we obtain that

$$\langle D_t\overline{u}(t), A_0(\overline{q})\overline{u}(t)\rangle + |A_0(\overline{q})\overline{u}(t)|^2 = \langle f(t), A_0(\overline{q})\overline{u}(t)\rangle, \qquad \text{a.e. } t > 0,$$

and therefore that

$$\frac{1}{2}D_t\|\overline{u}(t)\|_0^2 + |A_0(\overline{q})\overline{u}(t)|^2 \le |f(t)||A_0(\overline{q})\overline{u}(t)| \le \frac{1}{2}|f(t)|^2 + \frac{1}{2}|A_0(\overline{q})\overline{u}(t)|^2, \quad \text{a.e. } t > 0.$$

Integrating the above estimate from 0 to t and recalling (2.6), we find that

$$\|\overline{u}(t)\|_0^2 + \int_0^t |A_0(\overline{q})\overline{u}(s)|^2 ds \le \|u_0\|_0^2 + \int_0^t |f(s)|^2 ds \le \|u_0\|_0^2 + \|f\|^2_{L_2(0,\infty;H)}, \qquad t \ge 0.$$

It follows that $\|\overline{u}(t)\|$ is bounded uniformly in t for $t \ge 0$ and consequently, via (2.3) and assumption (A2), that $(\overline{q}, \overline{u})$ is a plant.

Let $(\overline{q}, \overline{u})$ be a plant, and assume that $\overline{u}$ is available and $\overline{q}$ is unknown. The on-line identification problem is to define a dynamical system which uses $\overline{u}$ to asymptotically estimate $\overline{q}$. Toward this end, we define an infinite-dimensional analogue of the finite-dimensional estimator treated in [22] and [24].

Let $A \in \mathcal{L}(V, V^*)$ satisfy the following two assumptions:

(A3) ($V \mapsto V^*$-boundedness) There exist $\alpha > 0$ for which $|\langle A\varphi, \psi\rangle| \le \alpha\|\varphi\|\|\psi\|$, $\varphi, \psi \in V$.

(A4) (V-coercivity) There exist $\beta > 0$ for which $\mathrm{Re}\,\langle A\varphi, \varphi\rangle \ge \beta\|\varphi\|^2$, $\varphi \in V$.

We define our estimator in the form of the initial value problem

$$D_t u(t) + Au(t) + B(\overline{u}(t))^* q(t) = f(t) + A\overline{u}(t) - A_2\overline{u}(t), \qquad \text{a.e. } t > 0, \tag{2.13}$$

$$D_t q(t) - B(\overline{u}(t))u(t) = -B(\overline{u}(t))\overline{u}(t), \qquad \text{a.e. } t > 0, \tag{2.14}$$

$$u(0) \in H, \qquad q(0) \in Q, \tag{2.15}$$

where for $\varphi \in V$, $B(\varphi)^* \in \mathcal{L}(Q, V^*)$ is the Banach space adjoint of $B(\varphi)$. That is, recalling (2.3), for $\varphi \in V$

$$\langle B(\varphi)^* p, \psi\rangle = \langle p, B(\varphi)\psi\rangle_Q = \langle A_1(p)\varphi, \psi\rangle, \qquad \psi \in V, \quad p \in Q. \tag{2.16}$$

To establish the well-posedness of the initial value problem (2.13)–(2.15), we let $X = H \times Q$ and $Y = V \times Q$. Endowing X and Y with the usual product topologies, X becomes a Hilbert space and Y a reflexive Banach space, and we have the dense and continuous embeddings $Y \hookrightarrow X \hookrightarrow Y^*$. For $t > 0$ define $\mathcal{A}(t) : Y \to Y^*$ by

$$\mathcal{A}(t) = \begin{bmatrix} A & B(\overline{u}(t))^* \\ -B(\overline{u}(t)) & 0 \end{bmatrix}, \tag{2.17}$$

and define $F(t) \in Y^*$ by

$$F(t) = \begin{bmatrix} f(t) + A\overline{u}(t) - A_2\overline{u}(t) \\ -B(\overline{u}(t))\overline{u}(t) \end{bmatrix}$$

for a.e. $t > 0$. The fact that $(\overline{q}, \overline{u})$ is a plant implies that $F \in L_2(0, T; Y^*)$ for all $T > 0$. Assumptions (A3) and (A4) together with $(\overline{q}, \overline{u})$ being a plant imply that $\mathcal{A}(t) \in \mathcal{L}(Y, Y^*)$, $t > 0$, and that for $t > 0$, Re $\langle \mathcal{A}(t)\varphi, \varphi \rangle_{Y^*,Y} + \rho|\varphi|_X^2 \geq \sigma\|\varphi\|_Y^2$, $\varphi \in Y$, where $|\cdot|_X$ and $\|\cdot\|_Y$ denote, respectively, the norms on X and Y, and ρ, $\sigma > 0$. It follows (see, for example, [19], [31], [32]) that the initial value problem

$$\begin{aligned} &D_t x(t) + \mathcal{A}(t)x(t) = F(t), \quad \text{a.e. } t > 0, \\ &x(0) \in X, \end{aligned}$$

admits a unique solution $x \in L_2(0, T; Y)$ with $D_t x \in L_2(0, T; Y^*)$, all $T > 0$. Consequently, the estimator (2.13)–(2.15) admits a unique solution $(q, u) \in L_2(0, T; Q) \times L_2(0, T; V)$ with $(D_t q, D_t u) \in L_2(0, T; Q) \times L_2(0, T; V^*)$, all $T > 0$. Moreover, for each $T > 0$, q and u agree almost everywhere with functions in $C([0, T]; Q)$ and $C([0, T]; H)$, respectively.

Let $e(t) = u(t) - \overline{u}(t)$ and $r(t) = q(t) - \overline{q}$, where $(\overline{q}, \overline{u})$ is a plant and (q, u) is a solution to the initial value problem (2.13)–(2.15). The functions e and r are solutions to the *error equations* given by

$$D_t e(t) + Ae(t) + B(\overline{u}(t))^* r(t) = 0, \quad \text{a.e. } t > 0, \tag{2.18}$$

$$D_t r(t) - B(\overline{u}(t))e(t) = 0, \quad \text{a.e. } t > 0, \tag{2.19}$$

or equivalently

$$D_t \begin{bmatrix} e(t) \\ r(t) \end{bmatrix} + \mathcal{A}(t) \begin{bmatrix} e(t) \\ r(t) \end{bmatrix} = 0, \quad \text{a.e. } t > 0, \tag{2.20}$$

where the operator $\mathcal{A}(t)$ is given by (2.17). In the next section we show that under appropriate hypotheses (i.e., that the plant $(\overline{q}, \overline{u})$ is *persistently excited*), the solution of (2.18), (2.19), or (2.20) with arbitrary initial data tends strongly to the trivial solution as $t \to \infty$. That is, in particular, for any $u(0) \in H$ and $q(0) \in Q$ we have $\lim_{t\to\infty} |u(t) - \overline{u}(t)| = \lim_{t\to\infty} |e(t)| = 0$ and $\lim_{t\to\infty} |q(t) - \overline{q}|_Q = \lim_{t\to\infty} |r(t)|_Q = 0$. (We in fact show that the convergence of the state estimator holds without any additional assumptions. The assumption of persistence of excitation is required only to establish parameter convergence).

3. Convergence. Throughout this section we assume that $(\overline{q}, \overline{u})$ is a plant. We begin by establishing the convergence of the state estimator. Define the function $E : [0, \infty) \to R^1$ by

$$E(t) = \frac{1}{2} \left| \begin{bmatrix} e(t) \\ r(t) \end{bmatrix} \right|_X^2 = \frac{1}{2} \left\{ |e(t)|^2 + |r(t)|_Q^2 \right\}, \qquad t \geq 0. \tag{3.1}$$

We require the following lemma.

LEMMA 3.1. *For all* $t \geq 0$

$$E(t) + \beta \int_0^t \|e(s)\|^2 ds \leq \xi, \tag{3.2}$$

where $\xi = E(0) = \frac{1}{2}\left\{|e(0)|^2 + |r(0)|_Q^2\right\}$ *and* β *is as defined in assumption* (A4).

Proof. From (2.18), (2.19), and assumption (A4) we find that for $s > 0$

$$\begin{aligned} D_s E(s) &= \operatorname{Re}\langle D_s e(s), e(s)\rangle + \operatorname{Re}\langle D_s r(s), r(s)\rangle_Q \\ &= -\operatorname{Re}\langle Ae(s), e(s)\rangle \\ &\leq -\beta\|e(s)\|^2. \end{aligned} \tag{3.3}$$

Integrating from 0 to t, we obtain the desired result. □

Using Lemma 3.1, we show that the state error, $e(t)$, converges to zero asymptotically as $t \to \infty$. The proof is in the spirit of the arguments used in [27] to verify a result known as Barbălat's lemma. A somewhat different proof of this result can be found in [30] (see also [2]).

THEOREM 3.2. *The function* E *given in* (3.1) *is nonincreasing and*

$$\lim_{t\to\infty} |e(t)| = 0.$$

Proof. That E is nonincreasing follows immediately from the estimate (3.3). For $t_2 > t_1$, (2.18), assumption (A4), Definition 2.1, and Lemma 3.1 (more precisely, (3.1) and (3.2)) yield

$$\begin{aligned} |e(t_2)|^2 - |e(t_1)|^2 &= \int_{t_1}^{t_2} D_t|e(t)|^2 dt \\ &= 2\int_{t_1}^{t_2} \operatorname{Re}\langle D_t e(t), e(t)\rangle dt \\ &= 2\int_{t_1}^{t_2} \{-\operatorname{Re}\langle Ae(t), e(t)\rangle - \operatorname{Re}\langle B(\overline{u}(t))^* r(t), e(t)\rangle\} dt \\ &\leq -2\beta\int_{t_1}^{t_2} \|e(t)\|^2 dt + 2\int_{t_1}^{t_2} |\langle B(\overline{u}(t))^* r(t), e(t)\rangle| dt \\ &\leq 2\gamma(\overline{u})\int_{t_1}^{t_2} \|e(t)\||r(t)|_Q dt \\ &\leq 2\gamma(\overline{u})\left\{\int_{t_1}^{t_2} |r(t)|_Q^2 dt\right\}^{\frac{1}{2}}\left\{\int_{t_1}^{t_2} \|e(t)\|^2 dt\right\}^{\frac{1}{2}} \\ &\leq \frac{2\sqrt{2}\gamma(\overline{u})\xi}{\sqrt{\beta}}(t_2 - t_1)^{\frac{1}{2}}. \end{aligned}$$

Note that the estimate (3.2) implies that for all $L > 0$

$$\lim_{t\to\infty} \int_{t-L}^{t} \|e(s)\|^2 ds = 0, \tag{3.4}$$

and suppose that $\lim_{t\to\infty} |e(t)|^2 \neq 0$. Then there exist $\epsilon > 0$ and a sequence $\{t_n\}_{n=1}^{\infty}$ with $t_n > 0$ and $\lim_{n\to\infty} t_n = \infty$ for which

$$|e(t_n)|^2 > \epsilon, \qquad n = 1, 2, \ldots. \tag{3.5}$$

It follows from (3.4) and (3.5) that for $\delta > 0$ and $n = 1, 2, \ldots$

$$\begin{aligned}\int_{t_n-\delta}^{t_n} |e(t)|^2 dt &= \int_{t_n-\delta}^{t_n} |e(t_n)|^2 dt - \int_{t_n-\delta}^{t_n} \{|e(t_n)|^2 - |e(t)|^2\} dt \\ &> \epsilon\delta - \frac{2\sqrt{2}\gamma(\overline{u})\xi}{\sqrt{\beta}} \int_{t_n-\delta}^{t_n} (t_n - t)^{\frac{1}{2}} dt \\ &= \epsilon\delta - \mu\delta^{\frac{3}{2}},\end{aligned}$$

where $\mu = \frac{4\sqrt{2}\gamma(\overline{u})\xi}{3\sqrt{\beta}}$. Choosing $\delta = \frac{\epsilon^2}{4\mu^2}$, we obtain that

$$\int_{t_n-\delta}^{t_n} |e(t)|^2 dt > \frac{\epsilon^3}{8\mu^2} = \frac{\epsilon\delta}{2}, \qquad n = 1, 2, \ldots. \tag{3.6}$$

The estimate (3.6) together with (2.2) implies that for $n = 1, 2, \ldots,$

$$\int_{t_n-\delta}^{t_n} \|e(t)\|^2 dt \geq K^{-2} \int_{t_n-\delta}^{t_n} |e(t)|^2 dt > \frac{\epsilon\delta}{2K^2}, \qquad n = 1, 2 \ldots.$$

However, this contradicts (3.4). Consequently, $\lim_{t\to\infty} |e(t)|^2 = 0$, and the proof is complete. ∎

To establish parameter convergence, an additional hypothesis is required. We extend the finite-dimensional notion of persistence of excitation to infinite dimensions and argue parameter convergence using ideas similar to those used in [22] (see also [23]) to study the uniform asymptotic stability of certain classes of linear nonautonomous finite dimensional systems.

DEFINITION 3.3. *A plant $(\overline{q}, \overline{u})$ is said to be persistently excited, or an input f is said to be persistently exciting for the plant $(\overline{q}, \overline{u})$, if there exist T_0, δ_0, $\epsilon_0 > 0$ such that for each $q \in Q$ with $|q|_Q = 1$ and each $t > 0$ sufficiently large, there exists a $\tilde{t} \in [t, t + T_0]$ such that*

$$\left\| \int_{\tilde{t}}^{\tilde{t}+\delta_0} B(\overline{u}(\tau))^* q d\tau \right\|_* \geq \epsilon_0,$$

where for $t \geq 0$, $B(\overline{u}(t))^* \in \mathcal{L}(Q, V^*)$ is the Banach space adjoint of the operator $B(\overline{u}(t))$ defined in (2.16).

THEOREM 3.4. *If the plant $(\overline{q}, \overline{u})$ is persistently excited, then* $\lim_{t\to\infty} |r(t)|_Q = 0$.

The proof of Theorem 3.4 is argued using two lemmas, which we now state and prove.

LEMMA 3.5. *Let $\delta > 0$ be given. If the plant $(\overline{q}, \overline{u})$ is persistently excited, then there exist positive numbers $\epsilon = \epsilon(\delta)$, $T_1 = T_1(\delta)$, and T such that for all $t_1 \geq T_1$, if $|r(t)|_Q \geq \delta$ for $t \in [t_1, t_1 + T]$, then there exists a $\hat{t} \in [t_1, t_1 + T]$ such that $|e(\hat{t})| \geq \epsilon$.*

Proof. Let T_0, δ_0, ϵ_0, and $\tilde{t}$ be as in Definition 3.3 with $t = t_1$ (t_1 assumed to be sufficiently large to apply the condition of persistence of excitation), and $q = p(t_1) = r(t_1)/|r(t_1)|_Q$. Set $T = T_0 + \delta_0$ and assume that $|r(t)|_Q \geq \delta$ for all $t \in [t_1, t_1 + T]$. Integrating (2.18) over the interval $[\tilde{t}, \tilde{t} + \delta_0]$, taking norms in V^*, and applying the triangle inequality we obtain

$$\|e(\tilde{t} + \delta_0)\|_* \geq \left\| \int_{\tilde{t}}^{\tilde{t}+\delta_0} B(\overline{u}(\tau))^* r(\tau) d\tau \right\|_* - \left\| e(\tilde{t}) - \int_{\tilde{t}}^{\tilde{t}+\delta_0} Ae(\tau) d\tau \right\|_*. \tag{3.7}$$

The second term on the right-hand side of (3.7) can be estimated using assumption (A3), (2.2), and the Cauchy–Schwarz inequality. Indeed

$$\left\| e(\tilde{t}) - \int_{\tilde{t}}^{\tilde{t}+\delta_0} Ae(\tau)d\tau \right\|_* \leq \|e(\tilde{t})\|_* + \alpha \int_{\tilde{t}}^{\tilde{t}+\delta_0} \|e(\tau)\|d\tau \leq K|e(\tilde{t})| + \alpha\sqrt{\delta_0}\sqrt{\int_{\tilde{t}}^{\tilde{t}+\delta_0} \|e(\tau)\|^2 d\tau}. \tag{3.8}$$

Applying the backward triangle inequality to the first term on the right-hand side of (3.7), we obtain

$$\left\| \int_{\tilde{t}}^{\tilde{t}+\delta_0} B(\overline{u}(\tau))^* r(\tau)d\tau \right\|_* \geq \left\| \int_{\tilde{t}}^{\tilde{t}+\delta_0} B(\overline{u}(\tau))^* p(t_1)|r(t_1)|_Q d\tau \right\|_* - \left\| \int_{\tilde{t}}^{\tilde{t}+\delta_0} B(\overline{u}(\tau))^* \left\{ p(t_1)|r(t_1)|_Q - r(\tau) \right\} d\tau \right\|_*. \tag{3.9}$$

Using the fact that $(\overline{q}, \overline{u})$ is a plant, integrating (2.19) over the interval $[t_1, \tau]$, for $\tau > t_1$, taking norms in Q, and applying the Cauchy–Schwarz inequality we obtain

$$\begin{aligned} |r(t_1) - r(\tau)|_Q &= \left| \int_{t_1}^{\tau} B(\overline{u}(t))e(t)dt \right|_Q \\ &\leq \int_{t_1}^{\tau} |B(\overline{u}(t))e(t)|_Q \, dt \\ &\leq \gamma(\overline{u}) \int_{t_1}^{\tau} \|e(t)\|dt \\ &\leq \gamma(\overline{u})(\tau - t_1)^{\frac{1}{2}} \sqrt{\int_{t_1}^{\tau} \|e(t)\|^2 dt}. \end{aligned} \tag{3.10}$$

Definition 2.1, $\tilde{t} \in [t_1, t_1 + T_0]$, and (3.10) then imply that

$$\begin{aligned} &\left\| \int_{\tilde{t}}^{\tilde{t}+\delta_0} B(\overline{u}(\tau))^* \left\{ p(t_1)|r(t_1)|_Q - r(\tau) \right\} d\tau \right\|_* \\ &\qquad \leq \int_{\tilde{t}}^{\tilde{t}+\delta_0} \|B(\overline{u}(\tau))^*\|_{\mathcal{L}(Q,V^*)} |r(t_1) - r(\tau)|_Q d\tau \\ &\qquad \leq \gamma(\overline{u})^2 \delta_0 \sqrt{\tilde{t} + \delta_0 - t_1} \sqrt{\int_{t_1}^{\tilde{t}+\delta_0} \|e(t)\|^2 dt} \\ &\qquad \leq \gamma(\overline{u})^2 \delta_0 \sqrt{T} \sqrt{\int_{t_1}^{t_1+T} \|e(t)\|^2 dt}. \end{aligned} \tag{3.11}$$

Since by assumption $|r(t)|_Q \geq \delta$, $t \in [t_1, t_1 + T]$, and $(\overline{q}, \overline{u})$ is persistently excited, (3.9) and (3.11) imply that

$$\left\| \int_{\tilde{t}}^{\tilde{t}+\delta_0} B(\overline{u}(\tau))^* r(\tau)d\tau \right\|_* \geq \delta\epsilon_0 - \gamma(\overline{u})^2 \delta_0 \sqrt{T} \sqrt{\int_{t_1}^{t_1+T} \|e(t)\|^2 dt}. \tag{3.12}$$

Then, from (3.7), (3.8), and (3.12), we obtain that

$$\|e(\tilde{t}+\delta_0)\|_* \geq \delta\epsilon_0 - K|e(\tilde{t})| - \left\{\alpha\sqrt{\delta_0} + \gamma(\overline{u})^2\sqrt{T}\delta_0\right\}\sqrt{\int_{t_1}^{t_1+T}\|e(t)\|^2 dt}. \tag{3.13}$$

Applying Lemma 3.1 and Theorem 3.2, let $T_1 = T_1(\delta)$ be so large that

$$|e(t)| \leq \frac{\delta\epsilon_0}{2K} \qquad \text{and} \qquad \sqrt{\int_t^{t+T}\|e(t)\|^2 dt} \leq \frac{\delta\epsilon_0}{4(\alpha\sqrt{\delta_0} + \gamma(\overline{u})^2\sqrt{T}\delta_0)} \tag{3.14}$$

for all $t \geq T_1$. It then follows from (3.13), (3.14), and (2.2) that $|e(\hat{t})| \geq \epsilon$, where $\hat{t} = \tilde{t} + \delta_0 \in [t_1 + \delta_0, t_1 + T_0 + \delta_0] \subset [t_1, t_1 + T]$ and $\epsilon = \frac{\delta\epsilon_0}{4K}$, and thus the lemma is proven. ∎

LEMMA 3.6. *Let $\delta > 0$ be given and $T_1 = T_1(\delta)$ and T be as they were defined in Lemma 3.5. If the plant $(\overline{q}, \overline{u})$ is persistently excited, then there exists $T_2 = T_2(\delta) > 0$ with $T_2 \geq T_1$ such that if $t_1 \geq T_2$, then there exists $t_2 \in [t_1, t_1 + T]$ such that $|r(t_2)|_Q < \delta$.*

Proof. Let $\epsilon = \epsilon(\delta) = \frac{\epsilon_0\delta}{4K}$ be as it was defined in the proof of Lemma 3.5. Theorem 3.2 implies that there exists $S = S(\epsilon)$ such that

$$|e(s)| < \epsilon = \frac{\epsilon_0\delta}{4K}, \qquad s \geq S. \tag{3.15}$$

Set $T_2 = T_2(\delta) = \max\{T_1(\delta), S(\epsilon)\}$. Now if the lemma were not true, there would exist a $t_1 \geq T_2$ such that $|r(t)|_Q \geq \delta$ for all $t \in [t_1, t_1 + T]$. But then Lemma 3.5 would imply that there exist $\hat{t} \in [t_1, t_1 + T]$ such that $|e(\hat{t})| \geq \epsilon$. Since $\hat{t} \geq t_1 \geq T_2 \geq S$, this contradicts (3.15), and the lemma is proven. ∎

We are now prepared to prove Theorem 3.4.

Proof of Theorem 3.4. We show that for any $\epsilon > 0$, there exists a $\hat{t}$ such that

$$E(\hat{t}) \leq \epsilon, \tag{3.16}$$

where the function E is given by (3.1). Since, by Theorem 3.2, E is nonincreasing, (3.16) implies that $\lim_{t\to\infty} E(t) = 0$ and, therefore, that $\lim_{t\to\infty} |r(t)|_Q = 0$.

To establish (3.16), first note that if $E(t_1) \leq \epsilon$, then we are finished. On the other hand, if $E(t_1) > \epsilon$, we show that there exist $M > 0$ and $\gamma \in (0, 1)$, both depending only on the estimator (2.13)–(2.15) and the plant $(\overline{q}, \overline{u})$ (i.e., A, T_0, δ_0, ϵ_0, f, etc.), such that there exists a $\hat{t}_1 \in [t_1, t_1 + M]$ for which

$$E(\hat{t}_1) \leq \gamma E(t_1) + \rho(t_1), \tag{3.17}$$

where ρ is such that $\rho(t) \geq 0$, $t \geq 0$, and $\lim_{t\to\infty} \rho(t) = 0$. It follows that there exists a positive integer $\overline{K}$ which depends only on ϵ such that $E(t_1 + \overline{K}M) \leq \epsilon$. Indeed, by repeating the argument that leads to (3.17), we obtain the difference inequality $E_{k+1} \leq \gamma E_k + \rho_k$, $k = 0, 1, 2, \ldots$, where $E_k = E(t_1 + kM)$, $k = 0, 1, 2, \ldots$, $\rho_0 = \rho(t_1)$, $\rho_k \geq 0$, $k = 0, 1, 2, \ldots$, and $\lim_{k\to\infty} \rho_k = 0$. It follows that

$$E_k \leq \gamma^k E_0 + \sum_{j=0}^{k-1} \gamma^{k-j-1}\rho_j, \qquad k = 0, 1, 2, \ldots.$$

Letting J be so large that $\rho_j \le \frac{\epsilon(1-\gamma)}{3}$, $j \ge J$, and choosing $\overline{K} > J$ so large that $\gamma^k E_0 = \gamma^k E(t_1) \le \frac{\epsilon}{3}$ and $\gamma^k \sum_{j=0}^{J} \gamma^{-(j+1)} \rho_j \le \frac{\epsilon}{3}$, $k \ge \overline{K}$, we obtain

$$\begin{aligned} E(t + \overline{K}M) = E_{\overline{K}} &\le \gamma^{\overline{K}} E_0 + \gamma^{\overline{K}} \sum_{j=0}^{J-1} \gamma^{-(j+1)} \rho_j + \sum_{j=J}^{\overline{K}-1} \gamma^{\overline{K}-j-1} \rho_j \\ &\le \frac{\epsilon}{3} + \frac{\epsilon}{3} + \frac{\epsilon}{3}(1 - \gamma^{\overline{K}-J}) \le \epsilon. \end{aligned} \tag{3.18}$$

Consequently, (3.18) yields (3.16) with $\hat{t} = t_1 + \overline{K}M$.

Let $\epsilon > 0$ be given and $c_1, c_2 > 0$ be chosen so that

$$0 < \beta K^{-2} c_2 \{(2 - c_1) - 2\gamma(\overline{u}) c_2\} < 1. \tag{3.19}$$

For example, set $c_1 = 1$ and $c_2 < \min\{\frac{1}{2\gamma(\overline{u})}, \frac{K^2}{\beta}\}$. Note that the values of c_1 and c_2 depend only on the plant and the estimator dynamics, A. Apply Lemma 3.6 with $\delta = \sqrt{\epsilon c_1}$ to obtain T_2. Let $t_1 \ge T_2$ and let $t_2 \in [t_1, t_1 + T]$ be such that $|r(t_2)|_Q < \delta$.

If $E(t_2) \le \epsilon$, we are finished. So assume that

$$E(t_2) > \epsilon. \tag{3.20}$$

Now

$$E(t_2) = \frac{1}{2}\left\{|e(t_2)|^2 + |r(t_2)|_Q^2\right\} \le \frac{1}{2}\left\{|e(t_2)|^2 + \delta^2\right\} = \frac{1}{2}\left\{|e(t_2)|^2 + \epsilon c_1\right\}.$$

But then (3.20) implies that

$$|e(t_2)|^2 \ge 2E(t_2) - \epsilon c_1 > (2 - c_1) E(t_2). \tag{3.21}$$

For $t \ge t_2$, (2.18), assumption (A3), Definition 2.1, and the Cauchy–Schwarz inequality imply that

$$\begin{aligned} |e(t_2)|^2 - |e(t)|^2 &\le \left| \int_{t_2}^{t} D_s |e(s)|^2 ds \right| = \left| 2 \int_{t_2}^{t} \operatorname{Re} \langle D_s e(s), e(s) \rangle ds \right| \\ &= \left| -2 \int_{t_2}^{t} \operatorname{Re} \langle Ae(s), e(s) \rangle ds - 2 \int_{t_2}^{t} \operatorname{Re} \langle B(\overline{u}(s))^* r(s), e(s) \rangle ds \right| \\ &\le 2\alpha \int_{t_2}^{t} \|e(s)\|^2 ds + 2\gamma(\overline{u}) \int_{t_2}^{t} |r(s)|_Q \|e(s)\| ds \\ &\le \{2\alpha + \gamma(\overline{u})\} \int_{t_2}^{t} \|e(s)\|^2 ds + \gamma(\overline{u}) \int_{t_2}^{t} |r(s)|_Q^2 ds. \end{aligned} \tag{3.22}$$

Recalling (3.1) and Theorem 3.2, (3.22) implies that for $t \in [t_2, t_2 + c_2]$ we have

$$|e(t_2)|^2 - |e(t)|^2 \le \{2\alpha + \gamma(\overline{u})\} \int_{t_2}^{t_2 + c_2} \|e(s)\|^2 ds + 2\gamma(\overline{u}) c_2 E(t_2). \tag{3.23}$$

Combining (3.21) and (3.23), we find that for $t \in [t_2, t_2 + c_2]$

$$|e(t)|^2 \ge \{(2 - c_1) - 2\gamma(\overline{u}) c_2\} E(t_2) - \{2\alpha + \gamma(\overline{u})\} \int_{t_2}^{t_2 + c_2} \|e(s)\|^2 ds. \tag{3.24}$$

Recalling (3.3), we have that

$$E(t) - E(t_2) \leq -\beta \int_{t_2}^{t} \|e(s)\|^2 ds. \tag{3.25}$$

Setting $t = t_2 + c_2$ in (3.25) and recalling (2.2), (3.24) implies that

$$\begin{aligned} &E(t_2) - E(t_2 + c_2) \geq \beta \int_{t_2}^{t_2+c_2} \|e(s)\|^2 ds \\ &\geq \beta K^{-2} \int_{t_2}^{t_2+c_2} |e(s)|^2 ds \\ &\geq \beta K^{-2} c_2 \left\{(2 - c_1) - 2\gamma(\overline{u})c_2\right\} E(t_2) - \beta K^{-2} c_2 \left\{2\alpha + \gamma(\overline{u})\right\} \int_{t_2}^{t_2+c_2} \|e(s)\|^2 ds, \end{aligned}$$

or

$$\begin{aligned} E(t_2 + c_2) &\leq \left\{1 - \beta K^{-2} c_2 \left\{(2 - c_1) - 2\gamma(\overline{u})c_2\right\}\right\} E(t_2) \\ &\quad + \beta K^{-2} c_2 \left\{2\alpha + \gamma(\overline{u})\right\} \int_{t_2}^{t_2+c_2} \|e(s)\|^2 ds \\ &\leq (1 - \gamma_0) E(t_1) + \rho(t_1), \end{aligned} \tag{3.26}$$

where $\gamma_0 = \beta K^{-2} c_2 \left\{(2 - c_1) - 2\gamma(\overline{u})c_2\right\}$ and $\rho(t) = \frac{\beta c_2 \{2\alpha + \gamma(\overline{u})\}}{K^2} \int_t^{t+T+c_2} \|e(s)\|^2 ds$. In the estimate (3.26) we have used the fact that $t_2 \in [t_1, t_1 + T]$ and, since E is nonincreasing and (3.19) implies that $\gamma_0 \in (0, 1)$, that $(1 - \gamma_0)E(t_2) \leq (1 - \gamma_0)E(t_1)$. Recalling Lemma 3.1, we have $\lim_{t\to\infty} \rho(t) = 0$. Thus (3.26) yields (3.17) with $\gamma = 1 - \gamma_0 \in (0, 1)$ and $\hat{t_1} = t_2 + c_2 \in [t_1, t_1 + M]$, where $M = T + c_2$. This proves the theorem. ∎

Considerable insight can be gained from the proofs of Lemma 3.5, Lemma 3.6, and Theorem 3.4. In particular, the arguments and estimates used in these proofs suggest how the persistence of excitation parameters T_0, δ_0, and ϵ_0 and the choice of the estimator dynamics A retard or accelerate convergence. The following observations can be made.

(i) As ϵ_0 increases, the value of ϵ in Lemma 3.5 increases, and therefore the value of T_2 in Lemma 3.6 decreases. Consequently, convergence will be more rapid.

(ii) As either T_0 or δ_0 decrease, the values of T and T_1 in Lemma 3.5, the value of T_2 in Lemma 3.6, and the value of M in Theorem 3.4 decrease as well. It follows that more rapid convergence results.

(iii) As the value of β in assumption (A4) increases, the convergence of $|e(t)|$ to zero as $t \to \infty$ is more rapid (see Theorem 3.2). Thus the value of T_1 in Lemma 3.5 and the value of T_2 in Lemma 3.6 decrease, and convergence will be more rapid. Also, in the proof of Theorem 3.4, if β increases, either the value of γ_0 will increase or the value of c_2 will decrease, and therefore either the value of $\gamma = 1 - \gamma_0$ will decrease or the value of M will decrease. In either case the rate of convergence will be enhanced.

One way to either increase ϵ_0 or decrease δ_0 or T_0 in Definition 3.3 is to increase the *gain* on the input f. Assuming that the plant is linear and initially at rest, the linearity of (2.5) implies that an increase in the gain on $\overline{u}$ will result, and therefore, it is likely that the value of $\gamma(\overline{u})$ will also increase. However, in the proof of Theorem 3.4, if $\gamma(\overline{u})$ increases for a fixed value of c_2 (and therefore M), γ_0 will decrease and consequently γ will increase, thus slowing convergence.

For $\tau > 0$, integrating (2.19) from t to $t+\tau$, taking norms, and using the fact that $(\overline{q}, \overline{u})$ is a plant, we find that

$$\begin{aligned} |r(t+\tau) - r(t)|_Q &= \left| \int_t^{t+\tau} B(\overline{u}(s))e(s)ds \right|_Q \\ &\leq \gamma(\overline{u}) \int_t^{t+\tau} \|e(s)\| ds \\ &\leq \gamma(\overline{u})\sqrt{\tau}\sqrt{\int_t^{t+\tau} \|e(s)\|^2 ds}, \qquad t \geq 0. \end{aligned}$$

It follows that

$$\left| \frac{\Delta_\tau r(t)}{\tau} \right|_Q \leq \gamma(\overline{u})\sqrt{\frac{1}{\tau}\int_t^{t+\tau} \|e(s)\|^2 ds}, \quad t \geq 0,$$

where $\Delta_\tau r(t) = r(t+\tau) - r(t)$, $t \geq 0$. Recalling Lemma 3.1, we have that for each $\tau > 0$

$$\lim_{t\to\infty} \left| \frac{\Delta_\tau r(t)}{\tau} \right|_Q = 0. \tag{3.27}$$

Moreover, from Lemma 3.1, and in particular (3.2), it follows that the rate of convergence in (3.27) increases with increasing β. Consequently, if the estimator dynamics, A, are chosen so that β is too large, the average rate of change in r, the parameter error, will tend to zero too rapidly. In effect, the estimator will be *overdamped* and sluggish parameter convergence will result.

The remarks above indicate that making an appropriate choice of an input, f, and the estimator dynamics, A, is delicate. One must balance those factors which tend to enhance convergence with those that tend to retard it. In [8] a careful study of this phenomenon was undertaken. By looking at a plant consisting of a one-dimensional heat equation with a monochromatic modal input, and an estimator whose dynamics are also described by a one-dimensional heat equation, it was observed that the error equations (2.18), (2.19), or (2.20), to first order, took the form of a damped linear harmonic oscillator. The damping was determined by the magnitude of β, and the stiffness was related to the value of $\gamma(\overline{u})^2$. If β was too large (relative to $\gamma(\overline{u})$), the system was overdamped and parameter convergence was slow. If, on the other hand, $\gamma(\overline{u})$ was too large (relative to β), the system was stiff and underdamped. Oscillations, which are particularly undesirable in a parameter estimator being used as a part of an indirect adaptive control algorithm, resulted. Choosing the estimator dynamics, A, and input f to optimize the performance of the estimator required finding an appropriate compromise between these two extremes.

It is possible to establish a parameter convergence result in the absence of persistence of excitation in the spirit of the treatment in [2] for the identification of second-order elliptic partial differential equations via an asymptotic embedding technique. The result which we will establish below also provides insight into a phenomenon which we refer to as *partial persistence of excitation*. That is, the plant is persistently excited with respect to some subset of the unknown parameters and is not, or is to a lesser degree, persistently excited with respect to the rest.

For ξ as defined in the statement of Lemma 3.1, let $B_\xi = \{q \in Q : |q|_Q \leq \sqrt{\xi}\}$ and

$$\hat{Q} = \left\{ q \in Q : \lim_{t\to\infty} \left| \int_t^{t+L} \langle B(\overline{u}(\tau))^* q, \varphi \rangle d\tau \right| = 0, \varphi \in V, L > 0 \right\}.$$

We assume that for $q \in Q$, $A_0(q) \in \mathcal{L}(V, V^*)$, and that in assumption (A1) $A_0(q) = A_1(q)$ (i.e., that $A_2\varphi = 0$, $\varphi \in V$). We assume further that $u_0 \in V$, $f \in C([0,\infty); V)$ is uniformly Hölder continuous, and there exists $f_\infty \in V^*$ such that $\lim_{t\to\infty} \|f(t) - f_\infty\|_* = 0$. Then, as was discussed in section 2, it follows that there exists $\overline{u}_\infty \in V$ such that

$$\lim_{t\to\infty} \|\overline{u}(t) - \overline{u}_\infty\| = 0. \tag{3.28}$$

Under these assumptions, we obtain the following theorem.

THEOREM 3.7. *For r satisfying* (2.19) *we have*

$$\lim_{t\to\infty} w - \operatorname{dist}(r(t), \hat{Q} \cap B_\xi) = 0, \tag{3.29}$$

where $w - \operatorname{dist}(\cdot,\cdot)$ denotes the distance function with respect to the weak topology on Q.

Proof. Suppose that (3.29) does not hold. Then there would exist a sequence, $\{t_n\}_{n=1}^\infty$, and $\eta > 0$ such that $\lim_{n\to\infty} t_n = \infty$ and $w - \operatorname{dist}(r(t_n), \hat{Q} \cap B_\xi) \geq \eta$. Since $Q \cap B_\xi$ is a bounded subset of the Hilbert space Q, it is weakly compact. Consequently there exists a subsequence which we will again denote by $\{t_n\}_{n=1}^\infty$, and $r_\infty \in B_\xi$ such that $w - \lim_{n\to\infty} r(t_n) = r_\infty$ and $w - \operatorname{dist}(r_\infty, \hat{Q} \cap B_\xi) \geq \eta$. It follows that $r_\infty \notin \hat{Q}$ and therefore that there exists $\tilde{\varphi} \in V$ and $\epsilon > 0$ such that

$$\overline{\lim_{n\to\infty}} \left| \int_{t_n}^{t_n+\epsilon} \langle B(\overline{u}(\tau))^* r_\infty, \tilde{\varphi} \rangle d\tau \right| = \delta > 0. \tag{3.30}$$

But (3.30) implies that there exists a subsequence, $\{t_{n_k}\}_{k=1}^\infty$, of $\{t_n\}_{n=1}^\infty$ such that

$$\left| \int_{t_{n_k}}^{t_{n_k}+\epsilon} \langle B(\overline{u}(\tau))^* r(t_{n_k}), \tilde{\varphi} \rangle d\tau \right| > \frac{\delta}{2} \tag{3.31}$$

for all $k = 1, 2, \ldots$. Indeed, if this were not the case, then there would exist N such that $|\int_{t_n}^{t_n+\epsilon} \langle B(\overline{u}(\tau))^* r(t_n), \tilde{\varphi} \rangle d\tau| \leq \frac{\delta}{2}$ for all $n > N$. It would then follow that for $n > N$

$$\begin{aligned}
&\left| \int_{t_n}^{t_n+\epsilon} \langle B(\overline{u}(\tau))^* r_\infty, \tilde{\varphi} \rangle d\tau \right| \\
&\quad \leq \left| \int_{t_n}^{t_n+\epsilon} \langle B(\overline{u}(\tau))^* \{r_\infty - r(t_n)\}, \tilde{\varphi} \rangle d\tau \right| + \left| \int_{t_n}^{t_n+\epsilon} \langle B(\overline{u}(\tau))^* r(t_n), \tilde{\varphi} \rangle d\tau \right| \\
&\quad \leq \left| \int_0^\epsilon \langle r(t_n) - r_\infty, B(\overline{u}(t_n+\tau))\tilde{\varphi} \rangle_Q d\tau \right| + \frac{\delta}{2} \\
&\quad \leq \left| \int_0^\epsilon \langle A_1(r(t_n) - r_\infty)\overline{u}(t_n+\tau) - A_1(r(t_n) - r_\infty)\overline{u}_\infty, \tilde{\varphi} \rangle d\tau \right| \\
&\qquad + \left| \int_0^\epsilon \langle A_1(r(t_n) - r_\infty)\overline{u}_\infty, \tilde{\varphi} \rangle d\tau \right| + \frac{\delta}{2} \\
&\quad \leq 2\alpha_1 \sqrt{\xi} \|\tilde{\varphi}\| \int_0^\epsilon \|\overline{u}(t_n+\tau) - \overline{u}_\infty\| d\tau + \epsilon \left| \langle r(t_n) - r_\infty, B(\overline{u}_\infty)\tilde{\varphi} \rangle_Q \right| + \frac{\delta}{2},
\end{aligned} \tag{3.32}$$

where the final two estimates above are consequences of (2.3), assumption (A2), and, recalling our findings in section 2, the fact that under the present assumptions $\|\overline{u}(t)\|$ is bounded for $t \in [0, \infty)$.

It follows from (3.28), together with the bounded convergence theorem, that the term involving the integral in the final estimate in (3.32) tends to zero as $n \to \infty$. Moreover, from the fact that $w - \lim_{n\to\infty} r(t_n) = r_\infty$, it also follows that $\overline{\lim_{n\to\infty}} | \int_{t_n}^{t_n+\epsilon} \langle B(\overline{u}(\tau))^* r_\infty, \tilde{\varphi} \rangle d\tau | \le \frac{\delta}{2}$, which contradicts (3.30).

Now (2.18), assumption (A3), and the Cauchy–Schwarz inequality imply that

$$\begin{aligned}\left\| \int_{t_{n_k}}^{t_{n_k}+\epsilon} B(\overline{u}(\tau))^* r(\tau) d\tau \right\|_* &= \left\| e(t_{n_k}) - e(t_{n_k}+\epsilon) - \int_{t_{n_k}}^{t_{n_k}+\epsilon} A e(\tau) d\tau \right\|_* \\ &\le K|e(t_{n_k})| + K|e(t_{n_k}+\epsilon)| + \alpha\sqrt{\epsilon}\sqrt{\int_{t_{n_k}}^{t_{n_k}+\epsilon} \|e(\tau)\|^2 d\tau},\end{aligned}$$

which, together with Theorem 3.2 and Lemma 3.1, yields that

$$\lim_{n\to\infty} \left\| \int_{t_{n_k}}^{t_{n_k}+\epsilon} B(\overline{u}(\tau))^* r(\tau) d\tau \right\|_* = 0. \tag{3.33}$$

Also, recalling the estimate (3.10), using (2.3) and assumption (A2), we find that

$$\begin{aligned}&\left| \left| \int_{t_{n_k}}^{t_{n_k}+\epsilon} \langle B(\overline{u}(\tau))^* r(\tau), \tilde{\varphi} \rangle d\tau \right| - \left| \int_{t_{n_k}}^{t_{n_k}+\epsilon} \langle B(\overline{u}(\tau))^* r(t_{n_k}), \tilde{\varphi} \rangle d\tau \right| \right| \\ &\le \left| \int_{t_{n_k}}^{t_{n_k}+\epsilon} \langle B(\overline{u}(\tau))^* \{r(\tau) - r(t_n)\}, \tilde{\varphi} \rangle d\tau \right| \\ &= \left| \int_{t_{n_k}}^{t_{n_k}+\epsilon} \langle A_1(r(\tau) - r(t_{n_k}))\overline{u}(\tau), \tilde{\varphi} \rangle d\tau \right| \\ &\le \alpha_1 \|\tilde{\varphi}\| \int_{t_{n_k}}^{t_{n_k}+\epsilon} |r(\tau) - r(t_{n_k})|_Q \|\overline{u}(\tau)\| d\tau \\ &\le \alpha_1 \|\tilde{\varphi}\| \gamma(\overline{u}) \sqrt{\epsilon} \sqrt{\int_{t_{n_k}}^{t_{n_k}+\epsilon} \|e(\tau)\|^2 d\tau} \int_{t_{n_k}}^{t_{n_k}+\epsilon} \|\overline{u}(\tau)\| d\tau.\end{aligned}$$

The fact that $\|\overline{u}(t)\|$ is bounded on $[0, \infty)$ (see section 2) and Lemma 3.1 then imply that

$$\lim_{k\to\infty} \left| \left| \int_{t_{n_k}}^{t_{n_k}+\epsilon} \langle B(\overline{u}(\tau))^* r(\tau), \tilde{\varphi} \rangle d\tau \right| - \left| \int_{t_{n_k}}^{t_{n_k}+\epsilon} \langle B(\overline{u}(\tau))^* r(t_{n_k}), \tilde{\varphi} \rangle d\tau \right| \right| = 0. \tag{3.34}$$

But (3.33) implies that

$$\lim_{k\to\infty} \left| \int_{t_{n_k}}^{t_{n_k}+\epsilon} \langle B(\overline{u}(\tau))^* r(\tau), \tilde{\varphi} \rangle d\tau \right| \le \lim_{k\to\infty} \left\| \int_{t_{n_k}}^{t_{n_k}+\epsilon} B(\overline{u}(\tau))^* r(\tau) d\tau \right\|_* \|\tilde{\varphi}\| = 0. \tag{3.35}$$

Combining (3.34) and (3.35) we obtain a contradiction to (3.31), and the theorem is proven. □

Theorem 3.7 yields the following corollary. Its proof, which is omitted, is exactly the same as the one given to verify Theorem 4.4 in [2] (see also [30]).

COROLLARY 3.8. *Under the same assumptions required to establish Theorem* 3.7, *we have* $\lim_{t\to\infty} w - \operatorname{dist}(q(t), \hat{P} \cap B_{\overline{\xi}}) = 0$, *where* $\overline{\xi} = \sqrt{\xi} + |\overline{q}|_Q$ *and* $\hat{P}$ *is the linear variety in* Q *given by* $\hat{P} = \overline{q} + \hat{Q}$. *Moreover* $w - \lim_{t\to\infty} q(t) = \overline{p} + P_{\hat{Q}} q(0)$, *where* $P_{\hat{Q}}$ *is the orthogonal projection of* Q *onto the closed linear subspace* $\hat{Q}$, *and* $\overline{p} = \overline{q} - P_{\hat{Q}}\overline{q}$ *is the unique element of minimum norm in* $\hat{P}$ (*see, for example,* [21]).

Note that the fact that $\hat{Q}$ is closed follows from the assumption that $(\overline{q}, \overline{u})$ is a plant.

If $\hat{Q} = \{0\}$, we shall say that the plant $(\overline{q}, \overline{u})$ is *weakly persistently excited*. When this is the case, we obtain weak parameter convergence. Indeed, Corollary 3.8 implies that $w - \lim_{t\to\infty} q(t) = \overline{q}$.

It is also possible to obtain what we shall call *weak partial parameter convergence* when the plant $(\overline{q}, \overline{u})$ is only partially weakly persistently excited. Suppose that $Q = Q_1 \oplus Q_1^{\perp}$ with $Q_1 \subset \hat{Q}^{\perp}$. Then Corollary 3.8 implies that $w - \lim_{t\to\infty} P_1 q(t) = P_1\overline{p} + P_1 P_{\hat{Q}} q(0) = P_1\overline{q} - P_1 P_{\hat{Q}}\overline{q} = P_1\overline{q}$, where P_1 denotes the orthogonal projection of Q onto Q_1. Thus if the plant $(\overline{q}, \overline{u})$ is weakly persistently excited with respect to some of the unknown parameters, the estimates for those parameters, $P_1 q(t)$, will converge weakly to the corresponding true plant parameters, $P_1\overline{q}$. An illustration of this phenomenon can be found in [8]. Note that when, as is frequently the case, Q is finite dimensional, the weak convergence discussed above becomes strong convergence.

Finally we note that persistence of excitation is sufficient to establish an identifiability result similar to the one in [5].

THEOREM 3.9. *If the plant* $(\overline{q}, \overline{u})$ *is persistently excited, then the parameter* $\overline{q}$ *is identifiable.*

Proof. Suppose not. That is, there exists $\overline{q}_1, \overline{q}_2 \in Q$ such that $\overline{u}$ is a solution to the initial value problem (2.5), (2.6) with either $q = \overline{q}_1$ or $q = \overline{q}_2$. Subtraction then yields that $\langle A_1(\overline{q}_1 - \overline{q}_2)\overline{u}(t), \varphi\rangle = 0$, a.e. $t > 0$, $\varphi \in V$, or, in light of (2.3), that $B(\overline{u}(t))^*\{\overline{q}_1 - \overline{q}_2\} = 0$, a.e. $t > 0$. This clearly contradicts Definition 3.3 unless $\overline{q}_1 = \overline{q}_2$, and the theorem is proven. ∎

4. Approximation theory. The estimator (2.13)–(2.15) is infinite dimensional. Its implementation requires finite-dimensional approximation. We consider Galerkin approximation and establish a convergence result.

For each $n = 1, 2, \ldots$, let H^n be a finite-dimensional subspace of H with $H^n \subset V$, and let Q^n be a finite-dimensional subspace of Q. The Galerkin equations corresponding to (2.13)–(2.15) are given by

$$\begin{aligned}\langle D_t u^n(t), \varphi^n\rangle + \langle Au^n(t), \varphi^n\rangle + \langle B(\overline{u}(t))^* q^n(t), \varphi^n\rangle = \langle f(t), \varphi^n\rangle \\ + \langle A\overline{u}(t), \varphi^n\rangle - \langle A_2\overline{u}(t), \varphi^n\rangle, \quad \varphi^n \in H^n, \text{ a.e. } t > 0,\end{aligned} \tag{4.1}$$

$$\begin{aligned}\langle D_t q^n(t), \psi^n\rangle_Q - \langle B(\overline{u}(t))u^n(t), \psi^n\rangle_Q \\ = -\langle B(\overline{u}(t))\overline{u}(t), \psi^n\rangle_Q, \ \psi^n \in Q^n, \text{ a.e. } t > 0,\end{aligned} \tag{4.2}$$

$$u^n(0) \in H^n, \qquad q^n(0) \in Q^n. \tag{4.3}$$

An argument similar to the one outlined previously in section 2 for the initial value problem (2.13)–(2.15) can be used to establish the existence of a unique solution, (q^n, u^n), to the initial value problem (4.1)–(4.3) for each $T > 0$ with $u^n \in H^1(0, T; H^n)$ and $q^n \in H^1(0, T; Q^n)$.

In order to establish convergence we require the following assumption:

(A5) For each fixed $T > 0$ and (q, u) the solution to the initial value problem (2.13)–(2.15), there exist functions $u_n \in H^1(0, T; H^n)$ and $q_n \in H^1(0, T; Q^n)$

such that $\lim_{n\to\infty} u_n = u$ in $L_2(0,T;V)$ and $C(0,T;H)$, $\lim_{n\to\infty} q_n = q$ in $C(0,T;Q)$, $\lim_{n\to\infty} D_t u_n = D_t u$ in $L_2(0,T;V^*)$, and $\lim_{n\to\infty} D_t q_n = D_t q$ in $L_2(0,T;Q)$.

THEOREM 4.1. *Assume that assumption* (A5) *holds, let* (q^n, u^n) *be the solution to the initial value problem* (4.1)–(4.3) *with* $u^n(0) = u_n(0)$ *and* $q^n(0) = q_n(0)$*, and let* (q,u) *be the solution to the initial value problem* (2.13)–(2.15)*. Then for each* $T > 0$*,* $\lim_{n\to\infty} u^n = u$ *in* $L_2(0,T;V)$ *and* $C(0,T;H)$ *and* $\lim_{n\to\infty} q^n = q$ *in* $C(0,T;Q)$*.*

Proof. Assumption (A5) and the triangle inequality imply that we need only show that

$$\lim_{n\to\infty} \int_0^T \|u_n(t) - u^n(t)\|^2 dt = 0, \qquad \lim_{n\to\infty} \sup_{t\in[0,T]} |u_n(t) - u^n(t)| = 0,$$

$$\text{and} \qquad \lim_{n\to\infty} \sup_{t\in[0,T]} |q_n(t) - q^n(t)|_Q = 0.$$

Toward this end, let $w^n = u^n - u_n$ and $p^n = q^n - q_n$. Then, using the fact that (q^n, u^n) satisfies (4.1), (4.2), and (q,u) satisfies (2.13), (2.14), we obtain the identity

$$\begin{aligned}
&D_t\{|w^n(t)|^2 + |p^n(t)|_Q^2\} = 2\mathrm{Re}\,\{\langle D_t w^n(t), w^n(t)\rangle + \langle D_t p^n(t), p^n(t)\rangle_Q\}\\
&= 2\mathrm{Re}\,\{-\langle A w^n(t), w^n(t)\rangle + \langle A\{u(t) - u_n(t)\}, w^n(t)\rangle\\
&+\langle B(\overline{u}(t))^*\{q(t) - q_n(t)\}, w^n(t)\rangle + \langle D_t\{u(t) - u_n(t)\}, w^n(t)\rangle\\
&+\langle B(\overline{u}(t))\{u_n(t) - u(t)\}, p^n(t)\rangle_Q + \langle D_t\{q(t) - q_n(t)\}, p^n(t)\rangle_Q\}, \qquad \text{a.e.} \quad t > 0.
\end{aligned}$$

Assumptions (A3) and (A4), the fact that $(\overline{q}, \overline{u})$ is a plant, and the well-known inequality

$$ab \le \frac{\epsilon}{2}a^2 + \frac{b^2}{2\epsilon}, \qquad a, b \in R,\ \epsilon > 0, \tag{4.4}$$

then yield that

$$\begin{aligned}
&D_t\{|w^n(t)|^2 + |p^n(t)|_Q^2\}\\
&\le -2\beta\|w^n(t)\|^2 + \frac{\alpha^2}{\epsilon}\|u(t) - u_n(t)\|^2 + \frac{\gamma(\overline{u})^2}{\epsilon}|q(t) - q_n(t)|_Q^2\\
&\quad + \frac{1}{\epsilon}\|D_t u(t) - D_t u_n(t)\|_*^2 + \gamma(\overline{u})^2\|u_n(t) - u(t)\|^2 + |D_t q(t) - D_t q_n(t)|_Q^2\\
&\quad + 3\epsilon\|w^n(t)\|^2 + 2|p^n(t)|_Q^2, \qquad \text{a.e.} \quad t > 0.
\end{aligned}$$

Choosing $\epsilon < \frac{2}{3}\beta$ and integrating from 0 to t, we obtain

$$\delta \int_0^t \|w^n(s)\|^2 ds + |w^n(t)|^2 + |p^n(t)|_Q^2 \le z^n(t) + 2\int_0^t |w^n(s)|^2 + |p^n(s)|_Q^2 ds, \tag{4.5}$$

where $\delta = 2\beta - 3\epsilon > 0$ and

$$\begin{aligned}
z^n(t) = \int_0^t &\left\{\frac{\alpha^2}{\epsilon} + \gamma(\overline{u})^2\right\}\|u_n(s) - u(s)\|^2 + \frac{1}{\epsilon}\|D_s u_n(s) - D_s u(s)\|_*^2\\
&+\frac{\gamma(\overline{u})^2}{\epsilon}|q_n(s) - q(s)|_Q^2 + |D_s q_n(s) - D_s q(s)|_Q^2 ds.
\end{aligned}$$

It follows from assumption (A5) that $\lim_{n\to\infty} z^n(t) = 0$ uniformly on $[0,T]$ for each $T > 0$. An application of the Gronwall lemma to the estimate (4.5) above yields the desired result. □

Since the state of the plant at each time t, $\overline{u}(t)$, is also in the infinite-dimensional space V, from an implementation point of view (i.e., sensor requirements), it may be desirable to replace $\overline{u}$ in (4.1)–(4.3) with a finite-dimensional approximation, $\overline{u}_n$. To establish a convergence result similar to the one given in Theorem 4.1 above, we require the following additional assumption.

(A6) For each fixed $T > 0$ and for the plant $(\overline{q}, \overline{u})$, there exists $\overline{u}_n \in C(0, T; H^n)$ such that $\overline{u}_n \to \overline{u}$ in $C(0, T; V)$.

THEOREM 4.2. *Assume that assumptions* (A5) *and* (A6) *hold, let* (q^n, u^n) *be the solution to the initial value problem* (4.1)–(4.3) *with* $u^n(0) = u_n(0)$ *and* $q^n(0) = q_n(0)$ *and* $\overline{u}$ *replaced by* $\overline{u}_n$*, and let* (q, u) *be the solution to the initial value problem* (2.13)–(2.15). *Then for each* $T > 0$, $\lim_{n\to\infty} u^n = u$ *in* $L_2(0, T; V)$ *and* $C(0, T; H)$ *and* $\lim_{n\to\infty} q^n = q$ *in* $C(0, T; Q)$.

Proof. Once again, letting $w^n = u^n - u_n$ and $p^n = q^n - q_n$, we show that $\lim_{n\to\infty} w^n = 0$ in $L_2(0, T; V)$ and $C(0, T; H)$ and that $\lim_{n\to\infty} p^n = 0$ in $C(0, T; Q)$. Using (2.13), (2.14), (4.1), (4.2), and (2.3) we obtain the identity

$$
\begin{aligned}
&D_t\{|w^n(t)|^2 + |p^n(t)|_Q^2\} = 2\mathrm{Re}\,\{\langle D_t w^n(t), w^n(t)\rangle + \langle D_t p^n(t), p^n(t)\rangle_Q\}\\
&= 2\mathrm{Re}\,\{-\langle Aw^n(t), w^n(t)\rangle + \langle A\{u(t) - u_n(t)\}, w^n(t)\rangle + \langle A_1(q(t) - q_n(t))\overline{u}_n(t), w^n(t)\rangle\\
&\qquad + \langle A_1(q(t))\overline{u}(t) - \overline{u}_n(t), w^n(t)\rangle + \langle A\{\overline{u}_n(t) - \overline{u}(t)\}, w^n(t)\rangle\\
&\qquad + \langle A_2\{\overline{u}_n(t) - \overline{u}(t)\}, w^n(t)\rangle + \overline{\langle A_1(p^n(t))\overline{u}_n(t), u_n(t) - u(t)\rangle}\\
&\qquad + \overline{\langle A_1(p^n(t))\overline{u}_n(t) - \overline{u}(t), u(t)\rangle} + \overline{\langle A_1(p^n(t))\overline{u}_n(t), \overline{u}(t) - \overline{u}_n(t)\rangle}\\
&\qquad + \overline{\langle A_1(p^n(t))\overline{u}(t) - \overline{u}_n(t), \overline{u}(t)\rangle} + \langle D_t\{u(t) - u_n(t)\}, w^n(t)\rangle\\
&\qquad + \langle D_t\{q(t) - q_n(t)\}, p^n(t)\rangle_Q\}, \qquad \text{a.e. } t > 0.
\end{aligned}
$$

Assumptions (A2), (A3), (A4), and (A6) then yield that

$$
\begin{aligned}
&D_t\{|w^n(t)|^2 + |p^n(t)|_Q^2\}\\
&\le -2\{\beta\|w^n(t)\|^2 + \alpha\|u(t) - u_n(t)\|\|w^n(t)\| + \alpha_1|q(t) - q_n(t)|_Q\|\overline{u}_n(t)\|\|w^n(t)\|\\
&\qquad + \alpha_1|q(t)|_Q\|\overline{u}(t) - \overline{u}_n(t)\|\|w^n(t)\| + \alpha\|\overline{u}_n(t) - \overline{u}(t)\|\|w^n(t)\|\\
&\qquad + \alpha_2\|\overline{u}_n(t) - \overline{u}(t)\|\|w^n(t)\| + \alpha_1|p^n(t)|_Q\|\overline{u}_n(t)\|\|u_n(t) - u(t)\|\\
&\qquad + \alpha_1|p^n(t)|_Q\|\overline{u}_n(t) - \overline{u}(t)\|\|u(t)\| + \alpha_1|p^n(t)|_Q\|\overline{u}_n(t)\|\|\overline{u}(t) - \overline{u}_n(t)\|\\
&\qquad + \alpha_1|p^n(t)|_Q\|\overline{u}(t) - \overline{u}_n(t)\|\|\overline{u}(t)\| + \|D_t\{u(t) - u_n(t)\}\|_*\|w^n(t)\|\\
&\qquad + |D_t\{q(t) - q_n(t)\}|_Q|p^n(t)|_Q\}, \qquad \text{a.e. } t > 0.
\end{aligned}
$$

Applying the inequality (4.4) and gathering like terms, we obtain that

$$
\begin{aligned}
&D_t\{|w^n(t)|^2 + |p^n(t)|_Q^2\}\\
&\le \{-2\beta + 6\epsilon\}\|w^n(t)\|^2 + \left\{\frac{\alpha_1^2}{\epsilon} + \alpha_1^2\|\overline{u}_n(t)\|^2\right\}\|u_n(t) - u(t)\|^2\\
&+ \left\{\frac{\alpha_1^2|q(t)|_Q^2}{\epsilon} + \frac{\alpha^2}{\epsilon} + \alpha_2^2 + \alpha_1^2\{\|u(t)\|^2 + \|\overline{u}_n(t)\|^2 + \|\overline{u}(t)\|^2\}\right\}\|\overline{u}(t) - \overline{u}_n(t)\|^2\\
&+ \frac{\alpha_1^2\|\overline{u}_n(t)\|^2}{\epsilon}|q(t) - q_n(t)|_Q^2 + \frac{1}{\epsilon}\|D_t\{u(t) - u_n(t)\}\|_*^2 + |D_t\{q(t) - q_n(t)\}|_Q^2 + 5|p^n(t)|_Q^2
\end{aligned}
$$

for a.e. $t > 0$. Choosing $\epsilon > 0$ so that $\epsilon < \frac{1}{3}\beta$ and then integrating both sides of the above estimate from 0 to t, we obtain

$$\delta \int_0^t \|w^n(s)\|^2 ds + |w^n(t)|^2 + |p^n(t)|_Q^2 \le z^n(t) + 5\int_0^t |w^n(s)|^2 + |p^n(s)|_Q^2 ds, \tag{4.6}$$

where $\delta = \beta - 3\epsilon > 0$ and

$$\begin{aligned} z^n(t) &= \int_0^t \left\{ \frac{\alpha^2}{\epsilon} + \alpha_1^2 \|\overline{u}_n(s)\|^2 \right\} \|u_n(s) - u(s)\|^2 \\ &+ \left[\frac{\alpha_1^2 |q(s)|_Q^2}{\epsilon} + \frac{\alpha^2}{\epsilon} + \alpha_2^2 + \alpha_1^2 \left(\|u(s)\|^2 + \|\overline{u}_n(s)\|^2 + \|\overline{u}(s)\|^2 \right) \right] \|\overline{u}(s) - \overline{u}_n(s)\|^2 \\ &+ \frac{\alpha_1^2 \|\overline{u}_n(s)\|^2}{\epsilon} |q(s) - q_n(s)|_Q^2 + \frac{\|D_s\{u(s) - u_n(s)\}\|_*^2}{\epsilon} + |D_s\{q(s) - q_n(s)\}|_Q^2 ds. \end{aligned}$$

Assumptions (A5) and (A6) imply that $\lim_{n\to\infty} z^n(t) = 0$, uniformly on $[0, T]$, for each fixed $T > 0$. Consequently, an application of the Gronwall lemma to the estimate (4.6) yields the desired result. □

5. Second-order systems. It is possible to use the framework developed in the previous three sections to identify unknown parameters in certain classes of strongly damped second-order, or abstract hyperbolic, systems on-line. We briefly outline the essential features of the requisite theory below. However, a more general and more versatile treatment of second-order systems can be found in [7] and [28].

Let H_0 be a Hilbert space with inner product $\langle \cdot, \cdot \rangle_0$ and V_0 be a reflexive Banach space with norm denoted by $\| \cdot \|_0$. We assume that V_0 is densely and continuously embedded in H_0. Let Q be a Hilbert space with inner product $\langle \cdot, \cdot \rangle_Q$, and for $i = 1, 2$, let $a^i(\cdot; \cdot, \cdot) : Q \times V_0 \times V_0 \to \mathbf{C}$ be a form satisfying the following assumptions:

(A7) (Q-linearity and symmetry) The map $q \to a^i(q; \cdot, \cdot)$ is linear from Q into the space of conjugate symmetric sesquilinear forms on V_0.

(A8) ($V_0 \times V_0$-boundedness) There exists $\alpha^i > 0$ for which

$$|a^i(q; \varphi, \psi)| \le \alpha^i |q|_Q \|\varphi\|_0 \|\psi\|_0, \quad \varphi, \psi \in V_0, \quad q \in Q.$$

(A9) (V_0-coercivity) There exists a subset $\tilde{Q} \subset Q$ such that for each $q^* \in \tilde{Q}$ there exists $\beta^i(q^*) > 0$ such that $a^i(q^*\varphi, \varphi) \ge \beta^i(q^*) \|\varphi\|_0^2$, $\varphi \in V_0$.

For $q \in Q$ and $i = 1, 2$, let $A^i(q) : V_0 \to V_0^*$ be the linear operator on V_0 determined by the form $a^i(q; \cdot, \cdot)$ via $\langle A^i(q)\varphi, \psi \rangle_0 = a^i(q; \varphi, \psi)$, $\varphi, \psi \in V_0$. Let $w_0 \in V_0$ and $w_1 \in H_0$, let $g \in L_2(0, T; V_0^*)$ for all $T > 0$, and for each $q \in Q$ consider the abstract second-order initial value problem given by

$$D_t^2 w(t) + A^2(q) D_t w(t) + A^1(q) w(t) = g(t), \qquad \text{a.e. } t > 0, \tag{5.1}$$

$$w(0) = w_0, \ \ D_t w(0) = w_1. \tag{5.2}$$

By a solution to the initial value problem (5.1), (5.2) on the interval $[0, T]$ for some $T > 0$, we mean a function $w \in L_2(0, T; V_0)$ with $D_t w \in L_2(0, T; V_0)$ and $D_t^2 w \in L_2(0, T; V_0^*)$ which satisfies (5.1) on a.e. $(0, T)$ as well as (5.2).

To apply the abstract theory developed in sections 2, 3, and 4 above, we effectively rewrite the initial value problem (5.1), (5.2) as an equivalent first-order system. Let

$q^* \in \tilde{Q}$ be fixed but arbitrary and H be the Hilbert space defined by $H = V_0 \times H_0$ with inner product given by

$$\langle \varphi, \psi \rangle = a^1(q^*; \varphi_1, \psi_1) + \langle \varphi_2, \psi_2 \rangle_0 \tag{5.3}$$

for $\varphi = (\varphi_1, \varphi_2)$, $\psi = (\psi_1, \psi_2) \in H$. Let V be the reflexive Banach space defined by $V = V_0 \times V_0$ with norm given by $\|\varphi\| = \left\{ \|\varphi_1\|_0^2 + \|\varphi_2\|_0^2 \right\}^{\frac{1}{2}}$ for $\varphi = (\varphi_1, \varphi_2) \in V$, and for $q \in Q$ define the operator $A_0(q) : V \to V^*$ by $\langle A_0(q)\varphi, \psi \rangle = -a^1(q^*; \varphi_2, \psi_1) + a^1(q; \varphi_1, \psi_2) + a^2(q; \varphi_2, \psi_2)$ for $\varphi = (\varphi_1, \varphi_2) \in V$ and $\psi = (\psi_1, \psi_2) \in V$. Assumptions (A7) and (A9) imply that the expression given in (5.3) is in fact an inner product on H. Assumption (A7) implies that for each $q \in Q$, the operator $A_0(q)$ satisfies assumption (A1) with the operator $A_1(q) : V \to V^*$ given by $\langle A_1(q)\varphi, \psi \rangle = a^1(q; \varphi_1, \psi_2) + a^2(q; \varphi_2, \psi_2)$ and the operator $A_2 : V \to V^*$ in assumption (A1) given by $\langle A_2\varphi, \psi \rangle = -a^1(q^*; \varphi_2, \psi_1)$. Assumption (A8) implies that assumption (A2) is satisfied with $\alpha_1 = 2\max\{\alpha^1, \alpha^2\}$ and $\alpha_2 = \alpha^1|q^*|_Q$. For any $\lambda > 0$, defining $A \in \mathcal{L}(V, V^*)$ by $\langle A\varphi, \psi \rangle = \langle A_0(q^*)\varphi + \lambda\varphi, \psi \rangle$, $\varphi, \psi \in V$, assumption (A8) implies that assumption (A3) holds, and assumption (A9) implies that assumption (A4) holds with $\beta = \min\{\lambda\beta^1(q^*), \beta^2(q^*)\}$.

For $\varphi \in V$, defining the operator $B(\varphi) : V \to Q$ as it was in (2.3) and setting $u_0 = (w_0, w_1) \in H$ and $f = (0, g) \in L_2(0, T; V^*)$, the second-order system (5.1), (5.2) and the first-order system (2.5), (2.6) are considered to be equivalent to $u \sim (w, D_t w)$.

For $\varphi = (\varphi_1, \varphi_2) \in V$, the operator $B(\varphi) : V \to Q$ defined in (2.3) is given by

$$\langle B(\varphi)\psi, q \rangle_Q = \overline{\langle A_1(q)\varphi, \psi \rangle} = \overline{a^1(q; \varphi_1, \psi_2)} + \overline{a^2(q; \varphi_2, \psi_2)}$$

for $q \in Q$ and $\psi = (\psi_1, \psi_2) \in V$. It then follows from Definition 2.1 that a pair $(\overline{q}, \overline{w})$ with $\overline{q} \in Q$ and $\overline{w}$ a solution to the initial value problem (5.1), (5.2) with $q = \overline{q}$ is a *plant* if there exists a constant $\gamma_0 = \gamma_0(\overline{w})$ such that

$$|\langle A^1(q)\overline{w}(t), \varphi \rangle_0 + \langle A^2(q) D_t \overline{w}(t), \varphi \rangle_0| \leq \gamma_0(\overline{w}) |q|_Q \|\varphi\|_0, \qquad \text{a.e. } t > 0,\ \varphi \in V_0.$$

It also follows from Definition 3.3 that the condition for a plant, $(\overline{q}, \overline{w})$, to be *persistently excited* is for there to exist T_0, δ_0, $\epsilon_0 > 0$ such that for each $q \in Q$ with $|q|_Q = 1$ and each $t > 0$ sufficiently large, there exists a $\tilde{t} \in [t, t + T_0]$ such that

$$\left\| \int_{\tilde{t}}^{\tilde{t}+\delta_0} A^1(q)\overline{w}(\tau) + A^2(q) D_\tau \overline{w}(\tau) d\tau \right\|_{V_0^*} \geq \epsilon_0.$$

The convergence results given in section 3 take the form $\lim_{t\to\infty} \|u_1(t) - \overline{w}(t)\|_0 = 0$, $\lim_{t\to\infty} |u_2(t) - D_t\overline{w}(t)|_0 = 0$, and if, in addition, the plant $(\overline{q}, \overline{w})$ is persistently excited, then $\lim_{t\to\infty} |q(t) - \overline{q}|_Q = 0$, where (q, u) with $u = (u_1, u_2)$ is the solution to the initial value problem (2.13)–(2.15).

It is also possible to restate our *partial persistence of excitation* and *partial* parameter convergence results, Theorem 3.7 and Corollary 3.8, in the context of second-order systems. In particular, note that the set $\hat{Q}$ takes the form

$$\hat{Q} = \left\{ q \in Q : \lim_{t\to\infty} \left| \int_t^{t+L} \langle A^1(q)\overline{w}(\tau) + A^2(q) D_\tau \overline{w}(\tau), \varphi \rangle_0 d\tau \right| = 0,\ \varphi \in V_0,\ L > 0 \right\}.$$

Finally we note that the appropriate modifications to the approximation theory presented in section 4 (i.e., assumptions (A5) and (A6) and Theorems 4.1 and 4.2), required to restate it in the context of second-order systems, should also be immediately clear.

6. Examples and numerical results. In this section we present and discuss a number of examples illustrating the application of the on-line estimation theory which was developed in the previous sections. We consider the estimation of both constant and functional (i.e., spatially varying) parameters in one-dimensional heat or diffusion equations, the estimation of constant damping and stiffness parameters in a one-dimensional wave equation with Kelvin–Voigt viscoelastic damping, and the estimation of the nonlinearity in a one-dimensional quasi-linear heat equation in which the thermal diffusivity is a function of the temperature gradient. The numerical studies for each example presented below were carried out via simulation of the plant. We also did not attempt to construct input signals which necessarily resulted in a persistently excited plant. Our concern here was to simply illustrate the feasibility of our approach. A detailed and complete numerical study of persistence of excitation and its effect on convergence has been carried out and is reported on elsewhere (see [8]). For simplicity, in the examples to follow, we have chosen all of the Hilbert and Banach spaces, H, V, and Q, to be real.

All of the computations described below were carried out on either a SUN SPARC-system 600 or a SPARCstation 10 in the Department of Mathematics at the University of Southern California. The finite-dimensional estimator equations, (4.1)–(4.3), were integrated using the stiff ODE solver from the Numerical Algorithms Group (NAG) Library, routine D02NBF. All required integrals were computed numerically via a composite two-point Gauss–Legendre quadrature rule.

6.1. Example 1. We consider the estimation of the parameters q_1, q_2, and q_3 in the one-dimensional heat or diffusion equation (with convective or advective and decay or growth terms) given by

$$\frac{\partial u}{\partial t}(t,x) = q_1 \frac{\partial^2 u}{\partial x^2}(t,x) - q_2 \frac{\partial u}{\partial x}(t,x) - q_3 u(t,x) + f(t,x), \quad t > 0,\ 0 < x < 1,$$

together with the Dirichlet boundary conditions $u(t,0) = 0 = u(t,1)$, $t > 0$. In this case we have $H = L_2(0,1)$ and $V = H_0^1(0,1)$ endowed with the usual inner products and corresponding induced norms. The embedding constant is $K = \pi^{-1}$ (see, for example, [29]). We take $Q = \mathbf{R}^3$ endowed with the weighted Euclidean inner product $\langle q, p\rangle = q^T \Omega p$, q, $p \in \mathbf{R}^3$, where Ω is the 3×3 diagonal matrix given by $\Omega = \text{Diagonal}(\omega_1, \omega_2, \omega_3)$, with $\omega_1, \omega_2, \omega_3 > 0$. We note that the weights ω_1, ω_2, and ω_3 serve as so-called *adaptive gains* or *tuning parameters* in the estimator. For $q \in Q$, the operator $A_0(q) = A_1(q) \in \mathcal{L}(V, V^*)$ is given by

$$\langle A_0(q)\varphi, \psi\rangle = q_1 \int_0^1 D\varphi(x) D\psi(x)dx + q_2 \int_0^1 D\varphi(x)\psi(x)dx + q_3 \int_0^1 \varphi(x)\psi(x)dx, \quad \varphi, \psi \in H_0^1(0,1).$$

It is easily verified that assumptions (A1) and (A2) are satisfied.

For the estimator dynamics, $A \in \mathcal{L}(V, V^*)$, we set $A = A_0(q^*)$ for an *appropriate* choice of $q^* \in Q$. It is immediately clear that $q^* \in Q$ can be chosen so that assumptions (A3) and (A4) hold. For example, let $q^* = (q_1^*, q_2^*, q_3^*)$, with $q_2^* = 0$, $q_1^* > 0$, and $q_3^* \geq 0$.

We approximate using linear B-splines. For $n = 1, 2, \ldots$, let $\{\varphi_j^n\}_{j=0}^n$ be the standard linear B-splines on the interval $[0,1]$ defined with respect to the uniform

mesh $\{0, \frac{1}{n}, \frac{2}{n}, \ldots, 1\}$. That is, for $i = 0, 1, 2, \ldots, n$

$$\varphi_i^n(x) = \begin{cases} 1 - |nx - i|, & x \in [\frac{i-1}{n}, \frac{i+1}{n}], \\ 0 & \text{elsewhere on } [0,1]. \end{cases} \tag{6.1}$$

Set $H^n = \operatorname{span}\{\varphi_j^n\}_{j=1}^{n-1}$. Since Q is finite dimensional, we simply set $Q^n = Q$, $n = 1, 2, \ldots$. For each $n = 1, 2, \ldots$, let P^n denote the orthogonal projection of H onto H^n, and setting $u_n = P^n u$, standard approximation results for spline functions (see [29]) can be used to establish that assumption (A5) is satisfied. For $n = 1, 2, \ldots$, let P_n denote the orthogonal projection of $V = H_0^1(0,1)$ onto H^n (with respect to the standard H_0^1 inner product), and set $\overline{u}_n = P_n \overline{u}$. If $\overline{u}$ is sufficiently smooth, it is not difficult to establish that assumption (A6) is satisfied as well. Thus the conclusions of Theorem 4.1 and Theorem 4.2 hold.

There is a practical advantage to using the V projection, P_n, to finite dimensionalize the plant. Indeed, if

$$\overline{u}_n(t) = P_n \overline{u}(t) = \sum_{j=1}^{n-1} \overline{U}_n(t)_j \varphi_j^n \tag{6.2}$$

(i.e., let $\overline{U}_n(t) \in \mathbf{R}^{n-1}$ be the coordinate vector for $\overline{u}_n(t)$ with respect to the basis $\{\varphi_j^n\}_{j=1}^{n-1}$), then

$$\overline{U}_n(t) = (K^n)^{-1} h_n(\overline{u}(t)), \tag{6.3}$$

where for $\varphi \in V$, $h_n(\varphi) \in \mathbf{R}^{n-1}$ is given by $h_n(\varphi)_j = \int_0^1 D\varphi(x) D\varphi_j^n(x)dx$, $j = 1, 2, \ldots, n-1$, and $K^n \in \mathbf{R}^{(n-1)\times(n-1)}$ is given by

$$K^n = \left[K_{ij}^n\right] = \left[\int_0^1 D\varphi_i^n(x) D\varphi_j^n(x)dx\right] = n \begin{bmatrix} 2 & -1 & 0 & & & 0 \\ -1 & 2 & \cdot & & & \\ 0 & \cdot & \cdot & \cdot & & \\ & & \cdot & \cdot & \cdot & 0 \\ & & & \cdot & \cdot & -1 \\ 0 & & & 0 & -1 & 2 \end{bmatrix}. \tag{6.4}$$

It is easily verified that for $\varphi \in V$

$$h_n(\varphi)_j = -n\Delta_{\frac{1}{n}}^2 \varphi(\frac{j-1}{n}) = -n\left\{\varphi\left(\frac{j+1}{n}\right) - 2\varphi\left(\frac{j}{n}\right) + \varphi\left(\frac{j-1}{n}\right)\right\}, \quad j = 1, \ldots, n-1.$$

Thus the approximating estimator (i.e., (4.1)–(4.3) with $\overline{u}$ replaced by $\overline{u}_n$) does not require spatially distributed data. For a given value of n, $\overline{u}$ need only be spatially sampled at the $n-1$ nodal points $\frac{1}{n}, \frac{2}{n}, \ldots, \frac{n-1}{n}$.

Let $U^n(t) \in \mathbf{R}^{n-1}$ be the coordinate vector for $u^n(t)$. That is,

$$u^n(t) = \sum_{j=1}^{n-1} U^n(t)_j \varphi_j^n. \tag{6.5}$$

Let M^n denote the Gram matrix corresponding to the basis $\{\varphi_j^n\}_{j=1}^{n-1}$. We have

$$M^n = \left[M_{ij}^n\right] = \left[\int_0^1 \varphi_i^n(x)\varphi_j^n(x)dx\right] = \frac{1}{6n} \begin{bmatrix} 4 & 1 & 0 & & & 0 \\ 1 & 4 & 1 & & & \\ 0 & \cdot & \cdot & \cdot & & \\ & & \cdot & \cdot & \cdot & 0 \\ & & & \cdot & \cdot & 1 \\ 0 & & & 0 & 1 & 4 \end{bmatrix}. \tag{6.6}$$

Also, let L^n be the $(n-1)\times(n-1)$ matrix given by

$$L^n = \left[L^n_{ij}\right] = \left[\int_0^1 \varphi_i^n(x) D\varphi_j^n(x)dx\right] = \frac{1}{2}\begin{bmatrix} 0 & 1 & 0 & & & 0 \\ -1 & 0 & 1 & & & \\ 0 & \cdot & \cdot & \cdot & & \\ & & \cdot & \cdot & \cdot & 0 \\ & & & \cdot & \cdot & 1 \\ 0 & & & 0 & -1 & 0 \end{bmatrix}.$$

The matrix form of the approximating estimator ((4.1)–(4.3) with $\overline{u}$ replaced by $\overline{u}_n$) is then given by

$$\begin{aligned} M^n\dot{U}^n(t) &+ q_1^* K^n U^n(t) + q_2^* L^n U^n(t) + q_3^* M^n U^n(t) \\ &+ q_1(t)K^n\overline{U}_n(t) + q_2(t)L^n\overline{U}_n(t) + q_3(t)M^n\overline{U}_n(t) \\ &= q_1^* K^n\overline{U}_n(t) + q_2^* L^n\overline{U}_n(t) + q_3^* M^n\overline{U}_n(t) + F^n(t), \quad t>0, \\ \omega_1\dot{q}_1^n(t) &+ \overline{U}_n(t)^T K^n\left\{\overline{U}_n(t) - U^n(t)\right\} = 0, \qquad t>0, \\ \omega_2\dot{q}_1^n(t) &+ \overline{U}_n(t)^T L^n\left\{\overline{U}_n(t) - U^n(t)\right\} = 0, \qquad t>0, \\ \omega_3\dot{q}_1^n(t) &+ \overline{U}_n(t)^T M^n\left\{\overline{U}_n(t) - U^n(t)\right\} = 0, \qquad t>0, \end{aligned}$$

where for $t>0$

$$F^n(t)_j = \int_0^1 f(t,x)\varphi_j^n(x)dx, \qquad j=1,2,\ldots,n-1. \tag{6.7}$$

To carry out our numerical studies, we set $\overline{q} = (\overline{q}_1,\overline{q}_2,\overline{q}_3)^T = (.1,0,.8)^T$ and let

$$f(t,x) = \left\{\sin(4\pi t) + 10^{-3}t^2\right\}\chi_{[.215,.315]}(x), \quad t>0, \quad 0<x<1, \tag{6.8}$$

where $\chi_{[a,b]}$ denotes the characteristic function corresponding to the interval $[a,b]$. We assume that the plant was initially at rest (i.e., $u_0=0$). Since $\overline{q}_2=0$, the mode shapes of the plant are $\varphi_j(x)=\sin(j\pi x)$, $j=1,2,\ldots$. To simulate the plant we used an N-dimensional truncated modal model with $N=65$. We set $q_1^*=.01$, $q_2^*=0$, and $q_3^*=0$. We also set $\omega_1=\omega_2=\omega_3=1.0$. We took the state estimate to be initially at rest and set $q_1^n(0)=0$, $q_2^n(0)=.25$, and $q_3^n(0)=-.15$, for all n. We integrated the estimator from $t=0$ to $t=100$. Our results for $n=4,8,16$, and 32 are plotted in Figures 6.1, 6.2, and 6.3. The estimates for q_1 are plotted in Figure 6.1; for q_2, in Figure 6.2; and for q_3, in Figure 6.3. The dotted line in each of the figures is the value for $\overline{q}_1$, $\overline{q}_2$, or $\overline{q}_3$.

It is clear from the figures that the asymptotic limits (with respect to time as opposed to n) of the approximating estimates for the unknown parameters approach the true values of the parameters as n increases. However, it is worth noting that reasonably good estimates are obtained for rather low values of n. This is valuable from a practical point of view. Indeed, the implication is that fewer data are required and that the estimator will be of relatively low dimension. Consequently fewer sensors are required, and the approximating estimator can be integrated more rapidly.

The oscillations which appear in the trajectories of the parameter estimates are a result of the relative levels of input excitation (i.e., f) and dissipation (i.e., q^*). The other tuning parameters (i.e., ω_i, $i=1,2,3$) also play a role in either amplifying or attenuating these oscillations. For an analysis and numerical study of these phenomena, see [8].

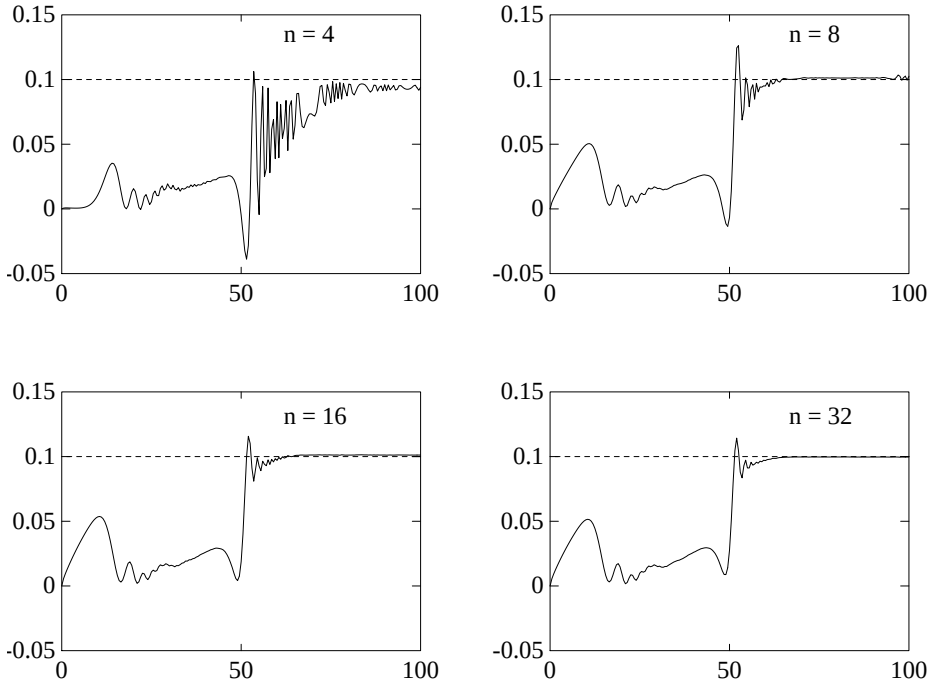


FIG. 6.1. *Estimates for* $\overline{q}_1$ *for* $n = 4, 8, 16, 32$.

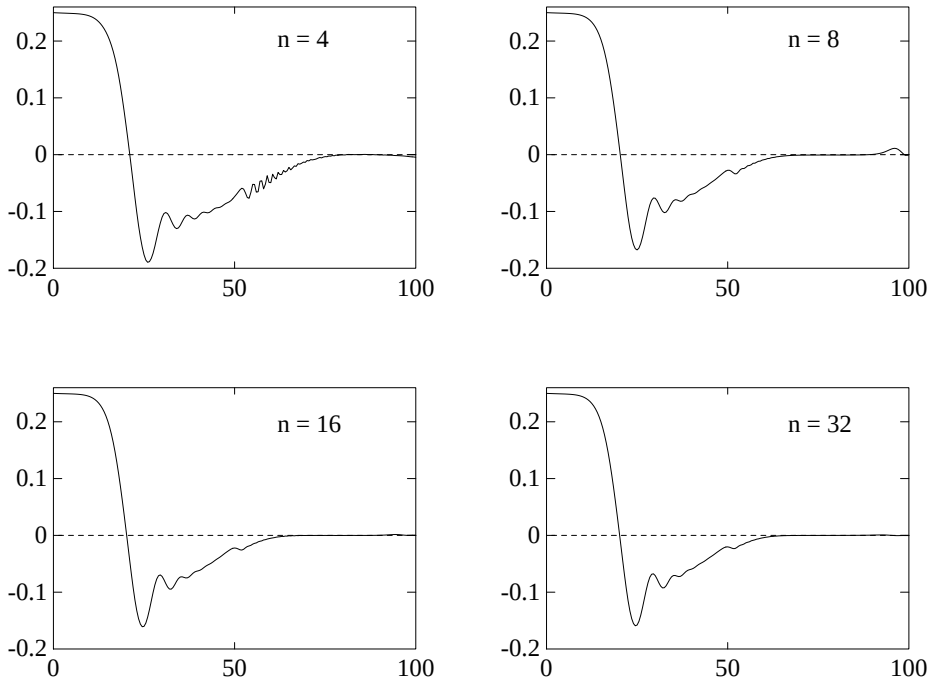


FIG. 6.2. *Estimates for* $\overline{q}_2$ *for* $n = 4, 8, 16, 32$.

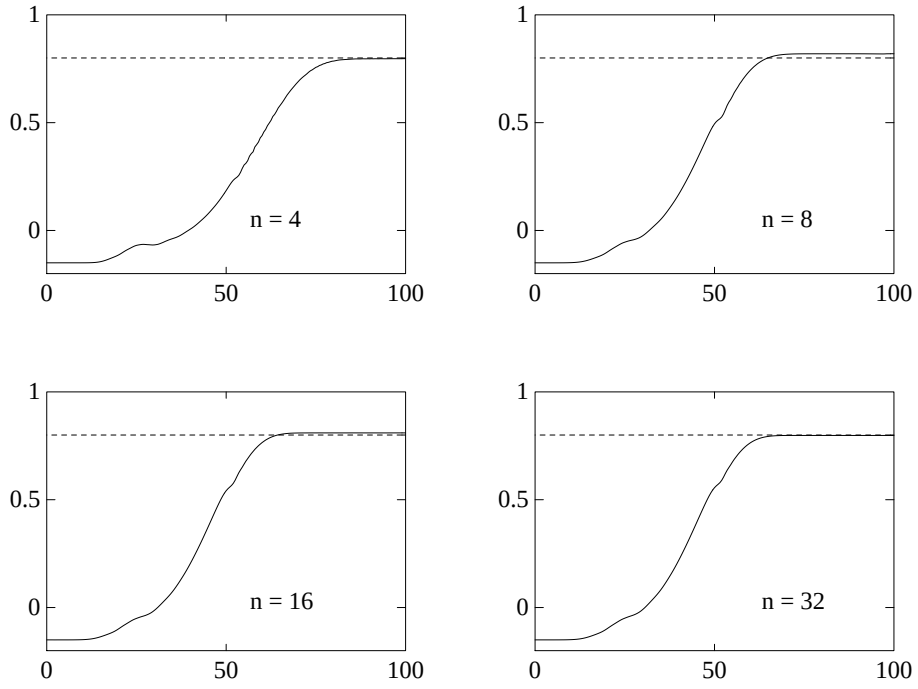


FIG. 6.3. *Estimates for $\overline{q}_3$ for $n = 4, 8, 16, 32$.*

6.2. Example 2. In this example we consider the estimation of the functional parameter q in the one-dimensional heat or diffusion equation

$$\frac{\partial u}{\partial t}(t,x) = \frac{\partial}{\partial x}\left\{q(x)\frac{\partial u}{\partial x}(t,x)\right\} + f(t,x), \quad t > 0, \quad 0 < x < 1,$$

together with the Dirichlet boundary conditions $u(t,0) = 0 = u(t,1)$, $t > 0$. Once again we set $H = L_2(0,1)$ and $V = H_0^1(0,1)$, each endowed with its usual inner product and corresponding induced norm. We let $Q = H^1(0,1)$ and take it to be endowed with the weighted inner product

$$\langle q, p\rangle_Q = \omega_1 \int_0^1 q(x)p(x)dx + \omega_2 \int_0^1 Dq(x)Dp(x)dx, \qquad p, q \in H^1(0,1),$$

where the weights ω_1 and ω_2 are assumed to be positive. When $\omega_2 = 0$, it is equivalent to taking $Q = L_2(0,1)$. For $q \in Q$, the operator $A_0(q) = A_1(q) \in \mathcal{L}(V, V^*)$ is given by

$$\langle A_0(q)\varphi, \psi\rangle = \int_0^1 q(x)D\varphi(x)D\psi(x)dx, \qquad \varphi, \psi \in H^1(0,1).$$

It is easily verified that assumptions (A1) and (A2) are satisfied.

Once again we choose the estimator dynamics $A \in \mathcal{L}(V, V^*)$, to be $A = A_0(q^*)$, for an appropriate choice of $q^* \in Q$. In particular for $x \in [0,1]$, we let $q^*(x) = q^* > 0$ (i.e., a constant function). For such a $q^* \in Q$, assumptions (A3) and (A4) are satisfied.

For $n = 1, 2, \ldots$, we choose the approximating subspaces for the state estimator, H^n, as they were in the previous example. We also use linear B-splines to discretize

the parameter space Q. For each $m = 1, 2, \ldots$, we set $Q^m = \text{span}\{\varphi_j^m\}_{j=0}^m$, where the linear spline basis, $\{\varphi_j^m\}_{j=0}^m$ is given by (6.1) with n replaced by m. Note that dim $H^n = n-1$ and dim $Q^m = m+1$. Consequently the dimension of the approximating estimator is $n-1+m+1 = n+m$.

Once again standard approximation results for linear splines yield that assumption (A5) is satisfied and consequently that the conclusion of Theorem 4.1 holds. If $\omega_2 > 0$, it is also easily verified that assumption (A6) is satisfied and therefore that the conclusion of Theorem 4.2 holds as well.

Define the family of $(n-1)\times(n-1)$ matrices $\{K_k^{n,m}\}_{k=0}^m$ by

$$K_k^{n,m} = [K_k^{n,m}]_{ij} = \int_0^1 \varphi_k^m(x) D\varphi_i^n(x) D\varphi_j^n(x)dx,$$
$$i,j = 0,1,\ldots,n-1, \qquad k = 0,1,\ldots,m.$$

It then follows that the matrix K^n given in (6.4) is given by $K^n = \sum_{k=0}^m K_k^{n,m}$. Let $p^m(t) = [p_0^m(t),\ldots,p_m^m(t)]^T \in \mathbf{R}^{m+1}$ denote the coordinate vector for the approximating estimate $q^m(t)$ with respect to the basis $\{\varphi_j^m\}_{j=0}^m$. That is, $q^m(t) = \sum_{j=0}^m p_j^m(t)\varphi_j^m$, $t > 0$. Taking, for simplicity, the tuning parameter $q^*(x) = q^*$, $x \in [0,1]$, to be constant, the matrix form of the approximating estimator (4.1)–(4.3) with $\overline{u}$ replaced by $\overline{u}_n$ ($\overline{u}_n$ as it was defined in the previous example) is given by

$$M^n\dot{U}^n(t) + q^*K^nU^n(t) + \sum_{k=0}^m p_k^m(t)K_k^{n,m}\overline{U}_n(t) = F^n(t) + q^*K^n\overline{U}_n(t), \qquad t > 0,$$

$$[\Omega^m\dot{p}^m(t)]_k + \overline{U}_n(t)^TK_k^{n,m}\left\{\overline{U}_n(t) - U^n(t)\right\} = 0, \qquad t > 0,\ k = 0,1,2,\ldots,m,$$

where the matrices M^n and K^n are given by (6.6) and (6.4), respectively, U^n is as defined in (6.5), $\overline{U}_n$ is as defined in (6.2) and (6.3), F^n is as given in (6.7), and the $(m+1)\times(m+1)$ matrix Ω^m is given by

$$\Omega^m = [\Omega^m]_{ij} = \frac{\omega_1}{6m}\begin{bmatrix} 2 & 1 & 0 & & & 0 \\ 1 & 4 & \cdot & & & \\ 0 & \cdot & \cdot & \cdot & & \\ & & \cdot & \cdot & \cdot & 0 \\ & & & \cdot & 4 & 1 \\ 0 & & & 0 & 1 & 2 \end{bmatrix} + \omega_2 m \begin{bmatrix} 1 & -1 & 0 & & & 0 \\ -1 & 2 & \cdot & & & \\ 0 & \cdot & \cdot & \cdot & & \\ & & \cdot & \cdot & \cdot & 0 \\ & & & \cdot & 2 & -1 \\ 0 & & & 0 & -1 & 1 \end{bmatrix}.$$

We set $\overline{q}(x) = .1 - .05\sin\{2\pi(x-.25)\}$, $0 < x < 1$, and took f to be as defined in (6.8). We assumed that the plant was initially at rest. To simulate the plant, we used a finite difference–based integrator for parabolic systems from the NAG Library, routine D03PAF. We set $q^* = .01$, $\omega_1 = 1$, and $\omega_2 = 0$. We note that strictly speaking assumption (A6) is not satisfied when $\omega_2 = 0$ (or equivalently when $Q = L_2(0,1)$). Therefore, in this case, Theorem 4.2 does not, in fact, apply. But nevertheless, we were still able to achieve convergence using approximating plant data.

We took the state estimator to be initially at rest and set $q(0,x) = .1$, $0 < x < 1$. In Figure 6.4 we plot estimate time trajectories for q for various values of n and m. The function $\overline{q}$ has also been plotted on the same sets of axes with a solid line. (The initial guess, $q(0,\cdot)$, is also plotted with a solid line.) Once again, reasonably accurate estimates are obtained with relatively low values of n and m. In Figure 6.5 we plot the approximating estimates of $\overline{q}$ at time $t = 100$ for $n = 64$ and $m = 8, 16, 24,$

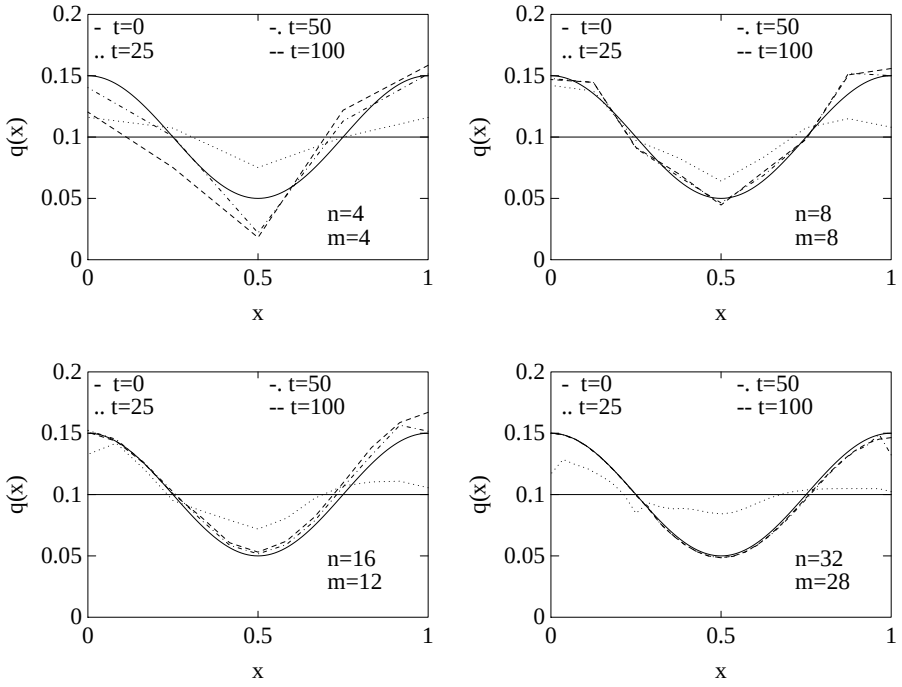


FIG. 6.4. *Estimate trajectories for $\overline{q}$ for various values of n and m.*

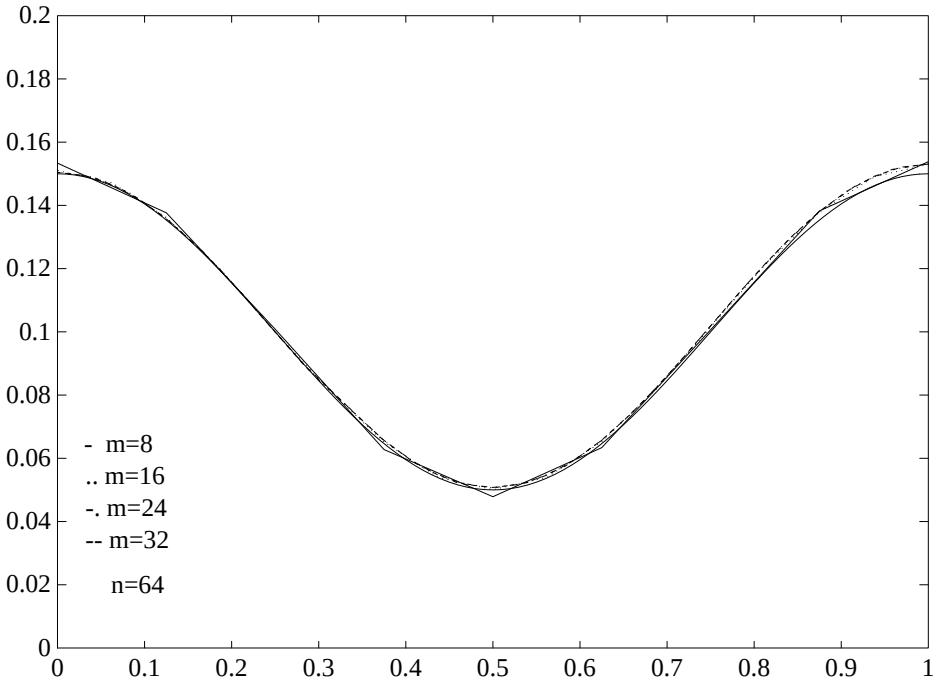


FIG. 6.5. *Estimate for $\overline{q}$ at time $t = 100$ with $n = 64$, $m = 8, 16, 24, 32$.*

and 32. Note that in the last case the dimension of the approximating estimator is $(n-1)+(m+1) = (64-1)+(32+1) = 96$. It is worth noting the high degree of stability exhibited by the scheme. Indeed, the problem of estimating functional coefficients in partial differential equations is well known to be, in general, ill posed. (see, for example, [2]). The instability usually becomes apparent as the value of m increases. For large values of n, we observed no evidence of instability until the value of m started to approach the value of n. We checked this for values of n as large as 64.

6.3. Example 3. In this example we consider the simultaneous estimation of constant stiffness and damping parameters in the one-dimensional wave equation with Kelvin–Voigt viscoelastic damping given by

$$\frac{\partial^2 w}{\partial t^2}(t,x) - q_2 \frac{\partial^2}{\partial x^2}\frac{\partial w}{\partial t}(t,x) - q_1 \frac{\partial^2 w}{\partial x^2} = g(t,x), \qquad t>0,\ 0<x<1, \tag{6.9}$$

with the Dirichlet (fixed endpoint) boundary conditions

$$w(t,0) = 0 = w(t,1), \qquad t>0. \tag{6.10}$$

Applying the theory developed in section 5, we set $H_0 = L_2(0,1)$ and $V_0 = H_0^1(0,1)$, each endowed with its respective usual inner product and corresponding induced norm. We let $Q = \mathbf{R}^2$ with the weighted inner product given by $\langle q,p\rangle = q^T\Omega p$, $q,p \in \mathbf{R}^2$, where Ω is the 2×2 diagonal matrix given by $\Omega = \text{Diagonal}(\omega_1,\omega_2)$, with $\omega_1,\omega_2>0$.

For $q=(q_1,q_2)^T \in Q$ and $i=1,2$, we define the forms $a^i(q;\cdot,\cdot): V_0\times V_0 \to R$ by

$$a^i(q;\varphi,\psi) = q_i \int_0^1 D\varphi(x) D\psi(x)dx, \qquad \varphi,\psi \in H_0^1(0,1).$$

Once again it is easily verified that assumptions (A7)–(A9) are satisfied for $i=1,2$, with the subset $\tilde{Q}$ of Q being the positive orthant of $\mathbf{R}^2$. That is, $\tilde{Q} = \{(q_1^*,q_2^*) \in \mathbf{R}^2 : q_i^*>0, i=1,2\}$.

We again approximate the state estimator using the linear spline basis given by (6.1). Let $H_0^n = \text{span}\{\varphi_j^n\}_{j=1}^{n-1}$, and set $u_1^n(t) = \sum_{j=1}^{n-1} U_1^n(t)_j \varphi_j^n$, $t\ge 0$, and $u_2^n(t) = \sum_{j=1}^{n-1} U_2^n(t)_j\varphi_j^n$, $t\ge 0$. Furthermore, let $\overline{W}_n(t) \in \mathbf{R}^{n-1}$ be the coordinate vector with respect to the basis $\{\varphi_j^n\}_{j=1}^{n-1}$ for the approximating plant $\overline{w}_n(t) = P_n\overline{w}(t)$, where P_n is the orthogonal projection of V_0 onto H_0^n with respect to $V_0 = H_0^1(0,1)$ inner product. The matrix form of the corresponding estimator is then given by

$$q_1^* K^n \dot{U}_1^n(t) - q_1^* K^n U_2^n(t) + \lambda q_1^* K^n U_1^n(t) = \lambda q_1^* K^n \overline{W}_n(t), \qquad t>0,$$

$$\begin{aligned} M^n\dot{U}_2^n(t) &+ q_2^* K^n U_2^n(t) + q_1^* K^n U_1^n(t) + \lambda M^n U_2^n(t) \\ &\quad + q_2^n(t) K^n \dot{\overline{W}}_n(t) + q_1^n(t) K^n \overline{W}_n(t) \\ &= q_2^* K^n \dot{\overline{W}}_n(t) + q_1^* K^n \overline{W}_n(t) + \lambda M^n \dot{\overline{W}}_n(t) + G^n(t), \qquad t>0, \end{aligned}$$

$$\omega_1 \dot{q}_1^n(t) + \overline{W}(t)^T K^n \left(\dot{\overline{W}}_n(t) - U_2^n(t)\right) = 0, \qquad t>0,$$

$$\omega_2 \dot{q}_1^n(t) + \dot{\overline{W}}(t)^T K^n \left(\dot{\overline{W}}_n(t) - U_2^n(t)\right) = 0, \qquad t>0,$$

where the matrices M^n and K^n are given by (6.6) and (6.4), respectively, $q_1^*, q_2^*, \lambda > 0$, and G^n is given by $G^n(t)_j = \int_0^1 g(t,x)\varphi_j^n(x)dx$, $j=1,2,\ldots,n-1$.

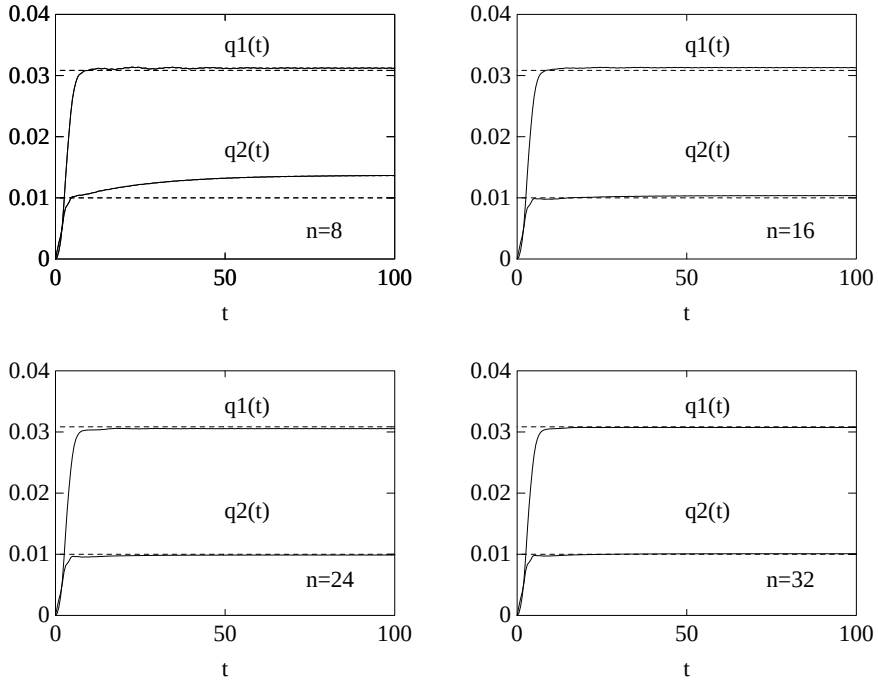


FIG. 6.6. *Estimator trajectories for* $\overline{q}_1$ *and* $\overline{q}_2$ *for* $n = 8, 16, 24,$ *and* 32.

To generate the numerical results that we present below, we set $\overline{q} = (\overline{q}_1, \overline{q}_2)^T = (.0308, .01)^T$, and let $g(t,x) = \{4\sin(4\pi t) + \cos(\pi t) + 2\}\,\chi_{[.215,.315]}(x)$, $t > 0$, $0 < x < 1$. We assumed that the plant was initially at rest and used the IMSL routine DMOLCH (a cubic Hermite polynomial method of lines solver for systems of partial differential equations) to integrate (6.9), (6.10) (together with zero initial data) with $q_i = \overline{q}_i$, $i = 1, 2$, to obtain $\overline{w}(t)$ and $\dot{\overline{w}}(t)$, for $t > 0$. We set $q_1^* = 2 \times 10^{-4}$, $q_2^* = .5$, $\omega_1 = \omega_2 = 53.334$, and $\lambda = 1$. We took the state estimator to be initially at rest and set $q_1^n(0) = q_2^n(0) = 0$ for all n. We integrated the estimator from $t = 0$ to $t = 100$. Our results for $n = 8, 16, 24,$ and 32 are plotted in Figure 6.6. The true values of the parameters, $\overline{q}_1$ and $\overline{q}_2$, are also plotted on the same axes with a dashed line. In Figure 6.7 we plot the Euclidean norm of the parameter error, $|r^n(t)| = \left\{(q_1^n(t) - \overline{q}_1)^2 + (q_2^n(t) - \overline{q}_2)^2\right\}^{\frac{1}{2}}$, from $t = 0$ to $t = 100$, for $n = 8, 16, 24,$ and 32. That convergence is achieved is immediately clear.

6.4. Example 4. In this example we consider the estimation of the thermal conductivity in a one-dimensional nonlinear (strictly speaking, quasi-linear) heat equation. More precisely, we consider the identification of the thermal conductivity, q, in the one-dimensional quasi-linear heat equation

$$\frac{\partial u}{\partial t}(t,x) - \frac{\partial}{\partial x}\left\{q\left(\left|\frac{\partial u}{\partial x}(t,x)\right|\right)\frac{\partial u}{\partial x}(t,x)\right\} = f(t,x), \qquad 0 < x < 1,\ t > 0, \tag{6.11}$$

together with the Dirichlet boundary conditions

$$u(t,0) = 0 \qquad \text{and} \qquad u(t,1) = 0, \qquad t > 0, \tag{6.12}$$

and the initial conditions

$$u(0,x) = u_0(x), \qquad 0 \le x \le 1. \tag{6.13}$$

We assume that $u_0 \in L_2(0,1)$ and $f(t,\cdot) \in L_2(0,1)$ for $t \ge 0$.

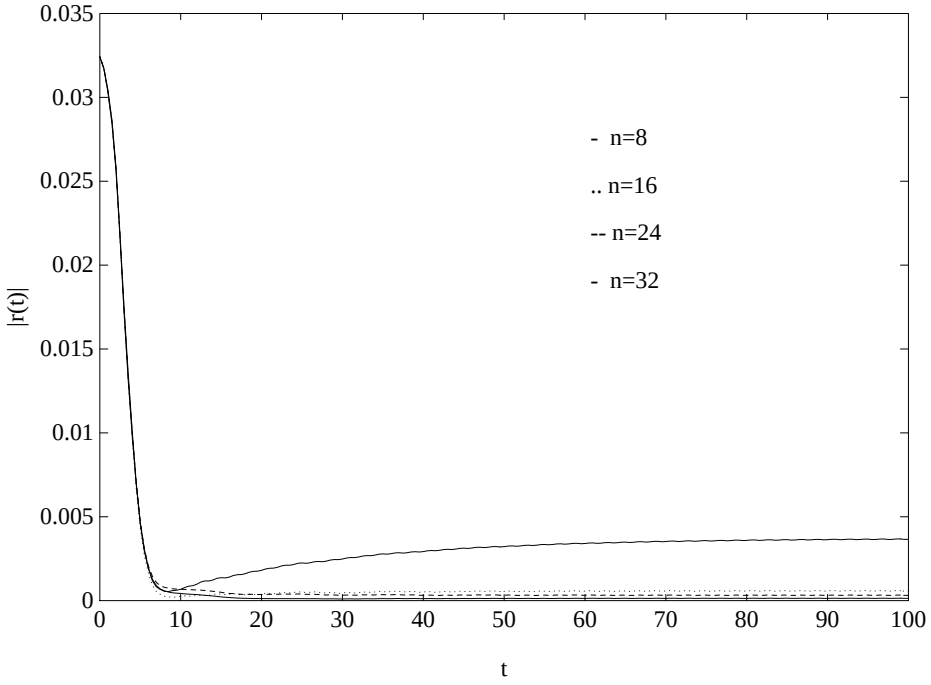


FIG. 6.7. *Euclidean norm of the parameter error for* $n = 8, 16, 24,$ *and* 32.

Let $H = L_2(0,1)$ be endowed with the standard inner product, let $V = H_0^1(0,1)$ be endowed with the usual norm, $\|\varphi\| = \{\int_0^1 |D\varphi(x)|^2 dx\}^{\frac{1}{2}}$, $\varphi \in H_0^1(0,1)$, and define the Hilbert space Q as follows. Let $\hat{Q} = \{\varphi : \varphi \in H^1_{Loc}(\mathbf{R}^+)$ and $\varphi, D\varphi \in L_\infty(\mathbf{R}^+)\}$. Define the inner product, $\langle \cdot, \cdot \rangle_Q$, on $\hat{Q}$ by

$$\langle \varphi, \psi \rangle_Q = \int_0^\infty \omega_0(\theta)\varphi(\theta)\psi(\theta)d\theta + \int_0^\infty \omega_1(\theta)D\varphi(\theta)D\psi(\theta)d\theta, \quad \varphi, \psi \in \hat{Q}, \tag{6.14}$$

where $\omega_0, \omega_1 \in L_1(\mathbf{R}^+)$ are positive weighting functions. Let $|\cdot|_Q$ denote the norm induced by the inner product given in (6.14), and define the Hilbert space Q to be the completion of the inner product space $\{\hat{Q}, \langle \cdot, \cdot \rangle_Q, |\cdot|_Q\}$. For $q \in Q$, let $A_0(q) : V \to V^*$ be given by

$$\langle A_0(q)\varphi, \psi \rangle = \int_0^1 q(|D\varphi(x)|)D\varphi(x)D\psi(x)dx, \qquad \varphi, \psi \in V.$$

It is not difficult to verify that assumptions (A1) and (A2) are satisfied.

For our estimator dynamics, we use a linear constant coefficient heat conduction operator with Dirichlet boundary conditions. That is, we define $A \in \mathcal{L}(V, V^*)$ by

$$\langle A\varphi, \psi \rangle = \alpha \int_0^1 D\varphi(x)D\psi(x)dx, \qquad \varphi, \psi \in V,$$

where $\alpha > 0$. It follows that for $\varphi, \psi \in V$ we have $|\langle A\varphi, \psi \rangle| \leq \alpha\|\varphi\|\|\psi\|$, and $\langle A\varphi, \varphi \rangle \geq \beta\|\varphi\|^2$, with $\beta = \alpha$. Consequently, assumptions (A3) and (A4) are satisfied.

We again approximate the state space using linear B-spline functions. Set $H^n = \text{span}\{\varphi_j^n\}_{j=1}^{n-1}$, where for each $n = 2, 3, \ldots$ and $j = 1, 2, \ldots n-1$, φ_j^n is given by (6.1). We again assume that (6.5) holds and that the plant, $\overline{u}$, is discretized as in (6.2).

We also use linear B-splines to discretize the parameter space Q. For each $m = 1, 2, \ldots$, and each $r > 0$, let $\{\hat{\psi}_j^{m,r}\}_{j=0}^m$ be the standard linear B-splines on the interval $[0, r]$ defined with respect to the uniform mesh $\{0, \frac{r}{m}, \frac{2r}{m}, \ldots, r\}$. Let $Q^{m,r} = \text{span}\{\psi_j^{m,r}\}_{j=0}^m$, where

$$\psi_j^{m,r} = \begin{cases} \hat{\psi}_j^{m,r}, & j = 0, 1, 2, \ldots, m-1, \\ \hat{\psi}_m^{m,r} + \chi_{[r,\infty)}, & j = m, \end{cases}$$

with χ_J denoting the characteristic function for the interval J. If we let $P_Q^{m,r}$ denote the orthogonal projection of Q onto $Q^{m,r}$, the requisite strong convergence to the identity which will ensure that assumption (A5) is satisfied can be demonstrated. Consequently, let $\Omega^{m,r}$ denote the $(m+1) \times (m+1)$ Gram matrix corresponding to the basis $\{\psi_j^{m,r}\}_{j=0}^m$. That is,

$$\Omega^{m,r} = [\Omega_{m,r}]_{i,j}$$

$$= \langle \psi_i^{m,r}, \psi_j^{m,r} \rangle_Q = \int_0^\infty \omega_0(\theta) \psi_i^{m,r}(\theta) \psi_j^{m,r}(\theta) d\theta + \int_0^\infty \omega_1(\theta) D\psi_i^{m,r}(\theta) D\psi_j^{m,r}(\theta) d\theta.$$

For $t \geq 0$, let $p^{m,r}(t) = [p^{m,r}(t)_0, \ldots, p^{m,r}(t)_m]^T \in \mathbf{R}^{m+1}$ denote the coordinate vector for the approximating parameter estimate $q^{m,r}(t)$ with respect to the basis $\{\psi_j^{m,r}\}_{j=0}^m$. That is, $q^{m,r}(t) = \sum_{j=0}^m p^{m,r}(t)_j \psi_j^{m,r}$, $t \geq 0$.

For each $t \geq 0$, define the $(n-1) \times (m+1)$ matrix $B^{n,m,r}(t)$ by

$$[B^{n,m,r}(t)]_{i,j}$$

$$= \int_0^1 \psi_j^{m,r}(|D_x \overline{u}_n(t,x)|) D_x \overline{u}_n(t,x) D\varphi_i^n(x) dx, \quad j = 0, 1, \ldots, m, \;\; i = 1, \ldots, n-1.$$

Using the fact that $\overline{u}_n(t) \in \text{span}\{\varphi_j^n\}_{j=1}^{n-1}$ and the fact that $D\varphi_j^n$ is piecewise constant, and adopting the convention that $\overline{U}_n(t)_0 = \overline{U}_n(t)_n = 0$, we obtain that

$$\begin{aligned} [B^{n,m,r}(t)]_{i,j} = {} & n\psi_j^{m,r}(n\{\overline{U}_n(t)_i - \overline{U}_n(t)_{i-1}\})\{\overline{U}_n(t)_i - \overline{U}_n(t)_{i-1}\} \\ & - n\psi_j^{m,r}(n\{\overline{U}_n(t)_{i+1} - \overline{U}_n(t)_i\})\{\overline{U}_n(t)_{i+1} - \overline{U}_n(t)_i\}, \\ & i = 1, 2, \ldots, n-1, \;\; j = 0, 1, \ldots, m. \end{aligned}$$

The matrix form of the approximating estimator, (4.1)–(4.3), is then given by

$$M^n \dot{U}^n(t) + K^n U^n(t) + B^{n,m,r}(t) p^{m,r}(t) = F^n(t) + K^n \overline{U}_n(t), \qquad t > 0, \tag{6.15}$$

$$\Omega^{m,r} \dot{p}^{m,r}(t) - B^{n,m,r}(t)^T U_n(t) = -B^{n,m,r}(t)^T \overline{U}_n(t), \qquad t > 0, \tag{6.16}$$

$$U_n(0) = (M^n)^{-1} U_0^n, \qquad p^{m,r}(0) = (\Omega^{m,r})^{-1} p_0^{m,r}, \tag{6.17}$$

where the $(n-1)$-vector U_0^n and the $(m+1)$-vector $p_0^{m,r}$ are given by

$$[U_0^n]_i = \int_0^1 u_0(x) \varphi_i^n(x) dx, \qquad i = 1, 2, \ldots, n-1,$$

and

$$[p_0^{m,r}]_i = \langle q_0, \psi_i^{m,r} \rangle_Q$$

$$= \int_0^\infty \omega_0(\theta) q_0(\theta) \psi_i^{m,r}(\theta) d\theta + \int_0^\infty \omega_1(\theta) Dq_0(\theta) D\psi_i^{m,r}(\theta) d\theta, \qquad i = 0, 1, 2, \ldots, m,$$

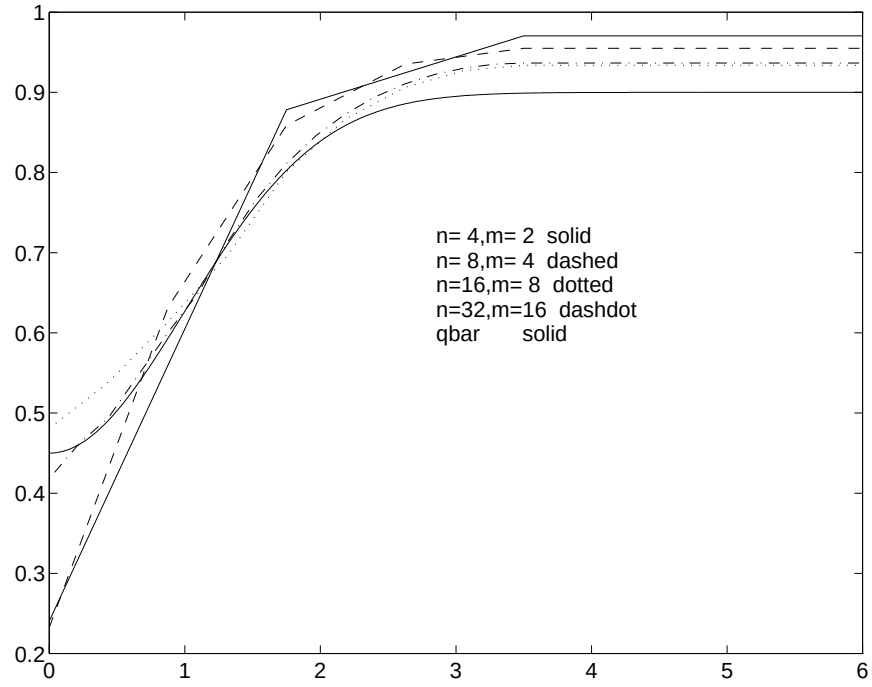


FIG. 6.8. *Final* ($t = 100$) *estimates for* $\overline{q}$ *for* $n = 2^{j+1}$, $m = 2^j$ *for* $j = 1, 2, 3, 4$.

respectively; F^n is given by (6.7); and the $(n-1) \times (n-1)$ matrices M^n and K^n are given by (6.6) and (6.4), respectively.

We set $\overline{q}(\theta) = .9(1 - \frac{1}{2}e^{-\frac{1}{2}\theta^2})$, $\theta \geq 0$, we let f be as it was given in (6.8), $f(t,x) = \{\sin(4\pi t) + 10^{-3}t^2\}\chi_{[.215,.315]}(x)$, $0 < x < 1$, $t > 0$, and set $\overline{u}_0(x) = 0$, $0 < x < 1$. We then proceeded to simulate the plant (i.e., (6.11)–(6.13)) using the IMSL routine DMOLCH, a double-precision Hermite polynomial–based method-of-lines partial differential equation solver. In our estimator, we set $\alpha = 10^{-2}$, $r = 3.5$,

$$\omega_0(\theta) = \omega_1(\theta) = \begin{cases} 1, & 0 \leq \theta < r, \\ \frac{1}{2}e^{-20\theta}, & r < \theta < \infty, \end{cases}$$

$u_0(x) = 0$, $0 < x < 1$, and $q_0(\theta) = 1$, $0 < \theta < \infty$. In Figure 6.8 we have plotted our final (i.e., at time $t = 100$) estimates for $\overline{q}$ for various values of n and m obtained by integrating the approximating estimator equations (6.15)–(6.17). In Figure 6.9 we have plotted the estimates for $\overline{q}$ at various times. These estimates were generated with $n = 32$ and $m = 16$.

7. Summary and concluding remarks. In this paper we have developed, analyzed, and tested an on-line, or adaptive, parameter identification scheme for abstract linear and nonlinear dynamical systems. Our estimator takes the form of an infinite–dimensional linear evolution system whose states consist of a state estimator and a parameter estimator. Using a standard Lyapunov estimate–based argument involving a variation of Barbălat's lemma, we were able to establish convergence of the state estimator. Under the additional assumption that the plant is sufficiently rich, or persistently excited, we were able to argue parameter convergence as well. Equivalently, when the plant is persistently excited, we were able to show that the

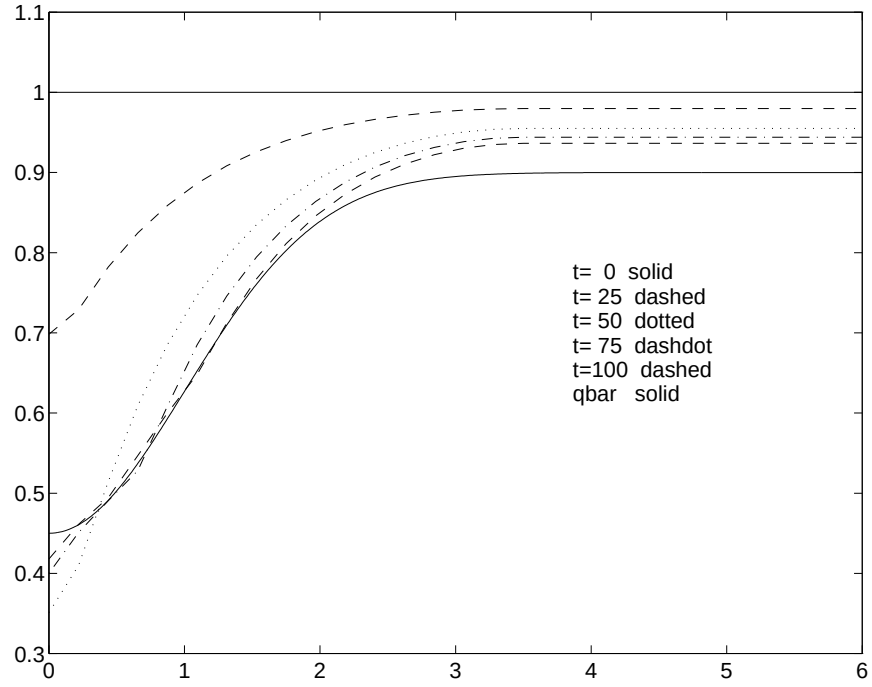


FIG. 6.9. *Estimates for $\bar{q}$ at various times generated with $n = 32$ and $m = 16$.*

solution to the error equations corresponding to the plant dynamics and the estimator with arbitrary initial data tends to the trivial solution as time tends to infinity. Our approach here represents an infinite-dimensional analogue, or extension, of some of the ideas and techniques found in the finite-dimensional treatment in [22]. We also developed a rather complete finite-dimensional approximation theory and established corresponding convergence results. We have considered the application of our general framework to certain classes of second-order systems and presented a number of examples (both first and second order, both linear and nonlinear, and involving both finite- and infinite-dimensional parameter spaces) and corresponding numerical results to demonstrate the feasibility of our schemes.

There are a number of significant extensions and applications of the results that we have presented here that we are currently pursuing. These include the development of a similar estimation theory for more general classes of distributed parameter systems, in particular, delay or hereditary systems and infinite-dimensional systems most appropriately formulated in a Banach space rather than Hilbert space setting. We are currently developing a rather general framework based upon either a single Hilbert space formulation (as opposed to the Gelfand triple approach taken here) or a Banach space formulation, which should be able to handle a significantly wider class of problems than does the treatment presented here. Extending our schemes to parameter estimation problems involving stochastic elements would also be quite useful. For example, these stochastic elements might take the form of noise in the plant measurement or the inclusion of a noise term in the plant dynamics.

A further and more significant extension of our results would involve the introduction of an observer for the purpose of eliminating the requirement that the full state be measured at each time. A modification of our scheme which does not require the entire state but rather only the output of a (finite rank) observation operator,

would represent a significant and valuable improvement. Indeed, even with the recent developments in sensor technology (for example, piezoceramics, fiber optics, and laser scanners) measuring the full state of an infinite-dimensional, or distributed, plant continues to present a substantial and, most likely, costly challenge. On the other hand, the analysis of a scheme such as ours coupled with an observer, is likely to present a significant mathematical challenge. This is because the observer would almost certainly destroy the overall linearity of the estimator. Consequently, a nonlinear stability and convergence analysis would now be required.

Finally, we are interested in using our on-line parameter estimator as a component in an indirect adaptive control algorithm for distributed parameter systems. For example, one approach that we are currently looking at involves using the evolving identified model (i.e., the output from the parameter estimator at any time instant) to design a linear quadratic controller. Such a treatment would be similar in spirit to the approach taken in [9], [10], [11], [12], [13], and [25] using either least squares or maximum likelihood estimators for the unknown parameters. A complete analysis of such a scheme is likely to be a nontrivial exercise as well.

REFERENCES

[1] H. W. Alt, K. H. Hoffmann, and J. Sprekels, *A numerical procedure to solve certain identification problems*, in Optimal Control of Partial Differential Equations, Internat. Schriftenreihe Numer. Math. 68, Birkhäuser, Basel, 1984, pp. 11–43.

[2] H. T. Banks and K. Kunisch, *Estimation Techniques for Distributed Parameter Systems*, Birkhäuser, Boston, 1989.

[3] V. Barbu, *Nonlinear Semigroups and Differential Equations in Banach Spaces*, Noordhoff, Leyden, the Netherlands, 1976.

[4] J. Baumeister and W. Scondo, *Adaptive methods for parameter identification*, in Optimization in Mathematical Physics, Methoden Verfahren Math. Phys., Peter Lang, Frankfurt am Main, 1987, pp. 87–116.

[5] J. Baumeister and W. Scondo, *Asymptotic embedding methods for parameter estimation*, in Proc. 26th IEEE Conf. on Decision and Control, 1987, pp. 170–174.

[6] M. A. Demetriou, *Adaptive Parameter Estimation of Abstract Parabolic and Hyperbolic Distributed Parameter Systems*, Ph.D. thesis, Departments of Electrical Engineering-Systems and Mathematics, University of Southern California, Los Angeles, CA, 1993.

[7] M. A. Demetriou and I. G. Rosen, *Adaptive identification of second order distributed parameter systems*, Inverse Problems, 10 (1994), pp. 261–294.

[8] M. A. Demetriou and I. G. Rosen, *On the persistence of excitation in the adaptive identification of distributed parameter systems*, IEEE Trans. Automat. Control, 39 (1994), pp. 1117–1123.

[9] T. E. Duncan, B. Maslowski, and B. Pasik-Duncan, *Adaptive boundary and point control of linear stochastic distributed parameter systems*, SIAM J. Control Optim., 32 (1994), pp. 648–672.

[10] T. E. Duncan and B. Pasik-Duncan, *Adaptive control of linear delay time systems*, Stochastics, 24 (1988), pp. 45–74.

[11] T. E. Duncan and B. Pasik-Duncan, *Adaptive control of continuous time linear stochastic systems*, Math. Control Signals Systems, 3 (1990), pp. 45–60.

[12] T. E. Duncan, B. Pasik-Duncan, and B. Goldys, *Adaptive control of linear stochastic evolution systems*, Stochastics Stochastics Rep., 36 (1991), pp. 71–90.

[13] T. E. Duncan, B. Pasik-Duncan, and B. Maslowski, *Some aspects of the adaptive boundary and point control of linear distributed parameter systems*, in Proc. 31st IEEE Conf. on Decision and Control, 1992, pp. 1077–1081.

[14] K. H. Hoffmann and J. Sprekels, *On the identification of coefficients of elliptic problems by asymptotic regularization*, Numer. Funct. Anal. Optim., 7 (1984–85), pp. 157–177.

[15] K. H. Hoffmann and J. Sprekels, *On the identification of parameters in general variational inequalities by asymptotic regularization*, SIAM J. Math. Anal., 17 (1986), pp. 1198–1217.

[16] K. S. Hong and J. Bentsman, *Application of averaging method for integro-differential equations to model reference adaptive control of parabolic systems*, Automatica J. IFAC, 30 (1994), pp. 1415–1419.

[17] K. S. Hong and J. Bentsman, *Direct adaptive control of parabolic systems: Algorithm synthesis, and convergence and stability analysis*, IEEE Trans. Automat. Control, 39 (1994), pp. 2018–2033.
[18] T. Kato, *Perturbation Theory for Linear Operators*, 2nd ed., Springer-Verlag, New York, 1984.
[19] J. L. Lions, *Optimal Control of Systems Governed by Partial Differential Equations*, Springer-Verlag, New York, 1971.
[20] J. L. Lions and E. Magenes, *Non-Homogeneous Boundary Value Problems*, Vol. I, Springer-Verlag, New York, 1972.
[21] D. G. Luenburger, *Optimization by Vector Space Methods*, John Wiley, New York, 1969.
[22] A. P. Morgan and K. S. Narendra, *On the stability of nonautonomous differential equations* $\dot{x} = [A + B(t)]x$, *with skew symmetric matrix* $B(t)^*$, SIAM J. Control Optim., 15 (1977), pp. 163–176.
[23] K. S. Narendra and A. M. Annaswamy, *Stable Adaptive Systems*, Prentice-Hall, Englewood Cliffs, NJ, 1989.
[24] K. S. Narendra and P. Kudva, *Stable adaptive schemes for system identification and control, parts* I *and* II, IEEE Transactions on Systems, Man and Cybernetics, SMC-4 (1974), pp. 542–560.
[25] B. Pasik-Duncan, *On the consistency of a least squares identification procedure in linear evolution systems*, Stochastics Stochastics Rep., 39 (1992), pp. 83–94.
[26] A. Pazy, *Semigroups of Linear Operators and Applications to Partial Differential Equations*, Springer-Verlag, New York, 1983.
[27] V. M. Popov, *Hyperstability of Control Systems*, Springer-Verlag, Berlin, Heidelberg, New York, 1973.
[28] I. G. Rosen and M. A. Demetriou, *An on-line parameter estimation scheme for flexible structures*, in Control of Flexible Structures, Fields Inst. Commun. 2, K. A. Morris, ed., The Fields Institute for Research in Mathematical Sciences, American Mathematical Society, Providence, RI, 1993, pp. 1–24.
[29] M. H. Schultz, *Spline Analysis*, Prentice-Hall, Englewood Cliffs, NJ, 1973.
[30] W. Scondo, *Ein Modellabgleichsverfahren zur adaptiven Parameteridentifikation in Evolutionsgleichungen*, Ph.D. thesis, Johann Wolfgang Goethe-Universitat zu Frankfurt am Main, Frankfurt am Main, Germany, 1987.
[31] R. E. Showalter, *Hilbert Space Methods for Partial Differential Equations*, Pitman, London, 1977.
[32] H. Tanabe, *Equations of Evolution*, Pitman, London, 1979.

SIAM J. CONTROL OPTIM.
Vol. 35, No. 3, pp. 715–743, May 1997

ASYMPTOTICALLY EFFICIENT ADAPTIVE CHOICE OF CONTROL LAWS IN CONTROLLED MARKOV CHAINS*

TODD L. GRAVES† AND TZE LEUNG LAI‡

Abstract. We consider a controlled Markov chain on a general state space whose transition probabilities are parameterized by an unknown parameter belonging to a compact metric space. There is a one-step reward associated with each pair of control and the following state of the process. Given a finite set of stationary control laws, under each of which the Markov chain is uniformly recurrent, an optimal control law in this set is one that maximizes the long-run average reward. In ignorance of the parameter value, we construct an adaptive control rule which uses the optimal control law(s) at a relative frequency of $1-O(n^{-1}\log n)$ and show that this relative frequency gives an asymptotically optimal balance between the control objective and the amount of information needed to learn about the unknown parameter. The basic idea underlying this construction is to introduce suitable "uncertainty adjustments" via sequential testing theory into the certainty-equivalence rule, thus resolving the apparent dilemma between control and information.

Key words. adaptive control of Markov chains, martingales, likelihood ratios, stationary distributions, certainty equivalence, sequential testing, multiarmed bandits

AMS subject classifications. 93C40, 93E20, 93E35, 60J20, 62L10

PII. S0363012994275440

1. Introduction and background.

We consider here a controlled Markov chain $\{X_n, n \geq 0\}$ on a measurable state space $(S, \mathcal{A})$, with a general control set U and a parametric family of transition density functions $p(x, y; u, \theta)$ with respect to some measure M on S, where θ is an unknown parameter taking values in a compact metric space Θ. Thus the transition probability measure under control action u and parameter θ is given by $P_\theta^u(X_{n+1} \in A | X_n = x) = \int_A p(x, y; u, \theta) dM(y)$. The initial distribution of X_0 under P_θ^u is also assumed to be absolutely continuous with respect to M. Let $G = \{g_1, \ldots, g_L\}$ be a finite set of stationary control laws $g_j : S \to U$ such that for every $g \in G$, the transition probability function $\{P_\theta^{g(x)}(x, A) : x \in S, A \in \mathcal{A}\}$ is irreducible with respect to some maximal irreducibility measure and has stationary distribution $\{\pi_\theta^g(A) : A \in \mathcal{A}\}$. Let $r(X_t, u_t)$ represent the one-step reward at time t, where $r : S \times U \to \mathbf{R}$, and define the long-run average reward

$$\mu_\theta(g) = \int r(x, g(x)) d\pi_\theta^g(x), \tag{1.1}$$

which will be assumed to be finite. If θ were known, then one would use the stationary control law $g_{j(\theta)}$ such that

$$\mu_\theta^* := \max_{g \in G} \mu_\theta(g) = \mu_\theta(g_{j(\theta)}). \tag{1.2}$$

*Received by the editors October 11, 1994; accepted for publication (in revised form) February 26, 1996.

http://www.siam.org/journals/sicon/35-3/27544.html

†National Institute of Statistical Sciences, P.O. Box 14162, Research Triangle Park, NC 27709-4162 (graves@niss.rti.org). The research of this author was supported by the National Science Foundation.

‡Department of Statistics, Stanford University, Stanford, CA 94305 (karola@playfair.stanford.edu). The research of this author was supported by the National Science Foundation and the National Security Agency.

In ignorance of θ, a certainty-equivalence control rule is to use the control law $g_{j(\hat{\theta}_t)}$ at time t, where $\widehat{\theta}_t$ is an estimate of θ based on the observed data $X_0, u_0, \ldots, X_{t-1}, u_{t-1}, X_t$ (in chronological order).

For the case of a finite state space S, Mandl [18] studied this certainty-equivalence rule in which $\widehat{\theta}_t$ is a minimum contrast estimate and showed that $\widehat{\theta}_t$ converges almost surely (a.s.) to θ under a restrictive "identifiability condition" and some other regularity conditions. Borkar and Varaiya [6] removed this identifiability condition and showed that when Θ is finite, the maximum likelihood estimate $\widehat{\theta}_t$ converges a.s. to a random variable θ^* such that

$$p(x, y; g_{j(\theta^*)}(x), \theta^*) = p(x, y, g_{j(\theta^*)}(x), \theta) \tag{1.3}$$

for all $x, y \in S$ (finite). They also gave an example for which $\theta^* \neq \theta$ with positive probability, showing that the certainty-equivalence rule can prematurely converge to a wrong parameter value so that it eventually uses only the suboptimal stationary control law $g_j(\theta^*)$ to the exclusion of other control laws.

In view of this difficulty with the certainty-equivalence rule, various modifications of the rule have appeared in the literature. Kumar [11] and Kumar and Varaiya [12] have provided comprehensive surveys of the developments up to the mid-1980s, which include (i) forced choice schemes that reserve some prespecified sparse set of times for experimentation with all stationary control laws in G, (ii) randomization schemes for which every $g \in G$ has a positive probability, whose value is to be determined adaptively from the past data, of being applied at each time, and (iii) using penalized (cost-biased) maximum likelihood estimators $\widehat{\theta}_t$.

1.1. Ideas from bandit theory. The past decade has witnessed other developments in a classical example of adaptive choice from a finite set of control actions, namely, the multiarmed bandit problem. In its simplest form, the problem can be described as follows. There are L statistical populations $\Pi_1, \ldots, \Pi_L$ with univariate density functions $p(y; \theta_1), \ldots, p(y; \theta_L)$ with respect to some measure M. At each time t we can sample from one of these populations, and the reward is the sampled value X_t. Thus the control set U is $\{1, \ldots, L\}$, where control action j refers to sampling from Π_j. An adaptive sampling rule consists of a sequence of random variables $u_1, u_2, \ldots$ taking values in $\{1, \ldots, k\}$ such that the event $\{u_t = j\}$ ("X_{t+1} is sampled from Π_j") belongs to the σ-field generated by $u_0, X_1, u_1, \ldots, X_{t-1}, u_{t-1}, X_t$. Let $\theta = (\theta_1, \ldots, \theta_L)$. If θ were known, then we would sample from the population $\Pi_{j(\theta)}$ with the largest mean; i.e., $\mu_\theta^* := \max_{1 \le j \le L} \mu_\theta(j) = \mu_\theta(j(\theta))$, where $\mu_\theta(j) = \int y p(y; \theta_j) dM(y)$ is assumed to be finite. In ignorance of θ, the problem is to sample $X_1, X_2, \ldots$ sequentially from the k populations to maximize $E_\theta(\sum_{i=1}^n X_i)$, or equivalently to minimize the regret

$$R_n(\theta) = n\mu_\theta^* - E_\theta\left(\sum_{i=1}^n X_i\right) = \sum_{j:\mu_\theta(j)<\mu_\theta^*} (\mu_\theta^* - \mu_\theta(j)) E_\theta T_n(j) \tag{1.4}$$

as $n \to \infty$, where $T_n(j) = \sum_{t=1}^n I_{\{u_{t-1}=j\}}$ and $I_A = 1$ if A occurs, $I_A = 0$ otherwise. Lai and Robbins [16] showed how to construct sampling rules for which $R_n(\theta) = O(\log n)$ at every θ. These rules are called "uniformly good." They also developed asymptotic lower bounds for the regret $R_n(\theta)$ of uniformly good rules and showed that the rules constructed actually attain these asymptotic lower bounds and are therefore asymptotically efficient. Specifically, they showed that under certain

regularity conditions

$$\liminf_{n\to\infty} R_n(\theta)/\log n \geq c(\theta) \tag{1.5}$$

for uniformly good rules and gave an explicit formula for $c(\theta)$ in terms of $\mu_\theta^* - \mu_\theta(j)$ and certain Kullback–Leibler information numbers. A more general representation of the lower bound $c(\theta)$ is given in section 2, where we extend this result on the multiarmed bandit problem to the general setting of adaptive choice of stationary control laws in controlled Markov chains.

Anantharam, Varaiya, and Walrand [5] generalized the results of [16] to the multiarmed bandit problem in which each Π_j represents an aperiodic, irreducible Markov chain on a finite state space S so that successive observations from Π_j are no longer independent but are governed by the Markov transition density $p(x, y; \theta_j)$. Assuming the successive observations from Π_j to be independent with a common density function $p(y; \theta_j)$, Agrawal, Hedge, and Teneketzis [1] incorporated an additional switching cost and showed that the sampling rules of [16] can be modified by sampling in blocks so that the asymptotic lower bound in (1.5) is still attained and the cumulative switching cost up to time n is of the order $o(\log n)$ when no more than one population has the largest mean μ_θ^*.

For the problem of adaptive choice of stationary control laws in controlled Markov chains, switching costs are particularly relevant since it usually takes time to change from a new control strategy to another. We shall assume no switching cost for switching among the (typically equivalent) optimal stationary control laws that attain the maximum in (1.2) and a cost $a(\theta)$ for each switch from one $g \in G$ to another $g' \in G$ when g and g' are not both optimal. An *adaptive control rule* ϕ is a sequence of random variables $\phi_1, \phi_2, \ldots$ taking values in G such that $\{\phi_t = g\} \in \mathcal{F}_t$ for all $g \in G$ and $t \geq 0$, where

$$\mathcal{F}_t = \sigma\text{-field generated by } X_0, \phi_0, \ldots, X_{t-1}, \phi_{t-1}, X_t. \tag{1.6}$$

Defining $\mu_\theta(g)$ and μ_θ^* by (1.1) and (1.2), we generalize (1.4) to controlled Markov chains by letting

$$R_n(\theta) = \sum_{g\in G:\mu_\theta(g)<\mu_\theta^*} (\mu_\theta^* - \mu_\theta(g)) E_\theta T_n(g), \text{ with } T_n(g) = \sum_{i=0}^{n-1} I_{\{\phi_i = g\}}. \tag{1.7}$$

In view of the additional switching cost $a(\theta)$ for each switch between two control laws in G, not both optimal, we define the overall regret to be $R_n(\theta) + a(\theta)S_n(\theta)$, where

$$S_n(\theta) = E_\theta \left(\sum_{i=1}^{n} I_{\{\phi_i \neq \phi_{i-1}, \min(\mu_\theta(\phi_i), \mu_\theta(\phi_{i-1})) < \mu_\theta^*\}} \right). \tag{1.8}$$

An adaptive control rule ϕ is said to be *uniformly good* if

$$R_n(\theta) = O(\log n) \text{ and } S_n(\theta) = o(\log n) \text{ for every } \theta \in \Theta. \tag{1.9}$$

In section 2 we develop an asymptotic lower bound for $R_n(\theta)$ among all uniformly good rules, and in section 3 we construct adaptive control rules that attain this lower bound. These results therefore generalize those of [16] on the multiarmed bandit problem to the setting of adaptive choice of control laws in controlled Markov chains.

A major technical difficulty in this generalization is that unlike Markovian bandit processes in which the state of Π_j is "frozen" until a new observation is sampled from Π_j, for controlled Markov chains X_{t+1} is governed by the immediately preceding state X_t and control action $\phi_t(X_t)$ irrespective of whether $\phi_{t+1} = \phi_t$ or not. We resolve this difficulty by using certain change-of-measure arguments in section 2 and some limit theorems for controlled Markov chains developed in section 4.

This difficulty disappears in the special case where the controlled Markov chain $\{X_t, t \geq 1\}$ is a sequence of independent random variables so that the conditional density of X_{t+1} given $u_t = u$ is $p(y; u, \theta)$. Assuming U and Θ to be finite, Agrawal, Teneketzis, and Anantharam [3] studied this special case by regarding each control action $u \in U$ as an arm and $r(X_1, u_1), r(X_2, u_2), \ldots$ as a sequence of rewards obtained by choosing the arms $u_1, u_2, \ldots$. They noted, however, another difficulty in reducing this problem to the multiarmed bandit problem because of the differences in how the parameter space Θ is defined in the two problems. In the controlled independent sequence problem, θ parameterizes all the arms $u \in U$, whereas in the multiarmed bandit problem $\theta = (\theta_1, \ldots, \theta_k)$ with each θ_j parameterizing the individual arm Π_j. They circumvented this difficulty by making use of the finiteness of Θ and introducing a finite set $B(\theta)$ of "bad" parameter values associated with θ. They thereby obtained an asymptotic lower bound for the regret (1.7) of uniformly good control rules and developed a rule that attains this bound. In section 2, without assuming Θ to be finite, we define the bad set $B(\theta)$ in the setting of controlled Markov chains with general state and parameter spaces. When the state space S, the control set U, and the parameter space Θ are all finite, Agrawal, Teneketzis, and Ananthanam [4] developed a "translation scheme" which together with the construction of an "extended probability space" enabled them to solve the controlled Markov chain problem by converting it to a form similar to that for the controlled independent sequence problem in [3]. This ingenious idea of translation schemes, however, depends heavily on the finiteness of S. Our development of an asymptotic lower bound for (1.7) in section 2 uses a different approach which involves large deviation probabilities for controlled Markov chains on general state spaces S satisfying certain uniform recurrence assumptions.

As a consequence of the translation scheme under their finiteness assumptions, Agrawal, Teneketzis, and Ananthanam [4] obtained the approximation

$$E_\theta \left\{ \sum_{i=0}^{n-1} r(X_i, \phi_i(X_i)) \right\} = \sum_{g \in G} \mu_\theta(g) E_\theta T_n(g) + O(1) \text{ as } n \to \infty. \tag{1.10}$$

Hence in this case (1.7) can be expressed as

$$R_n(\theta) = \widetilde{R}_n(\theta) + O(1), \text{ where } \widetilde{R}_n(\theta) = n\mu_\theta^* - E_\theta \left\{ \sum_{i=0}^{n-1} r(X_i, \phi_i(X_i)) \right\}. \tag{1.11}$$

Note that $\widetilde{R}_n(\theta)$ is the shortfall between the long-run cumulative reward using the optimal stationary control law $g_{j(\theta)}$ and the cumulative reward of the adaptive control rule ϕ. Moreover, by making use of the translation scheme in the development of their asymptotic lower bound for $R_n(\theta)$, Agrawal, Teneketzis, and Ananthanam [4] did not need to impose the constraint on the expected number of switches in (1.9) for uniformly good rules. However, for general state spaces, (1.10) and (1.11) need no longer hold, and there may even exist adaptive control rules for which $\lim_{n\to\infty} \widetilde{R}_n(\theta) = -\infty$ at certain values of θ. This difficulty arises because in the absence of (1.10), the long-run average optimality property (1.2) of the stationary control law $g_{j(\theta)}$ no longer

ensures it to be asymptotically optimal among adaptive control rules that can switch freely among stationary control laws in G. Note that G does not contain such adaptive control rules which are not stationary. We therefore have to put some constraint on the expected number of switches in the adaptive control rules to compare them with the optimal stationary control law $g_{j(\theta)}$ (which makes no switch in G). This can be regarded as a "complexity constraint," consistent with our basic assumption of a *finite* set of *stationary* control policies to reduce the complexity of the Markov control problem. Under the switching constraint that $S_n(\theta) = o(\log n)$ and assuming the transition probability function $\{P_\theta^{g(x)}(x, A) : x \in S, A \in \mathcal{A}\}$ to be uniformly recurrent for every $g \in G$, it is shown in [14] that the "reward regret" $\widetilde{R}_n(\theta)$ is asymptotically equivalent to the more tractable weighted sum (1.7) of expected frequencies of using suboptimal stationary control laws; i.e.,

$$\widetilde{R}_n(\theta) = R_n(\theta) + o(\log n) \text{ as } n \to \infty. \tag{1.12}$$

The constraint on $S_n(\theta)$ in (1.9) relates only to switches between two stationary control laws which are not both optimal when θ is the true parameter. We do not impose the $o(\log n)$ constraint on the expected number of switches between two optimal stationary control laws. In fact, since one cannot infer from the past data which of these optimal stationary control laws is significantly inferior, one is expected to keep switching among them to learn their performance, as in [16] for the multiarmed bandit problem.

1.2. Uncertainty adjustments to the certainty-equivalence rule via sequential testing theory. Lai [13] pointed out the usefulness of sequential testing theory in making uncertainty adjustments of the certainty-equivalence rule, leading to asymptotically optimal rules when the control set is finite. To illustrate this, he considered the following bivariate bandit problem. Let Π_1, Π_2, Π_3 be three bivariate normal populations with respective mean vectors $(\mu_1, \xi), (\mu_2, \mu_3)$, and $(\mu_3, \mu_2 + \xi)$ and with a common known covariance matrix which is equal to the identity matrix. Here $\theta = (\mu_1, \mu_2, \mu_3, \xi)$ is the unknown parameter vector and the problem is to sample $X_1, X_2, \ldots$ sequentially from the three populations in order to maximize the expected value of the first component of $\sum_{i=1}^n X_i$ as $n \to \infty$. The relevant information we need for optimal control can be represented by the three hypotheses $H_j : \mu_j = \max(\mu_1, \mu_2, \mu_3),\ j = 1, 2, 3$. In other words, we do not need to know the actual values of μ_1, μ_2, μ_3, ξ but need only to determine which of μ_1, μ_2, μ_3 is the largest. While information about μ_1 can only be obtained by sampling from Π_1, information about μ_2 and μ_3 can be obtained by sampling from Π_2 alone or from Π_3 and Π_1. Using results from sequential testing theory, Lai [13] constructed an asymptotically optimal rule whose regret (1.7) satisfies $R_n(\theta) = O(1)$ if $\mu_1 = \max(\mu_2, \mu_3)$ and $R_n(\theta) \sim c(\theta) \log n$ otherwise, where

$$\begin{aligned} c(\theta) &= 2/\{\mu_1 - \max(\mu_2, \mu_3)\} \text{ if } \mu_1 > \max(\mu_2, \mu_3) \\ &= 2/(\mu_2 - \mu_1) \text{ if } \mu_2 > \max(\mu_1, \mu_3) \\ &= 2/(\mu_3 - \mu_1) \text{ if } \mu_3 = \mu_2 > \mu_1 \text{ or } \mu_3 > \mu_1 \geq \mu_2 \\ &= 2/(\mu_3 - \mu_1) + 2/(\mu_3 - \mu_2) - 2(\mu_3 - \mu_2)/(\mu_3 - \mu_1)^2 \text{ if } \mu_3 > \mu_2 > \mu_1. \end{aligned}$$

In section 3 we use sequential testing theory to construct asymptotically efficient adaptive control rules in controlled Markov chains. These rules are considerably

simpler than those in [4] which require finiteness of S and Θ for their implementation and for the analysis that shows their regret $R_n(\theta)$ to be of the order $O(\log n)$. The rules in section 3 are applicable to general state spaces S and compact metric spaces Θ, and we prove in section 4 that they attain the asymptotic lower bound $(c(\theta)+o(1))\log n$ for the regret established in section 2.

In summary, by making use of bandit theory and sequential testing methodology, we generalize herein previous work of Agrawal, Tenektetzis, and Ananthanam [4] from the case of finite Θ and S to compact Θ and general state spaces S while still assuming finiteness of G, which is crucial for both the asymptotic lower bound in section 2 and the rules proposed in section 3. This generalization requires certain constraints on the expected number of switches among the stationary control laws in G and uniform recurrence assumptions on the transition probability functions. We construct in section 3 adaptive control rules with regret $R_n(\theta)$ having the asymptotically minimal order $c(\theta)\log n$, where the constant $c(\theta)$ is given in section 2. Using nonparametric sequential testing theory instead of the parametric likelihood ratio approach here and assuming G to be countable instead of finite, Lai and Yakowitz [17] removed the parametric and related assumptions herein and developed adaptive control rules with regret $R_n(\theta)=O(\alpha_n\log n)$ for any given nondecreasing sequence of positive numbers $\alpha_n\to\infty$ and $\alpha_{2n}=O(\alpha_n)$. Earlier, Agrawal and Teneketzis [2] also used a nonparametric approach to construct adaptive control rules with regret $R_n(\theta)=O((\log n))^{1+\delta})$ for any given $\delta>0$ in the case of finite G, Θ, and S so that the translation scheme of Agrawal, Teneketzis, and Ananthanam [4] is applicable.

2. Decomposition of the parameter space and an asymptotic lower bound for the regret of uniformly good rules. Using the same notation as that introduced at the beginning of section 1, define for $g\in G$ the Kullback–Leibler information number

$$I^g(\theta,\lambda)=\int\int\left\{\log\frac{p(x,y;g(x),\theta)}{p(x,y;g(x),\lambda)}\right\}p(x,y;g(x),\theta)dM(y)d\pi_\theta^g(x), \tag{2.1}$$

which will be assumed to be finite for all $\theta,\lambda\in\Theta$. We shall decompose Θ as the union of L subsets: $\Theta=\Theta_1\cup\cdots\cup\Theta_L$, where

$$\Theta_j=\{\theta\in\Theta:\mu_\theta(g_j)=\max_{g\in G}\mu_\theta(g)\}; \tag{2.2}$$

i.e., g_j is an optimal stationary control law if $\theta\in\Theta_j$. For $\theta\in\Theta$, let

$$J(\theta)=\{1\le j\le L:\mu_\theta(g_j)=\max_{g\in G}\mu_\theta(g)(=\mu_\theta^*)\}, \tag{2.3}$$

$$B(\theta)=\left\{\lambda\in\Theta:\ \lambda\notin\bigcup_{j\in J(\theta)}\Theta_j \text{ and } I^{g_j}(\theta,\lambda)=0 \text{ for all } j\in J(\theta)\right\}, \tag{2.4}$$

(2.5)

$$c(\theta)=\inf\left\{\sum_{j\notin J(\theta)}c_j[\mu_\theta^*-\mu_\theta(g_j)]:c_j\in[0,\infty),\ \inf_{\lambda\in B(\theta)}\sum_{j\notin J(\theta)}c_jI^{g_j}(\theta,\lambda)\ge 1\right\}(\inf\emptyset=\infty).$$

Thus, $\{g_j, j\in J(\theta)\}$ is the set of all optimal stationary control laws when θ is the true parameter value, and $B(\theta)$ consists of all "bad" parameter values $\lambda\notin\cup_{j\in J(\theta)}\Theta_j$

which are statistically indistinguishable from θ if one only uses the optimal control laws $g_j, j \in J(\theta)$, because $I^{g_j}(\theta, \lambda) = 0$. Theorem 1 below shows that under certain regularity conditions $(c(\theta) + o(1)) \log n$ is an asymptotic lower bound for the regret (1.7) of uniformly good rules. Note that (2.5) can also be expressed as

(2.6)
$$c(\theta) = \inf \left\{ \frac{\sum_{j \notin J(\theta)} \alpha_j [\mu_\theta^* - \mu_\theta(g_j)]}{\inf_{\lambda \in B(\theta)} \sum_{j \notin J(\theta)} \alpha_j I^{g_j}(\theta, \lambda)} \ : \ \alpha_j \geq 0, \ \sum_{j \notin J(\theta)} \alpha_j = 1 \right\} \ (\inf \emptyset = \infty).$$

This alternative form of the asymptotic lower bound of $R_n(\theta)/\log n$ was obtained by Agrawal, Teneketzis, and Ananthanam [4] when the state space S and the parameter space Θ are finite. Theorem 1 uses a different argument which involves the equivalent form (2.5) of (2.6) to establish the result for general state spaces and compact parameter spaces. We first give some examples to illustrate the computation of $c(\theta)$.

Example 1. Consider the multiarmed bandit problem of section 1.1. Here $\theta = (\theta_1, \cdots, \theta_L)$, $g_j = j$ ("sample from Π_j") and $I^j(\theta, \lambda) = I(\theta_j, \lambda_j)$, where

$$I(a, b) = \int p(y; a) \log[p(y; a)/p(y; b)] dM(y), \ \mu(a) = \int y p(y; a) dM(y). \tag{2.7}$$

Assume that $I(a, b) < \infty$ and that $I(a, b) = 0$ iff $\mu(a) = \mu(b)$, analogous to the assumptions (1.6) and (1.7) of Lai and Robbins [16].

(i) Suppose $L = 2, \Theta = \{(\alpha, \beta), (\beta, \alpha)\}$, and $\mu(\alpha) \neq \mu(\beta)$. Thus, it is known that one population has a specified parameter value α and the other has parameter value β, but it is not known whether Π_1 or Π_2 is associated with α. This is the two-armed bandit problem studied by Feldman [7]. Here $I^1((\alpha, \beta), (\beta, \alpha)) = I(\alpha, \beta) > 0, I^2((\alpha, \beta), (\beta, \alpha)) = I(\beta, \alpha) > 0$, and therefore $B(\theta) = \emptyset, c(\theta) = 0$ for $\theta \in \Theta$. In fact, Feldman's procedure has regret $R_n(\theta) = O(1)$. Lai and Robbins [15] considered more general k-armed bandit problems in which $B(\theta) = \emptyset$ for all $\theta \in \Theta$ and developed sampling rules with $R_n(\theta) = O(1)$.

(ii) Suppose $\Theta = \Delta^L$, where Δ is a compact metric space. For $\theta = (\theta_1, \ldots, \theta_L)$, let $\theta^* \in \{\theta_1, \ldots, \theta_L\}$ be such that $\mu(\theta^*) = \max_{1 \leq i \leq L} \mu(\theta_i)$. Then $J(\theta) = \{1 \leq j \leq L : \mu(\theta_j) = \mu(\theta^*)\}$ and

(2.8)
$$B(\theta) = \left\{ (\lambda_1, \ldots, \lambda_L) \in \Theta : \mu(\lambda_j) = \mu(\theta^*) \text{ for all } j \in J(\theta), \max_{1 \leq i \leq L} \mu(\lambda_i) > \mu(\theta^*) \right\}$$

since $I(a, b) = 0$ iff $\mu(a) = \mu(b)$ by assumption. Assume as in (1.7) of [16] that

$$I(\theta_j, b) \to I(\theta_j, \theta^*) \quad \text{as} \quad \mu(b) \downarrow \mu(\theta^*). \tag{2.9}$$

Consider the minimization problem in (2.5) which reduces here to finding nonnegative numbers $c_j, j \notin J(\theta)$, to minimize $\sum_{j \notin J(\theta)} c_j(\mu(\theta^*) - \mu(\theta_j))$ subject to the constraints

$$\inf_{\lambda \in B_i(\theta)} \sum_{j \notin J(\theta)} c_j I(\theta_j, \lambda_j) \geq 1, \ \ i \notin J(\theta), \tag{2.10}$$

where $B_i(\theta) = \{\lambda \in B(\theta) : \mu(\lambda_i) = \max_{1 \leq s \leq L} \mu(\lambda_s)\}$. For fixed $i \notin J(\theta)$, since $(\theta_1, \ldots, \theta_{i-1}, b, \theta_{i+1}, \ldots, \theta_L) \in B_i(\theta)$ for any $b \in \Delta$ with $\mu(b) > \mu(\theta^*)$, (2.9) and

(2.10) imply that $c_i \geq 1/I(\theta_i, \theta^*)$. Hence $c(\theta) \geq \sum_{j \notin J(\theta)} (\mu(\theta^*) - \mu(\theta_j))/I(\theta_j, \theta^*)$, which is the asymptotic lower bound for $R_n(\theta)/\log n$ given in [16], where sampling rules that attain this lower bound are also constructed for certain parametric families having the monotonicity property

$$I(a, b) \geq I(a, \theta^*) \quad \text{whenever} \quad \mu(b) \geq \mu(\theta^*) \geq \mu(a). \tag{2.11}$$

Under (2.11), $I(\theta_i, \lambda_i)/I(\theta_i, \theta^*) \geq 1$ for all $\lambda \in B_i(\theta)$, and therefore the constraint (2.10) holds with $c_j = 1/I(\theta_j, \theta^*)$, $j \notin J(\theta)$. This choice of c_j therefore solves the minimization problem in (2.5) under the assumptions (2.9) and (2.11), yielding $c(\theta) = \sum_{j \notin J(\theta)} (\mu(\theta^*) - \mu(\theta_j))/I(\theta_j, \theta^*)$.

Example 2. Consider the following variant of Example 1. Let Π_1, Π_2, Π_3 be three univariate normal populations with respective means $\gamma, \xi + 1$, and ξ^2 and common variance 1, where γ and ξ are unknown parameters. Here $\Theta = \{\theta = (\gamma, \xi) : -\infty < \gamma < \infty, -\infty < \xi < \infty\}$, and the problem is to sample $X_1, X_2, \ldots$ sequentially from the three populations to maximize the expected value of $\sum_{i=1}^n X_i$ as $n \to \infty$. Therefore, as in Example 1, $g_j = j$ ("sample from Π_j"), $I^1((\gamma, \xi), (\widetilde{\gamma}, \widetilde{\xi})) = (\gamma - \widetilde{\gamma})^2/2, I^2((\gamma, \xi), (\widetilde{\gamma}, \widetilde{\xi})) = (\xi - \widetilde{\xi})^2/2, I^3((\gamma, \xi), (\widetilde{\gamma}, \widetilde{\xi})) = (\xi^2 - \widetilde{\xi}^2)^2/2$, and

$$\begin{aligned} B(\gamma, \xi) &= \{(\widetilde{\gamma}, \widetilde{\xi}) : \max(\widetilde{\xi} + 1, \widetilde{\xi}^2) > \gamma\} \text{ if } \gamma > \max(\xi + 1, \xi^2) \\ &= \{(\widetilde{\gamma}, \xi) : \widetilde{\gamma} > \xi + 1\} \text{ if } \xi + 1 > \max(\gamma, \xi^2) \\ &= \{(\widetilde{\gamma}, \widetilde{\xi}) : |\widetilde{\xi}| = |\xi|, \max(\widetilde{\gamma}, \widetilde{\xi} + 1) > \xi^2\} \text{ if } \xi^2 > \max(\gamma, \xi + 1). \end{aligned}$$

To compute $c(\theta)$, we can use arguments similar to those in Example 1 to show that

$$c(\gamma, \xi) = 2/(\xi + 1 - \gamma) \text{ if } \xi + 1 > \max(\gamma, \xi^2). \tag{2.12}$$

The case $\gamma > \max(\xi + 1, \xi^2)$ is considerably more complicated, and it is more convenient to use the representation (2.6), which reduces to

$$c(\gamma, \xi) = \inf_{0 \leq \pi \leq 1} \frac{\pi(\gamma - \xi - 1) + (1 - \pi)(\gamma - \xi^2)}{\inf_{\widetilde{\xi}: \widetilde{\xi} + 1 > \gamma \text{ or } \widetilde{\xi}^2 > \gamma} \pi(\xi - \widetilde{\xi})^2/2 + (1 - \pi)(\xi^2 - \widetilde{\xi}^2)^2/2}. \tag{2.13}$$

To solve the minimization problem in (2.13), first fix $\pi \in [0, 1]$ and find $\widetilde{\xi}_\pi$ to minimize $\psi_\pi(\widetilde{\xi}) := \pi(\xi - \widetilde{\xi})^2/2 + (1 - \pi)(\xi^2 - \widetilde{\xi}^2)^2/2$ subject to $\widetilde{\xi} \geq \gamma - 1$ or $|\widetilde{\xi}| \geq \sqrt{\gamma}$. Then find $\pi(\gamma, \xi) \in [0, 1]$ that minimizes $\{\pi(\gamma - \xi - 1) + (1 - \pi)(\gamma - \xi^2)\}/\psi_\pi(\widetilde{\xi}_\pi)$. Note that $d\psi_\pi/d\widetilde{\xi} = -(\xi - \widetilde{\xi})\{\pi + 2(1 - \pi)\widetilde{\xi}(\xi + \widetilde{\xi})\}$, which has zeroes at $\widetilde{\xi} = \xi$ and $\widetilde{\xi} = \frac{1}{2}\{-\gamma + [\gamma^2 - 2\pi/(1 - \pi)]^{1/2}\}$. For example, if $(\gamma, \xi) = (1.69, -1)$, then $\pi(\gamma, \xi) = 0.112$. This is in sharp contrast to (2.12) or Example 1, for which the optimizing π is always 0 or 1 if we use the representation (2.6) to evaluate $c(\theta)$. In section 3 we shall consider the case $\xi^2 > \max(\gamma, \xi + 1)$.

In the following theorem we use the same notation as that introduced at the beginning of section 1. We shall assume that the transition probability function $\{P_\theta^{g(x)}(x, A) : x \in S, A \in \mathcal{A}\}$ is uniformly recurrent for every $\theta \in \Theta$ and $g \in G$; i.e., there exist positive constants $a_\theta^g < b_\theta^g$ such that

$$a_\theta^g \leq p(x, y; g(x), \theta) \leq b_\theta^g \text{ for almost every (with respect to } M) \ x \text{ and } y \tag{2.14}$$

(cf. [9]). This implies that for every $g \in G, \theta \in \Theta$, and $\lambda \in B(\theta)$, there exist positive constants $\alpha_{\theta,\lambda}^g < \beta_{\theta,\lambda}^g$ such that

$$\alpha_{\theta,\lambda}^g \leq p(x, y; g(x), \theta)/p(x, y; g(x), \lambda) \leq \beta_{\theta,\lambda}^g \text{ for } (M\text{-})\text{almost every } x \text{ and } y. \tag{2.15}$$

We consider situations where there are switching costs, for which "uniformly good" rules are defined by (1.9). The following theorem gives an asymptotic lower bound for the regret (1.7) of uniformly good rules.

THEOREM 1. *Under* (2.14), *for any uniformly good rule* ϕ,

$$\liminf_{n\to\infty} \sum_{j\notin J(\theta)} I^{g_j}(\theta,\lambda) E_\theta T_n(g_j)/\log n \geq 1 \text{ for every } \lambda \in B(\theta) \tag{2.16}$$

and therefore

$$\liminf_{n\to\infty} R_n(\theta)/\log n \geq c(\theta) \tag{2.17}$$

for every $\theta \in \Theta$.

Proof. Since $R_n(\theta) = \sum_{j\notin J(\theta)} [\mu_\theta^* - \mu_\theta(g_j)] E_\theta T_n(g_j)$ by (1.7), (2.17) follows from (2.5) and (2.16) (writing $E_\theta T_n(g_j) = c_{j,n} \log n$ and noting that $\inf \emptyset = \infty$). To prove (2.16), it suffices to show that for every $\lambda \in B(\theta)$ and $\epsilon > 0$,

$$\lim_{n\to\infty} P_\theta \left\{ \sum_{j\notin J(\theta)} I^{g_j}(\theta,\lambda) T_n(g_j) \geq (1-\epsilon)\log n \right\} = 1. \tag{2.18}$$

The proof of (2.18) uses a change-of-measure argument similar to that in the proof of Theorem 2 of Lai and Robbins [16] on the multiarmed bandit problem. Since $\lambda \in B(\theta)$, $\lambda \notin \cup_{j\in J(\theta)} \Theta_j$ and therefore $J(\lambda) \cap J(\theta) = \emptyset$. Since ϕ is uniformly good, $E_\lambda\{n - \sum_{i\in J(\lambda)} T_n(g_i)\} = O(\log n)$ by (1.9). For $g \in G$, if

$$0 = I^g(\theta,\lambda) = \int\int p(x,y;g(x),\theta) \log[p(x,y;g(x),\theta)/p(x,y;g(x),\lambda)] dM(y) d\pi_\theta^g(x), \tag{2.19}$$

then $p(x,y;g(x),\theta) = p(x,y;g(x),\lambda)$ for M-almost everywhere (a.e.) x and y (noting that $d\pi_\theta^g/dM > 0$ a.e.$[M]$ by (2.14)), and therefore $\mu_\theta(g) = \mu_\lambda(g)$. Since $\lambda \in B(\theta), \mu_\lambda^* > \mu_\lambda(g_i)$ and $I^{g_i}(\theta,\lambda) = 0$ for all $i \in J(\theta)$. Hence $\mu_\theta(g_i) = \mu_\lambda(g_i) < \mu_\lambda^*$ for all $i \in J(\theta)$, implying that $\mu_\lambda^* > \mu_\theta^*$. For $j \in J(\lambda), \mu_\lambda(g_j) = \mu_\lambda^* > \mu_\theta^* \geq \mu_\theta(g_j)$ and therefore $I^{g_j}(\theta,\lambda) > 0$. It then follows that for all large n,

$$\begin{aligned}
&P_\lambda \left\{ \sum_{j\notin J(\theta)} I^{g_j}(\theta,\lambda) T_n(g_j) < (1-\epsilon)\log n \right\} \\
&\leq P_\lambda \left\{ \sum_{j\in J(\lambda)} I^{g_j}(\theta,\lambda) T_n(g_j) < (1-\epsilon)\log n \right\} \\
&\leq P_\lambda \left\{ \sum_{j\in J(\lambda)} T_n(g_j) \leq n/2 \right\} = P_\lambda \left\{ n - \sum_{j\in J(\lambda)} T_n(g_j) \geq n/2 \right\} \\
&\leq 2n^{-1} E_\lambda \left\{ n - \sum_{j\in J(\lambda)} T_n(g_j) \right\} = O(n^{-1}\log n).
\end{aligned} \tag{2.20}$$

Let $L_n = \sum_{i=0}^{n-1} \log[p(X_i, X_{i+1}, \phi_i(X_i), \theta)/p(X_i, X_{i+1}; \phi_i(X_i), \lambda)]$ and let $0 < \delta < \epsilon/2$. Note that

(2.21)
$$\begin{aligned} &P_\lambda \left\{ \sum_{j \notin J(\theta)} I^{g_j}(\theta, \lambda) T_n(g_j) < (1-\epsilon)\log n, L_n \le (1-\delta)\log n \right\} \\ &= \int_{\left\{\sum_{j \notin J(\theta)} I^{g_j}(\theta,\lambda) T_n(g_j) < (1-\epsilon)\log n, L_n \le (1-\delta)\log n\right\}} e^{-L_n} dP_\theta \\ &\ge e^{-(1-\delta)\log n} P_\theta \left\{ \sum_{j \notin J(\theta)} I^{g_j}(\theta, \lambda) T_n(g_j) < (1-\epsilon)\log n, L_n \le (1-\delta)\log n \right\}. \end{aligned}$$

Combining (2.20) and (2.21) yields

$$\lim_{n\to\infty} P_\theta \left\{ \sum_{j \notin J(\theta)} I^{g_j}(\theta, \lambda) T_n(g_j) < (1-\epsilon)\log n, L_n \le (1-\delta)\log n \right\} = 0. \tag{2.22}$$

Let $h_g(x, y) = \log[p(x, y; g(x), \theta)/p(x, y; g(x), \lambda)]$. For $-\infty < \alpha < \infty$, define the measure

$$\widehat{P}_{x,\alpha,g}(A) = \int_A e^{\alpha h_g(x,y)} p(x, y; g(x), \theta) M(dy), \quad A \in \mathcal{A},$$

and define the linear operator $\widehat{P}_g(\alpha)$ on the space of bounded measurable functions $f : S \to \mathbf{R}$ by $\widehat{P}_g(\alpha) f(x) = \int f(y) \widehat{P}_{x,\alpha,g}(dy)$. In view of (2.14) and (2.15), $\widehat{P}_g(\alpha)$ has a maximal simple real eigenvalue $\rho_g(\alpha)$, with associated right eigenfunction $r_g(\cdot; \alpha) : S \to (0, \infty)$ and left eigenmeasure $\ell_g(\cdot; \alpha) : \mathcal{A} \to [0, \infty)$ normalized so that $\int r_g(x; \alpha) \ell_g(dx; \alpha) = 1$; moreover, $r_g(\cdot; \alpha)$ is bounded and uniformly positive for every fixed α (cf. [10]). For $j \in J(\theta)$, since $\lambda \in B(\theta)$, it follows that $I^{g_j}(\theta, \lambda) = 0$, and therefore by (2.19), $p(x, y; g_j(x), \theta) = p(x, y; g_j(x), \lambda)$; i.e., $h_g(x, y) = 0$, for M-a.e. x and y. Hence

$$L_n = \sum_{i=0}^{n-1} I_{\{\phi_i \notin G_J\}} \log[p(X_i, X_{i+1}; \phi_i(X_i), \theta)/p(X_i, X_{i+1}; \phi_i(X_i), \lambda)] \quad \text{a.s. } [P_\theta],$$

where $G_J = \{g_j : j \in J(\theta)\}$, recalling that the initial distribution of X_0 under P_θ is assumed to be absolutely continuous with respect to M.

Let $\Lambda_g(\alpha) = \log \rho_g(\alpha)$ and define a new probability measure Q_α on the controlled Markov chain by the "twisting" transformation (cf. [9], [10]):

$$Q_\alpha(B) = E_\theta \left\{ I_B \prod_{0 \le i < n : \phi_i \notin G_J} \frac{r_{\phi_i}(X_{i+1}; \alpha)}{r_{\phi_i}(X_i; \alpha)} e^{-\Lambda_{\phi_i}(\alpha) + \alpha h_{\phi_i}(X_i, X_{i+1})} \right\}, \quad B \in \mathcal{F}_n,$$

where $\Pi_{i \in \emptyset} = 1$ and $\mathcal{F}_n$ is the σ-field defined in (1.6). Letting

$$\begin{aligned} B = \Bigg\{ &\sum_{j \notin J(\theta)} I^{g_j}(\theta, \lambda) T_n(g_j) < (1-\epsilon)\log n, \quad L_n > (1-\delta)\log n, \text{ and} \\ &\sum_{i=1}^{n} I_{\{\phi_i \ne \phi_{i-1}, \phi_i \notin G_J\}} \le \delta \log n \Bigg\}, \end{aligned} \tag{2.23}$$

and noting that $L_n = \sum_{0\le i<n:\phi_i\notin G_J} h_{\phi_i}(X_i, X_{i+1})$, it then follows that for $\alpha > 0$,

(2.24)
$$\begin{aligned} &P_\theta(B) \\ &= \int_B \left\{ \prod_{0\le i<n:\phi_i\notin G_J} \frac{r_{\phi_i}(X_i;\alpha)}{r_{\phi_i}(X_{i+1};\alpha)} \right\} \exp\left\{ -\alpha L_n + \sum_{0\le i<n:\phi_i\notin G_J} \Lambda_{\phi_i}(\alpha) \right\} dQ_\alpha \\ &\le e^{-\alpha(1-\delta)\log n} \int_B \left\{ \prod_{0\le i<n:\phi_i\notin G_J} \frac{r_{\phi_i}(X_i;\alpha)}{r_{\phi_i}(X_{i+1};\alpha)} \right\} \exp\left\{ \sum_{j\notin J(\theta)} \Lambda_{g_j}(\alpha) T_n(g_j) \right\} dQ_\alpha \end{aligned}$$

since $L_n > (1-\delta)\log n$ on B. For $j \notin J(\theta)$,

$$\Lambda_{g_j}(0) = 0,\ (d\Lambda_{g_j}/d\alpha)(0) = \int\int h_{g_j}(x,y) p(x,y;g_j(x),\theta) M(dy) \pi_\theta^{g_j}(dx) = I^{g_j}(\theta,\lambda).$$

Therefore, we can choose $\alpha > 0$ sufficiently small so that $\Lambda_{g_j}(\alpha)/\alpha \le (1+\epsilon/2) I^{g_j}(\theta,\lambda)$. Since $\sum_{j\notin J(\theta)} I^{g_j}(\theta,\lambda) T_n(g_j) < (1-\epsilon)\log n$ on B, it then follows that

$$\sum_{j\notin J(\theta)} \Lambda_{g_j}(\alpha) T_n(g_j) < \alpha(1+\epsilon/2)(1-\epsilon)\log n < \alpha(1-\epsilon/2)\log n \text{ on } B. \tag{2.25}$$

Noting that $C := \max_{g\in G} \sup_{x\in S} r_g(x;\alpha) < \infty, D := \min_{g\in G} \inf_{x\in S} r_g(x;\alpha) > 0$ and that $\sum_{i=1}^n I_{\{\phi_i\ne\phi_{i-1},\phi_i\notin G_J\}} \le \delta \log n$ on B, we obtain that

$$\prod_{0\le i<n:\phi_i\notin G_J} \{r_{\phi_i}(X_i,\alpha)/r_{\phi_i}(X_{i+1},\alpha)\} \le (C/D)^{\delta\log n+1} \quad \text{on} \quad B. \tag{2.26}$$

Indeed, letting $i_1 < \cdots < i_m$ denote the elements of $\{1 \le i \le n : \phi_i \ne \phi_{i-1}, \phi_i \notin G_J\}$, we have $\phi_0 = \phi_1 = \cdots = \phi_{i_1-1}, \phi_{i_1} = \phi_{i_1+1} = \cdots = \phi_{i_2-1}, \ldots, \phi_{i_m} = \phi_{i_m+1} = \cdots = \phi_n$, and therefore the left-hand side of (2.26) can be expressed as

$$\{r_{\phi_0}(X_0,\alpha)/r_{\phi_{n-1}}(X_n,\alpha)\} \prod_{t=1}^m \{r_{\phi_{i_t}}(X_{i_t},\alpha)/r_{\phi_{i_t-1}}(X_{i_t},\alpha)\},$$

from which (2.26) follows since $m \le \delta \log n$ on B.

From (2.24)–(2.26), it follows that by choosing δ sufficiently small,

(2.27)
$$P_\theta(B) \le CD^{-1} \exp(\{-\alpha(1-\delta) + \alpha(1-\epsilon/2) + \delta\log(C/D)\}\log n) \to 0 \text{ as } n\to\infty.$$

Since ϕ is uniformly good, (1.9) yields that

(2.28)
$$\begin{aligned} &P_\theta\left\{ \sum_{i=1}^n I_{\{\phi_i\ne\phi_{i-1},\phi_i\notin G_J\}} > \delta\log n \right\} \\ &\le E_\theta\left(\sum_{i=1}^n I_{\{\phi_i\ne\phi_{i-1},\phi_i\notin G_J \text{ or } \phi_{i-1}\notin G_J\}} \right) / (\delta\log n) \to 0 \end{aligned}$$

as $n\to\infty$. From (2.22), (2.23), (2.27), and (2.28), the desired conclusion (2.18) follows.

3. Construction of asymptotically efficient rules. The main idea behind the adaptive control rule ϕ^* presented in this section is to introduce suitable "uncertainty adjustments" into the certainty-equivalence rule that uses the control law $g_{j(\widehat{\theta}_t)}$ at time t, where $j(\theta)$ is defined in (1.2) and $\widehat{\theta}_n$ is the following weighted maximum likelihood estimate of θ at time n:

$$\widehat{\theta}_n = \arg\max_{\theta\in\Theta} L_n(\theta),$$
$$L_n(\theta) = \sum_{g\in G}(T_n(g))^{-1} \sum_{1\le t\le n, \phi^*_{t-1}=g} \log p(X_{t-1}, X_t; g(X_{t-1}), \theta), \tag{3.1}$$

if the maximizer in (3.1) exists, as is the case when p is a continuous function of θ, since Θ is assumed to be compact. If the maximizer in (3.1) does not exist, then we define $\widehat{\theta}_n$ as an ϵ_n-maximizer of $L_n(\theta)$ in the sense that $L_n(\widehat{\theta}_n) \ge \sup_{\theta\in\Theta} L_n(\theta) - \epsilon_n$, where ϵ_n are positive numbers such that $\lim_{n\to\infty}\epsilon_n = 0$. The asymptotic lower bounds in section 2 provide valuable insights into how the uncertainty adjustments should be made and quantify the need for experimentation with the inferior control laws. In particular, they suggest that the total amount of experimentation with an inferior control law g_j up to time n should be at least of the order $\{c_j(\theta)+o(1)\}\log n$, where the $c_j(\theta)$ solve the minimization problem that defines $c(\theta)$ in (2.5); i.e.,

$$c(\theta) = \sum_{j\notin J(\theta)} c_j(\theta)[\mu^*_\theta - \mu_\theta(g_j)] \text{ and } \inf_{\lambda\in B(\theta)} \sum_{j\notin J(\theta)} c_j(\theta) I^{g_j}(\theta,\lambda) = 1 \text{ if } B(\theta)\neq\emptyset, \tag{3.2}$$
$$c_j(\theta) = 0 \text{ for all } j\notin J(\theta) \text{ if } B(\theta)=\emptyset.$$

Example 1 (continuation). We have shown in the multiarmed bandit problem of Example 1(ii) that $c_j(\theta) = 1/I(\theta_j,\theta^*)$ if $j\notin J(\theta)$. This suggests that to achieve the asymptotic lower bound $(c(\theta)+o(1))\log n$ for the regret, the sampling rule should take $(1/I(\theta_j,\theta^*)+o(1))\log n$ observations, up to stage n, from an inferior population Π_j to determine whether it is indeed inferior. If the sampling rule should indeed attain the asymptotic lower bound, then it would take $n - O(\log n)$ observations, up to stage n, from the population with mean $\mu(\theta^*)$, so we can regard the value of $\mu(\theta^*)$ as known with relatively negligible uncertainty in this case. The problem of determining whether Π_j has a larger mean than $\mu(\theta^*)$ then becomes that of testing the null hypothesis $H_j : \mu(\theta_j) > \mu(\theta^*)$. The theory of optimal stopping and sequential analysis shows that subject to the constraint that the probability of rejecting H_j when it is true be $\le\alpha$, the expected number of observations from Π_j of a sequential test under the alternative hypothesis is at least $\{1/I(\theta_j,\theta^*)+o(1)\}|\log\alpha|$ as $\alpha\to 0$, and there are sequential tests based on generalized likelihood ratio statistics or mixture likelihood ratio statistics that attain this asymptotic lower bound for the expected sample size. The construction of asymptotically efficient sampling rules in section 4 of [16] uses this sequential testing theory, with $|\log\alpha| \sim |\log n|$, although the procedure is described there in terms of certain "upper confidence bounds."

Example 2 (continuation). Here the $c_j(\theta)$ are considerably more complicated than those in Example 1. The means of the normal populations Π_1,Π_2,Π_3 are $\gamma, \xi+1$, and ξ^2, involving only two unknown parameters γ (which has to be learned from Π_1) and ξ (which can be learned from Π_2 or Π_3). If $\xi+1>\gamma$ and $\xi+1\ge\xi^2$, sampling from the superior population Π_2 would give information about the means of both Π_2 and Π_3, and therefore the same argument as that in Example 1 yields $c_1(\gamma,\xi) = 2/(\xi+1-\gamma)^2$, $c_3(\gamma,\xi)=0$, which in turn gives (2.12) as the solution of the minimization problem (2.5).

In the case $\xi^2 > \gamma$ and $\xi^2 \geq \xi+1$, sampling from the superior population Π_3 would give information about $|\xi|$ but not about the sign of ξ, which has to be learned from Π_2. If $\xi^2 \geq \xi + 1$ and $\xi^2 \geq -\xi + 1$ or, equivalently, if $\xi \notin ((-1-\sqrt{5})/2, (1-\sqrt{5})/2)$, then $B(\gamma,\xi) = \{(\widetilde{\gamma},\widetilde{\xi}) : |\widetilde{\xi}| = |\xi|, \max(\widetilde{\gamma},\widetilde{\xi}+1) > \xi^2\} = \{(\widetilde{\gamma},\xi) : \widetilde{\gamma} > \xi^2\}$, and the same argument as in Example 1 yields $c(\gamma,\xi) = 2/(\xi^2-\gamma)$, $c_1(\gamma,\xi) = 2/(\xi^2-\gamma)^2$, $c_2(\gamma,\xi) = 0$. On the other hand, if $-\xi+1 > \xi^2$, then $B(\gamma,\xi) = \{(\widetilde{\gamma},\widetilde{\xi}) : |\widetilde{\xi}| = |\xi|, \widetilde{\gamma} > \xi^2$ or $\widetilde{\xi}+1 > \xi^2\}$, and putting this in (2.5) yields

$$c_1(\gamma,\xi) = \frac{1}{(\xi^2-\gamma)^2/2}, \quad c_2(\gamma,\xi) = \frac{1}{(2\xi)^2/2}, \quad c(\gamma,\xi) = \frac{2}{\xi^2-\gamma} + \frac{\xi^2-\xi-1}{2\xi^2}. \tag{3.3}$$

In the case $\gamma > \max(\xi+1, \xi^2)$, (2.13) yields

$$c_2(\gamma,\xi) = \pi(\gamma,\xi)/\psi_{\pi(\gamma,\xi)}(\widetilde{\xi}_{\pi(\gamma,\xi)}), \quad c_3(\gamma,\xi) = (1-\pi(\gamma,\xi))/\psi_{\pi(\gamma,\xi)}(\widetilde{\xi}_{\pi(\gamma,\xi)}), \tag{3.4}$$

where $\psi_\pi(\widetilde{\xi})$, $\widetilde{\xi}_\pi$, and $\pi(\gamma,\xi)$ are defined in the two sentences following (2.13). For the case $\gamma = \xi + 1 \geq \xi^2$, sampling from Π_2 will give information about ξ, from which one can learn that Π_3 has mean ξ^2, and $B(\gamma,\xi) = \emptyset$ in this case. If $\gamma = \xi^2 \geq \xi + 1$ with $\xi \notin (-(1-\sqrt{5})/2, (1-\sqrt{5})/2)$, then $\xi^2 \geq \xi + 1$ and $\xi^2 \geq -\xi + 1$, and knowledge of ξ^2 will show that Π_2 does not have a larger mean than Π_3, so $B(\gamma,\xi) = \emptyset$ in this case. If $\gamma = \xi^2 > \xi + 1$ and $(-1-\sqrt{5})/2 < \xi < (1-\sqrt{5})/2$, then $-\xi + 1 > \xi^2$ and $B(\gamma,\xi) = \{(\gamma,-\xi)\}$, so putting this in (2.5) yields

$$J(\gamma,\xi) = \{1,3\}, \quad c_2(\gamma,\xi) = 2/(2\xi)^2, \quad c(\gamma,\xi) = (\xi^2-\xi-1)/(2\xi^2). \tag{3.5}$$

The main idea behind the uncertainty adjustments, presented below, to certainty-equivalence rules in controlled Markov chains is to apply sequential testing theory to assess whether an inferior-looking control law is indeed inferior on the basis of all the current and past observations. We shall use sequential likelihood ratio tests of composite hypotheses in general stochastic systems to test the null hypothesis that θ belongs to Θ_i, with prescribed error probability of wrongly rejecting the null hypothesis when it is true and with asymptotically minimal expected waiting time to reject the null hypothesis when it is false. In the present context, the "waiting time" has to be interpreted broadly as a weighted sum of the number of times that an inferior control law g_j is used. Because of switching costs and because of the technical difficulties in controlled Markov chains due to the change of the transition probability function P^g whenever the control law is changed, we shall designate blocks of successive times to use control law g_j for an entire block if the sequential likelihood ratio test performed at the beginning of the block does not reject the hypothesis that θ belongs to Θ_j.

Since θ is unknown, it is natural to replace $c_j(\theta)$ by $c_j(\widehat{\theta}_t)$, where $\widehat{\theta}_t$ is the weighted maximum likelihood estimate defined in (3.1), as in the "certainty-equivalent" testing phase of the control scheme described below. This certainty equivalence approach raises the question concerning how well $c_j(\widehat{\theta}_t)$ approximates $c_j(\theta)$. When one does not have enough information to estimate θ well, an alternative approach is to ignore the constants $c_j(\theta)$ and to experiment equally with each stationary control law, as is done in the following control scheme during its "evenly allocated" testing phase. The control scheme takes an integer $a \geq 2$ and initializes with a common control law for times $1, \ldots, a$.

Outline of control scheme between times a^i and a^{i+1}. Let n_i be positive integers such that

$$n_i \sim i/\log i \quad \text{as} \quad i \to \infty. \tag{3.6}$$

For fixed i, we now describe our control scheme at times $n \in \{a^i+1, \ldots, a^{i+1}\}$, which we partition into $m(i) := [(a^{i+1}-a^i)/n_i]$ blocks of consecutive integers, each block of length n_i except possibly the last one whose length may range from n_i to $2n_i - 1$. Label these blocks as $B^i_1, \ldots, B^i_{m(i)}$ so that the mth block begins at time $\nu_i(m) := a^i + 1 + (m-1)n_i$ for $1 \le m \le m(i)$. To begin with, at time a^i compute the weighted maximum likelihood estimate $\widehat{\theta}_{a^i}$ of θ. To this estimate corresponds a set $\{g_j : j \in J(\widehat{\theta}_{a^i})\}$ of apparently optimal stationary control laws, where $J(\theta)$ is defined in (2.3). We use $c_j(\widehat{\theta}_{a^i})$ to define below the "certainty-equivalent" testing phase (during the period from time a^i+1 to a^{i+1}), whose objective is to test sequentially whether $\theta \notin \cup_{j \in J(\widehat{\theta}_{a^i})} \Theta_j$. The certainty-equivalent testing phase is continued until we either (i) switch to the "evenly allocated" testing phase or (ii) terminate testing and use the same (apparently optimal) stationary control law up to time a^{i+1}. This adaptive control rule will be denoted by ϕ^*. Let $\mathcal{C}_i$ denote the set of times belonging to all those blocks B^i_m that begin with certainty-equivalent testing (at $\nu_i(m)$). Let

$$\mathcal{C} = \bigcup_{s=1}^{\infty} \mathcal{C}_s, \quad \tau_n(g) = \sum_{t=0}^{n-1} I_{\{t \in \mathcal{C}, \phi^*_t = g\}} \quad \text{for} \quad g \in G. \tag{3.7}$$

Thus, $\tau_n(g)$ is the total number of times $t < n$, within these certainty-equivalent-tested blocks, that use the control law $g \in G$. For $t \in \mathcal{C}$, let

$$\widehat{G}_{J,t} = \{g_j : j \in J(\widehat{\theta}_{a^s})\} \quad \text{if} \quad t \in \mathcal{C}_s, \tag{3.8}$$

which is the set of apparently optimal stationary control laws used for the certainty-equivalent test (3.11) below.

The certainty-equivalent testing phase. During the first L or fewer blocks of the certainty-equivalent testing phase, we use in succession the stationary control laws $g_j (j \in L)$ with $\tau_{a^i}(g_j) < n_i$. Suppose $\{1 \le j \le L : j \notin J(\widehat{\theta}_{a^i})$ and $\tau_{a^i}(g_j) < (\log a^i)[c_j(\widehat{\theta}_{a^i}) \wedge \log i]\} = \{j_1, \ldots, j_N\}$. The next blocks of stages use g_{j_1} until time $\nu_i(m_1) - 1$ and then use g_{j_2} until time $\nu_i(m_2) - 1$, etc., where

$$m_k = \inf\{m > m_{k-1} : \tau_{\nu_i(m)}(g_{j_k}) \ge (\log a^i)[c_{j_k}(\widehat{\theta}_{a^i}) \wedge \log i]\}, \quad k = 1, \ldots, N. \tag{3.9}$$

For $m \ge m_N$, alternate using the stationary control laws g_j (one for each block of consecutive times) that satisfy either (i) $j \notin J(\widehat{\theta}_{a^i})$ and

$$\tau_{\nu_i(m)}(g_j) \le (2\log a^i)\{c_j(\widehat{\theta}_{a^i}) \wedge \log i\} + n_i \ \text{ and } \ g_j \ \text{ has not been eliminated} \tag{3.10a}$$

or (ii) $j \in J(\widehat{\theta}_{a^i})$ and

$$\tau_{\nu_i}(m)(g_j) \le (2\log a^i)\log i + n_i. \tag{3.10b}$$

Sequential testing of the hypotheses $H_j : \theta \in \Theta_j$, $j \notin J(\widehat{\theta}_{a^i})$, is performed at times $\nu_i(m)$ with $m \ge m_N$, and we eliminate g_j from further use through time a^{i+1} once the hypothesis H_j is rejected. Rejection of H_j occurs at the first time $n = \nu_i(m)$ with $m \ge m_N$ when

$$\inf_{\lambda \in \Theta_j} \max \left\{ \frac{\int \Pi_{1 \le t \le n, t-1 \in \mathcal{C}, \phi^*_{t-1} \notin \widehat{G}_{J,t}} p(X_{t-1}, X_t; \phi^*_{t-1}(X_{t-1}), \theta) dF(\theta)}{\Pi_{1 \le t \le n, t-1 \in \mathcal{C}, \phi^*_{t-1} \notin \widehat{G}_{J,t}} p(X_{t-1}, X_t; \phi^*_{t-1}(X_{t-1}), \lambda)}, \right.$$
$$\left. \frac{\int \Pi_{1 \le t \le n, T_{t-1}(\phi^*_{t-1}) \ge a^{i-1}/L} \, p(X_{t-1}, X_t; \phi^*_{t-1}(X_{t-1}), \theta) dF(\theta)}{\Pi_{1 \le t \le n, T_{t-1}(\phi^*_{t-1}) \ge a^{i-1}/L} \, p(X_{t-1}, X_t; \phi^*_{t-1}(X_{t-1}), \lambda)} \right\} \ge ia^i, \tag{3.11}$$

where $\inf \emptyset = \infty$, $\Pi_{t=\emptyset} = 1$, F is a probability measure on Θ such that $F(A) > 0$ for all open subsets A of Θ, and $\mathcal{C}$ and $\widehat{G}_{J,t}$ are defined in (3.7) and (3.8). Certainty-equivalent testing is terminated when (3.10a) fails for all $j \notin J(\widehat{\theta}_{a^i})$. If only one stationary control law g_{j^*} is not eliminated at the termination of certainty-equivalent testing, we use g_{j^*} up to time a^{i+1}. Otherwise we switch to the evenly allocated testing phase.

The evenly allocated testing phase. This testing phase does not use the maximum likelihood estimate $\widehat{\theta}_{a^i}$, its associated set $J(\widehat{\theta}_{a^i})$, and the estimates $c_j(\widehat{\theta}_{a^i})$ that have been used in (3.9)–(3.11). Sequential testing of the hypotheses $H_j : \theta \in \Theta_j$ is performed at the times $\nu_i(m)$ for those g_j not yet eliminated (during the times $\nu_i(m')$ between a^i and a^{i+1} with $m' < m$, which include the times of certainty-equivalent testing) in succession in ascending order of j, and we eliminate g_j from further use through time a^{i+1} once the hypothesis H_j is rejected. We reject H_j at the test time $n = \nu_i(m)$ if

$$\inf_{\lambda\in\Theta_j} \max\left\{ \frac{\int \Pi_{1\le t\le n,\phi^*_{t-1}=g_j} p(X_{t-1},X_t;\phi^*_{t-1}(X_{t-1}),\theta)dF(\theta)}{\Pi_{1\le t\le n,\phi^*_{t-1}=g_j} p(X_{t-1},X_t;\phi^*_{t-1}(X_{t-1}),\lambda)}\right. , \tag{3.12}$$

$$\left. \frac{\int \Pi_{1\le t\le n,T_{t-1}(\phi^*_{t-1})\ge a^{i-1}/L}\ p(X_{t-1},X_t;\phi^*_{t-1}(X_{t-1}),\theta)dF(\theta)}{\Pi_{1\le t\le n,T_{t-1}(\phi^*_{t-1})\ge a^{i-1}/L}\ p(X_{t-1},X_t;\phi^*_{t-1}(X_{t-1}),\lambda)} \right\} \ge ia^i.$$

The "even" allocation rule alternates using the stationary control laws that have not been eliminated, one for each block B^i_m of consecutive times. The testing phase terminates as soon as all except one stationary control law have been eliminated, and we use the remaining stationary control law up to time a^{i+1}. For example, a typical pattern of the evenly allocated phase, sampling from four control laws labeled $1, 2, 3, 4$, is

$$1\cdots1\ 2\cdots2\ 3\cdots3\ 4\cdots4\ 1\cdots1\ 2\cdots2_{\uparrow}3\cdots3\ 4\cdots4\ 1\cdots1\ 3\cdots3\ 4\cdots4_{\downarrow}1\cdots1\ 3\cdots3\ 1\cdots1_{*},$$

where $\uparrow$ denotes the time at which 2 is eliminated, $\downarrow$ denotes the time at which 4 is eliminated, and $*$ denotes the time at which the testing phase terminates with the elimination of 1, leaving behind only the control law 3.

In the certainty-equivalent testing phase, we consider the maximum of two mixture likelihood ratio statistics instead of combining them into a single mixture likelihood ratio because we have different roles in mind for the two statistics. One of them has the form

$$\frac{\int \prod_{1\le t\le n,T_{t-1}(\phi^*_{t-1})\ge a^{i-1}/L} p(X_{t-1},X_t;\phi^*_{t-1}(X_{t-1}),\theta)dF(\theta)}{\prod_{1\le t\le n,T_{t-1}(\phi^*_{t-1})\ge a^{i-1}/L} p(X_{t-1},X_t;\phi^*_{t-1}(X_{t-1}),\lambda)}. \tag{3.13}$$

Controls from G that have been used most often as in (3.13) would thus provide accurate information about many of the characteristics of the unknown parameter but are incapable of distinguishing the true parameter θ_0 from the candidate values in $B(\theta_0)$, and therefore may not be able to settle whether $H_j : \theta \in \Theta_j$ is true. The controls in the complement of $\widehat{G}_{J,t}$, which make up the other mixture likelihood ratio statistic, are therefore needed to distinguish θ_0 from $B(\theta_0)$, but they would be used relatively infrequently. Similar reasoning has led us to replace the usual maximum likelihood estimate by a weighted version with weights $(T_n(g))^{-1}$ in (3.1). Since $\inf_{\lambda\in\Theta_j} = 1/\sup_{\lambda\in\Theta_j}$, the $\inf_{\lambda\in\Theta_j}$ in (3.11) is essentially tantamount to taking the

supremum of the denominator of (3.13) over $\lambda \in \Theta_j$, which is typically done in generalized likelihood ratio tests of the composite null hypothesis $H_j : \theta \in \Theta_j$. Our modification of the usual generalized likelihood ratio statistics consists of replacing $\sup_{\theta\in\Theta}$ by an integral with respect to a probability measure on Θ in the numerator of (3.13) and replacing a single likelihood ratio statistic by the maximum of two likelihood ratio statistics. The evenly allocated testing phase does not make use of $\widehat{\theta}_{a^i}$ and $\widehat{G}_{J,t}$. To test $H_j : \theta \in \Theta_j$, it uses the maximum of (3.13) and another mixture likelihood ratio statistic, which has the form (3.13) but with $T_{t-1}(\phi^*_{t-1}) \geq a^{i-1}/L$ replaced by $\phi^*_{t-1} = g_j$ and which is therefore based on data generated by the stationary control law g_j. For the special case of controlled independently and identically distributed (i.i.d.) processes, further details of the statistical ideas behind the modifications (3.11) and (3.12) of the classical generalized likelihood ratio statistics, together with illustrative examples and some variants of the adaptive control rule ϕ^*, are given in [8].

Throughout what follows we shall let θ_0 denote the true value of the unknown parameter. We shall also let $\mathbf{E}^\phi_x$ denote expectation under the probability measure $\mathbf{P}^\phi_x$ of the controlled Markov chain starting at x and using control rule ϕ, assuming the true value θ_0 of the parameter. Theorem 2 below shows that the regret $R_n(\theta_0)$ of ϕ^* satisfies

$$R_n(\theta_0) \sim \sum_{j\notin J(\theta_0)} c_j(\theta_0)[\mu^*_{\theta_0} - \mu_{\theta_0}(g_j)] \log n \tag{3.14}$$

under regularity conditions (C1)–(C5) in addition to those assumed at the beginning of section 1. Although we still use the notations (2.1)–(2.4) and define the $c_j(\theta)$ by (3.2) in Theorem 2, we do not assume condition (2.14) of Theorem 1. Our objective here is to establish (3.14), irrespective of whether it is an asymptotic lower bound for the regret of uniformly good rules. For $\delta > 0$, let $B_\delta(\theta_0)$ denote the open δ-neighborhood of $B(\theta_0)$; i.e., $B_\delta(\theta_0) = \{\theta \in \Theta : \inf_{\lambda\in B(\theta_0)} \rho(\theta, \lambda) < \delta\}$ ($B_\delta(\theta_0) = \emptyset$ if $B(\theta_0) = \emptyset$), where ρ denotes the metric of the compact metric space Θ.

(C1) For every $\epsilon > 0$, there exists $\delta > 0$ such that if $\rho(\theta_0, \theta) \leq \delta$ then $J(\theta) \subset J(\theta_0)$ and $\max_{j\notin J(\theta_0)} |c_j(\theta) - c_j(\theta_0)| < \epsilon$. Moreover, there exist ξ and $\delta^* > 0$ such that $\max_{j\notin J(\theta)} |c_j(\theta)| \leq \xi$ if $\rho(\theta_0, \theta) \leq \delta^*$.

(C2) $I^g(\theta_0, \theta)$ is a continuous function of θ for every $g \in G$.

(C3) $\max_{g\in G} I^g(\theta_0, \lambda) > 0$ for all $\lambda \in \cup_{j\in J(\theta_0)}\Theta_j - \{\theta_0\}$, $\inf_{\lambda\in\Theta_i\cap B(\theta_0)} I^{g_i}(\theta_0, \lambda) > 0$ and $\inf_{\lambda\in\Theta_i\setminus B_\delta(\theta_0)} \max_{j\in J(\theta_0)} I^{g_j}(\theta_0, \lambda) > 0$ for all $i \notin J(\theta_0)$ and $\delta > 0$ ($\inf \emptyset = \infty$).

(C4) For every $\theta \in \Theta$ and $g \in G$, there exist $\delta_\theta > 0$ and $r_\theta > 2$ such that

$$\sup_{x\in S} \mathbf{E}^g_x \left\{ \sup_{\lambda:\rho(\theta,\lambda)\leq\delta_\theta} \left| \log \frac{p(x, X_1; g(x), \theta_0)}{p(x, X_1; g(x), \lambda)} \right|^{r_\theta} \right\} < \infty$$

and, as $\delta \to 0$,

$$\sup_{x\in S} \mathbf{P}^g_x \left\{ \sup_{\lambda:\rho(\theta,\lambda)\leq\delta} \left| \frac{p(x, X_1; g(x), \theta)}{p(x, X_1; g(x), \lambda)} - 1 \right| \geq \epsilon \right\} \to 0 \text{ for all } \epsilon > 0.$$

(C5) For every $\theta \in \Theta$ and $g \in G$, as $n \to \infty$,

$$\sup_{x\in S} \left| \mathbf{E}^g_x \left\{ \frac{1}{n} \sum_{i=1}^n \log \frac{p(X_{t-1}, X_t; g(X_{t-1}), \theta_0)}{p(X_{t-1}, X_t; g(X_{t-1}), \theta)} \right\} - I^g(\theta_0, \theta) \right| \to 0.$$

Conditions (C1) and (C2) are continuity assumptions on $c_j(\theta)$ at $\theta = \theta_0$ and on $I^g(\theta_0, \theta)$. (C3) ensures that under ϕ^* one can estimate θ_0 consistently by the method of maximum likelihood, as will be shown in the proof of Theorem 2. From the definition (2.4) of $B(\theta_0)$, it follows that $\max_{j\in J(\theta_0)} I^{g_j}(\theta_0, \lambda) > 0$ if $\lambda \notin B(\theta_0) \cup (\cup_{j\in J(\theta_0)}\Theta_j)$, and the last inequality of (C3) requires this to be uniformly bounded away from zero for $\lambda \in \Theta_i \backslash B_\delta(\theta_0)$ with $i \notin J(\theta_0)$. Conditions (C4) and (C5) are natural moment and ergodicity assumptions on the log-likelihood ratio statistics. The uniformity over $x \in S$ in these assumptions enables us to get around difficulties with controlled Markov chains whose transition probability function $P_\theta^{g(x)}(x, A)$ is changed when the control law is changed. In section 4 we make use of martingale theory and uniform integrability to analyze the adaptive control rule ϕ^* in the proof of the following.

THEOREM 2. *Under* (C1)–(C5), *for every* $x \in S$ *and* $j \notin J(\theta_0)$,

$$\lim_{n\to\infty} \mathbf{E}_x^{\phi^*}(T_n(g_j))/\log n = c_j(\theta_0),$$

and therefore the regret $R_n(\theta_0)$ *of the rule* ϕ^* *satisfies* (3.14). *Moreover,* $S_n(\theta_0) = o(\log n)$, *where* $S_n(\theta_0)$ *is the expected number* (1.8) *of times that* ϕ^* *switches from one control law in* G *to another, not both optimal, up to stage* n.

The $c_j(\theta)$ are obtained by solving a constrained optimization problem in (3.2), which may be quite difficult in certain cases. Although the certainty-equivalent testing phase involves $J(\theta)$ and $c_j(\theta)$, these quantities are not used in the evenly allocated testing phase. In cases where the $c_j(\theta)$ are difficult to determine or fail to satisfy the continuity assumption (C1), an obvious modification of the adaptive control rule ϕ^* is to abandon the certainty-equivalent testing phase. Thus, partitioning $\{a^i + 1, \ldots, a^{i+1}\}$ into $m(i)$ blocks of consecutive integers so that the mth block B_m^i begins at time $\nu_i(m) := a^i + 1 + (m-1)n_i$, this modified rule $\widetilde{\phi}$ performs sequential testing of the hypotheses $H_j : \theta \in \Theta_j$ at the times $\nu_i(m)$ for those g_j not yet eliminated (during the times $\nu_i(m')$ with $m' < m$), in succession in ascending order of j, and eliminates g_j from further use through time a^{i+1} once the hypothesis H_j is rejected. Rejection of H_j occurs at the test time $n = \nu_i(m)$ if (3.12) holds. The rule $\widetilde{\phi}$ alternates using the stationary control laws that have not been eliminated, one for each block B_m^i of consecutive times. If all except one stationary control law have been eliminated, then $\widetilde{\phi}$ uses the remaining stationary control law up to time a^{i+1}. The following theorem shows that although this simpler rule $\widetilde{\phi}$ may be less efficient than ϕ^*, which attains the asymptotic lower bound $(c(\theta_0) + o(1))\log n$ for the regret, $\widetilde{\phi}$ still has a regret of the order $O(\log n)$.

THEOREM 3. *Under* (C2)–(C5), *the rule* $\widetilde{\phi}$ *satisfies* $R_n(\theta_0) = O(\log n)$ *and* $S_n(\theta_0) = o(\log n)$.

4. Martingale inequalities, uniform integrability, and proof of Theorems 2 and 3. We first consider some simple implications of conditions (C1)–(C5). Let $\epsilon > 0$ and take $2 < r'_\theta < r_\theta$. By (C2) together with (C4) for every $\theta \in \Theta$ we can choose $0 < \delta'_\theta \le \delta_\theta$ such that

$$\sup_{\lambda:\rho(\theta,\lambda)\le\delta'_\theta} |I^g(\theta_0, \theta) - I^g(\theta_0, \lambda)| < \epsilon \text{ for all } g \in G \text{ and} \tag{4.1}$$

$$\sup_{x\in S, g\in G} \mathbf{E}_x^g \left\{ \sup_{\lambda:\rho(\theta,\lambda)\le\delta'_\theta} \left| \log \frac{p(x, X_1; g(x), \theta_0)}{p(x, X_1; g(x), \lambda)} - \log \frac{p(x, X_1; g(x), \theta_0)}{p(x, X_1; g(x), \theta)} \right|^{r'_\theta} \right\} \le \epsilon^{r'_\theta}, \tag{4.2}$$

as will be explained in the next paragraph. Since Θ is compact, there exist finitely many points $\theta_1, \ldots, \theta_K$ such that

$$\Theta = \cup_{k=0}^{K} \{\lambda : \rho(\theta_k, \lambda) < \delta'_{\theta_k}\}. \tag{4.3}$$

By (C5) we can choose a positive integer D large enough so that

(4.4)

$$\sup_{x \in S, g \in G, k \le K} \left| \mathbf{E}_x^g \left\{ \frac{1}{n} \sum_{t=1}^{n} \log \frac{p(X_{t-1}, X_t; g(X_{t-1}), \theta_0)}{p(X_{t-1}, X_t; g(X_{t-1}), \theta_k)} \right\} - I^g(\theta_0, \theta_k) \right| \le \epsilon \text{ for all } n \ge D.$$

Concerning (4.2), first note that by (C4), $\sup_{x \in S} \mathbf{P}_x^g(\Omega(\delta, \eta; x)) \to 0$ as $\delta \to 0$ for every fixed $\eta > 0$, where $\Omega(\delta, \eta; x) = \{\sup_{\lambda : \rho(\theta, \lambda) \le \delta} |p(x, X_1; g(x), \theta)/p(x, X_1; g(x), \lambda) - 1| \ge \eta\}$. For sufficiently small η, $|\log y| < 2\eta$ if $|y - 1| < \eta$, and therefore, for $0 < \delta < \delta_\theta$,

$$\begin{aligned}
&\sup_{x \in S} \mathbf{E}_x^g \left\{ \sup_{\lambda : \rho(\theta, \lambda) \le \delta} \left| \log \frac{p(x, X_1; g(x), \theta)}{p(x, X_1; g(x), \lambda)} \right|^{r'_\theta} \right\} \\
&\le (2\eta)^{r'_\theta} + \sup_{x \in S} \mathbf{E}_x^g \left\{ \sup_{\lambda : \rho(\theta, \lambda) \le \delta} \left| \log \frac{p(x, X_1; g(x), \theta)}{p(x, X_1; g(x), \lambda)} \right|^{r'_\theta} I_{\Omega(\delta, \eta; x)} \right\} \\
&\le (2\eta)^{r'_\theta} \\
&\quad + \sup_{x \in S} \left\{ \mathbf{E}_x^g \left[\sup_{\lambda : \rho(\theta, \lambda) \le \delta_\theta} \left| \log \frac{p(x, X_1; g(x), \theta_0)}{p(x, X_1; g(x), \lambda)} - \log \frac{p(x, X_1; g(x), \theta_0)}{p(x, X_1; g(x), \theta)} \right|^{r_\theta} \right] \right\}^{r'_\theta / r_\theta} \\
&\quad \times \{\mathbf{P}_x^g(\Omega(\delta, \eta; x))\}^{1 - r'_\theta / r_\theta} \le (2\eta)^{r'_\theta} + o(1) \text{ as } \delta \to 0
\end{aligned}$$

by (C4), where we have used Hölder's inequality to obtain the second inequality.

We next make use of martingale theory to analyze the log-likelihood ratio statistics from the controlled Markov chain using the adaptive control rule ϕ^*. For $t \ge 0$, let $\mathcal{F}_t$ be the σ-field defined by (1.6). The control rule ϕ^* uses the same stationary control law for an entire block of stages $\nu_i(m), \ldots, \nu_i(m+1) - 1$, with the choice of the control law determined at the beginning of the block. Define a sequence of positive integers h_s such that $D \le h_s - h_{s-1} \le 2D - 1$ for $s > 1$ and all the $\nu_i(m)$ with $i \ge D$ belong to the sequence, where D is given by (4.4). For example, take $h_0 = 0, h_1 = a^D$, and for $s > 1$ let $h_s = h_{s-1} + D$ except when $h_s = \nu_i(m)$, for which we may change the above recursive definition of h_s to $h_s = h_{s-1} + D + r$, with $0 \le r < D$ being the remainder obtained when $\nu_i(m) - \nu_i(m-1)$ is divided by D. Since ϕ^* uses the same stationary control law for $h_{s-1}, \ldots, h_s - 1$ on the basis of observations prior to h_{s-1}, it follows that for $h_{s-1} < t \le h_s$ and $g \in G$, $\{\phi^*_{t-1} = g\} \in \mathcal{F}_{h_{s-1}}$. Therefore, by (4.4),

(4.5)

$$\begin{aligned}
\sup_{n \ge 1, g \in G, k \le K} \Bigg| &\frac{1}{T_{h_n}(g)} \sum_{s=1}^{n} \sum_{h_{s-1} < t \le h_s} \mathbf{E}_x^{\phi^*} \left\{ \log \frac{p(X_{t-1}, X_t; \phi^*_{t-1}(X_{t-1}), \theta_0)}{p(X_{t-1}, X_t; \phi^*_{t-1}(X_{t-1}), \theta_k)} \middle| \mathcal{F}_{h_{s-1}} \right\} \\
&\cdot I_{\{\phi^*_{t-1} = g\}} - I^g(\theta_0, \theta_k) \Bigg| I_{\{T_{h_n}(g) \ge D\}} \le \epsilon,
\end{aligned}$$

noting that $T_{h_n}(g) = \sum_{s=1}^{n} \sum_{h_{s-1} \le i < h_s} I_{\{\phi_i^* = g\}}$ and

$$\sum_{t=h_{s-1}+1}^{h_s} \mathbf{E}_x^{\phi^*} \left\{ \log \frac{p(X_{t-1}, X_t; \phi_{t-1}^*(X_{t-1}), \theta_0)}{p(X_{t-1}, X_t; \phi_{t-1}^*(X_{t-1}), \theta_k)} \middle| \mathcal{F}_{h_{s-1}} \right\} I_{\{\phi_{t-1}^* = g\}} \tag{4.6}$$

$$= \mathbf{E}_y^g \left\{ \sum_{t=1}^{h_s - h_{s-1}} \log \frac{p(X_{t-1}, X_t; g(X_{t-1}), \theta_0)}{p(X_{t-1}, X_t; g(X_{t-1}), \theta_k)} \right\} \text{ on } \{\phi_{h_{s-1}}^* = g, X_{h_{s-1}} = y\}.$$

Let

$$\ell_t(\theta) = \log[p(X_{t-1}, X_t; \phi_{t-1}^*(X_{t-1}), \theta_0) / p(X_{t-1}, X_t; \phi_{t-1}^*(X_{t-1}), \theta)]. \tag{4.7}$$

Lemma 1 states a result of [14], and Lemmas 2, 3, and 4 use it to approximate $\sum_{t=1}^{h_n} \ell_t(\theta)\ \chi_t$, where the χ_t are indicator variables (taking values in $\{0, 1\}$).

LEMMA 1. *Let $\{Z_n\}$ be a martingale difference sequence with respect to an increasing sequence of σ-fields $\{\mathcal{B}_n\}$ such that $\sup_n E(|Z_n|^\beta | \mathcal{B}_{n-1}) \le C$ a.s. for some nonrandom constants $\beta > 2$ and $C < \infty$. Let χ_n be $\mathcal{B}_{n-1}$-measurable variables taking values in $\{0, 1\}$ and let $\#_n = \sum_{t=1}^{n} \chi_t$. Then there exists a universal constant A depending only on C and β (and not on the distribution of $\{(\chi_n, Z_n) : n \ge 1\}$) such that for any $\eta > 0$ and $m \ge 1$,*

$$P\left\{ \sup_{n: \#_n \ge m} \left| \sum_{t=1}^{n} Z_t \chi_t \right| / \#_n > \eta \right\} \le A\{m^{-(\beta-1)} \eta^{-\beta} + (\eta^2 m)^{-(\beta-1)}\},$$

$$P\left\{ \max_{1 \le n \le m} \left| \sum_{t=1}^{n} Z_t \chi_t \right| > \eta m \right\} \le A\{m^{-(\beta-1)} \eta^{-\beta} + (\eta^2 m)^{-(\beta-1)}\}.$$

LEMMA 2. *Let $\beta = \min_{0 \le k \le K} r'_{\theta_k} (> 2)$, where the r'_θ are given by (4.2). Define $\ell_t(\theta)$ by (4.7) and $\mathcal{C}, \tau_n(g)$ by (3.7). Then for any $g \in G$ and $0 \le k \le K$,*

$$\mathbf{P}_x^{\phi^*} \left\{ \sup_{n: T_{h_n}(g) \ge m} \left| (T_{h_n}(g))^{-1} \sum_{t=1}^{h_n} \ell_t(\theta_k) I_{\{\phi_{t-1}^* = g\}} - I^g(\theta_0, \theta_k) \right| > 2\epsilon \right\} = O(m^{-(\beta-1)}),$$

$$\mathbf{P}_x^{\phi^*} \left\{ \sup_{n: \tau_{h_n}(g) \ge m} \left| (\tau_{h_n}(g))^{-1} \sum_{t=1}^{h_n} \ell_t(\theta_k) I_{\{t-1 \in \mathcal{C}, \phi_{t-1}^* = g\}} - I^g(\theta_0, \theta_k) \right| > 2\epsilon \right\}$$
$$= O(m^{-(\beta-1)}).$$

Proof. Let $Z_s = \sum_{h_{s-1} < t \le h_s} \ell_t(\theta_k) - \mathbf{E}_x^{\phi^*} \{\sum_{h_{s-1} < t \le h_s} \ell_t(\theta_k) | \mathcal{F}_{h_{s-1}}\}$. Then $\{Z_s, \mathcal{B}_s, s \ge 1\}$ is a martingale difference sequence, where $\mathcal{B}_s = \mathcal{F}_{h_s}$. Moreover, since $h_s - h_{s-1} \le 2D - 1$,

$$\sup_{s \ge 1} \mathbf{E}_x^{\phi^*} (|Z_s|^\beta | \mathcal{B}_{s-1}) \le A_{\beta, D} \sup_{y \in S, g \in G} \sum_{t=1}^{2D-1} \mathbf{E}_y^g |\ell_t(\theta_k)|^\beta$$

for some positive constant $A_{\beta,D}$ that depends only on β and D. For fixed $g \in G$, let $\chi_s = I_{\{\phi_{h_{s-1}}^* = g\}}$, which is $\mathcal{B}_{s-1}$ measurable, and note that

$$\sum_{h_{s-1} < t \le h_s} \ell_t(\theta_k) I_{\{\phi_{t-1}^* = g\}} = \chi_s \sum_{h_{s-1} < t \le h_s} \ell_t(\theta_k), \tag{4.8}$$

since ϕ^* uses the same stationary control law at the times $h_{s-1}, \ldots, h_s - 1$. Moreover,

$$D\sum_{s=1}^{n}\chi_s \le \sum_{s=1}^{n}(h_s - h_{s-1})\chi_s = T_{h_n}(g) \le (2D-1)\sum_{s=1}^{n}\chi_s.$$

Therefore it follows from Lemma 1 that

$$\mathbf{P}_x^{\phi^*}\left\{\sup_{n:T_{h_n}(g)\ge m}\left|\sum_{s=1}^{n} Z_s\chi_s\right| / T_{h_n}(g) > \epsilon\right\} = O(m^{-(\beta-1)}). \tag{4.9}$$

Since $\sum_{t=1}^{h_n}\ell_t(\theta_k)I_{\{\phi^*_{t-1}=g\}} = \sum_{s=1}^{n} Z_s\chi_s + \sum_{s=1}^{n}\sum_{h_{s-1}<t\le h_s}\mathbf{E}_x^{\phi^*}\{\ell_t(\theta_k)|\mathcal{F}_{h_{s-1}}\}I_{\{\phi^*_{t-1}=g\}}$ by (4.8), the desired conclusion for $T_{h_n}(g)$ follows from (4.5) and (4.9). The conclusion for $\tau_{h_n}(g)$ can be proved similarly, noting that for $h_{s-1} < t \le h_s$, $\{t-1 \in \mathcal{C}, \phi^*_{t-1} = g\} \in \mathcal{F}_{h_{s-1}}$.

LEMMA 3. *With β defined in Lemma 2 and δ'_θ given by (4.1) and (4.2), let χ_n be $\mathcal{F}_{n-1}$-measurable random variables taking values in $\{0,1\}$ and let $\#_n = \sum_{t=1}^{n}\chi_t$. Then for any $0 \le k \le K$,*

$$\mathbf{P}_x^{\phi^*}\left\{\sup_{n:\#_n\ge m}\left[\sup_{\theta:\rho(\theta_k,\theta)\le\delta'_{\theta_k}}\sum_{t=1}^{n}|\ell_t(\theta)-\ell_t(\theta_k)|\chi_t\right]/\#_n > 2\epsilon\right\} = O(m^{-(\beta-1)}).$$

Proof. Let $\Gamma_k = \{\theta : \rho(\theta_k,\theta) \le \delta'_{\theta_k}\}$. In view of (C4), applying Lemma 1 to $Z_t = \sup_{\theta\in\Gamma_k}|\ell_t(\theta)-\ell_t(\theta_k)| - \mathbf{E}_x^{\phi^*}\{\sup_{\theta\in\Gamma_k}|\ell_t(\theta)-\ell_t(\theta_k)||\mathcal{F}_{t-1}\}$ yields

$$\mathbf{P}_x^{\phi^*}\left\{\sup_{n:\#_n\ge m}\left(\sum_{t=1}^{n} Z_t\chi_t\right)/\#_n > \epsilon\right\} = O(m^{-(\beta-1)}). \tag{4.10}$$

Moreover, by (4.2),

$$\begin{aligned}\sum_{t=1}^{n}\chi_t\mathbf{E}_x^{\phi^*}&\left\{\sup_{\theta\in\Gamma_k}|\ell_t(\theta)-\ell_t(\theta_k)|\,\middle|\,\mathcal{F}_{t-1}\right\}\\ &\le\left(\sum_{t=1}^{n}\chi_t\right)\sup_{y\in S,g\in G}\mathbf{E}_y^g\left(\sup_{\lambda\in\Gamma_k}|\ell_t(\lambda)-\ell_t(\theta_k)|\right)\\ &\le\epsilon\#_n.\end{aligned} \tag{4.11}$$

Since $\sup_{\theta\in\Gamma_k}\sum_{t=1}^{n}|\ell_t(\theta)-\ell_t(\theta_k)|\chi_t \le \sum_{t=1}^{n} Z_t\chi_t + \sum_{t=1}^{n}\chi_t\mathbf{E}_x^{\phi^*}\{\sup_{\theta\in\Gamma_k}|\ell_t(\theta)-\ell_t(\theta_k)||\mathcal{F}_{t-1}\}$, the desired conclusion follows from (4.10) and (4.11).

LEMMA 4. *With the same notation as in Lemma 3, for any $0 \le k \le K$,*

$$\mathbf{P}_x^{\phi^*}\left\{\max_{1\le n\le m}\sup_{\theta:\rho(\theta_k,\theta)\le\delta'_{\theta_k}}\sum_{t=1}^{n}|\ell_t(\theta)-\ell_t(\theta_k)|\chi_t > 2\epsilon m\right\} = O(m^{-(\beta-1)}).$$

Moreover, for any $g \in G$ and $0 \le k \le K$,

$$\sup_{\tau\in\mathcal{T}}\mathbf{P}_x^{\phi^*}\left\{\max_{1\le n\le m}\left|\sum_{t=1}^{h_n}(\ell_t(\theta_k)-I^g(\theta_0,\theta_k))I_{\{t>\tau,\phi^*_{t-1}=g\}}\right| > 3\epsilon h_m\right\} = O(m^{-(\beta-1)}),$$

where $\mathcal{T}$ denotes the class of all stopping times (with respect to $\{\mathcal{F}_n\}$).

Proof. By using the second instead of the first inequality of Lemma 1, we can proceed as in the proof of Lemma 3 to obtain the first conclusion. To prove the second conclusion, let τ be a stopping time and let $\sigma = \inf\{s : h_s \geq \tau\}$. Define Z_s and $\mathcal{B}_s$ as in the proof of Lemma 2 but change the definition of χ_s there to $\chi_s = I_{\{\phi^*_{h_{s-1}}=g, s>\sigma\}}$. Since

$$\{s > \sigma\} = \{s - 1 \geq \sigma\} = \{h_{s-1} \geq \tau\} \in \mathcal{F}_{h_{s-1}} = \mathcal{B}_{s-1},$$

χ_s is $\mathcal{B}_{s-1}$ measurable. Therefore, by Lemma 1 there exists a constant A (which does not depend on σ) such that, for all $m \geq 1$,

$$\mathbf{P}_x^{\phi^*}\left\{\max_{1\leq n\leq m}\left|\sum_{s=1}^{n} Z_s\chi_s\right| > \epsilon m\right\} \leq A\{m^{-(\beta-1)}\epsilon^{-\beta} + (\epsilon^2 m)^{-(\beta-1)}\}.$$

Note that for $n \geq \sigma$, $\sum_{t=1}^{h_n} \ell_t(\theta_k) I_{\{\phi^*_{t-1}=g, t>\tau\}}$ can be written as

$$\sum_{\tau+1\leq t\leq h_\sigma} \ell_t(\theta_k) I_{\{\phi^*_{t-1}=g\}} + \sum_{s=1}^{n} \chi_s \mathbf{E}_x^{\phi^*}\left\{\sum_{h_{s-1}<t\leq h_s} \ell_t(\theta_k)\,\middle|\, \mathcal{F}_{h_{s-1}}\right\} + \sum_{s=1}^{n} Z_s\chi_s.$$

Since $h_s - h_{s-1} \geq D$, it follows from (4.4) that

$$\max_{1\leq s\leq m}\left|\sum_{s=1}^{n} \chi_s \sum_{h_{s-1}<t\leq h_s} \{\mathbf{E}_x^{\phi^*}[\ell_t(\theta_k)|\mathcal{F}_{h_{s-1}}] - I^g(\theta_0,\theta_k)\}\right| \leq \epsilon h_m.$$

Since $h_{\sigma-1} < \tau \leq h_\sigma$ and $h_\sigma - h_{\sigma-1} \leq 2D - 1$, the strong Markov property implies that

$$\mathbf{E}_x^{\phi^*}\left\{\left(\sum_{\tau+1\leq t\leq h_\sigma} |\ell_t(\theta_k)| I_{\{\phi^*_{t-1}=g\}}\right)^{\beta}\,\middle|\, \mathcal{F}_\tau\right\} \leq A_{\beta,D} \sup_{y\in S} \sum_{t=1}^{2D-1} \mathbf{E}_y^g |\ell_t(\theta_k)|^\beta$$

for some positive constant $A_{\beta,D}$ that depends only on β and D. Hence by the Markov inequality

$$\mathbf{P}_x^{\phi^*}\left\{\sum_{\tau+1\leq t\leq h_\sigma} |\ell_t(\theta_k)| I_{\{\phi^*_{t-1}=g\}} > \epsilon m\right\} \leq C\epsilon^{-\beta} m^{-\beta},$$

where $C = A_{\beta,D} \sup_{y\in S} \sum_{t=1}^{2D-1} \mathbf{E}_y^g |\ell_t(\theta_k)|^\beta < \infty$. Since the same constants A and C in the above probability bounds hold for all stopping times τ, these bounds yield the second conclusion of the lemma.

We shall make use of Lemmas 2–4 to prove the following two lemmas from which Theorems 2 and 3 follow easily. Recall that $G_J = \{g_j : j \in J(\theta_0)\}$.

LEMMA 5. *With $\beta > 2$ defined in Lemma 2, for every $\eta > 0$,*

$$\mathbf{P}_x^{\phi^*}\{\rho(\theta_0, \widehat{\theta}_{h_n}) \geq \eta \text{ for some } h_n \geq a^i\} = O((i/\log i)^{-(\beta-1)}). \tag{4.12}$$

Moreover, for any $j \notin J(\theta_0)$, as $n \to \infty$,

$$T_n(g_j)/\log n \to c_j(\theta_0),\ \sum_{i=1}^{n} I_{\{\phi_i^* \neq \phi_{i-1}^*, \phi_i^* \notin G_J \text{ or } \phi_{i-1}^* \notin G_J\}}/\log n \to 0 \text{ a.s. } [\mathbf{P}_x^{\phi^*}]. \tag{4.13}$$

LEMMA 6. *For any $j \notin J(\theta_0)$, $\{T_n(g_j)/\log n, n \geq 2\}$ is uniformly integrable under $\mathbf{P}_x^{\phi^*}$ or $\mathbf{P}_x^{\hat{\phi}}$.*

Proof of Theorem 2. From (4.13) and Lemma 6, it follows that $\mathbf{E}_x^{\phi^*} T_n(g_j)/\log n \to c_j(\theta_0)$ for any $j \notin J(\theta_0)$. This and (3.2) imply (3.14). The uniform integrability of

$$\sum_{i=1}^{n} I_{\{\phi_i^* \neq \phi_{i-1}^*, \phi_i^* \notin G_J \text{ or } \phi_{i-1}^* \notin G_J\}}/\log n,$$

which is $\leq 2\{1 + \sum_{j \notin J(\theta_0)} T_n(g_j)\}/\log n$, follows from Lemma 6. Therefore $S_n(\theta_0) = o(\log n)$ by (4.13).

Proof of Theorem 3. The desired conclusion on $R_n(\theta_0)$ follows from Lemma 6, and that on $S_n(\theta_0)$ can be proved by an argument similar to the proof of the second convergence in (4.13) and the associated uniform integrability in Theorem 2.

The proof of Lemmas 5 and 6 makes use of the following lemma, which applies martingale inequalities to analyze boundary crossing probabilities associated with (3.11) and (3.12).

LEMMA 7. *As in* (3.11) *and* (3.12), *let F be a probability measure on Θ such that $F(A) > 0$ for all open subsets A of Θ. For $a^i < n \leq a^{i+1}$, let*

$$U_n(\lambda) = \frac{\int \prod_{1 \leq t \leq n, t-1 \in \mathcal{C}, \phi_{t-1}^* \notin \hat{G}_{J,t}} p(X_{t-1}, X_t; \phi_{t-1}^*(X_{t-1}), \theta) dF(\theta)}{\prod_{1 \leq t \leq n, t-1 \in \mathcal{C}, \phi_{t-1}^* \notin \hat{G}_{J,t}} p(X_{t-1}, X_t; \phi_{t-1}^*(X_{t-1}), \lambda)}, \tag{4.14}$$

$$\widetilde{W}_{n,j}(\lambda) = \frac{\int \prod_{1 \leq t \leq n, \phi_{t-1}^* = g_j} p(X_{t-1}, X_t; \phi_{t-1}^*(X_{t-1}), \theta) dF(\theta)}{\prod_{1 \leq t \leq n, \phi_{t-1}^* = g_j} p(X_{t-1}, X_t; \phi_{t-1}^*(X_{t-1}), \lambda)}, \tag{4.15}$$

$$W_n(\lambda) = \frac{\int \prod_{1 \leq t \leq n, T_{t-1}(\phi_{t-1}^*) \geq a^{i-1}/L} p(X_{t-1}, X_t; \phi_{t-1}^*(X_{t-1}), \theta) dF(\theta)}{\prod_{1 \leq t \leq n, T_{t-1}(\phi_{t-1}^*) \geq a^{i-1}/L} p(X_{t-1}, X_t; \phi_{t-1}^*(X_{t-1}), \lambda)}. \tag{4.16}$$

If $\theta_0 \in \Theta_j$, then

$$\mathbf{P}_x^{\phi^*}\left\{\inf_{\lambda \in \Theta_j} \max(U_n(\lambda), W_n(\lambda)) \geq ia^i \text{ for some } n > a^i\right\} \leq 2(ia^i)^{-1},$$

$$\mathbf{P}_x^{\phi^*}\left\{\inf_{\lambda \in \Theta_j} \max(\widetilde{W}_{n,j}(\lambda), W_n(\lambda)) \geq ia^i \text{ for some } n > a^i\right\} \leq 2(ia^i)^{-1}.$$

Proof. Note that $\{t-1 \in \mathcal{C}, \phi_{t-1}^* \notin \widehat{G}_{J,t}\} \in \mathcal{F}_{t-1}$ by (3.7) and (3.8). Hence $U_n(\theta_0), W_n(\theta_0)$ and $\widetilde{W}_{n,j}(\theta_0), n > a^i$, are nonnegative martingales with common mean 1. Therefore, if $\theta_0 \in \Theta_j$, then

$$\begin{aligned}
&\mathbf{P}_x^{\phi^*}\left\{\inf_{\lambda \in \Theta_j} \max(U_n(\lambda), W_n(\lambda)) \geq ia^i \text{ for some } n > a^i\right\} \\
&\leq \mathbf{P}_x^{\phi^*}\{U_n(\theta_0) \geq ia^i \text{ for some } n > a^i\} + \mathbf{P}_x^{\phi^*}\{W_n(\theta_0) \geq ia^i \text{ for some } n > a^i\} \\
&\leq \{EU_{a^i+1}(\theta_0) + EW_{a^i+1}(\theta_0)\}/(ia^i).
\end{aligned}$$

Replacing $U_n(\lambda)$ by $\widetilde{W}_{n,j}(\lambda)$ in the above argument proves the second inequality.

Proof of Lemma 5. We first prove (4.12). By Lemma 3 (with $\chi_t = I_{\{\phi_{t-1}^* = g\}}$), for every $g \in G$ and $0 \leq k \leq K$,

(4.17)
$$\mathbf{P}_x^{\phi^*}\left\{\sup_{\theta\in\Gamma_k}\sum_{t=1}^{n}|\ell_t(\theta)-\ell_t(\theta_k)|I_{\{\phi^*_{t-1}=g\}}\ge 2\epsilon T_n(g)\text{ for some }T_n(g)\ge m\right\}=O(m^{-(\beta-1)}).$$

From (4.17) and Lemma 2, it follows that

(4.18)
$$\mathbf{P}_x^{\phi^*}\left\{\inf_{\theta\in\Gamma_k}(T_{h_n}(g))^{-1}\sum_{t=1}^{h_n}\ell_t(\theta)I_{\{\phi^*_{t-1}=g\}}\ge I^g(\theta_0,\theta_k)-4\epsilon\text{ for all }T_{h_n}(g)\ge m\right.$$
$$\left.\text{and every } g\in G \text{ and } 0\le k\le K\right\}\ge 1-O(m^{-(\beta-1)}).$$

By (4.1) and the compactness of Θ, for every $\epsilon>0$, there exists $\delta>0$ such that

(4.19)
$$\sup_{g\in G}|I^g(\theta_0,\lambda)-I^g(\theta_0,\lambda')|<\epsilon \text{ if } \rho(\lambda,\lambda')\le\delta.$$

For $i\notin J(\theta_0)$, since $\inf_{\lambda\in\Theta_i\cap B(\theta_0)}I^{g_i}(\theta_0,\lambda)>0$ by (C3), it follows from (4.19) (with ϵ sufficiently small) that $\inf_{\lambda\in\Theta_i\cap B_\delta(\theta_0)}I^{g_i}(\theta_0,\lambda)>0$ for some $\delta>0$. This and (C3) imply that $\max_{g\in G}I^g(\theta_0,\lambda)>0$ for all $\lambda\ne\theta_0$. Hence given $\eta>0$, we can choose ϵ sufficiently small so that

(4.20)
$$\max_{g\in G}I^g(\theta_0,\theta)\ge 5L\epsilon \quad\text{if}\quad \rho(\theta_0,\theta)\ge\eta,$$

in view of (C2) and the compactness of Θ. Since

$$L_n(\theta_0)-L_n(\theta)=\sum_{g\in G}(T_n(g))^{-1}\sum_{t=1}^{n}\ell_t(\theta)I_{\{\phi^*_{t-1}=g\}}$$

by (3.1), and since $\ell_t(\theta_0)=0$ and $I^g(\theta_0,\lambda)\ge 0$ for all $g\in G$ and $\lambda\in\Theta$, it follows from (4.18) and (4.20) that

(4.21)
$$\mathbf{P}_x^{\phi^*}\left\{\sup_{\theta:\rho(\theta_0,\theta)\ge\eta}L_{h_n}(\theta)<L_{h_n}(\theta_0)-\epsilon\text{ for all }h_n\ge a^i\right\}\ge 1-O((i^{-1}\log i)^{\beta-1}),$$

noting that $T_{a^i}(g)\ge\tau_{a^i}(g)\ge n_{i-1}(\sim i/\log i)$ because at least n_{i-1} stages in the certainty-equivalent testing phase between the times a^{i-1} and a^i use g if $\tau_{a^{i-1}}(g)<n_{i-1}$ for every $g\in G$. From (4.21), (4.12) follows.

Combining (4.12) with (C1) yields $\mathbf{P}_x^{\phi^*}(\cap_{i\ge t}A_i)\ge 1-O((t^{-1}\log t)^{\beta-1})$, where

(4.22)
$$A_i=\left\{J(\widehat{\theta}_{a^i})\subset J(\theta_0),\ \max_{j\notin J(\theta_0)}|c_j(\widehat{\theta}_{a^i})-c_j(\theta_0)|<\epsilon,\ \max_{j\in J(\theta_0)-J(\widehat{\theta}_{a^i})}|c_j(\widehat{\theta}_{a^i})|\le\xi\right\}.$$

Let $\Gamma_k=\{\theta:\rho(\theta_k,\theta)\le\delta'_{\theta_k}\}$ and $w_k=F(\Gamma_k)(>0)$ for $0\le k\le K$. By Lemma 3, $\mathbf{P}_x^{\phi^*}(\cap_{i\ge t}C_i)\ge 1-O((t^{-1}\log t)^{\beta-1})$, where

(4.23)
$$C_i=\left\{\max_{0\le k\le K}\sup_{\theta\in\Gamma_k}\sum_{t=1}^{n}|\ell_t(\theta)-\ell_t(\theta_k)|I_{\{t-1\in\mathcal{C},\phi^*_{t-1}=g\}}\le 2\epsilon\tau_n(g)\right.$$
$$\left.\text{for all } a^i<n\le a^{i+1}\text{ and } g\in G\right\},$$

since $\tau_{a^i}(g) \geq n_{i-1} \sim i/\log i$ for every $g \in G$. Note that $\ell_t(\theta_0) = 0$ and that

$$\begin{aligned}
&\int \prod_{1\leq t\leq n, t-1\in\mathcal{C}, \phi^*_{t-1}\notin \widehat{G}_{J,t}} p(X_{t-1}, X_t; \phi^*_{t-1}(X_{t-1}), \theta) dF(\theta) \\
&\geq \left\{ \prod_{1\leq t\leq n, t-1\in\mathcal{C}, \phi^*_{t-1}\notin \widehat{G}_{J,t}} p(X_{t-1}, X_t; \phi^*_{t-1}(X_{t-1}), \theta_0) \right\} \\
&\quad \cdot w_0 \inf_{\theta\in\Gamma_0} \exp\left(- \sum_{1\leq t\leq n, t-1\in\mathcal{C}, \phi^*_{t-1}\notin \widehat{G}_{J,t}} \ell_t(\theta) \right).
\end{aligned} \tag{4.24}$$

Suppose $\Theta_j \cap B_\delta(\theta_0) \neq \emptyset$. Let $\chi_t = I_{\{t-1\in\mathcal{C}, \phi^*_{t-1}\notin \widehat{G}_{J,t}\}}$. On $A_i \cap C_i$, if $a^i < n \leq a^{i+1}$, then

$$\inf_{\lambda\in\Gamma_k} \exp\left(\sum_1^n \chi_t \ell_t(\lambda) \right) \geq \exp\left\{ \sum_1^n \chi_t \ell_t(\theta_k) - 2\epsilon \sum_1^n \chi_t \right\},$$

$$\inf_{\lambda\in\Gamma_0} \exp\left(-\sum_1^n \chi_t \ell_t(\lambda) \right) \geq \exp\left(-2\epsilon \sum_1^n \chi_t \right),$$

so it follows from (4.7), (4.14), and (4.24) that

$$\inf_{\lambda\in\Theta_j\cap B_\delta(\theta_0)} U_n(\lambda) \geq w_0 \exp\left\{ \inf_{0\leq k\leq K:\Gamma_k\cap\Theta_j\cap B_\delta(\theta_0)\neq\emptyset} \sum_{t=1}^n (\ell_t(\theta_k) - 4\epsilon)\chi_t \right\}. \tag{4.25}$$

Since $\tau_{a^i}(g) \geq n_{i-1}$, it follows from Lemma 2 that $\mathbf{P}_x^{\phi^*}(\cap_{i\geq t} D_i) \geq 1 - O((t^{-1}\log t)^{\beta-1})$, where

$$\begin{aligned}
D_i = \Bigg\{ &\max_{g\in G} \max_{0\leq k\leq K} \left| (\tau_{h_n}(g))^{-1} \sum_{t=1}^{h_n} \ell_t(\theta_k) I_{\{t-1\in\mathcal{C}, \phi^*_{t-1}=g\}} - I^g(\theta_0, \theta_k) \right| \leq 2\epsilon \\
&\text{for all } a^i < h_n \leq a^{i+1} \Bigg\}.
\end{aligned} \tag{4.26}$$

Let $\Omega = \cup_{t=1}^\infty \cap_{i\geq t} (A_i \cap C_i \cap D_i)$. Then $\mathbf{P}_x^{\phi^*}(\Omega) = \lim_{t\to\infty} \mathbf{P}_x^{\phi^*}\{\cap_{i\geq t}(A_i\cap C_i\cap D_i)\} = 1$. On Ω, for all large i and at the times h_n during the certainty-equivalent testing phase of ϕ^* between a^i and a^{i+1}, it follows from (3.9), (3.10a), and (4.22) that $J(\widehat{\theta}_{a^i}) \subset J(\theta_0)$ and

$$(\log a^i)(c_\ell(\theta_0) - \epsilon) \leq \tau_{h_n}(g_\ell) \leq (2\log a^i)(c_\ell(\theta_0) + \epsilon) + 3n_i \text{ if } \ell \notin J(\theta_0), \tag{4.27}$$

$$\tau_{h_n}(g_\ell) \leq 3\xi \log a^i \text{ if } \ell \in J(\theta_0) - J(\widehat{\theta}_{a^i}), \tag{4.28}$$

and from (4.25), (4.26), and (4.28) that

(4.29)
$$\inf_{\lambda\in\Theta_j\cap B_\delta(\theta_0)} \log U_{h_n}(\lambda)$$
$$\geq \inf_{0\leq k\leq K:\Gamma_k\cap\Theta_j\cap B_\delta(\theta_0)\neq\emptyset} \sum_{\ell\notin J(\theta_0)} \{I^{g_\ell}(\theta_0,\theta_k)-6\epsilon\}\tau_{h_n}(g_\ell) - 6L\epsilon(3\xi\log a^i)+O(1),$$

noting that $I^g(\theta_0,\lambda)\geq 0$ for all $g\in G$ and $\lambda\in\Theta$. From (4.1) and (4.29), it follows that on Ω

(4.30)
$$\begin{aligned}&\inf_{\lambda\in\Theta_j\cap B_\delta(\theta_0)} \log U_{h_n}(\lambda)\\ &\geq \inf_{\lambda\in\Theta_j\cap B_\delta(\theta_0)} \sum_{\ell\notin J(\theta_0)} \{I^{g_\ell}(\theta_0,\lambda)-7\epsilon\}\tau_{h_n}(g_\ell) - 6L\epsilon(3\xi\log a^i)+O(1)\end{aligned}$$

at the times h_n during the certainty-equivalent phase of ϕ^* between a^i and a^{i+1} for all large i.

In view of (3.11), (3.12), and Lemma 7, for every $\ell\in J(\theta_0)$,

(4.31) $\mathbf{P}_x^{\phi^*}\{g_\ell \text{ is eliminated at some testing time between } a^i \text{ and } a^{i+1}\}\leq 4(ia^i)^{-1}$.

Hence by the Borel–Cantelli lemma, $\mathbf{P}_x^{\phi^*}\{\Omega_i \text{ for all large } i\}=1$, where

(4.32)
$$\begin{aligned}&\Omega_i\\ &=\{g_\ell \text{ is not eliminated during all test times between } a^{i-2} \text{ and } a^{i+1},\ \text{for all } \ell\in J(\theta_0)\}.\end{aligned}$$

In the event Ω_i, since $1-a^{-2}\geq 3/4$ and $3/5>a^{-1}$, we have the following for all sufficiently large i:

(4.33) $$T_{a^i-1}(g_\ell)\geq (4/5)\{(a^i-1-a^{i-2})/L\}>a^{i-1}/L \text{ for all } \ell\in J(\theta_0).$$

Note that $\{T_{t-1}(g)\geq a^{i-1}/L\}=\{t-1\geq\tau^{(i)}\}=\{t>\tau^{(i)}\}$, where $\tau^{(i)}=\inf\{s: T_s(g)\geq a^{i-1}/L\}$ is a stopping time. Let $N_n(g)=\sum_{t=1}^n I_{\{T_{t-1}(g)\geq a^{i-1}/L,\phi^*_{t-1}=g\}}$, and define

(4.34)
$$\begin{aligned}\Lambda_i=\Bigg\{&\max_{0\leq k\leq K}\sup_{\theta\in\Gamma_k}\sum_{t=1}^{h_n}|\ell_t(\theta)-\ell_t(\theta_k)|I_{\{T_{t-1}(g)\geq a^{i-1}/L,\phi^*_{t-1}=g\}}\leq 2\epsilon a^{i+1} \text{ and}\\ &\max_{0\leq k\leq K}\left|\sum_{t=1}^{h_n}\ell_t(\theta_k)I_{\{T_{t-1}(g)\geq a^{i-1}/L,\phi^*_{t-1}=g\}}-I^g(\theta_0,\theta_k)N_{h_n}(g)\right|\leq 2\epsilon a^{i+1}\\ &\text{for all } a^i<h_n\leq a^{i+1} \text{ and all } g\in G\Bigg\}.\end{aligned}$$

Letting Λ^c denote the complement of an event Λ, it follows from Lemma 4 that

(4.35) $$\mathbf{P}_x^{\phi^*}(\Lambda_i^c)=O(a^{-i(\beta-1)}).$$

Therefore by the Borel–Cantelli lemma, $\mathbf{P}_x^{\phi^*}\{\Lambda_i \text{ for all large } i\}=1$.

Suppose $\Theta_j \backslash B_\delta(\theta_0) \neq \emptyset$. Then we can use (4.33) and an argument similar to that leading to (4.25) and (4.30) to show that, for all large i, on $\Omega_i \cap \Lambda_i$,

$$\begin{aligned} &\min_{1\le m\le m(i)} \inf_{\lambda\in\Theta_j\backslash B_\delta(\theta)} \log W_{\nu_i(m)}(\lambda) \\ &\ge \inf_{\lambda\in\Theta_j\backslash B_\delta(\theta_0)} \sum_{\ell\in J(\theta_0)} I^{g_\ell}(\theta_0,\lambda) a^{i-1}/L - 7\epsilon a^{i+1} + \log w_0, \end{aligned} \tag{4.36}$$

noting that $I^g(\theta_0,\lambda) \ge 0$. Let $\Omega_* = \cup_{t=1}^\infty \cap_{i\ge t} (\Omega_i \cap \Lambda_i)$. Since

$$K := \inf_{\lambda\in\Theta_j\backslash B_\delta(\theta_0)} \sum_{\ell\in J(\theta_0)} I^{g_\ell}(\theta_0,\lambda) > 0$$

by (C3), (4.36) with $\epsilon > 0$ sufficiently small implies that on Ω_*, for all large i and $a^i < h_n \le a^{i+1}$,

$$\inf_{\lambda\in\Theta_j\backslash B_\delta(\theta_0)} \log W_{h_n}(\lambda) > K a^{i-1}/(2L). \tag{4.37}$$

Note that

$$\inf_{\lambda\in\Theta_j} \max\{U_{h_n}(\lambda), W_{h_n}(\lambda)\} \ge \min\{\inf_{\lambda\in\Theta_j\cap B_\delta(\theta_0)} U_{h_n}(\lambda), \inf_{\lambda\in\Theta_j\backslash B_\delta(\theta_0)} W_{h_n}(\lambda)\}.$$

By (4.19), for sufficiently small ϵ,

$$0 \le \inf_{\lambda\in\Theta_j\cap B(\theta_0)} \sum_{\ell\notin J(\theta_0)} c_\ell(\theta_0) I^{g_\ell}(\theta_0,\lambda) - \inf_{\lambda\in\Theta_j\cap B_\delta(\theta_0)} \sum_{\ell\notin J(\theta_0)} c_\ell(\theta_0) I^{g_\ell}(\theta_0,\lambda) \le \epsilon^{3/4}. \tag{4.38}$$

From (4.27), (4.30), (4.37), and (4.38) with ϵ sufficiently small, it follows that on $\Omega\cap\Omega_*$, for all large i, the certainty-equivalent testing phase between times a^i and a^{i+1} rejects H_j at time $\nu_i(m)$ with $(c_j(\theta_0)-\epsilon)\log a^i \le \tau_{\nu_i(m)}(g_j) \le (c_j(\theta_0)+\sqrt{\epsilon})\log a^i$ for every $j \notin J(\theta_0)$ (or equivalently $\theta_0 \notin \Theta_j$). In particular, the upper bound for $\tau_{\nu_i(m)}(g_j)$ follows from the lower bound in (4.27) together with (3.11), (4.37), and (4.30), noting that the $\epsilon^{3/4}$ in (4.38) is much smaller than $\sqrt{\epsilon}$ if ϵ is sufficiently small and that $\inf_{\lambda\in\Theta_j\cap B(\theta_0)} \sum_{\ell\notin J(\theta_0)} c_\ell(\theta_0) I^{g_\ell}(\theta_0,\lambda) \ge 1$ if $B(\theta_0) \neq \emptyset$ by (3.2). Hence on $\Omega\cap\Omega_*$, for all large i, the evenly allocated testing phase between times a^i and a^{i+1} is applied only to controls g_ℓ with $\ell \in J(\theta_0)$. Since ϵ can be arbitrarily small, this implies that $T_n(g_j) = \tau_n(g_j) + O(1)$ and that $T_n(g_j)/\log n \to c_j(\theta_0)$ a.s. $[\mathbf{P}_x^{\phi^*}]$ for every $j \notin J(\theta_0)$. This also implies that with probability 1, for all large i, ϕ^* only uses rules from G_J after certainty-equivalent testing between times a^i and a^{i+1}. Hence in view of (4.27) and the use of the same stationary control law for an entire block (of stages) B_m^i, of size $\ge n_i \sim i/\log i$, the desired conclusion on $\sum_1^n I_{\{\phi_i^*\neq\phi_{i-1}^*, \phi_i^*\notin G_J \text{ or } \phi_{i-1}^*\notin G_J\}}$ follows.

Proof of Lemma 6. Fix $j \notin J(\theta_0)$. The evenly allocated testing phase of ϕ^*, which was shown to use eventually only controls from G_J in the proof of the a.s. convergence of $T_n(g_j)/\log n$ in Lemma 5, will play a crucial role here in establishing uniform integrability of $T_n(g_j)/\log n$. Also the assumption $\beta > 2$ will be important here. Let $\tau_t = \tau_{a^{t+1}}(g_j)$ and $\widetilde{\tau}_t = T_{a^{t+1}}(g_j) - \tau_t$. It suffices to show that

$$\{\tau_t/t, t\ge 1\} \text{ and } \{\widetilde{\tau}_t/t, t\ge 1\} \text{ are uniformly integrable under } \mathbf{P}_x^{\phi^*}. \tag{4.39}$$

Take any $\epsilon > 0$ and define A_i as in (4.22). In the line above (4.22) we have shown that

$$\mathbf{P}_x^{\phi^*}(\Delta_t) \geq 1 - O(t^{-(\beta-1)}(\log t)^{2\beta}), \text{ where } \Delta_t = \cap_{i \geq [t/\log t]} A_i. \tag{4.40}$$

From the constraint (3.10a) or (3.10b) in the certainty-equivalent testing phase, it follows that

$$\tau_t \leq (2 \log a^t) \log t + 3n_t. \tag{4.41}$$

Moreover, in view of (4.41) together with (3.10a,b) and (4.22) we have, for all large t,

$$\tau_{[t/\log t]} < 3t \log a \text{ and } \tau_i < 3t(c_j(\theta_0) + \epsilon) \log a \text{ for all } [t/\log t] < i \leq t \text{ on } \Delta_t.$$

Hence for all large t,

$$\tau_t/t \leq (3 \log a)(\log t) I_{\Delta_t^c} + (3 \log a)(c_j(\theta_0) + \epsilon) I_{\Delta_t}. \tag{4.42}$$

Since $\mathbf{P}_x^{\phi^*}(\Delta_t^c) = O(t^{-(\beta-1)}(\log t)^{2\beta})$ by (4.40), the uniform integrability of $\{\tau_t/t, t \geq 1\}$ follows from (4.42).

Labeling the elements of $\{\nu_i(m) : i \geq 1, 1 \leq m \leq m(i)\}$ as $s_1 < s_2 < \cdots$ (test times), define

$$\sigma_t = \sup \left\{ s_n \leq a^{t+1} : \inf_{\lambda \in \Theta_j \cap B_\delta(\theta_0)} \widetilde{W}_{s_n,j}(\lambda) < t a^t \right\}, \quad \#_{t,1} = T_{\sigma_t}(g_j), \tag{4.43}$$

$$\#_{t,2} = \sum_{i=1}^{t} I_{\Omega_i \cap \Lambda_i} \sum_{m=1}^{m(i)} 2n_i I_{\{\inf_{\lambda \in \Theta_j \setminus B_\delta(\theta_0)} W_{\nu_i(m)}(\lambda) < i a^i\}}, \tag{4.44}$$

$$\#_{t,3} = \sum_{i=1}^{t} (a^{i+1} - a^i)(I_{\Omega_i^c} + I_{\Lambda_i^c}), \tag{4.45}$$

where Ω_i and Λ_i are defined in (4.32) and (4.34). Since $\nu_i(m) - \nu_i(m-1) \leq 2n_i$ and since

$$\inf_{\lambda \in \Theta_j} \max(\widetilde{W}_{n,j}(\lambda), W_n(\lambda)) \geq \min \left\{ \inf_{\lambda \in \Theta_j \cap B_\delta(\theta_0)} \widetilde{W}_{n,j}(\lambda), \inf_{\lambda \in \Theta_j \setminus B_\delta(\theta_0)} W_n(\lambda) \right\},$$

it follows from (3.12), (4.15), and (4.16) that

$$\widetilde{\tau}_t (= T_{a^{t+1}}(g_j) - \tau_t) \leq \#_{t,1} + \#_{t,2} + \#_{t,3}. \tag{4.46}$$

By (4.31) and (4.35) with $\beta > 2$, $\mathbf{E}_x^{\phi^*}(\#_{t,3}) = \sum_{i=1}^{t} O(i^{-1}) = O(\log t)$. Since $\#_{t,3} \geq 0$ and $\mathbf{E}_x^{\phi^*}(t^{-1}\#_{t,3}) \to 0$ as $t \to \infty$, it follows that $\{\#_{t,3}/t, t \geq 1\}$ is uniformly integrable under $\mathbf{P}_x^{\phi^*}$.

To prove that $\{\#_{t,1}/t, t \geq 1\}$ is uniformly integrable under $\mathbf{P}_x^{\phi^*}$, take $\epsilon > 0$, choose δ by (4.19) and define $\theta_1, \ldots, \theta_K$ as in (4.3). Letting $\Gamma_k = \{\theta : \rho(\theta_k, \theta) \leq \delta'_{\theta_k}\}$, we shall modify the use of (4.23) and (4.26) in the proof of Lemma 5 by introducing

$$\begin{aligned} \sigma^* = \sup \left\{ T_{s_n}(g_j) : \max_{0 \leq k \leq K} \sup_{\theta \in \Gamma_k} \sum_{s=1}^{s_n} |\ell_s(\theta) - \ell_s(\theta_k)| I_{\{\phi^*_{s-1} = g_j\}} > 2\epsilon T_{s_n}(g_j) \right\} \\ \vee \sup \left\{ T_{s_n}(g_j) : \max_{0 \leq k \leq K} \left| (T_{s_n}(g_j))^{-1} \sum_{s=1}^{s_n} \ell_s(\theta_k) I_{\{\phi^*_{s-1} = g_j\}} - I^{g_j}(\theta_0, \theta_k) \right| > 2\epsilon \right\}. \end{aligned} \tag{4.47}$$

For $T_{s_n}(g_j) > \sigma^*$, we have

$$\max_{0\le k\le K} \sup_{\theta\in\Gamma_k} \sum_{s=1}^{s_n} |\ell_s(\theta) - \ell_s(\theta_k)| I_{\{\phi^*_{s-1}=g_j\}} \le 2\epsilon T_{s_n}(g_j)$$

and

$$\max_{0\le k\le K} \left| T_{s_n}(g_j))^{-1} \sum_{s=1}^{s_n} \ell_s(\theta_k) I_{\{\phi^*_{s-1}=g_j\}} - I^{g_j}(\theta_0,\theta_k) \right| \le 2\epsilon.$$

Hence an argument similar to that used to derive (4.25) and (4.30) can be used to show that if $T_{s_n}(g_j) > \sigma^*$ then

$$\inf_{\lambda\in\Theta_j\cap B_\delta(\theta_0)} \log \widetilde{W}_{s_n,j}(\lambda) \ge \left\{ \inf_{\lambda\in\Theta_j\cap B_\delta(\theta_0)} I^{g_j}(\theta_0,\lambda) - 7\epsilon \right\} T_{s_n}(g_j) + \log w_0, \tag{4.48}$$

where $w_0 = F(\Gamma_0)$. From (4.48) and (4.43) it follows that

$$\#_{t,1}(= T_{\sigma_t}(g_j)) \le \max\{\sigma^*, 1 + \log(ta^t/w_0)/[\inf_{\lambda\in\Theta_j\cap B_\delta(\theta_0)} I^{g_j}(\theta_0,\lambda) - 7\epsilon]\}, \tag{4.49}$$

noting that $\inf_{\lambda\in\Theta_j\cap B_\delta(\theta_0)} I^{g_j}(\theta_0,\lambda) - 7\epsilon > 0$ by (C3) and (4.19), provided that ϵ is chosen sufficiently small. By (4.47) and Lemmas 2 and 3, $\sum_{m=1}^\infty \mathbf{P}_x^{\phi^*}\{\sigma^* \ge m\} = \sum_{m=1}^\infty O(m^{-(\beta-1)}) < \infty$ since $\beta > 2$. Therefore $\mathbf{E}_x^{\phi^*}(\sigma^*) < \infty$ and the uniform integrability of $\{\#_{t,1}/t, t \ge 1\}$ follows from (4.49).

To prove the uniform integrability of $\{\#_{t,2}/t, t \ge 1\}$ under $\mathbf{P}_x^{\phi^*}$, recall that (4.36) holds on $\Omega_i \cap \Lambda_i$ for all large i. By (C3) and choosing ϵ sufficiently small, this implies that $\{\#_{t,2}, t \ge 1\}$ is uniformly bounded by some constant.

The case where ϕ^* is replaced by $\widetilde{\phi}$ is even simpler and is similar to the preceding proof of the uniform integrability of $\{\widetilde{\tau}_t/t,\ t \ge 1\}$.

REFERENCES

[1] R. Agrawal, M. Hedge, and D. Teneketzis, *Asymptotically efficient adaptive allocation rules for the multiarmed bandit problem with switching cost*, IEEE Trans. Automat. Control, AC-33 (1988), pp. 899–906.

[2] R. Agrawal and D. Teneketzis, *Certainty equivalence control with forcing: Revisited*, Systems Control Lett., 13 (1989), pp. 405–412.

[3] R. Agrawal, D. Teneketzis, and V. Anantharam, *Asymptotically efficient adaptive allocation schemes for controlled I.I.D. process: Finite parameter space*, IEEE Trans. Automat. Control, AC-34 (1989), pp. 258–267.

[4] R. Agrawal, D. Teneketzis, and V. Anantharam, *Asymptotically efficient adaptive allocation schemes for controlled Markov chains: Finite parameter space*, IEEE Trans. Automat. Control, AC-34 (1989), pp. 1249–1259.

[5] V. Anantharam, P. Varaiya, and J. Walrand, *Asymptotically efficient allocation rules for multiarmed bandit problem with multiple plays. Part* II: *Markovian rewards*, IEEE Trans. Automat. Control, AC-32 (1987), pp. 975–982.

[6] V. Borkar and P. Varaiya, *Adaptive control of Markov chains,* I: *Finite parameter set*, IEEE Trans. Automat. Control, AC-24 (1979), pp. 953–958.

[7] D. Feldman, *Contributions to the "two-armed bandit" problem*, Ann. Math. Statist., 33 (1962), pp. 847–856.

[8] T. L. Graves, *Comparison of Treatments Under Adaptive Treatment Allocation in Clinical Trials and Stochastic Adaptive Control*, Ph.D. dissertation, Department of Statistics, Stanford University, Stanford, CA, 1995.

[9] I. Iscoe, P. Ney, and E. Nummelin, *Large deviations of uniformly recurrent Markov additive processes*, Adv. Appl. Math., 6 (1985), pp. 373–412.

[10] J. L. Jensen, *Saddlepoint expansions for sums of Markov dependent variables on a continuous state space*, Probab. Theory Related Fields, 89 (1991), pp. 181–199.
[11] P. R. Kumar, *A survey of some results in stochastic adaptive control*, SIAM J. Control Optim., 23 (1985), pp. 329–380.
[12] P. R. Kumar and P. Varaiya, *Stochastic Systems: Estimation, Identification and Adaptive Control*, Prentice-Hall, Englewood Cliffs, NJ, 1986.
[13] T. L. Lai, *Certainty equivalence with uncertainty adjustments in stochastic adaptive control*, in Stochastic Theory and Adaptive Control, T. Duncan and B. Pasik-Duncan, eds., Springer-Verlag, New York, 1992, pp. 270–284.
[14] T. L. Lai, *Tail Probability Bounds for Martingales and Markov Random Walks with Applications to Sequential Analysis and Stochastic Control*, Tech. Report, Department of Statistics, Stanford University, Stanford, CA, 1994.
[15] T. L. Lai and H. Robbins, *Asymptotically optimal allocation of treatments in sequential experiments*, in Design of Experiments, T. J. Santner and A. C. Tamhane, eds., Marcel Dekker, New York, 1984, pp. 127–142.
[16] T. L. Lai and H. Robbins, *Asymptotically efficient adaptive allocation rules*, Adv. Appl. Math., 6 (1985), pp. 4–22.
[17] T. L. Lai and S. Yakowitz, *Machine learning and nonparametric bandit theory*, IEEE Trans. Automat. Control, AC-40 (1995), pp. 1199–1209.
[18] P. Mandl, *Estimation and control of Markov chains*, Adv. in Appl. Probab., 6 (1974), pp. 40–60.

SIAM J. CONTROL OPTIM.
Vol. 35, No. 3, pp. 744–765, May 1997

002

BLOCK TRIANGULAR DECOUPLING FOR LINEAR SYSTEMS OVER PRINCIPAL IDEAL DOMAINS*

NAOHARU ITO† AND HIROSHI INABA†

Abstract. This paper studies in the framework of the so-called geometric approach the block triangular decoupling problem with state feedback for linear systems defined over a principal ideal domain with identity. First, various properties of feedback reachability submodules are discussed, and then under certain assumptions necessary and sufficient conditions for its solvability are obtained. Further, the pole assignability for decoupled systems is investigated. Finally, a simple example is given to illustrate the results.

Key words. block triangular decoupling, linear systems over rings, feedback reachability submodules, pole assignability

AMS subject classifications. 93B05, 93B27, 93D15

PII. S0363012995281109

1. Introduction.

In the so-called geometric approach, various block decoupling problems for linear systems over the field of real numbers were first studied in [9] and [13]. In particular, they studied the block decoupling problem with certain assumptions [13] as well as the block triangular decoupling problem [9] and obtained necessary and sufficient conditions for each problem to be solvable. The results were given in terms of the largest elements in some families of reachability subspaces satisfying certain conditions.

It seems quite interesting and useful to investigate the corresponding decoupling problems for systems over rings also in the framework of the geometric approach because systems over rings are a natural generalization of systems over the field of real numbers and are used as abstract descriptions of, for instance, systems with parameters or with time-delay operators. In fact, such systems have been extensively studied in the last two decades (see, e.g., [1], [3], [7], [11], and [10]).

The present authors [6] and with Munaka [5] have already shown that the results for the block decoupling problem given in [13] are essentially valid for the corresponding problem for systems over principal ideal domains. So, the purpose of the present investigation is to discuss the corresponding triangular decoupling problem for systems over principal ideal domains and to present necessary and sufficient conditions for the problem to be solvable. Although this problem brings a number of new difficulties to be resolved, it turns out that the results obtained in [9] are still essentially valid for this new problem. Furthermore, the pole assignability of decoupled systems is also studied in some detail.

The outline of the paper is as follows. Section 2 gives preliminaries and notions of reachability submodules and their important properties. In section 3, the block triangular decoupling problem is formulated in the framework of the geometric approach, and necessary and sufficient conditions for its solvability are obtained. Section 4 con-

*Received by the editors February 3, 1995; accepted for publication (in revised form) February 29, 1996. This research was supported in part by Grant-Aid for General Scientific Research C-04650386 of the Japanese Ministry of Education, Science and Culture and in part by Research Center for Technology of Tokyo Denki University grant Q92-S71.

http://www.siam.org/journals/sicon/35-3/28110.html

†Department of Information Sciences, Tokyo Denki University, Hatoyama-machi, Hiki-gun, Saitama 350-03, Japan (naoharu@j.dendai.ac.jp, inaba@j.dendai.ac.jp).

siders the pole assignability of a decoupled system and shows that the reachability of the original system implies the pole assignability. In section 5 an example is presented to demonstrate our results, and finally in section 6 some concluding remarks are given.

2. Preliminaries and reachability submodules. Throughout this paper we consider a system (A, B, C) over a commutative principal ideal domain $\boldsymbol{\mathcal{K}}$ with the identity 1, where $A \in \boldsymbol{\mathcal{K}}^{n\times n}$, $B \in \boldsymbol{\mathcal{K}}^{n\times m}$, and $C \in \boldsymbol{\mathcal{K}}^{l\times n}$ denote the system matrix, the input matrix, and the output matrix, respectively. When the output matrix C is immaterial, we will simply write (A, B). Let $X := \boldsymbol{\mathcal{K}}^n$, $U := \boldsymbol{\mathcal{K}}^m$, and $Y := \boldsymbol{\mathcal{K}}^l$ be free $\boldsymbol{\mathcal{K}}$-modules. Then, system (A, B, C) defines a discrete-time system described by

$$x(t+1) = Ax(t) + Bu(t), \quad y(t) = Cx(t), \tag{2.1}$$

where $x(t) \in X$, $u(t) \in U$, and $y(t) \in Y$ are the state, the input, and the output, respectively. For a given input sequence $u = (u(t))_{t=0}^{\infty}$ and an initial state $x(0) = x_0 \in X$, we denote by $x(t; x_0, u)$ the state at time t resulting via (2.1). The triple (A, B, C) may also be considered to represent a continuous-time system of the form

$$\frac{dx(t)}{dt} = Ax(t) + Bu(t), \quad y(t) = Cx(t) \tag{2.2}$$

provided that the derivative $dx(t)/dt$ can be suitably defined. However, the development to follow in this paper does not depend on whether system (A, B, C) represents a discrete-time or a continuous-time system, so we may simply consider it as a discrete-time system. Throughout this paper, the field of real numbers will be denoted by $\boldsymbol{R}$.

Systems over principal ideal domains can describe various systems appearing applications, such as parametrized systems, systems over the integer ring, time-delay systems, etc.; see, e.g., [1], [7], [11], [10].

Example 2.1. Linearize around the equilibrium point $(x_1, x_2) = (\lambda, 0)$ the nonlinear system over $\boldsymbol{R}$ given by

$$\left\{ \begin{array}{rcl} x_1(t+1) & = & x_1(t) + x_1^2(t) + u_1(t), \\ x_2(t+1) & = & 2x_2(t) + u_2(t), \\ y(t) & = & x_1(t)x_2(t) + x_1(t), \end{array} \right. \tag{2.3}$$

where λ is an arbitrary real constant number. Then, the resulting linearized system is described by

$$x(t+1) = A(\lambda)x(t) + B(\lambda)u(t), \quad y = C(\lambda)x(t), \tag{2.4}$$

where

$$A(\lambda) = \begin{bmatrix} 1 & \lambda^2 \\ 0 & 2 \end{bmatrix}, \quad B(\lambda) = \begin{bmatrix} 1 & 0 \\ 0 & 1 \end{bmatrix}, \quad C(\lambda) = \begin{bmatrix} 1 & \lambda \end{bmatrix},$$

which are considered to be matrices over $\boldsymbol{R}[\lambda]$, the ring of polynomials in the indeterminate λ with coefficients in $\boldsymbol{R}$. Since $\boldsymbol{R}[\lambda]$ is a principal ideal domain, the triple $(A(\lambda), B(\lambda), C(\lambda))$ of (2.4) is regarded as a linear system over the principal ideal domain $\boldsymbol{R}[\lambda]$. □

Example 2.2. Consider the linear time-delay system described by

$$\frac{dx(t)}{dt} = \sum_{i=0}^{p} A_i x(t - ih) + \sum_{i=0}^{q} B_i u(t - ih), \quad y(t) = \sum_{i=0}^{r} C_i x(t - ih), \tag{2.5}$$

where h is a fixed positive number, A_i, B_i, and C_i are $n \times n, n \times m$, and $l \times n$ matrices over $\boldsymbol{R}$, respectively. Now, let σ denote the time-delay operator defined by $(\sigma x)(t) := x(t-h)$. Then (2.5) can be written in the form

$$\frac{dx}{dt} = A(\sigma)x + B(\sigma)u, \qquad y(t) = C(\sigma)x, \tag{2.6}$$

where $A(\sigma), B(\sigma)$, and $C(\sigma)$ are $n \times n, n \times m$, and $l \times n$ matrices over $\boldsymbol{R}[\sigma]$. Thus system (2.6) can be described as a linear system over the principal ideal domain $\boldsymbol{R}[\sigma]$. □

Consider a system (A, B) over $\boldsymbol{\mathcal{K}}$ with $A \in \boldsymbol{\mathcal{K}}^{n\times n}$ and $B \in \boldsymbol{\mathcal{K}}^{n\times m}$. As in the case of systems over $\boldsymbol{R}$ [12], the *reachable submodule* for system (A, B) is given by

$$\langle A|\operatorname{Im} B\rangle := \operatorname{Im} B + A(\operatorname{Im} B) + \cdots + A^{n-1}(\operatorname{Im} B) \subset X,$$

where $\operatorname{Im} B$ stands for the image of B. That $\langle A|\operatorname{Im} B\rangle$ is called the reachable submodule comes from the fact that for every $\tilde{x} \in \langle A|\operatorname{Im} B\rangle$ there exist a time $T \geq 0$ and an input u such that $x(T; 0, u) = \tilde{x}$. Thus, the pair (A, B) is called *reachable* if $\langle A|\operatorname{Im} B\rangle = X$.

The following definition gives the basic notions in the geometric approach for linear systems [13], [12].

DEFINITION 2.3. *Let φ be a submodule of X.*

(i) *φ is said to be* (A, B)-invariant *if $A\varphi \subset \varphi + \operatorname{Im} B$.*

(ii) *φ is said to be* feedback (A, B)-invariant *if there exists $F \in \boldsymbol{\mathcal{K}}^{m\times n}$ such that $(A + BF)\varphi \subset \varphi$. The set of all those F's satisfying such inclusion is denoted by $\mathbf{F}(\varphi; A, B)$. (Note that φ is feedback (A, B)-invariant if and only if $\mathbf{F}(\varphi; A, B) \neq \emptyset$.)* □

An (A, B)-invariant submodule φ has the property that for every $x(0) \in \varphi$ there exists an input sequence u such that $x(t; x(0), u) \in \varphi$ for $t = 0, 1, 2, \ldots$. A feedback (A, B)-invariant submodule φ has the property that there exists a matrix $F \in \boldsymbol{\mathcal{K}}^{m\times n}$ such that by defining an input sequence $u := (Fx(t))_{t=0}^{\infty}$ for every $x(0) \in \varphi$, $x(t; x(0), u)$ belongs to φ for $t = 0, 1, 2, \ldots$.

A submodule $\varphi \subset X$ is called a *direct summand* of X if there exists a submodule $\psi \subset X$ such that $X = \varphi + \psi$ and $\varphi \cap \psi = \{0\}$. Then we write $X = \varphi \oplus \psi$. We remark that subspaces of a linear space are always direct summands but submodules are not necessarily. As will be seen later, this difference causes the main difficulty in studying systems over $\boldsymbol{\mathcal{K}}$. So the following notion plays an important role in what follows.

DEFINITION 2.4. *Let $\psi \subset \varphi$ be submodules of X. The* closure $\operatorname{Cl}_\varphi(\psi)$ *of ψ in φ is defined to be the submodule given by*

$$\operatorname{Cl}_\varphi(\psi) := \{x \in \varphi \text{ such that } ax \in \psi \text{ for some nonzero } a \in \boldsymbol{\mathcal{K}}\}.$$

ψ is said to be closed *in φ if* $\operatorname{Cl}_\varphi(\psi) = \psi$. □

Remark 2.5. From [1] and [4] we recall the following facts.

(i) A submodule of X is closed in X if and only if it is a direct summand of X.

(ii) If a submodule of X is feedback (A, B)-invariant, then so is its closure in X. However, (A, B)-invariance does not necessarily imply (A, B)-invariance of its closure.

(iii) Feedback (A, B)-invariance implies (A, B)-invariance, but the converse is not generally true.

(iv) The sum of two (A, B)-invariant submodules is again an (A, B)-invariant submodule. Therefore, the family of all (A, B)-invariant submodules contained in

a given submodule has a unique largest (A, B)-invariant submodule. However, this property does not hold true for the case of feedback (A, B)-invariant submodules.

(v) The largest (A, B)-invariant submodule contained in a given closed submodule of X is feedback (A, B)-invariant if and only if it is closed in X. □

Now, the following two notions of reachability submodules are introduced.

DEFINITION 2.6. *Let φ be a submodule of X.*

(i) *φ is said to be a* reachability submodule *for (A, B) if for each $\tilde{x} \in \varphi$ there exist a time $T \geq 0$ and an input sequence $u = \{u(t)\}_{t=0}^{\infty}$ such that $x(t; 0, u)$ belongs to φ for $t = 0, 1, 2, \ldots$ and $x(T; 0, u) = \tilde{x}$.*

(ii) *φ is called a* feedback reachability submodule *for (A, B) if there exist $F \in \boldsymbol{\mathcal{K}}^{m\times n}$ and $G \in \boldsymbol{\mathcal{K}}^{m\times m}$ such that $\varphi = \langle A + BF | \operatorname{Im}(BG)\rangle$.* □

Remark 2.7. What was called a reachability submodule in the papers [6] and [5] is renamed in this paper a *feedback reachability submodule*. Conte and Perdon [2] introduced the notion of precontrollability submodules for linear systems over a commutative ring. It is possible to verify that this notion is equivalent to that of reachability submodules given in Definition 2.6. □

A feedback reachability submodule φ for (A, B) is the reachable submodule for the new system $(A + BF, BG)$ resulting from applying a state feedback of the form $u(t) = Fx(t) + Gv(t)$ to (2.1), where v denotes the new input. In other words, for each $\tilde{x} \in \varphi$ there exist a time $T \geq 0$ and an input sequence $v = (v(t))_{t=0}^{\infty}$ such that, defining an input sequence $u := (Fx(t) + Gv(t))_{t=0}^{\infty}$, we have $x(t; 0, u) \in \varphi$ for $t = 0, 1, 2, \ldots$ and $x(T; 0, u) = \tilde{x}$. Therefore, if a submodule of X is a feedback reachability submodule then it is a reachability submodule. However, a reachability submodule is not necessarily a feedback reachability submodule (see Remark 2.11), although these two notions are equivalent for linear systems over the field $\boldsymbol{R}$. In fact, Conte and Perdon [2] showed that a submodule is a feedback reachability submodule if and only if it is a reachability submodule and a feedback (A, B)-invariant submodule.

For feedback reachability submodules, the following lemma can be proved in the same manner as in [13], except for the sufficiency part of (i).

LEMMA 2.8. *Let φ be a submodule of X.*

(i) *φ is a feedback reachability submodule for (A, B) if and only if there exists $F \in \boldsymbol{\mathcal{K}}^{m\times n}$ such that $\varphi = \langle A + BF | \operatorname{Im} B \cap \varphi\rangle$.*

(ii) *If φ is a feedback reachability submodule for (A, B), then*

$$\varphi = \langle A + BF | \operatorname{Im} B \cap \varphi\rangle$$

for all $F \in \mathbf{F}(\varphi; A, B)$.

Proof. Only a proof of the sufficiency of (i) will be given. Since $\operatorname{Im} B \cap \varphi$ is a submodule of the free $\boldsymbol{\mathcal{K}}$-module X, it is free. Hence, if the rank of $\operatorname{Im} B \cap \varphi$ is p, there is a basis $\{b_1, b_2, \ldots, b_p\} \subset X$ of $\operatorname{Im} B \cap \varphi$. For each b_i $(i = 1, 2, \ldots, p)$ there exists $u_i \in U$ $(i = 1, 2, \ldots, p)$ such that $b_i = Bu_i$. Defining $G := [u_1, \ldots, u_p, 0, \ldots, 0] \in \boldsymbol{\mathcal{K}}^{m\times m}$, we obtain $\operatorname{Im} B \cap \varphi = \operatorname{Im}(BG)$, showing that φ is a feedback reachability submodule for (A, B). □

It should be remarked that if φ is a feedback reachability submodule, then φ is feedback (A, B)-invariant and hence the set $\mathbf{F}(\varphi; A, B)$ is not empty. Furthermore, a feedback reachability submodule is not necessarily closed in X, and the closure of a feedback reachability submodule is not always a feedback reachability submodule, as is seen from the following example.

Example 2.9. Let $\boldsymbol{\mathcal{K}}:=\boldsymbol{R}[\sigma]$ and $X := \boldsymbol{\mathcal{K}}^2$, and set

$$A := \begin{bmatrix} 1 & 1 \\ 0 & 1 \end{bmatrix}, \quad B := \begin{bmatrix} \sigma & 0 \\ 0 & 1 \end{bmatrix}, \quad \varphi := \operatorname{Im} \begin{bmatrix} \sigma \\ 0 \end{bmatrix}.$$

Then, it is easy to see that φ is A-invariant and $\varphi \cap \operatorname{Im} B = \varphi$. Therefore, $\varphi = \langle A | \operatorname{Im} B \cap \varphi \rangle$, and hence φ is a feedback reachability submodule for (A, B) by Lemma 2.8. The closure is given by

$$\operatorname{Cl}_X(\varphi) = \operatorname{Im} \begin{bmatrix} 1 \\ 0 \end{bmatrix} \neq \varphi,$$

and hence φ is not closed in X. Moreover, $\operatorname{Cl}_X(\varphi)$ is not a feedback reachability submodule for (A, B). In fact, since $\operatorname{Cl}_X(\varphi)$ is A-invariant and hence satisfies $0 \in \mathbf{F}(\operatorname{Cl}_X(\varphi); A, B)$, Lemma 2.8 and the relation

$$\langle A | \operatorname{Im} B \cap \operatorname{Cl}_X(\varphi) \rangle = \operatorname{Im} \begin{bmatrix} \sigma \\ 0 \end{bmatrix} \neq \operatorname{Cl}_X(\varphi)$$

imply that $\operatorname{Cl}_X(\varphi)$ is not a feedback reachability submodule. □

It is well known that for systems over $\boldsymbol{R}$ the sum of two feedback reachability subspaces is again a feedback reachability subspace. However, this statement is not true for systems over $\boldsymbol{\mathcal{K}}$. Thus there is no guarantee that a largest feedback reachability submodule in the family of feedback reachability submodules contained in a given submodule exists. The next example demonstrates that the sum of two feedback reachability submodules is not a feedback reachability submodule.

Example 2.10. Let $\boldsymbol{\mathcal{K}}:=\boldsymbol{R}[\sigma]$ and $X := \boldsymbol{\mathcal{K}}^3$, and set

$$A := \begin{bmatrix} 1 & 1 & 0 \\ 0 & 1 & 0 \\ 0 & 0 & 1 \end{bmatrix}, \quad B := \begin{bmatrix} \sigma & 0 & 0 \\ 0 & \sigma & 0 \\ 0 & 1 & 1 \end{bmatrix},$$

$$\varphi_1 := \operatorname{Im} \begin{bmatrix} 0 \\ \sigma \\ 1 \end{bmatrix}, \quad \varphi_2 := \operatorname{Im} \begin{bmatrix} 0 \\ 0 \\ 1 \end{bmatrix}.$$

Then, it can easily be checked that $\varphi_1 \cap \operatorname{Im} B = \varphi_1$ and $\varphi_2 \cap \operatorname{Im} B = \varphi_2$, and hence that $\langle A + BF_1 | \varphi_1 \cap \operatorname{Im} B \rangle = \varphi_1$ and $\langle A + BF_2 | \varphi_2 \cap \operatorname{Im} B \rangle = \varphi_2$, where

$$F_1 = \begin{bmatrix} 0 & 0 & -1 \\ 0 & 0 & 0 \\ 0 & 0 & 0 \end{bmatrix} \in \mathbf{F}(\varphi_1; A, B), \qquad F_2 = \begin{bmatrix} 0 & 0 & 0 \\ 0 & 0 & 0 \\ 0 & 0 & 0 \end{bmatrix} \in \mathbf{F}(\varphi_2; A, B).$$

Therefore, by Lemma 2.8 φ_1 and φ_2 are feedback reachability submodules. On the other hand, the closure of the submodule $\varphi_1 + \varphi_2$, i.e., the closed submodule

$$\operatorname{Cl}_X(\varphi_1 + \varphi_2) = \operatorname{Im} \begin{bmatrix} 0 & 0 \\ 1 & 0 \\ 0 & 1 \end{bmatrix},$$

is not (A, B)-invariant because for $[0\ 1\ 0]^\top \in \operatorname{Cl}_X(\varphi_1 + \varphi_2)$

$$A \begin{bmatrix} 0 \\ 1 \\ 0 \end{bmatrix} = \begin{bmatrix} 1 \\ 1 \\ 0 \end{bmatrix} \notin \operatorname{Cl}_X(\varphi_1 + \varphi_2) + \operatorname{Im} B = \operatorname{Im} \begin{bmatrix} \sigma & 0 & 0 \\ 0 & 1 & 0 \\ 0 & 0 & 1 \end{bmatrix}.$$

Thus, by virtue of Remark 2.5, $\mathrm{Cl}_X(\varphi_1+\varphi_2)$ is not feedback (A,B)-invariant, and hence $\varphi_1+\varphi_2$ is not feedback (A,B)-invariant. Therefore, $\varphi_1+\varphi_2$ is not a feedback reachability submodule for (A,B). □

Remark 2.11. In Example 2.10, $\varphi_1+\varphi_2$ is a reachability submodule for (A,B) but not a feedback reachability submodule for (A,B). □

As will be seen later, in investigating systems over $\boldsymbol{\mathcal{K}}$ it is vital to look for conditions under which the largest feedback reachability submodule contained in a given submodule exists. To this purpose, we first introduce the following algorithm.

ALGORITHM 2.12. For a given submodule ψ of X, compute the sequence $(R^{(i)})$ of submodules of X for (A,B) as

$$\begin{cases} R^{(0)}(\psi) := 0, \\ R^{(i)}(\psi) := \psi \cap (AR^{(i-1)}(\psi) + \operatorname{Im} B) \quad (i = 1,2,\ldots). \end{cases} \tag{2.7}$$

Then, it is obvious that the sequence $(R^{(i)})$ is nondecreasing. Thus, since $\boldsymbol{\mathcal{K}}$ is a principal ideal domain, there exists an integer $q \geq 0$ such that $R^{(i)}(\psi) = R^{(i+1)}(\psi)$ for $i \geq q$. So set $\mathcal{R}(\psi) := R^{(q)}(\psi)$. □

Now, the following proposition will be proved.

PROPOSITION 2.13. *Let S be a submodule of X and V^*, the largest (A,B)-invariant submodule contained in S. If $\mathcal{R}(V^*)$ is feedback (A,B)-invariant, then $\mathcal{R}(V^*)$ is the largest feedback reachability submodule for (A,B) contained in S, and it is given as*

$$\mathcal{R}(V^*) = \langle A + BF | \operatorname{Im} B \cap V^* \rangle,$$

where F is an arbitrary element in $\mathbf{F}(\mathcal{R}(V^*); A, B)$.

Proof. Assume that $\mathcal{R}(V^*)$ is feedback (A,B)-invariant. Hence, it follows from Definition 2.3 that $\mathbf{F}(\mathcal{R}(V^*); A, B) \neq \emptyset$. Then, it is claimed that for any $F \in \mathbf{F}(\mathcal{R}(V^*); A, B)$

$$R^{(i)}(V^*) = \sum_{j=1}^{i} (A+BF)^{j-1}(V^* \cap \operatorname{Im} B) \quad (i = 1,2,\ldots). \tag{2.8}$$

To verify this claim, first notice that

$$(A+BF)R^{(i)}(V^*) \subset \mathcal{R}(V^*) \subset V^* \quad (i = 1,2,\ldots). \tag{2.9}$$

It is clear that (2.8) is true for $i = 1$. Next, assuming that (2.8) holds true for $i = p$, and using (2.9) and (2.7), one obtains

$$\begin{aligned} \sum_{j=1}^{p+1} (A+BF)^{j-1}(V^* \cap \operatorname{Im} B) &= V^* \cap \operatorname{Im} B + (A+BF)R^{(p)}(V^*) \\ &= V^* \cap [(A+BF)R^{(p)}(V^*) + \operatorname{Im} B] \\ &= V^* \cap (AR^{(p)}(V^*) + \operatorname{Im} B) \\ &= R^{(p+1)}(V^*). \end{aligned}$$

Thus (2.8) is true for all i, and hence the claim is verified.

Now, it is easily seen from (2.8) and the Cayley–Hamilton theorem that

$$\mathcal{R}(V^*) = R^{(n)}(V^*) = \langle A + BF | \operatorname{Im} B \cap V^* \rangle.$$

As in the proof of the sufficiency of (i) of Lemma 2.8, it is possible to construct a matrix $G \in \mathcal{K}^{m\times m}$ such that $\operatorname{Im} B \cap V^* = \operatorname{Im}(BG)$. Hence, $\mathcal{R}(V^*)$ is a feedback reachability submodule for (A, B). Moreover, it follows from Algorithm 2.12 and the inclusion $V^* \subset S$ that

$$\mathcal{R}(V^*) = R^{(n)}(V^*) \subset V^* \subset S.$$

Therefore, $\mathcal{R}(V^*)$ is a feedback reachability submodule for (A, B) contained in S.

To see that $\mathcal{R}(V^*)$ is the largest feedback reachability submodule contained in S, let φ be any feedback reachability submodule for (A, B) contained in S. First, note that $\mathbf{F}(\varphi; A, B) \neq \emptyset$ because φ is feedback (A, B)-invariant, and so take any $F \in \mathbf{F}(\varphi; A, B)$. Then, it is claimed that

$$R^{(i)}(\varphi) = \sum_{j=1}^{i} (A + BF)^{j-1}(\varphi \cap \operatorname{Im} B) \quad (i = 1, 2, \ldots). \tag{2.10}$$

In fact, this can be shown in the same manner as verifying (2.8). Further, it is easily seen from (2.10) and the Cayley–Hamilton theorem that

$$R^{(n)}(\varphi) = \langle A + BF | \operatorname{Im} B \cap \varphi \rangle = \varphi.$$

Now, since φ is an (A, B)-invariant submodule contained in S, one has $\varphi \subset V^*$, and hence

$$\varphi = R^{(n)}(\varphi) \subset R^{(n)}(V^*) = \mathcal{R}(V^*),$$

showing that $\mathcal{R}(V^*)$ is the largest feedback reachability submodule for (A, B) contained in S. □

COROLLARY 2.14. *Let S be a submodule of X and V^* be the largest (A, B)-invariant submodule contained in S. If V^* is closed in X, then $\mathcal{R}(V^*)$ is the largest feedback reachability submodule for (A, B) contained in S, and it is given as*

$$\mathcal{R}(V^*) = \langle A + BF | \operatorname{Im} B \cap V^* \rangle,$$

where F is an arbitrary element in $\mathbf{F}(\mathcal{R}(V^*); A, B)$. Furthermore, the inclusion $\mathbf{F}(V^*; A, B) \subset \mathbf{F}(\mathcal{R}(V^*); A, B)$ is satisfied.

Proof. First, we will show that V^* is feedback (A, B)-invariant. Since V^* is closed in X, there exists a basis $\{x_1, x_2, \ldots, x_n\}$ of X such that $\{x_1, x_2, \ldots, x_r\}$ is a basis of V^*, where $r(\leq n)$ is the rank of V^*. Further, since V^* is (A, B)-invariant, there exist $z_1, z_2, \ldots, z_r \in V^*$ and $u_1, u_2, \ldots, u_r \in U := \mathcal{K}^m$ such that

$$Ax_i = z_i - Bu_i \quad (i = 1, 2, \ldots, r).$$

Now, letting $T := [x_1, x_2, \ldots, x_n]$ and defining

$$F := [u_1, u_2, \ldots, u_r, 0_{m\times(n-r)}]T^{-1},$$

where $0_{m\times(n-r)}$ denotes the $m \times (n-r)$ zero matrix, it is easily seen that $(A+BF)x \in V^*$ for all $x \in V^*$. Thus, V^* is feedback (A, B)-invariant and $F \in \mathbf{F}(V^*; A, B)$.

Next, take any $F \in \mathbf{F}(V^*; A, B)$. Then for any $x \in \mathcal{R}(V^*) \subset V^*$ one has $(A + BF)x \in V^*$, and further

$$(A + BF)x = Ax + BFx \in A\mathcal{R}(V^*) + \operatorname{Im} B.$$

Thus, with the help of Algorithm 2.12,

$$(A+BF)x \in V^* \cap (A\mathcal{R}(V^*) + \operatorname{Im} B) = \mathcal{R}(V^*),$$

which shows that $\mathcal{R}(V^*)$ is feedback (A,B)-invariant and that

$$\mathbf{F}(V^*;A,B) \subset \mathbf{F}(\mathcal{R}(V^*);A,B).$$

Therefore, it follows from Proposition 2.13 that $\mathcal{R}(V^*)$ is the largest feedback reachability submodule for (A,B) contained in S, and it is given as

$$\mathcal{R}(V^*) = \langle A+BF|\operatorname{Im} B \cap V^*\rangle,$$

where F is an arbitrary element in $\mathbf{F}(\mathcal{R}(V^*);A,B)$. □

3. Block triangular decoupling. Consider a system (A,B,C) over $\boldsymbol{\mathcal{K}}$, and suppose that the output matrix C is partitioned into k blocks as

$$C = \begin{bmatrix} C_1 \\ C_2 \\ \vdots \\ C_k \end{bmatrix},$$

where $C_i \in \boldsymbol{\mathcal{K}}^{l_i \times n}$ and $l_1 + l_2 + \cdots + l_k = l$. Then, system (A,B,C) is represented as $(A,B,\{C_i\}_{i=1}^k)$, which defines the following system:

$$\begin{cases} x(t+1) &= Ax(t) + Bu(t), \\ y_i(t) &= C_i x(t) \quad (i = 1,2,\ldots,k). \end{cases}$$

Now, apply to this system a state feedback control of the form

$$u(t) = Fx(t) + \sum_{i=1}^{k} G_i v_i(t), \tag{3.1}$$

where $F \in \boldsymbol{\mathcal{K}}^{m\times n}$, $G_i \in \boldsymbol{\mathcal{K}}^{m\times m}$, and $v_i(t)$ are the new inputs. Then, the resulting closed loop system is easily seen to be $(A+BF, B[G_1,G_2,\ldots,G_k], \{C_i\}_{i=1}^k)$; i.e.,

$$\begin{cases} x(t+1) &= (A+BF)x(t) + \sum_{i=1}^{k} BG_i v_i(t), \\ y_i(t) &= C_i x(t) \quad (i = 1,2,\ldots,k). \end{cases}$$

Accordingly, the feedback reachability submodule φ_i generated by the input v_i is given by

$$\varphi_i = \langle A+BF|\operatorname{Im}(BG_i)\rangle.$$

Roughly speaking, the *block triangular decoupling problem* for $(A,B,\{C_i\}_{i=1}^k)$ can be stated as follows [9]: find a state feedback (3.1) such that the resulting closed loop system $(A+BF, B[G_1,G_2,\ldots,G_k], \{C_i\}_{i=1}^k)$ should control the output $y_1, y_2, \ldots, y_k$ sequentially; that is to say, v_1 controls y_1, possibly changing the values of $y_2, y_3, \ldots, y_k$, then v_2 controls y_2, possibly changing the values of $y_3, y_4, \ldots, y_k$ without allowing influence on y_1, and so forth, with v_k controlling y_k without allowing influence on

$y_1, y_2, \ldots, y_{k-1}$. As in the case of the field $\boldsymbol{R}$ [9], this problem can be described more precisely as follows: given system $(A, B, \{C_i\}_{i=1}^k)$, the problem is to find, if possible, $F \in \boldsymbol{\mathcal{K}}^{m\times n}$ and $G_i \in \boldsymbol{\mathcal{K}}^{m\times m}$ $(i = 1, 2, \ldots, k)$ such that the feedback reachability submodules given by

$$\varphi_i = \langle A + BF | \operatorname{Im}(BG_i) \rangle \quad (i = 1, 2, \ldots, k) \tag{3.2}$$

satisfy the following two conditions: the first is

$$\varphi_1 \subset X =: \Omega_1, \tag{3.3}$$

$$\varphi_i \subset \bigcap_{j=1}^{i-1} \operatorname{Ker} C_j =: \Omega_i \quad (i = 2, 3, \ldots, k), \tag{3.4}$$

which requires that v_i does not affect the outputs $y_1, \ldots, y_{i-1}$, where $\operatorname{Ker} C_j$ stands for the kernel of C_j, and the second is

$$\varphi_i + \operatorname{Ker} C_i = X \quad (i = 1, 2, \ldots, k), \tag{3.5}$$

which requires that v_i controls the corresponding output y_i completely. Furthermore, using Lemma 2.8, this problem can be rephrased in a more compact form as follows.

Problem 3.1. Given system $(A, B, \{C_i\}_{i=1}^k)$ over $\boldsymbol{\mathcal{K}}$, the block triangular decoupling problem is to find, if possible, a set $\{\varphi_1, \varphi_2, \ldots, \varphi_k\}$ of feedback reachability submodules for (A, B) such that

$$\bigcap_{i=1}^{k} \mathbf{F}(\varphi_i; A, B) \neq \emptyset, \tag{3.6}$$

$$\varphi_i \subset \Omega_i, \tag{3.7}$$

$$\varphi_i + \operatorname{Ker} C_i = X \quad (i = 1, 2, \ldots, k). \tag{3.8}$$

Such a set $\{\varphi_1, \varphi_2, \ldots, \varphi_k\}$ is called a *solution* of the problem. □

For systems over the field $\boldsymbol{R}$ of real numbers, this problem was first studied by Morse and Wonham [9] where it was shown that necessary and sufficient conditions for its solvability are given by

$$\varphi_i^* + \operatorname{Ker} C_i = X \quad (i = 1, 2, \ldots, k), \tag{3.9}$$

where φ_i^* is the largest feedback reachability subspace contained in Ω_i, and further that in this case $\{\varphi_1^*, \varphi_2^*, \ldots, \varphi_k^*\}$ is a solution of the problem. Accordingly, if Problem 3.1 with $\boldsymbol{\mathcal{K}} = \boldsymbol{R}$ is solvable then $\{\varphi_1^*, \varphi_2^*, \ldots, \varphi_k^*\}$ is always a solution of it, and vice versa. Of course, it does not mean that $\{\varphi_1^*, \varphi_2^*, \ldots, \varphi_k^*\}$ is the only solution. Although the largest feedback reachability subspaces φ_i^* always exist for systems over $\boldsymbol{R}$, this property does not hold in general for systems over $\boldsymbol{\mathcal{K}}$ as stated earlier. However, Proposition 2.13 and Corollary 2.14 give some sufficient conditions for existence of such largest submodules for systems over $\boldsymbol{\mathcal{K}}$.

Therefore, in investigating Problem 3.1 in the geometric approach, it is inevitable to make the assumption that such largest submodules exist, and our main concern is whether (3.9) also gives necessary and sufficient conditions for Problem 3.1 to be solvable under this assumption. The answer is yes. However, its proof involves a number of detailed technical verifications which are not needed for the case of systems over the field $\boldsymbol{R}$. First, we make the following assumption.

Assumption 3.2. For system $(A, B, \{C_i\}_{i=1}^k)$ over $\boldsymbol{\mathcal{K}}$, it is assumed that there exists the largest feedback reachability submodule φ_i^* for (A, B) contained in Ω_i $(i = 1, 2, \ldots, k)$. □

It should be remarked from Corollary 2.14 and Proposition 2.13 that this assumption is satisfied either if every largest (A, B)-invariant submodule V_i^* contained in Ω_i is closed in X or if each submodule $\mathcal{R}(V_i^*)$ is feedback (A, B)-invariant. Further, it should be remarked that $\varphi_i^* = \mathcal{R}(V_i^*)$ if Assumption 3.2 is satisfied. Finally, we note that $\varphi_1^*, \varphi_2^*, \ldots, \varphi_k^*$ satisfy

$$\varphi_k^* \subset \varphi_{k-1}^* \subset \cdots \subset \varphi_1^*. \tag{3.10}$$

Before proving our main theorem, the following lemma will be shown.

LEMMA 3.3. *If $\varphi \subset X$ is a feedback (A, B)-invariant submodule, then*

$$\mathbf{F}(\varphi; A, B) \subset \mathbf{F}(\mathrm{Cl}_X(\varphi); A, B).$$

Proof. Take any $F \in \mathbf{F}(\varphi; A, B)$. Then, for any $x \in \mathrm{Cl}_X(\varphi)$ there exists a nonzero element $\alpha \in \boldsymbol{\mathcal{K}}$ such that $\alpha x \in \varphi$, and hence $\alpha(A + BF)x = (A + BF)(\alpha x) \in \varphi$. Therefore, one has $(A + BF)x \in \mathrm{Cl}_X(\varphi)$, showing $F \in \mathbf{F}(\mathrm{Cl}_X(\varphi); A, B)$. □

Now, we are ready to prove our main theorem.

THEOREM 3.4. *Suppose that system $(A, B, \{C_i\}_{i=1}^k)$ satisfies Assumption* 3.2. *Then the block triangular decoupling Problem* 3.1 *is solvable if and only if*

$$\varphi_i^* + \mathrm{Ker}\, C_i = X \quad (i = 1, 2, \ldots, k). \tag{3.11}$$

Moreover, in this case the set $\{\varphi_1^*, \varphi_2^*, \ldots, \varphi_k^*\}$ is a solution of the problem.

Proof. First, the sufficiency will be proved. So assume that (3.11) is satisfied. Then the definition of φ_i^* and (3.11) imply that the set $\{\varphi_1^*, \varphi_2^*, \ldots, \varphi_k^*\}$ satisfies (3.7) and (3.8) with φ_i replaced by φ_i^*. Therefore, if $\{\varphi_1^*, \varphi_2^*, \ldots, \varphi_k^*\}$ is shown to satisfy (3.6), then it is a solution of the problem; hence both the sufficiency and the second assertion in the theorem are proved simultaneously.

To verify that $\{\varphi_1^*, \varphi_2^*, \ldots, \varphi_k^*\}$ satisfies (3.6), first recall the relation of (3.10), and choose submodules $N_0, N_1, \ldots, N_k$ through the equations

$$\begin{aligned} &X = N_0 \oplus \mathrm{Cl}_X(\varphi_1^*), \\ &\mathrm{Cl}_X(\varphi_i^*) = N_i \oplus \mathrm{Cl}_X(\varphi_{i+1}^*) \quad (i = 1, \ldots, k-1), \\ &N_k := \mathrm{Cl}_X(\varphi_k^*). \end{aligned} \tag{3.12}$$

Then it is easy to obtain the relation

$$X = N_0 \oplus N_1 \oplus \cdots \oplus N_k. \tag{3.13}$$

Since each N_i is free, it has a basis $\{x_{i1}, x_{i2}, \ldots, x_{ir_i}\}$, where r_i denotes the rank of N_i. So, if matrices $T_i \in \boldsymbol{\mathcal{K}}^{n \times r_i}$ and $T \in \boldsymbol{\mathcal{K}}^{n \times n}$ are defined by

$$\begin{aligned} T_i &:= [x_{i1}, x_{i2}, \ldots, x_{ir_i}] \quad (i = 0, 1, \ldots, k), \\ T &:= [T_0, T_1, \ldots, T_k], \end{aligned}$$

then it is easy to see from (3.13) that T is invertible over $\boldsymbol{\mathcal{K}}$.

Next, since each φ_i^* is a feedback reachability submodule for (A, B), it is feedback (A, B)-invariant and hence satisfies

$$\mathbf{F}(\varphi_i^*; A, B) \neq \emptyset.$$

So, choosing an $F_i \in \mathbf{F}(\varphi_i^*; A, B)$ and defining $F \in \boldsymbol{\mathcal{K}}^{m\times n}$ by

$$F := [0_{m\times r_0}, F_1T_1, F_2T_2, \ldots, F_kT_k]T^{-1},$$

it is claimed that $F \in \bigcap_{i=1}^{k} \mathbf{F}(\varphi_i^*; A, B)$, i.e., that $\{\varphi_1^*, \varphi_2^*, \ldots, \varphi_k^*\}$ satisfies (3.6). To verify this, introduce the following feedback reachability submodules for (A, B):

$$\widetilde{\varphi}_i := \langle A + BF | \operatorname{Im} B \cap \varphi_i^* \rangle \quad (i = 1, 2, \ldots, k), \tag{3.14}$$

and we will first show that

$$\widetilde{\varphi}_i \subset \Omega_i \quad (i = 1, 2, \ldots, k). \tag{3.15}$$

To show (3.15), first notice that

$$F = \begin{cases} 0 & \text{on} \quad N_0, \\ F_i & \text{on} \quad N_i \end{cases} \quad (i = 1, 2, \ldots, k) \tag{3.16}$$

and further that by Lemma 3.3

$$F_i \in \mathbf{F}(\operatorname{Cl}_X(\varphi_i^*); A, B). \tag{3.17}$$

Now, (3.12), (3.16), and (3.17) imply that

$$(A + BF)\operatorname{Cl}_X(\varphi_k^*) = (A + BF_k)\operatorname{Cl}_X(\varphi_k^*) \subset \operatorname{Cl}_X(\varphi_k^*), \tag{3.18}$$

showing $F \in \mathbf{F}(\operatorname{Cl}_X(\varphi_k^*); A, B)$. Next, suppose that for some j $(2 \le j \le k)$

$$F \in \mathbf{F}(\operatorname{Cl}_X(\varphi_j^*); A, B),$$

or equivalently

$$(A + BF)\operatorname{Cl}_X(\varphi_j^*) \subset \operatorname{Cl}_X(\varphi_j^*). \tag{3.19}$$

Then, it follows from (3.12), (3.16), (3.17), and (3.19) that

$$\begin{aligned} (A + BF)\operatorname{Cl}_X(\varphi_{j-1}^*) &= (A + BF)(N_{j-1} + \operatorname{Cl}_X(\varphi_j^*)) \\ &\subset (A + BF_{j-1})N_{j-1} + (A + BF)\operatorname{Cl}_X(\varphi_j^*) \\ &\subset \operatorname{Cl}_X(\varphi_{j-1}^*) + \operatorname{Cl}_X(\varphi_j^*) \\ &= \operatorname{Cl}_X(\varphi_{j-1}^*). \end{aligned} \tag{3.20}$$

Hence, (3.20) together with (3.18) and (3.19) implies that

$$(A + BF)\operatorname{Cl}_X(\varphi_i^*) \subset \operatorname{Cl}_X(\varphi_i^*) \quad (i = 1, 2, \ldots, k).$$

Thus, noticing that $\varphi_i^* \subset \Omega_i$ and that Ω_i are closed in X, one obtains

$$\begin{gathered} (A + BF)^{j-1}\varphi_i^* \subset (A + BF)^{j-1}\operatorname{Cl}_X(\varphi_i^*) \subset \operatorname{Cl}_X(\Omega_i) = \Omega_i, \\ (j = 1, 2, \ldots, n; i = 1, 2, \ldots, k), \end{gathered}$$

which together with (3.14) imply that (3.15) holds.

Next, we will prove equalities $\varphi_i^* = \widetilde{\varphi}_i$ $(i = 1, 2, \ldots, k)$. First, note that since $\widetilde{\varphi}_i$ is a feedback reachability submodule contained in Ω_i and φ_i^* is the largest feedback reachability submodule in Ω_i, one has

$$\widetilde{\varphi}_i \subset \varphi_i^* \quad (i = 1, 2, \ldots, k). \tag{3.21}$$

To show the converse inclusion, fix i $(1 \leq i \leq k)$ and suppose that for some j $(1 \leq j \leq n-1)$

$$(A+BF_i)^{j-1}(\operatorname{Im} B \cap \varphi_i^*) \subset \widetilde{\varphi}_i. \tag{3.22}$$

Then,

$$\begin{aligned}(A+BF_i)^{j}(\operatorname{Im} B \cap \varphi_i^*) &\subset (A+BF_i)\widetilde{\varphi}_i \\ &= [A+BF+B(F_i-F)]\widetilde{\varphi}_i \\ &\subset (A+BF)\widetilde{\varphi}_i + B(F_i-F)\widetilde{\varphi}_i. \end{aligned} \tag{3.23}$$

Now, for $x \in \widetilde{\varphi}_i$ one has

$$B(F_i-F)x \in \operatorname{Im} B$$

and furthermore by noticing $F_i \in \mathbf{F}(\varphi_i^*; A, B)$ and using (3.21)

$$\begin{aligned} B(F_i-F)x &= (A+BF_i)x - (A+BF)x \\ &\in \varphi_i^* + \widetilde{\varphi}_i = \varphi_i^*, \end{aligned} \tag{3.24}$$

and hence $B(F_i-F)\widetilde{\varphi}_i \subset \operatorname{Im} B \cap \varphi_i^*$. Thus, it follows from (3.23) and (3.14) that

$$(A+BF_i)^{j}(\operatorname{Im} B \cap \varphi_i^*) \subset (A+BF)\widetilde{\varphi}_i + \operatorname{Im} B \cap \varphi_i^* \subset \widetilde{\varphi}_i.$$

Hence, it has been shown that (3.22) holds true for all $j = 1, 2, \ldots, n$ because by (3.14) the inclusion (3.22) is satisfied for $j = 1$. Now, using Lemma 2.8 one has

$$\begin{aligned}\varphi_i^* &= \langle A+BF_i | \operatorname{Im} B \cap \varphi_i^* \rangle \\ &= \sum_{j=1}^{n} (A+BF_i)^{j-1}(\operatorname{Im} B \cap \varphi_i^*) \subset \widetilde{\varphi}_i \quad (i = 1, 2, \ldots, k),\end{aligned}$$

which together with (3.21) yields the desired result

$$\varphi_i^* = \widetilde{\varphi}_i \ (i = 1, \ldots, k). \tag{3.25}$$

Now, (3.14) and (3.25) lead us to the identities

$$\varphi_i^* = \langle A+BF | \operatorname{Im} B \cap \varphi_i^* \rangle \quad (i = 1, 2, \ldots, k),$$

which imply that

$$F \in \bigcap_{i=1}^{k} \mathbf{F}(\varphi_i^*; A, B).$$

Thus, $\{\varphi_1^*, \varphi_2^*, \ldots, \varphi_k^*\}$ satisfies (3.6).

Finally, to prove the necessity of the theorem, assume Problem 3.1 is solvable and let $\{\varphi_1, \varphi_2, \ldots, \varphi_k\}$ be an arbitrary solution of the problem. Then, since every feedback reachability submodule φ_i satisfies (3.7), one has

$$\varphi_i \subset \varphi_i^* \quad (i = 1, 2, \ldots, k).$$

Hence, this inclusion with (3.8) implies that

$$X = \varphi_i + \operatorname{Ker} C_i \subset \varphi_i^* + \operatorname{Ker} C_i \subset X \quad (i = 1, 2, \ldots, k),$$

showing that (3.11) holds true. This completes the proof of Theorem 3.4. □

4. Pole assignability. The present section deals with the pole assignability problem for decoupled systems. As in the case of systems over $\boldsymbol{R}$, a system (A, B) over $\boldsymbol{\mathcal{K}}$ is said to be *pole assignable* if for any $\beta_1, \ldots, \beta_n \in \boldsymbol{\mathcal{K}}$ there exists an $F \in \boldsymbol{\mathcal{K}}^{m\times n}$ such that

$$\det(sI_n - A - BF) = \prod_{i=1}^{n}(s - \beta_i),$$

where I_n is the $n \times n$ identity matrix and $\det(\cdot)$ means determinant. It is well known [8] that a principal ideal domain is a pole-assignable ring; that is to say, (A, B) is pole assignable if and only if (A, B) is reachable.

It is also well known [12, Theorem 5.1] that if (A, B) is a system over $\boldsymbol{R}$ and if φ is a feedback reachability subspace for (A, B) with dimension $r \geq 1$ then for an arbitrary symmetric set $\{\beta_1, \ldots, \beta_r\}$ of complex numbers there exists an $F \in \mathbf{F}(\varphi; A, B)$ such that

$$\det(sI_r - (A + BF)|\varphi) = \prod_{i=1}^{r}(s - \beta_i),$$

where $(A + BF)|\varphi$ denotes the restriction of $A + BF$ onto φ as mappings. However, this fact does not hold true for systems over $\boldsymbol{\mathcal{K}}$ as seen from the following simple example.

Example 4.1. Take $\boldsymbol{\mathcal{K}}{:=}\boldsymbol{R}[\sigma]$, $X := \boldsymbol{\mathcal{K}}^2$, and

$$A := \begin{bmatrix} 1 & 1 \\ 0 & 1 \end{bmatrix}, \quad B := \begin{bmatrix} 1 & 0 \\ 1 & \sigma \end{bmatrix}, \quad \varphi = \operatorname{Im} \begin{bmatrix} 0 \\ \sigma \end{bmatrix}.$$

Then, it is easy to check that system (A, B) is reachable and that φ is not closed. Simple computation shows that φ is feedback (A, B)-invariant and any $F \in \mathbf{F}(\varphi; A, B)$ must have the form

$$F := \begin{bmatrix} f_1 & -1 \\ f_2 & f_3 \end{bmatrix}, \quad f_i \in \boldsymbol{\mathcal{K}} \ (i = 1, 2, 3).$$

Furthermore, it is not difficult to verify the relation $\langle A + BF | \operatorname{Im} B \cap \varphi \rangle = \varphi$, which together with Lemma 2.8 implies that φ is a feedback reachability submodule.

Now, since

$$A + BF = \begin{bmatrix} f_1 + 1 & 0 \\ f_1 + \sigma f_2 & \sigma f_3 \end{bmatrix},$$

it follows that $(A + BF)|\varphi = \sigma f_3$. Therefore, there is no $F \in \mathbf{F}(\varphi; A, B)$ such that $\det(sI_1 - (A + BF)|\varphi) = s + 1$, showing that -1 cannot be assigned as a pole. □

Before going into the detailed discussion, some remarks are in order. First, the meaning of restriction $(A + BF)|\varphi$ needs to be made more precise. Let $H \in \boldsymbol{\mathcal{K}}^{n\times n}$ and $\psi \subset X := \boldsymbol{\mathcal{K}}^n$ be an H-invariant submodule with rank $r \geq 1$. Then, the matrix representation H_0 of $H|\psi$ with respect to a basis $\{w_1, w_2, \ldots, w_r\}$ of ψ is given to be the matrix $H_0 \in \boldsymbol{\mathcal{K}}^{r\times r}$ uniquely determined through the equation $H[w_1\ w_2\ \cdots\ w_r] = [w_1\ w_2\ \cdots\ w_r]H_0$. Next, matrices over $\boldsymbol{\mathcal{K}}$ and various matrix operations in $\boldsymbol{\mathcal{K}}$ may be considered to be those in the quotient field of $\boldsymbol{\mathcal{K}}$ whenever it is necessary and allowable.

First, the following two lemmas are proved, which will play vital roles in what follows.

LEMMA 4.2. *Let $\varphi \subset X$ be a feedback (A,B)-invariant submodule with rank $r \geq 1$. Then, for any $F \in \mathbf{F}(\varphi; A, B)$,*

$$\det(sI_r - (A+BF)|\varphi) = \det(sI_r - (A+BF)|\operatorname{Cl}_X(\varphi)).$$

Proof. Take a basis $\{x_1, \ldots, x_r, x_{r+1}, \ldots, x_n\}$ of X and a set $\{\alpha_1, \alpha_2, \ldots, \alpha_r\} \subset \boldsymbol{\mathcal{K}}$ with $\alpha_i \neq 0$ $(i = 1, 2, \ldots, r)$ such that $\{\alpha_1 x_1, \alpha_2 x_2, \ldots, \alpha_r x_r\}$ and $\{x_1, x_2, \ldots, x_r\}$ are bases of φ and $\operatorname{Cl}_X(\varphi)$, respectively. Then for $F \in \mathbf{F}(\varphi; A, B)$, the matrix representation of $(A+BF)|\varphi$ with respect to the basis $\{\alpha_1 x_1, \alpha_2 x_2, \ldots, \alpha_r x_r\}$ is characterized by

$$(A+BF)[\alpha_1 x_1, \ldots, \alpha_r x_r] = [\alpha_1 x_1, \ldots, \alpha_r x_r](A+BF)|\varphi.$$

Now, considering all matrices over $\boldsymbol{\mathcal{K}}$ and matrix operations in $\boldsymbol{\mathcal{K}}$ as those in the quotient field of $\boldsymbol{\mathcal{K}}$, one obtains

$$\begin{aligned}
&(A+BF)[x_1, \ldots, x_r] \\
&= [x_1, \ldots, x_r] \begin{bmatrix} \alpha_1 & & 0 \\ & \ddots & \\ 0 & & \alpha_r \end{bmatrix} (A+BF)|\varphi \begin{bmatrix} \alpha_1^{-1} & & 0 \\ & \ddots & \\ 0 & & \alpha_r^{-1} \end{bmatrix},
\end{aligned}$$

which leads to the matrix representation

$$(A+BF)|\operatorname{Cl}_X(\varphi) = \begin{bmatrix} \alpha_1 & & 0 \\ & \ddots & \\ 0 & & \alpha_r \end{bmatrix} (A+BF)|\varphi \begin{bmatrix} \alpha_1^{-1} & & 0 \\ & \ddots & \\ 0 & & \alpha_r^{-1} \end{bmatrix}.$$

So, finally one obtains the desired result as follows:

$$\begin{aligned}
&\det(sI_r - (A+BF)|\operatorname{Cl}_X(\varphi)) \\
&= \det\left(sI_r - \begin{bmatrix} \alpha_1 & & 0 \\ & \ddots & \\ 0 & & \alpha_r \end{bmatrix} (A+BF)|\varphi \begin{bmatrix} \alpha_1^{-1} & & 0 \\ & \ddots & \\ 0 & & \alpha_r^{-1} \end{bmatrix} \right) \\
&= \det\left(\begin{bmatrix} \alpha_1 & & 0 \\ & \ddots & \\ 0 & & \alpha_r \end{bmatrix} (sI_r - (A+BF)|\varphi) \begin{bmatrix} \alpha_1^{-1} & & 0 \\ & \ddots & \\ 0 & & \alpha_r^{-1} \end{bmatrix} \right) \\
&= \det(sI_r - (A+BF)|\varphi). \qquad \square
\end{aligned}$$

LEMMA 4.3. *Let $\varphi \subset X$ be a feedback (A,B)-invariant submodule with its rank $r \geq 1$. If for arbitrary $\beta_1, \beta_2, \ldots, \beta_r \in \boldsymbol{\mathcal{K}}$ there exists $F \in \mathbf{F}(\varphi; A, B)$ such that*

$$\det(sI_r - (A+BF)|\varphi) = \prod_{i=1}^{r} (s - \beta_i), \tag{4.1}$$

then both φ and $\operatorname{Cl}_X(\varphi)$ are feedback reachability submodules for (A,B).

Proof. First, we will show that under the given hypothesis φ is a feedback reachability submodule for (A,B). Choose an $F_0 \in \mathbf{F}(\varphi; A, B)$ and a basis $\{x_1, x_2, \ldots, x_r\}$ of φ, and set $A_0 := (A+BF_0)|\varphi$, or equivalently, let A_0 be the $r \times r$ matrix over $\boldsymbol{\mathcal{K}}$ determined by

$$(A+BF_0)[x_1, x_2, \ldots, x_r] = [x_1, x_2, \ldots, x_r]A_0. \tag{4.2}$$

Further, letting $\widehat{B} \in \boldsymbol{\mathcal{K}}^{n\times m}$ be such that $\operatorname{Im}\widehat{B} = \operatorname{Im} B \cap \varphi$, define $B_0 \in \boldsymbol{\mathcal{K}}^{r\times m}$ so as to satisfy

$$\widehat{B} = [x_1, x_2, \ldots, x_r]B_0.$$

Now, note that $B(F - F_0)\varphi \subset \operatorname{Im} B \cap \varphi$ for any $F \in \mathbf{F}(\varphi; A, B)$. Hence, for each $F \in \mathbf{F}(\varphi; A, B)$ there exists an $F_1 \in \boldsymbol{\mathcal{K}}^{m\times r}$ such that

$$B(F - F_0)[x_1, x_2, \ldots, x_r] = \widehat{B}F_1 = [x_1, x_2, \ldots, x_r]B_0F_1. \tag{4.3}$$

Thus, (4.2) and (4.3) lead to

$$\begin{aligned}(A + BF)[x_1, x_2, \ldots, x_r] &= ((A + BF_0) + B(F - F_0))[x_1, x_2, \ldots, x_r]\\ &= [x_1, x_2, \ldots, x_r]A_0 + [x_1, x_2, \ldots, x_r]B_0F_1\\ &= [x_1, x_2, \ldots, x_r](A_0 + B_0F_1),\end{aligned}$$

showing that the matrix representation of $(A+BF)|\varphi$ is given by $A_0+B_0F_1$. Therefore, (4.1) implies that for arbitrary $\beta_1, \beta_2, \ldots, \beta_r \in \boldsymbol{\mathcal{K}}$ there exists an $F_1 \in \boldsymbol{\mathcal{K}}^{m\times r}$ such that

$$\det(sI_r - (A_0 + B_0F_1)) = \prod_{i=1}^{r}(s - \beta_i).$$

Therefore, the system (A_0, B_0) with its state module $\boldsymbol{\mathcal{K}}^r$ is reachable; that is,

$$\langle A_0 | \operatorname{Im} B_0\rangle = \boldsymbol{\mathcal{K}}^r.$$

Now, using (4.2) one obtains

$$\begin{aligned}&\langle A + BF_0 | \operatorname{Im} B \cap \varphi\rangle\\ &= \operatorname{Im}\widehat{B} + (A + BF_0)(\operatorname{Im}\widehat{B}) + \cdots + (A + BF_0)^{n-1}(\operatorname{Im}\widehat{B})\\ &= [x_1, \ldots, x_r](\operatorname{Im} B_0) + (A + BF_0)[x_1, \ldots, x_r](\operatorname{Im} B_0) + \cdots\\ &\qquad +(A + BF_0)^{n-1}[x_1, \ldots, x_r](\operatorname{Im} B_0)\\ &= [x_1, \ldots, x_r](\operatorname{Im} B_0) + [x_1, \ldots, x_r]A_0(\operatorname{Im} B_0) + \cdots\\ &\qquad +[x_1, \ldots, x_r]A_0^{n-1}(\operatorname{Im} B_0)\\ &= [x_1, \ldots, x_r]\langle A_0 | \operatorname{Im} B_0\rangle\\ &= \varphi.\end{aligned}$$

So, it follows from Lemma 2.8 that φ is a feedback reachability submodule for (A, B).

Next, to show that $\mathrm{Cl}_X(\varphi)$ is a feedback reachability submodule for (A, B), first note from Lemma 4.2 that $\det(sI_r - (A + BF)|\varphi) = \det(sI_r - (A + BF)|\mathrm{Cl}_X(\varphi))$. Therefore, replacing φ with $\mathrm{Cl}_X(\varphi)$ in the previous proof easily leads to the desired conclusion. □

The next theorem will play a key role to study the pole assignability problem for decoupled systems.

THEOREM 4.4. *Let φ be a feedback reachability submodule for (A, B) with its rank $r \geq 1$. Then for arbitrary $\beta_1, \beta_2, \ldots, \beta_r \in \boldsymbol{\mathcal{K}}$ there exists an $F \in \mathbf{F}(\varphi; A, B)$ such that*

$$\det(sI_r - (A + BF)|\varphi) = \prod_{i=1}^{r}(s - \beta_i) \tag{4.4}$$

if and only if $\mathrm{Cl}_X(\varphi)$ is a feedback reachability submodule for (A, B).

Proof. The necessity follows from a direct application of Lemma 4.3, and therefore only the sufficiency will be proved. First choose an $F_0 \in \mathbf{F}(\varphi; A, B)$. Then Lemmas 2.8 and 3.3 imply that there exist $G,\ \widehat{G} \in \boldsymbol{\mathcal{K}}^{m\times m}$ such that

$$\begin{aligned}\varphi &= \langle A + BF_0 | \operatorname{Im}(BG)\rangle,\\ \operatorname{Cl}_X(\varphi) &= \langle A + BF_0 | \operatorname{Im}(B\widehat{G})\rangle.\end{aligned}$$

Since $\operatorname{Cl}_X(\varphi)$ is closed in X, there exists a basis $\{x_1, \dots, x_r, x_{r+1}, \dots, x_n\}$ of X such that $\{x_1, \dots, x_r\}$ is a basis of $\operatorname{Cl}_X(\varphi)$. Noticing that $(A + BF_0)\operatorname{Cl}_X(\varphi) \subset \operatorname{Cl}_X(\varphi)$ and $\operatorname{Im}(B\widehat{G}) \subset \operatorname{Cl}_X(\varphi)$, introduce the matrices $A_0 \in \boldsymbol{\mathcal{K}}^{r\times r}$ and $B_0 \in \boldsymbol{\mathcal{K}}^{r\times m}$ uniquely determined by

$$(A + BF_0)[x_1, \dots, x_r] = [x_1, \dots, x_r]A_0, \quad B\widehat{G} = [x_1, \dots, x_r]B_0.$$

Then,

$$\begin{aligned}\operatorname{Cl}_X(\varphi) &= \langle A + BF_0 | \operatorname{Im}(B\widehat{G})\rangle\\ &= \operatorname{Im}(B\widehat{G}) + (A + BF_0)(\operatorname{Im}(B\widehat{G})) + \cdots\\ &\qquad +(A + BF_0)^{n-1}(\operatorname{Im}(B\widehat{G}))\\ &= [x_1, \dots, x_r](\operatorname{Im} B_0) + (A + BF_0)[x_1, \dots, x_r](\operatorname{Im} B_0) + \cdots\\ &\qquad +(A + BF_0)^{n-1}[x_1, \dots, x_r](\operatorname{Im} B_0)\\ &= [x_1, \dots, x_r](\operatorname{Im} B_0) + [x_1, \dots, x_r]A_0(\operatorname{Im} B_0) + \cdots\\ &\qquad +[x_1, \dots, x_r]A_0^{n-1}(\operatorname{Im} B_0)\\ &= [x_1, \dots, x_r]\langle A_0 | \operatorname{Im} B_0\rangle.\end{aligned}$$

Therefore, noticing that $\{x_1, x_2, \dots, x_r\}$ is a basis of $\operatorname{Cl}_X(\varphi)$ one sees that

$$\langle A_0 | \operatorname{Im} B_0\rangle = \boldsymbol{\mathcal{K}}^r$$

and hence that the system (A_0, B_0) with its state module $\boldsymbol{\mathcal{K}}^r$ is reachable. So, for any $\beta_1, \beta_2, \dots, \beta_r \in \boldsymbol{\mathcal{K}}$ there exists an $F_1 \in \boldsymbol{\mathcal{K}}^{r\times r}$ such that

$$\det(sI_r - A_0 - B_0F_1) = \prod_{i=1}^{r}(s - \beta_i). \tag{4.5}$$

Next, let matrices $T_1 \in \boldsymbol{\mathcal{K}}^{n\times r}$ and $T \in \boldsymbol{\mathcal{K}}^{n\times n}$ be given by

$$T_1 := [x_1, \dots, x_r], \quad T := [x_1, \dots, x_r, x_{r+1}, \dots, x_n],$$

respectively, and noticing that T is invertible over $\boldsymbol{\mathcal{K}}$, define

$$F := F_0 + G[F_1, 0_{n\times(n-r)}]T^{-1}.$$

Then, we claim that this F satisfies (4.4). To clarify this, first notice that $F_0 \in \mathbf{F}(\varphi; A, B)$ and $\operatorname{Im}(BG) \subset \varphi$ and hence that

$$(A + BF)\varphi \subset (A + BF_0)\varphi + BG[F_1, 0_{n\times(n-r)}]T^{-1}\varphi \subset \varphi.$$

Further, noticing that

$$\begin{aligned}(A + BF)T_1 &= ((A + BF_0) + B\widehat{G}[F_1, 0_{n\times(n-r)}]T^{-1})T_1\\ &= (A + BF_0)T_1 + B\widehat{G}F_1\\ &= T_1A_0 + T_1B_0F_1\\ &= T_1(A_0 + B_0F_1),\end{aligned}$$

one obtains the equality $(A+BF)|\mathrm{Cl}_X(\varphi) = A_0 + B_0F_1$. Therefore, (4.5) implies that

$$\det(sI_r - (A+BF)|\,\mathrm{Cl}_X(\varphi)) = \prod_{i=1}^{r}(s-\beta_i).$$

Finally, by virtue of Lemma 4.2 this F satisfies (4.4). This completes the proof of Theorem 4.4. □

Now, we are ready to show the pole assignability of block triangular decoupled systems.

THEOREM 4.5. *Suppose that system* $(A, B, \{C_i\}_{i=1}^k)$ *satisfies Assumption* 3.2 *and is reachable and further that the block triangular decoupling Problem* 3.1 *is solvable. Then, the decoupled system is pole assignable; in other words, for arbitrary* $\beta_1, \beta_2, \ldots, \beta_n \in \boldsymbol{\mathcal{K}}$ *there exists an* $F \in \bigcap_{i=1}^k \mathbf{F}(\varphi_i^*; A, B)$ *such that*

$$\det(sI_n - A - BF) = \prod_{i=1}^{n}(s-\beta_i) \tag{4.6}$$

if and only if all φ_i^* *are closed in* X.

Proof. First, notice that since $\{\varphi_1^*, \varphi_2^*, \ldots, \varphi_k^*\}$ is a solution of Problem 3.1 by Theorem 3.4, for any $F \in \bigcap_{i=1}^k \mathbf{F}(\varphi_i^*; A, B)$ there exist $G_1, G_2, \ldots, G_k \in \boldsymbol{\mathcal{K}}^{m\times m}$ such that the state feedback control law $(F, \{G_i\}_{i=1}^k)$ of (3.1) achieves the block triangular decoupling.

To prove the necessity, fix $i \in \{1, 2, \ldots, k\}$ and take a basis $\{x_1, x_2, \ldots, x_r, \ldots, x_n\}$ of X such that $\{x_1, x_2, \ldots, x_r\}$ is a basis of $\mathrm{Cl}_X(\varphi_i^*)$, where r is the rank of $\mathrm{Cl}_X(\varphi_i^*)$. Note that the matrix $T := [x_1, \ldots, x_n]$ is invertible over $\boldsymbol{\mathcal{K}}$. Let $\beta_1, \beta_2, \ldots, \beta_n \in \boldsymbol{\mathcal{K}}$ be arbitrary, and choose $F \in \bigcap_{i=1}^k \mathbf{F}(\varphi_i^*; A, B)$ to satisfy (4.6). Then, since by Lemma 3.3 $F \in \mathbf{F}(\mathrm{Cl}_X(\varphi_i^*); A, B)$, one obtains

$$A + BF = T \begin{bmatrix} (A+BF)|\,\mathrm{Cl}_X(\varphi_i^*) & * \\ 0 & * \end{bmatrix} T^{-1},$$

where $*$ indicates a suitable matrix. Hence, there exists a subset $\{q(1), \ldots, q(r)\} \subset \{1, 2, \ldots, n\}$ such that

$$\det(sI_r - (A+BF)|\,\mathrm{Cl}_X(\varphi_i^*)) = \prod_{j=1}^{r}(s-\beta_{q(j)}).$$

Since $\{\beta_1, \beta_2, \ldots, \beta_n\}$ was arbitrary, Lemmas 4.2 and 4.3 imply that $\mathrm{Cl}_X(\varphi_i^*)$ is a feedback reachability submodule for (A, B). Moreover, noticing that Ω_i is closed and $\varphi_i^* \subset \Omega_i$, one has

$$\varphi_i^* \subset \mathrm{Cl}_X(\varphi_i^*) \subset \mathrm{Cl}_X(\Omega_i) = \Omega_i.$$

Since $\mathrm{Cl}_X(\varphi_i^*)$ is a feedback reachability submodule contained in Ω_i, the supremality of φ_i^* implies $\mathrm{Cl}_X(\varphi_i^*) \subset \varphi_i^*$. Therefore, $\mathrm{Cl}_X(\varphi_i^*) = \varphi_i^*$, showing that φ_i^* is closed in X.

Next, to show the sufficiency of the theorem, first recall the relation $\varphi_1^* \supset \varphi_2^* \supset \cdots \supset \varphi_k^*$ of (3.10). Noticing that each φ_i^* is closed by the hypothesis, introduce submodules M_i of X such that

$$\begin{aligned} &\varphi_i^* = M_i \oplus \varphi_{i+1}^* \quad (i = 1, 2, \ldots, k-1), \\ &M_k := \varphi_k^*. \end{aligned}$$

Since (A, B) is reachable and φ_i^* is supremal, one has $\varphi_1^* = X$ and hence

$$X = M_1 \oplus M_2 \oplus \cdots \oplus M_k.$$

Each M_i is free, so that there exists a basis $\{x_{i1}, x_{i2}, \ldots, x_{ir_i}\}$ of M_i, where r_i denotes the rank of M_i. Now, define

$$\begin{aligned} T_i &:= [x_{i1}, x_{i2}, \ldots, x_{ir_i}] \quad (i = 1, 2, \ldots, k), \\ T &:= [T_1, T_2, \ldots, T_k], \end{aligned}$$

and note that T is invertible over $\boldsymbol{\mathcal{K}}$ and $r_1 + \cdots + r_k = n$. Recall $\bigcap_{i=1}^k \mathbf{F}(\varphi_i^*; A, B) \neq \emptyset$ because $\{\varphi_1^*, \ldots, \varphi_k^*\}$ is a solution of Problem 3.1. So choose any $F_0 \in \bigcap_{i=1}^k \mathbf{F}(\varphi_i^*; A, B)$ and set $A_0 := A + BF_0$. Then, by Lemma 2.8 there exists $G_i \in \boldsymbol{\mathcal{K}}^{m \times m}$ such that

$$\varphi_i^* = \langle A_0 | \operatorname{Im}(BG_i) \rangle \quad (i = 1, 2, \ldots, k),$$

and hence the resulting closed loop system $(A_0, B[G_1, G_2, \ldots, G_k], \{C_i\}_{i-1}^k)$ is a triangularly decoupled system as desired.

As the next step, we will show that the decoupled system $(A_0, B[G_1, G_2, \ldots, G_k], \{C_i\}_{i=1}^k)$ is arbitrarily pole assignable without destroying its decoupledness. To show this, first note that matrices $T^{-1}A_0T$ and $T^{-1}BG_i$ have the following forms:

$$T^{-1}A_0T = \begin{bmatrix} A_1 & 0 & \cdots & 0 \\ * & A_2 & \ddots & \vdots \\ \vdots & \ddots & \ddots & 0 \\ * & \cdots & * & A_k \end{bmatrix},$$

$$T^{-1}BG_i = \begin{bmatrix} 0_{(r_1 + \cdots + r_{i-1}) \times m} \\ B_i \\ *_{(r_{i+1} + \cdots + r_k) \times m} \end{bmatrix} \quad (i = 1, 2, \ldots, k),$$

where $A_i \in \boldsymbol{\mathcal{K}}^{r_i \times r_i}$, $B_i \in \boldsymbol{\mathcal{K}}^{r_i \times m}$, and $*_{(r_{i+1} + \cdots + r_k) \times m}$ denotes a suitable $(r_{i+1} + \cdots + r_k) \times m$ matrix. Since $F_0 \in \bigcap_{i=1}^k \mathbf{F}(\varphi_i^*; A, B)$, the system $(T^{-1}A_0T, T^{-1}B[G_1, \ldots, G_k], \{C_iT\}_{i=1}^k)$ is also block triangularly decoupled, and hence each subsystem (A_i, B_i) with its state module $\boldsymbol{\mathcal{K}}^{r_i}$ $(i = 1, \ldots, k)$ is reachable. Therefore, letting $\{\beta_1, \ldots, \beta_n\} \subset \boldsymbol{\mathcal{K}}$ be an arbitrary set and dividing the set into k subsets $\{\beta_{i1}, \ldots, \beta_{ir_i}\}$ $(i = 1, \ldots, k)$ in an arbitrary way, there exist $F_i \in \boldsymbol{\mathcal{K}}^{m \times r_i}$ such that

$$\det(sI_{r_i} - (A_i + B_iF_i)) = \prod_{j=1}^{r_i} (s - \beta_{ij}) \quad (i = 1, 2, \ldots, k).$$

Now, define

$$F := F_0 + [G_1F_1, G_2F_2, \ldots, G_kF_k]T^{-1}. \tag{4.7}$$

Then one obtains

$$\begin{aligned} (A + BF)\varphi_i^* &= (A + BF_0)\varphi_i^* + B[G_1F_1, G_2F_2, \ldots, G_kF_k]T^{-1}\varphi_i^* \\ &\subset \varphi_i^* + \operatorname{Im}(BG_i) \subset \varphi_i^* \quad (i = 1, 2, \ldots, k), \end{aligned}$$

and hence $F \in \bigcap_{i=1}^{k} \mathbf{F}(\varphi_i^*; A, B)$, implying that $(A+BF, B[G_1, G_2, \ldots, G_k], \{C_i\}_{i=1}^k)$ is still decoupled as desired. Further, since

$$
\begin{aligned}
&T^{-1}(A+BF)T \\
&= T^{-1}(A+BF_0)T + T^{-1}B[G_1F_1, G_2F_2, \ldots, G_kF_k] \\
&= \begin{bmatrix} A_1 & 0 & \cdots & 0 \\ * & A_2 & \ddots & \vdots \\ \vdots & \ddots & \ddots & 0 \\ * & \cdots & * & A_k \end{bmatrix} + \begin{bmatrix} B_1F_1 & 0 & \cdots & 0 \\ * & B_2F_2 & \ddots & \vdots \\ \vdots & \ddots & \ddots & 0 \\ * & \cdots & * & B_kF_k \end{bmatrix} \\
&= \begin{bmatrix} A_1+B_1F_1 & & & 0 \\ * & A_2+B_2F_2 & & \\ \vdots & \ddots & \ddots & \\ * & \cdots & * & A_k+B_kF_k \end{bmatrix},
\end{aligned}
$$

one obtains that

$$
\begin{aligned}
\det(sI_n - (A+BF)) &= \det(sI_n - T^{-1}(A+BF)T) \\
&= \prod_{i=1}^{k} \det(sI_{r_i} - (A_i + B_iF_i)) \\
&= \prod_{i=1}^{k} \prod_{j=1}^{r_i} (s - \beta_{ij}) \\
&= \prod_{i=1}^{n} (s - \beta_i).
\end{aligned}
$$

Therefore, the matrix F given by (4.7) satisfies (4.6), and hence the decoupled system is arbitrarily pole assignable. This completes the proof of the theorem. □

COROLLARY 4.6. *Suppose that system* $(A, B, \{C_i\}_{i=1}^k)$ *satisfies Assumption* 3.2 *and the reachable submodule* $\langle A|\operatorname{Im} B\rangle$ *has its rank* $r < n$ *and further that the block triangular decoupling Problem* 3.1 *is solvable. Then* r *poles of the decoupled system corresponding to the solution* $\{\varphi_1^*, \varphi_2^*, \ldots, \varphi_k^*\}$ *are arbitrarily assignable; in other words, for arbitrary* $\beta_1, \beta_2, \ldots, \beta_r \in \boldsymbol{\mathcal{K}}$ *there exists an* $F \in \bigcap_{i=1}^{k} \mathbf{F}(\varphi_i^*; A, B)$ *such that*

$$
\det(sI_r - (A+BF)|\langle A|\operatorname{Im} B\rangle) = \prod_{i=1}^{r} (s - \beta_i)
$$

if and only if all φ_i^* *are closed in* X.

Proof. This corollary easily follows from Theorems 4.5 and 4.4. In fact, for the sufficiency, noticing from the hypothesis that $\varphi_1^* = \langle A|\operatorname{Im} B\rangle$ is closed in X, consider the reachable subsystem of $(A, B, \{C_i\}_{i=1}^k)$ and apply Theorem 4.5 to this subsystem to complete the proof. For the necessity, first use Theorem 4.4 to see that $\mathrm{Cl}_X(\langle A|\operatorname{Im} B\rangle) = \mathrm{Cl}_X(\varphi_1^*)$ is a feedback reachability submodule, and then noticing that $\mathrm{Cl}_X(\varphi_1^*) \subset \Omega_1 = X$ and φ_1^* is the largest feedback reachability submodule in Ω_1, conclude that $\mathrm{Cl}_X(\varphi_1^*) = \varphi_1^*$; i.e., φ_1^* is closed in X. Finally, again considering the reachable subsystem, apply Theorem 4.5 to this subsystem to show that all φ_i^* are closed in X. □

5. Example. As an illustrative example, we will consider block triangular decoupling in Problem 3.1 for the following simple system $(A, B, \{C_i\}_{i=1}^2)$ over $\boldsymbol{\mathcal{K}} := \boldsymbol{R}[\lambda]$:

$$A := \begin{bmatrix} 0 & \lambda & -1 & 0 \\ 0 & -1 & 1 & 1 \\ 0 & -1 & 1 & 0 \\ 0 & 0 & 0 & 0 \end{bmatrix}, \quad B := \begin{bmatrix} 0 & 0 \\ 0 & 0 \\ 1 & 0 \\ 0 & 1 \end{bmatrix},$$

$$C_1 := \begin{bmatrix} 1 & 0 & 0 & 0 \end{bmatrix}, \quad C_2 := \begin{bmatrix} 0 & 1 & 0 & 0 \\ 0 & 0 & 0 & 1 \end{bmatrix}.$$

By simple computation, we see that the system (A, B) is reachable, and that the largest (A, B)-invariant submodules ψ_1^* and ψ_2^* contained, respectively, in $\Omega_1 = \boldsymbol{\mathcal{K}}^4$ and $\Omega_2 = \operatorname{Ker} C_1$ are given as

$$\psi_1^* = \boldsymbol{\mathcal{K}}^4, \quad \psi_2^* = \operatorname{Im} \begin{bmatrix} 0 & 0 \\ 0 & 1 \\ 0 & \lambda \\ 1 & 0 \end{bmatrix}.$$

Clearly ψ_1^* and ψ_2^* are closed in $\boldsymbol{\mathcal{K}}^4$, and hence by Corollary 2.14 this system satisfies Assumption 3.2; that is, there exist the largest feedback reachability submodules φ_1^* and φ_2^* contained, respectively, in Ω_1 and Ω_2. Using Algorithm 2.12, they can be computed as

$$\varphi_1^* = \mathcal{R}(\psi_1^*) = \psi_1^*, \quad \varphi_2^* = \mathcal{R}(\psi_2^*) = \psi_2^*.$$

Furthermore, it is easy to check that $\{\varphi_1^*, \varphi_2^*\}$ satisfies (3.11) of Theorem 3.4 and, hence, that the given system $(A, B, \{C_i\}_{i=1}^2)$ can be block triangularly decoupled. Next, a decoupling state feedback $(F_0, \{G_i\}_{i=1}^2)$ of the form

$$u(t) = F_0 x(t) + \sum_{i=1}^{2} G_i v_i(t)$$

will be computed. First, it is not difficult to see that such an $F_0 \in \mathbf{F}(\varphi_1^*; A, B) \cap \mathbf{F}(\varphi_2^*; A, B)$ can be chosen as

$$F_0 := \begin{bmatrix} 0 & \lambda^2 - 2\lambda + 1 & 0 & \lambda \\ 0 & 0 & 0 & 0 \end{bmatrix},$$

and as in the proof of (i) of Lemma 2.8, matrices $G_i \in \boldsymbol{\mathcal{K}}^{2\times 2}$ satisfying $\varphi_i^* = \langle A + BF_0 | \operatorname{Im}(BG_i) \rangle$ $(i = 1, 2)$ can be computed as

$$G_1 := \begin{bmatrix} 1 & 0 \\ 0 & 1 \end{bmatrix}, \quad G_2 := \begin{bmatrix} 0 & 0 \\ 0 & 1 \end{bmatrix}.$$

Now, note that the characteristic polynomial for the block triangularly decoupled system is

$$\det(sI_4 - A - BF_0) = s^2(s - \lambda + 1)(s + \lambda - 1).$$

So, following the proof of the sufficiency of Theorem 4.5, all the poles for the block triangularly decoupled system will be assigned to -1. First, note that a submodule $M_1 \subset X$ satisfying $\varphi_1^* = M_1 \oplus \varphi_2^*$ is given by

$$M_1 := \mathrm{Im} \begin{bmatrix} 1 & 0 \\ 0 & 0 \\ 0 & 1 \\ 0 & 0 \end{bmatrix}.$$

Further, set $M_2 := \varphi_2^* = \psi_2^*$. Then clearly $\boldsymbol{\mathcal{K}}^4 = M_1 \oplus M_2$. Now, defining

$$T := \begin{bmatrix} 1 & 0 & 0 & 0 \\ 0 & 0 & 0 & 1 \\ 0 & 1 & 0 & \lambda \\ 0 & 0 & 1 & 0 \end{bmatrix},$$

one easily obtains

$$T^{-1}(A + BF_0)T = \left[\begin{array}{cc|cc} 0 & -1 & 0 & 0 \\ 0 & -\lambda+1 & 0 & 0 \\ \hline 0 & 0 & 0 & 0 \\ 0 & 1 & 1 & \lambda-1 \end{array}\right],$$

$$T^{-1}BG_1 = \left[\begin{array}{cc} 0 & 0 \\ 1 & 0 \\ \hline 0 & 1 \\ 0 & 0 \end{array}\right], \quad T^{-1}BG_2 = \left[\begin{array}{cc} 0 & 0 \\ 0 & 0 \\ \hline 0 & 1 \\ 0 & 0 \end{array}\right].$$

Based on this equivalence transformation, introduce the two subsystems (A_1, B_1) and (A_2, B_2), where

$$A_1 := \begin{bmatrix} 0 & -1 \\ 0 & -\lambda+1 \end{bmatrix}, \quad A_2 := \begin{bmatrix} 0 & 0 \\ 1 & \lambda-1 \end{bmatrix},$$

$$B_1 := \begin{bmatrix} 0 & 0 \\ 1 & 0 \end{bmatrix}, \quad B_2 := \begin{bmatrix} 0 & 1 \\ 0 & 0 \end{bmatrix}.$$

Then since (A_1, B_1) and (A_2, B_2) are reachable, we can construct matrices F_i satisfying $\det(sI_2 - A_i - B_iF_i) = (s+1)^2$ as

$$F_1 := \begin{bmatrix} 1 & \lambda-3 \\ 0 & 0 \end{bmatrix}, \quad F_2 := \begin{bmatrix} 0 & 0 \\ -\lambda-1 & -\lambda^2 \end{bmatrix}.$$

Finally, using (4.7) a desired matrix $F \in \mathbf{F}(\varphi_1^*; A, B) \cap \mathbf{F}(\varphi_2^*; A, B)$ can be obtained as

$$F := F_0 + [G_1F_1, G_2F_2]T^{-1} = \begin{bmatrix} 1 & \lambda+1 & \lambda-3 & \lambda \\ 0 & -\lambda^2 & 0 & -\lambda-1 \end{bmatrix},$$

which gives $\det(sI_4 - A - BF) = (s+1)^4$. Hence, $(F, \{G_i\}_{i=1}^2)$ is a desired state feedback control law that achieves block triangular decoupling and simultaneously assigns all the poles of the decoupled system to be -1.

6. Conclusions. This paper studied in the framework of geometric approach the block triangular decoupling problem for linear systems defined over principal ideal domains. First, various properties of feedback reachability submodules were given, and then necessary and sufficient conditions for the problem to be solvable were obtained under the assumption that the largest feedback reachability submodules contained in some given submodules exist. Finally, the pole assignability for block triangularly decoupled systems was investigated.

Acknowledgments. The authors thank the referees for their helpful remarks.

REFERENCES

[1] G. Conte and A. M. Perdon, *Systems over a principal ideal domain. a polynomial model approach*, SIAM J. Control Optim., 20 (1982), pp. 112–124.
[2] G. Conte and A. M. Perdon, *The decoupling problem for systems over a ring*, Proc. 34th IEEE Conference on Decision and Control, New Orleans, LA, 1995, pp. 2041–2045.
[3] K. B. Datta and M. L. J. Hautus, *Decoupling of multivariable control systems over unique factorization domains*, SIAM J. Control Optim., 22 (1984), pp. 28–39.
[4] M. L. J. Hautus, *Controlled Invariance in Systems over Rings*, Lecture Notes in Control and Information Sciences 39, Springer-Verlag, New York, 1982, pp. 107–122.
[5] H. Inaba, N. Ito, and T. Munaka, *Decoupling and pole assignment for linear systems defined over principal ideal domains*, in Linear Circuits, Systems and Signal Processing: Theory and Application, C. I. Byrnes, C. F. Martin, and R. E. Saeks, eds., North–Holland, Amsterdam, 1988, pp. 55–62.
[6] N. Ito and H. Inaba, *Block decoupling of linear systems over principal ideal domains with the assumption that the output matrix is fullrank*, Trans. Inst. Systems, Control and Information Engineers, 3 (1990), pp. 121–127 (in Japanese).
[7] M. Kono, *Decoupling and arbitrary coefficient assignment in time-delay systems*, Systems Control Lett., 3 (1983), pp. 349–354.
[8] A. S. Morse, *Ring models for delay differential systems*, Automatica, 2 (1976), pp. 529–531.
[9] A. S. Morse and W. M. Wonham, *Triangular decoupling of linear multivariable systems*, IEEE Trans. Automat. Control, 15 (1970), pp. 447–449.
[10] E. D. Sontag, *Linear systems over commutative rings. a survey*, Ricerche di Automatica, 7 (1976), pp. 1–34.
[11] E. D. Sontag, *An introduction to the stabilization problem for parametrized families of linear systems*, in Linear Algebra and its Role in Systems Theory, Contemporary Mathematics 47, AMS, Providence, RI, 1985, pp. 369–400.
[12] W. M. Wonham, *Linear Multivariable Control: A Geometric Approach*, 3rd ed., Springer-Verlag, New York, 1985.
[13] W. M. Wonham and A. S. Morse, *Decoupling and pole-assignment in linear multivariable systems. A geometric approach*, SIAM J. Control Optim., 8 (1970), pp. 1–18.

SIAM J. CONTROL OPTIM.
Vol. 35, No. 3, pp. 766–790, May 1997

003

CONFIGURATION CONTROLLABILITY OF SIMPLE MECHANICAL CONTROL SYSTEMS*

ANDREW D. LEWIS† AND RICHARD M. MURRAY†

Abstract. In this paper we present a definition of "configuration controllability" for mechanical systems whose Lagrangian is kinetic energy with respect to a Riemannian metric minus potential energy. A computable test for this new version of controllability is derived. This condition involves an object which we call the *symmetric product*. Of particular interest is a definition of "equilibrium controllability" for which we are able to derive computable sufficient conditions. Examples illustrate the theory.

Key words. mechanics, Riemannian geometry, controllability, symmetric product

AMS subject classifications. 53B20, 70H35, 70Q05, 93B03, 93B03, 93B29

PII. S0363012995287155

1. Introduction.

Mechanical systems form a large subset of control systems which have many diverse applications. These systems are characterized by a rich structure which has been underexploited in the current controls literature. In this paper we utilize the structure of a specific class of mechanical systems to obtain controllability results which are meaningful for these systems. These results are important in two respects. First, they identify the structure of mechanical systems which lends to controllability of these systems. Second, the results provide computable checks for useful notions of controllability. One important aspect of our work is that the computations for checking controllability are performed on the configuration space and not on the phase space. This is important since the phase space has twice the dimension of the configuration space for mechanical systems.

Much of the previous work in the area of mechanical control systems has relied on specific structure of these systems. Bloch and Crouch [2] study mechanical systems on Riemannian manifolds. Under suitable hypotheses on the inputs and assuming some group symmetries for the systems under investigation, the authors are able to use a result of San Martin and Crouch [10] to arrive at a controllability result. Mechanical systems with nonholonomic constraints are studied by Bloch, Reyhanoglu, and McClamroch [3]. In this paper the authors are able to show that the systems considered are controllable if the inputs span a complement to the constraint forces. In both of the above papers, the results are limited by the hypotheses placed on the system: symmetries in the first case, and constraints in the second. In this paper we attempt to develop control theoretic tools for *mechanical* control systems. We emphasize mechanical because it is our intent to use the mechanical structure to advantage in the control problem rather than any additional structure imposed on the system.

In section 2 we motivate the development of the paper by posing various controllability questions for a simple example. In this section we also preview the results of the paper by stating a simplified form of the most general results. In section 3 we present enough background from the theory of free Lie algebras and Riemannian

*Received by the editors June 5, 1995; accepted for publication (in revised form) March 13, 1996. This research was supported in part by the Powell Foundation.
http://www.siam.org/journals/sicon/35-3/28715.html
†Division of Engineering and Applied Science 104–44, California Institute of Technology, Pasadena, CA 91125 (andrew@indra.caltech.edu, murray@indra.caltech.edu).

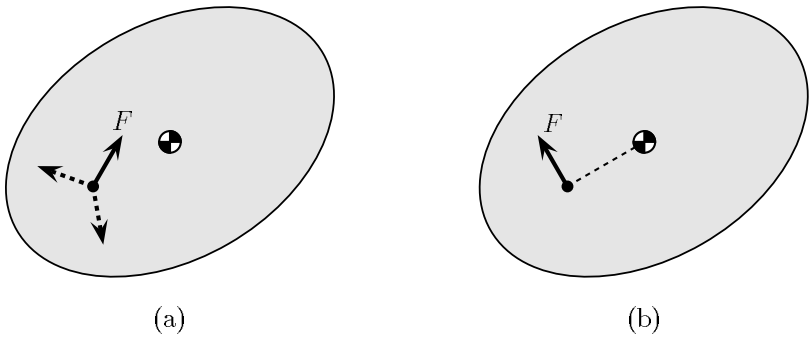


FIG. 1. *A planar rigid body with a variable direction thruster* (a) *and a fixed directional thruster* (b).

geometry that we can use these ideas in section 5. In section 3.2 we introduce the notion of a symmetric algebra which is new and will be particularly interesting to us. The *symmetric product* is defined in section 3.3. This is an interesting object whose geometric meaning is not fully utilized in this paper. However, it proves to be a useful computational tool for expressing our controllability results. In section 4 we state a result of Sussmann [13] which we shall use to prove some controllability results in section 5. The main results of the paper are stated in section 5. Illustrative examples are given in section 6.

2. Preliminary statement of results. It is possible to state a subset of the results of the paper without going through all of the formality needed to state the most general results. In this section we give some idea of the questions that we answer in the paper as well as state the results in the case when no potential energy is present.

Consider the planar rigid body system of Figure 1. On this body we consider two possible sets of forces. In one case we are able to apply a force in any direction to the body at a point away from the center of mass (case (a) in the figure). In the other case, we can apply a force which only is in a direction perpendicular to the line joining the point of application of the force with the center of mass (case (b) in the figure). The reader may wish to consider the former case as corresponding to having a thruster on the body whose direction may be varied, while in the second case the thruster can only provide thrust in one direction. In each of these cases one may ask certain questions about the controllability of this system. We list some of these questions below and in parentheses give the name of the general notion corresponding to this question.

(1) Starting from rest at a given configuration, is it possible to reach an open set of configurations? (local configuration accessibility)

(2) Starting from rest in a given configuration, is it possible to reach a neighborhood of the initial configuration? (local configuration controllability)

(3) Is it possible to get to these configurations with zero velocity? (equilibrium controllability)

It is exactly these questions which we address in this paper. Observe that the above controllability questions have the feature that the initial velocity is assumed to be zero. This turns out to greatly simplify the controllability computations. We observe that for this example the linearization is not controllable, so if the system is controllable, nonlinear tools must be employed.

Although we delay answering the above questions for the planar rigid body until section 6.2, we may state general results for a class of systems smaller than the general class we consider in the sequel. Let us consider, for the moment, mechanical systems whose Lagrangian is kinetic energy with respect to a Riemannian metric g on the configuration manifold Q. Suppose that the inputs are modeled by vector fields $\mathcal{Y} = \{Y_1, \dots, Y_m\}$. We may define the *symmetric product* between two vector fields on Q by

$$\langle X : Y \rangle = \nabla_X Y + \nabla_Y X,$$

where $\nabla_X Y$ is the *covariant derivative* of Y with respect to X. If $\mathfrak{X}(Q)$ denotes the set of vector fields on Q and if $\mathcal{V} \subset \mathfrak{X}(Q)$, we denote by $\overline{\text{Sym}}(\mathcal{V})$ the set of vector fields on Q obtained by taking iterated symmetric products of vector fields from $\mathcal{V}$. The usual involutive closure of $\mathcal{V}$ will be denoted $\overline{\text{Lie}}(\mathcal{V})$. We shall say that a symmetric product from $\overline{\text{Sym}}(\mathcal{Y})$ is *bad* if it contains an even number of each of the vector fields in $\mathcal{Y}$. Otherwise we shall call a symmetric product from $\overline{\text{Sym}}(\mathcal{Y})$ *good*. The *degree* of an iterated symmetric product of factors from $\mathcal{Y}$ will denote the total number of factors.

Notice that with the Lagrangian given by just kinetic energy, all configurations with zero velocity are equilibrium points for the unforced mechanical system. We shall say the system is *locally configuration accessible* at $q \in Q$ if the set of points reachable starting from q at zero velocity is open in Q. We shall say the system is *equilibrium controllable* if, starting from a given configuration at zero velocity, we can reach an open set of final configurations at zero velocity. Now we may state two results.

THEOREM 2.1. *Consider the mechanical control system on the configuration manifold Q whose Lagrangian is the kinetic energy with respect to a Riemannian metric g and whose input vector fields are $\mathcal{Y} = \{Y_1, \dots, Y_m\}$. Then*

(i) *the system is locally configuration accessible at q if the distribution defined by $\overline{\text{Lie}}(\overline{\text{Sym}}(\mathcal{Y}))$ has maximal rank at q;*

(ii) *the system is equilibrium controllable if it is locally configuration accessible and if every bad symmetric product is a linear combination of good symmetric products of lower degree.*

To prove this result, one basically proceeds as follows. Compute the accessibility distribution on TQ for the mechanical control system and evaluate at zero velocity. This will describe the set of *states* accessible from points of zero velocity. However, since we are interested in controllability of the *configurations*, we can project the accessibility distribution to Q with $T\tau_Q$, the derivative of the tangent bundle projection. It turns out that this is exactly the distribution $\overline{\text{Lie}}(\overline{\text{Sym}}(\mathcal{Y}))$. In this way we see that the conditions in (i) give local configuration accessibility. To prove (ii), we appeal to the controllability results of Sussmann [13] on local controllability. An application of Sussmann's results to the systems we are considering yields (ii).

The sections which follow formalize the above definitions and results and also generalize them to the case where the system has potential energy.

3. Mathematical preliminaries. In this section we present the necessary mathematical ideas that we shall need for our exposition of section 5.

3.1. Free Lie algebras and families of vector fields. In this section we recall some ideas for Lie algebras as presented by Serre [11]. These ideas will be important in our adaptation of Sussmann's conditions for small-time local controllability [13] as well as for some bracket calculations in section 5.1.

Let X be a set and let $A(X)$ be the free algebra of associative but not necessarily commutative products of elements in X. Let I be the two-sided ideal of $A(X)$ generated by elements of the form $a \cdot a$ and $a \cdot (b \cdot c) + c \cdot (a \cdot b) + b \cdot (c \cdot a)$. The algebra $L(X) = A(X)/I$ is called the *free Lie algebra* generated by X. The inherited product on this algebra satisfies the usual Lie bracket properties of a Lie algebra. We denote by $\mathrm{Br}(X)$ the subset of $L(X)$ consisting of brackets whose elements are in X. This subset generates $L(X)$ as a real vector space. In fact, the following proposition, whose proof may be found in Jacobson [6], gives a subset of $\mathrm{Br}(X)$ which generates $L(X)$.

PROPOSITION 3.1. *Every element of $L(X)$ is a linear combination of repeated brackets of the form*

$$[X_k, [X_{k-1}, [\ldots, [X_2, X_1], \ldots]]] \tag{1}$$

where $X_i \in X, i = 1, \ldots, k$.

We will need the notion of what we shall call the components of an element $u \in L(X)$. Every such element u has a unique decomposition as $u = [u_1, u_2]$. In turn, each of u_1 and u_2 may be uniquely expressed as $u_1 = [u_{11}, u_{12}]$ and $u_2 = [u_{21}, u_{22}]$. This process may be continued until we end up with elements which are not decomposable. All such elements $u_{i_1,\ldots,i_m}$, $i_a \in \{1, 2\}$, shall be called *components* of u.

If $\boldsymbol{X} = \{X_0, \ldots, X_l\}$, for $B \in \mathrm{Br}(\boldsymbol{X})$ we define $\delta_a(B)$, $a = 0, \ldots, l$, to be the number of times that X_a occurs in B. The sum of the δ_a's we shall call the *degree* of B.

Given a family of vector fields on a manifold M, $\mathcal{V} \subset \mathfrak{X}(M)$, we may define a distribution on M by

$$D_{\mathcal{V}}(x) = \mathrm{span}_{\mathbb{R}}\{X(x) \mid \quad X \in \mathcal{V}\}.$$

Since $\mathfrak{X}(M)$ is a Lie algebra, we may ask for the smallest Lie subalgebra of $\mathfrak{X}(M)$ which contains a family of vector fields $\mathcal{V}$. It is convenient to describe this subalgebra using the ideas from free Lie algebras presented above.

Let X be a set which is bijective to $\mathcal{V}$ with bijection ϕ. Thus, with each element of X we associate a vector field in $\mathcal{V}$. We establish a Lie algebra homomorphism, $\mathrm{Ev}(\phi) : L(X) \to \mathfrak{X}(M)$, in a natural manner. Thus we define $\mathrm{Ev}(\phi)$ so that $[\mathrm{Ev}(\phi)(B_1), \mathrm{Ev}(\phi)(B_2)] = \mathrm{Ev}(\phi)([B_1, B_2])$ for $B_1, B_2 \in \mathrm{Br}(X)$ and then extend this to $L(X)$ by $\mathbb{R}$-linearity. The smallest Lie subalgebra of $\mathfrak{X}(M)$ which contains $\mathcal{V}$ may now be stated in a simple manner. It is simply the image of $L(X)$ under the homomorphism $\mathrm{Ev}(\phi)$. We shall denote this subalgebra by $\overline{\mathrm{Lie}}(\mathcal{V})$ and call it the *involutive closure* of $\mathcal{V}$.

For $x \in M$ we define the map $\mathrm{Ev}_x(\phi) : L(X) \to T_xM$ by

$$\mathrm{Ev}_x(\phi)(u) = (\mathrm{Ev}(\phi)(u))(x).$$

We shall say that $\mathcal{V}$ satisfies the *Lie algebra rank condition* (LARC) at x if $\mathrm{Ev}_x(\phi)(L(X)) = T_xM$.

It is possible to talk about the involutive closure and the LARC without using free Lie algebras. However, since we will have to use free Lie algebras later in the paper, using them here provides us an opportunity to introduce the ideas in a more straightforward setting.

3.2. Symmetric algebra. As far as we know, the idea of a symmetric algebra does not appear in the literature. However, the concept is a natural one and shall be useful to us. A *symmetric algebra* is an algebra, A, where the multiplication (which

we shall denote by $(u, v) \mapsto \langle u : v \rangle$) satisfies $\langle u : v \rangle = \langle v : u \rangle$ for $u, v \in A$. A map $\sigma : A \to A'$ between symmetric algebras is called a *symmetric algebra homomorphism* if $\sigma(\langle u : v \rangle) = \langle \sigma(u) : \sigma(v) \rangle$ for each $u, v \in A$.

We now construct a symmetric algebra which is generated by a given set X. To construct this algebra, let X be a set, and recall that $A(X)$ is the free algebra on X. The *free symmetric algebra* on X, denoted $S(X)$, is the quotient algebra obtained by taking the quotient of $A(X)$ by the two-sided ideal generated by all elements of the form $a \cdot b - b \cdot a$, where $a, b \in A(X)$. We shall denote the product in $S(X)$ by $\langle u : v \rangle$. Note that, by construction, $\langle u : v \rangle = \langle v : u \rangle$ for every $u, v \in S(X)$. We denote by $\mathrm{Pr}(X)$ the subset of $S(X)$ consisting of the symmetric products whose elements are in X.

As with free Lie algebras, the finitely generated case is the most interesting to us. Let $\boldsymbol{Y} = \{X_1, \ldots, X_{l+1}\}$ (the reason for the slightly unusual enumeration will become clear in section 5.5). For $P \in \mathrm{Pr}(\boldsymbol{Y})$ define $\gamma_a(P)$ to be the number of times the element X_a occurs in $P \in \mathrm{Pr}(\boldsymbol{Y})$ for $a = 1, \ldots, l+1$. We shall call the sum of the γ_a's the *degree* of P.

3.3. Some Riemannian geometry. The kinetic energy of a mechanical system may be regarded as being determined by a *Riemannian metric* on the configuration space. A Riemannian metric g on a manifold M is simply a smooth assignment of an inner product for each tangent space of the manifold. In a set of coordinates $(x^1, \ldots, x^n)$ for M, the components of the metric are given by $g_{ij} = g(\frac{\partial}{\partial x^i}, \frac{\partial}{\partial x^j})$. For each $x \in M$, we may define isomorphisms $\sharp$: $T_x^*M \to T_xM$ and $\flat$: $T_xM \to T_x^*M$ in the usual manner (see [7]). These maps naturally extend to isomorphisms from $\mathfrak{X}(M)$, the set of vector fields on M, to $\Omega^1(M)$, the set of one-forms on M. In this case, given a function $f \in C^\infty(M)$, we define $\mathrm{grad} f = (\boldsymbol{d}f)^\sharp$.

A Riemannian manifold is endowed with a unique *affine connection* (called the Levi–Cività connection), which is characterized by being torsion free and by its parallel transportation being metric preserving (see [7]). This affine connection defines $\nabla_X Y$, which is called the *covariant derivative* of Y with respect to X. In coordinates we have

$$\nabla_X Y = \left(\frac{\partial Y^i}{\partial x^j} X^j + \Gamma^i_{jk} X^j Y^k \right) \frac{\partial}{\partial x^i}.$$

The Γ^i_{jk} are the *Christoffel symbols* and are given by

$$\Gamma^i_{jk} = \frac{1}{2} g^{il} \left(\frac{\partial g_{lj}}{\partial x^k} + \frac{\partial g_{lk}}{\partial x^j} - \frac{\partial g_{jk}}{\partial x^l} \right).$$

Here g^{ij} is the inverse of the matrix g_{ij}. On TM we may define a second-order vector field called the *geodesic spray*, which we denote by Z_g. This vector field is characterized by the fact that the projection to M of the integral curves of Z_g by the tangent bundle projection are *geodesics*. In coordinates we have

$$Z_g = v^i \frac{\partial}{\partial x^i} - \Gamma^i_{jk} v^j v^k \frac{\partial}{\partial v^i}.$$

Here we are denoting by $(x^1, \ldots, x^n, v^1, \ldots, v^n)$ the natural coordinates for TM corresponding to coordinates $(x^1, \ldots, x^n)$ for M.

We shall need the concept of a "symmetric subalgebra" of $\mathfrak{X}(M)$, which is generated by a family of vector fields $\mathcal{V} \subset \mathfrak{X}(M)$. This construction relies on the covariant

derivative discussed above. We may make $\mathfrak{X}(M)$ into a symmetric algebra by defining the symmetric product

$$\langle X : Y \rangle = \nabla_X Y + \nabla_Y X.$$

We remark that this product first appeared in the work of Crouch [4] on gradient dynamical systems. Let $\mathcal{V}$ be a family of vector fields on M and X be a set which is bijective to $\mathcal{V}$ with bijection ψ: $X \to \mathcal{V}$. As in section 3.2, let $S(X)$ be the free symmetric algebra on X and $\mathrm{Pr}(X)$ be the symmetric products with elements in X. We may define a symmetric algebra homomorphism from $S(X)$ to $\mathfrak{X}(M)$ by extending ψ in the natural way much as we did for Lie brackets in section 3.1. We denote the resulting map from $S(X)$ to $\mathfrak{X}(M)$ by $\mathrm{Ev}(\psi)$. We also define $\mathrm{Ev}_x(\psi)(P) = (\mathrm{Ev}(\psi)(P))(x)$ for $x \in M$. We denote by $\overline{\mathrm{Sym}}(\mathcal{V})$ the image of $S(X)$ under this homomorphism and call this the *symmetric closure* of $\mathcal{V}$.

4. Sufficient conditions for small-time local controllability. Sussmann [13] gives a general result concerning so-called small-time local controllability. We are interested in a version of Sussmann's result and so will present only as much background as is necessary to state this result. We consider control systems of the form

$$\dot{x} = X(x) + u^a Y_a(x) \tag{2}$$

on a manifold M, where $X, Y_1, \ldots, Y_m$ are analytic. (Here and in what follows, when we write $u^a Y_a$, there will be an implied sum over a from 1 to m.) We shall consider inputs from the set $\mathcal{U}$ of piecewise constant inputs. Let $x_0 \in M$, let V be a neighborhood of x_0, and let $T > 0$. We denote by $\mathcal{R}^V(x_0, T)$ the set of points which can be reached from x_0 in time T while remaining in V using inputs from $\mathcal{U}$. We also denote $\mathcal{R}^V(x_0, \leq T) = \cup_{t=0}^{T} \mathcal{R}^V(x_0, t)$. We say that the system (2) is *locally accessible* at x_0 if $\mathcal{R}^V(x_0, \leq T)$ contains an open subset of M for each V and for each T sufficiently small. Furthermore, we say that (2) is *small-time locally controllable* (STLC) if it is locally accessible and if x_0 is in the interior of $\mathcal{R}^V(x_0, \leq T)$ for each V and for each T sufficiently small.

Let $\boldsymbol{X} = \{X_0, \ldots, X_m\}$. An element $B \in \mathrm{Br}(\boldsymbol{X})$ is said to be *bad* if $\delta_0(B)$ is odd and $\delta_a(B)$ is even for each $a = 1, \ldots, m$. A bracket is *good* if it is not bad. Let S_m denote the permutation group on m symbols. For $\pi \in S_m$ and $B \in \mathrm{Br}(\boldsymbol{X})$, define $\bar{\pi}(B)$ to be the bracket obtained by fixing X_0 and sending X_a to $X_{\pi(a)}$ for $a = 1, \ldots, m$. Now define

$$\beta(B) = \sum_{\pi \in S_m} \bar{\pi}(B).$$

We may state sufficient conditions for STLC.

THEOREM 4.1 (see Sussmann [13]). *Consider the bijection ϕ: $\boldsymbol{X} \to \{X, Y_1, \ldots, Y_m\}$ which sends X_0 to X and X_a to Y_a for $a = 1, \ldots, m$. Suppose that* (2) *is such that every bad bracket $B \in \mathrm{Br}(\boldsymbol{X})$ has the property that*

$$\mathrm{Ev}_x(\phi)(\beta(B)) = \sum_{a=1}^{m} \xi^a \mathrm{Ev}_x(\phi)(C_a),$$

where C_a are good brackets in $\mathrm{Br}(\boldsymbol{X})$ of lower degree than B and $\xi^a \in \mathbb{R}$ for $a = 1, \ldots, m$. Also suppose that (2) *satisfies the LARC at x. Then* (2) *is STLC at x.*

Sussmann [13] gives this result as a corollary of a special case originally conjectured by Hermes [5] and proven by Sussmann [12].

5. Controllability of simple mechanical control systems. In this section we present the main results of the paper. First we make explicit the class of control systems that we are considering. All problem data will be assumed to be analytic so that we may use piecewise constant inputs. The data for the systems that we consider are an n-dimensional configuration manifold Q; a Riemannian metric g on Q, which represents the kinetic energy; an $\mathbb{R}$-valued function V on Q, which represents the potential energy; and m linearly independent one-forms, $F^1, \dots, F^m$, on Q, which represent the input forces for the system. We call a system described by this data a *simple mechanical control system*. Although the one-forms $F^1, \dots, F^m$ describe the forces for the problem, it is the vector fields $Y_a = (F^a)^\sharp$, $a = 1, \dots, m$, which will appear in the computations. Nevertheless, it is the one-forms which are basic in the problem description.

Given a vector field X on Q, we define the *vertical lift* (see [1]) of X as the vector field on TQ defined by

$$X^{lift}(v) = \left.\frac{d}{dt}\right|_{t=0} v + tX(\tau_Q(v))$$

for $v \in TQ$ and where $\tau_Q\colon TQ \to Q$. If $(q^1, \dots, q^n)$ are coordinates for Q, we shall denote the corresponding natural coordinates for TQ by $(q^1, \dots, q^n, v^1, \dots, v^n)$. In coordinates we have

$$X^{lift}(v_q) = X^i(q)\frac{\partial}{\partial v^i}$$

for $v_q \in T_qQ$. We may now define the vector field X_L on TQ by $X_L = Z_g - \mathrm{grad}V^{lift}$, where we recall that Z_g is the geodesic spray introduced in section 3.3. With this notation, the Euler–Lagrange equations for the forced system may be shown to be equivalent to the first-order system

$$\dot{v} = X_L + u^a Y_a^{lift} \tag{3}$$

on TQ. Thus the *drift vector field* for the system is X_L, and the control vector fields are $Y_1^{lift}, \dots, Y_m^{lift}$. It is this first-order affine control system which we study in this section. We are particularly interested in the following problem.

Problem Statement. *Describe the set of configurations reachable from a given configuration when starting at rest.*

Observe that we place no restriction on the final velocities of the system. The reader will further observe that this problem statement involves only *configurations* and not velocities. It would be desirable, therefore, to derive an answer to this problem in terms of quantities on the *configuration space*. As we shall see, this can in fact be done and is one of the more compelling aspects of this approach.

Since the computations in this section are quite involved, let us outline them here before we begin. The main goal of the computations is to describe the accessibility distribution for (3) at points of zero velocity in TQ. Thus we need to compute the involutive closure of the family of vector fields $\mathcal{V}' = \{X_L, Y_1^{lift}, \dots, Y_m^{lift}\}$. Observe that since $X_L = Z_g - \mathrm{grad}V^{lift}$, we may write vector fields in $\overline{\mathrm{Lie}}(\mathcal{V}')$ as $\mathbb{R}$-linear combinations of vector fields in $\mathcal{V} = \{Z_g, Y_1^{lift}, \dots, Y_m^{lift}, \mathrm{grad}V^{lift}\}$. This is made precise by using free Lie algebras in section 5.1. When we evaluate the brackets which are used in the computation of the accessibility distribution at zero velocity, only a small number of them make a contribution, and the rest vanish. The brackets which

vanish do so in one of two ways. Either they are identically zero or they are polynomial in the velocity coordinates and so go to zero when the velocity goes to zero. Therefore, we have three possible classes of brackets: one class which is nonzero when the velocity is zero, one class which is identically zero, and one class which is not identically zero but is zero when the velocity is zero. In section 5.1 we categorize these three types of brackets. There we shall see that the brackets which make a contribution to the accessibility distribution at zero velocity may be written as linear combinations of special brackets which we call *primitive* brackets. The computations in section 5.1 are done at the level of free Lie algebras since this provides a rigorous way to perform the necessary computations. In section 5.2 we give expressions for primitive brackets in terms of the geometry of the problem. It is here that the symmetric product introduced in section 3.3 makes its appearance. In section 5.3 we assemble the results of sections 5.1 and 5.2 to arrive at the form of the accessibility distribution for (3) at points of zero velocity. In section 5.4 we provide a precise statement of the types of controllability we consider, and in section 5.5 we provide computable conditions for these versions of controllability.

We remark that most of the complexity of this section is a consequence of including potential energy in the formulation. In [9] the authors provide sufficient conditions for controllability when there is no potential energy function. Due to space considerations, some of the free Lie algebra proofs from section 5.1 are omitted. We refer the reader to the dissertation [8] for these proofs.

5.1. Computations with free Lie algebras. In this section we perform some calculations with a pair of free Lie algebras which are suited to our purposes. Rather than just using a generating set which is in one-to-one correspondence with the set $\mathcal{V}' = \{X_L, Y_1^{lift}, \ldots, Y_m^{lift}\}$ of control vector fields and the drift vector field, we also use a generating set which is in one-to-one correspondence with the set $\mathcal{V} = \{Z_g, Y_1^{lift}, \ldots, Y_m^{lift}, \mathrm{grad}V^{lift}\}$. The reason for this is that vector fields in $\mathcal{V}'$ are $\mathbb{R}$-linear combinations of vector fields in $\mathcal{V}$, and as we shall see in section 5.3, it is comparatively easy to describe the involutive closure of $\mathcal{V}$.

Let $\boldsymbol{X} = \{X_0, \ldots, X_{m+1}\}$, and let $L(\boldsymbol{X})$ be the free Lie algebra generated by the set $\boldsymbol{X}$. We can simplify many of our computations for the controllability analysis of (3) by making simplifications to a set of generators for $L(\boldsymbol{X})$. We first need some notation. Let

$$\mathrm{Br}^k(\boldsymbol{X}) = \{B \in \mathrm{Br}(\boldsymbol{X}) \mid \text{ the degree of } B \text{ is } k\},$$

$$\mathrm{Br}_k(\boldsymbol{X}) = \left\{ B \in \mathrm{Br}(\boldsymbol{X}) \mid \ \delta_0(B) - \sum_{a=1}^{m+1} \delta_a(B) = k \right\}.$$

We shall see in section 5.2 that, when we restrict ourselves to zero velocities, only a small subset of $\mathrm{Br}(\boldsymbol{X})$ will evaluate to something nonzero. In turn, these brackets will be seen to be linear combinations of a special class of brackets which we shall call *primitive* brackets. Recall from section 3.1 the notion of components in $L(\boldsymbol{X})$.

DEFINITION 5.1. *Let $B \in \mathrm{Br}_0(\boldsymbol{X}) \cup \mathrm{Br}_{-1}(\boldsymbol{X})$, and let $B_1, B_2, B_{11}, B_{12}, B_{21}, B_{22}, \ldots$ be the decomposition of B into its components. We shall say that B is* primitive *if each of its components is in $\mathrm{Br}_{-1}(\boldsymbol{X}) \cup \mathrm{Br}_0(\boldsymbol{X}) \cup \{X_0\}$.*

The relevant observations that need to be made regarding primitive brackets are as follows:

Prim1. If $B \in \mathrm{Br}_{-1}(\boldsymbol{X})$ is primitive, then, up to sign, we may write $B = [B_1, B_2]$ with $B_1 \in \mathrm{Br}_{-1}(\boldsymbol{X})$ and $B_2 \in \mathrm{Br}_0(\boldsymbol{X})$ both primitive.

Prim2. If $B \in \mathrm{Br}_0(\boldsymbol{X})$ is primitive, then, up to sign, B may have one of two forms. Either $B = [X_0, B_1]$ with $B_1 \in \mathrm{Br}_{-1}(\boldsymbol{X})$ primitive or $B = [B_1, B_2]$ with $B_1, B_2 \in \mathrm{Br}_0(\boldsymbol{X})$ both primitive.

Using these two rules, it is possible to construct primitive brackets of any degree. For example, the primitive brackets of degrees one through four are, up to sign,

$$\begin{aligned}
&\text{Degree 1}: \ \{X_a \mid \ a = 1, \dots, m\},\\
&\text{Degree 2}: \ \{[X_0, X_a] \mid \ a = 1, \dots, m\},\\
&\text{Degree 3}: \ \{[X_a, [X_0, X_b]] \mid \ a, b = 1, \dots, m\},\\
&\text{Degree 4}: \ \{[X_0, [X_a, [X_0, X_b]]] \mid \ a, b = 1, \dots, m\}\\
&\qquad\qquad \cup \{[[X_0, X_a], [X_0, X_b]] \mid \ a, b = 1, \dots, m\}.
\end{aligned}$$

From Proposition 3.1 we know that to generate $L(\boldsymbol{X})$ we need only look at brackets of the form

$$[X_{a_k}, [X_{a_{k-1}}, \dots, [X_{a_2}, X_{a_1}]]], \tag{4}$$

where $a_i \in \{0, \dots, m+1\}$ for $i = 1, \dots, k$. We shall see in section 5.2 that brackets from $\mathrm{Br}_j(\boldsymbol{X})$, where $j \geq 1$ or $j \leq -2$, will not be of interest to us. In particular, we shall see that when $j \leq -2$ these brackets evaluate identically to zero. Therefore, in this section we concentrate our attention on brackets in $\mathrm{Br}_0(\boldsymbol{X}) \cup \mathrm{Br}_{-1}(\boldsymbol{X})$ which satisfy certain requirements. We state the form of these brackets in the following lemma.

LEMMA 5.2. *Let us impose the condition on elements of* $\mathrm{Br}(\boldsymbol{X})$ *that we shall consider a bracket to be zero if any of its components is in* $\mathrm{Br}_{-j}(\boldsymbol{X})$ *for* $j \geq 2$. *Let* $B \in \mathrm{Br}_0(\boldsymbol{X}) \cup \mathrm{Br}_{-1}(\boldsymbol{X})$. *Then we may write* B *as a finite sum of primitive brackets.*

The inductive proof is straightforward, and we refer the interested reader to [8]. However, in lieu of a proof an example is illustrative.

Example 5.3. Consider the bracket $B = [X_{m+1}, [X_0, [X_0, X_a]]] \in \mathrm{Br}_0(\boldsymbol{X})$. This bracket is in $\mathrm{Br}_0(\boldsymbol{X})$ but is not primitive. However, by Lemma 5.2, we may write B as a finite sum of primitive brackets. Indeed, by Jacobi's identity we have

$$\begin{aligned}
B &= [X_{m+1}, [X_0, [X_0, X_a]]] = -[[X_0, X_a], [X_{m+1}, X_0]] - [X_0, [[X_0, X_a], X_{m+1}]]\\
&= [[X_0, X_a], [X_0, X_{m+1}]] + [X_0, [X_{m+1}, [X_0, X_a]]].
\end{aligned}$$

The proof of Lemma 5.2 is essentially a generalization of this example.

Now we relate the free Lie algebra $L(\boldsymbol{X})$ with a free Lie algebra which corresponds to the set $\mathcal{V}' = \{X_L, Y_1^{lift}, \dots, Y_m^{lift}\}$. As we mentioned above, the reason why we wish to do this is that the vector fields in $\mathcal{V}'$ are $\mathbb{R}$-linear combinations of vector fields in $\mathcal{V} = \{Z_g, Y_1^{lift}, \dots, Y_m^{lift}, \mathrm{grad} V^{lift}\}$, the latter family of vector fields being bijective with the set $\boldsymbol{X}$. Let $\boldsymbol{X}' = \{X_0', \dots, X_m'\}$. We formally set $X_0' = X_0 - X_{m+1}$ and $X_a' = X_a$ for $a = 1, \dots, m$. We may now write brackets in $\mathrm{Br}(\boldsymbol{X}')$ as linear combinations of brackets in $\mathrm{Br}(\boldsymbol{X})$ by $\mathbb{R}$-linearity of the bracket. We may, in fact, be even more precise about this.

Let $B' \in \mathrm{Br}(\boldsymbol{X}')$. We define a subset $\boldsymbol{S}(B')$ of $\mathrm{Br}(\boldsymbol{X})$ by saying that $B \in \boldsymbol{S}(B')$ if each occurrence of X_a' in B' is replaced with X_a for $a = 1, \dots, m$ and if each occurrence of X_0' in B' is replaced with *either* X_0 *or* X_{m+1}. An example is illustrative. Suppose that

$$B' = [[X_0', X_1'], [X_2', [X_0', X_3']]].$$

Then

$$\boldsymbol{S}(B') = \{[[X_0, X_1], [X_2, [X_0, X_3]]], [[X_0, X_1], [X_2, [X_{m+1}, X_3]]], \\ [[X_{m+1}, X_1], [X_2, [X_0, X_3]]], [[X_{m+1}, X_1], [X_2, [X_{m+1}, X_3]]]\}.$$

Now we may precisely state how we write brackets in $\mathrm{Br}(\boldsymbol{X}')$.

LEMMA 5.4. *Let $B' \in \mathrm{Br}(\boldsymbol{X}')$. Then*

$$B' = \sum_{B \in \boldsymbol{S}(B')} (-1)^{\delta_{m+1}(B)} B.$$

The proof is by induction and may be found in [8].

We shall be interested only in terms in the above decomposition of B' which are in $\mathrm{Br}_0(\boldsymbol{X}) \cup \mathrm{Br}_{-1}(\boldsymbol{X})$ since, as we shall see in section 5.2, these are the only ones which will contribute to $\mathrm{Ev}_{0_q}(\phi')(B')$. Here 0_q is the zero vector in T_qQ.

5.2. Distribution computations for simple mechanical control systems. In this section we use the simplifications of section 5.1 to get a complete description of the brackets which contribute to the accessibility distribution for (3) restricted to $Z(TQ)$, the zero section of TQ. Note that we restrict ourselves to $Z(TQ)$ because we are interested in determining the reachable points starting with zero initial velocity. To make the correspondence between the free Lie algebra $L(\boldsymbol{X})$ used in section 5.1 and the accessibility algebra for (3), we use the family of vector fields $\mathcal{V} = \{Z_g, Y_1^{lift}, \ldots, Y_m^{lift}, \mathrm{grad} V^{lift}\}$ and establish a bijection ϕ from $\boldsymbol{X}$ to $\mathcal{V}$ by mapping X_0 to X_L, X_a to Y_a^{lift} for $a = 1, \ldots, m$, and X_{m+1} to $\mathrm{grad}\, V^{lift}$. Please note that $\mathcal{V}$ is *not* the family of vector fields which generates the accessibility algebra. The accessibility algebra is generated by the family $\mathcal{V}' = \{X_L, Y_1^{lift}, \ldots, Y_m^{lift}\}$. We establish a bijection ϕ' from $\boldsymbol{X}'$ to $\mathcal{V}'$ by mapping X_0' to X_L and X_a' to Y_a^{lift} for $a = 1, \ldots, m$. By Lemma 5.4, each vector field in $\overline{\mathrm{Lie}}(\mathcal{V}')$ is a $\mathbb{R}$-linear sum of vector fields in $\overline{\mathrm{Lie}}(\mathcal{V})$. That lemma also completely describes the sum.

Now we shall show that it is possible to compute the brackets from $\mathrm{Br}(\boldsymbol{X})$ in terms of the problem data. We first present a lemma which gives the basic structure of primitive brackets. In this lemma we see that a large number of brackets are computable in terms of quantities defined on Q. This is worth noting since the vector fields themselves are defined on TQ. Of particular interest in the lemma is the appearance of the symmetric product which was introduced in section 3.3.

We need to say a few words about the structure of TQ. We denote by $Z(TQ)$ the zero section of TQ. Since Q is naturally diffeomorphic to $Z(TQ)$, there is a natural inclusion of T_qQ into $T_{0_q}TQ$ for each $q \in Q$. We shall call the image of this inclusion in $T_{0_q}TQ$ the *horizontal subspace*. We shall call the subspace of $T_{0_q}TQ$ which is tangent to the fiber of TQ at q the *vertical subspace* and denote it by $V_{0_q}TQ$. We have $T_{0_q}TQ = T_qQ \oplus V_{0_q}TQ$ for each $q \in Q$. We mention that this notion of vertical is valid at any point in TQ. However, the definition of horizontal is valid only on $Z(TQ)$.

LEMMA 5.5. *Suppose that $B \in \mathrm{Br}^k(\boldsymbol{X})$ is primitive.*

(i) *If $B \in \mathrm{Br}_{-1}(\boldsymbol{X})$, then $\mathrm{Ev}(\phi)(B)$ is the vertical lift of a vector field on Q.*

(ii) *If $B \in \mathrm{Br}_0(\boldsymbol{X})$, then $U = \mathrm{Ev}(\phi)(B)$ has the property that, when expressed in a local chart, the vertical components of U are linear in the fiber coordinates v and the horizontal components are independent of v. In particular, we may define a vector field on Q by $U_Q \colon q \mapsto U(0_q) \in T_qQ \subset T_{0_q}TQ$. There are two cases to consider.*

(a) *$B = [X_0, B_1]$ with $B_1 \in \mathrm{Br}_{-1}(\boldsymbol{X})$: Define U_1 to be the vector field on Q such that $\mathrm{Ev}(\phi)(B_1) = U_1^{lift}$. Then $U(0_q) = \mathrm{Ev}(\phi)(B)(0_q) = -U_1(q)$. Let $U_2 \in \mathfrak{X}(Q)$. Then $[U_2^{lift}, U] = (\nabla_{U_1} U_2 + \nabla_{U_2} U_1)^{lift}$.*

(b) *$B = [B_1, B_2]$ with $B_1, B_2 \in \mathrm{Br}_0(\boldsymbol{X})$: Define $U_{1,Q}, U_{2,Q}$ to be the vector fields on Q corresponding to $\mathrm{Ev}(\phi)(B_1), \mathrm{Ev}(\phi)(B_2)$, respectively. Then $\mathrm{Ev}(\phi)(B)(0_q) = [U_{1,Q}, U_{2,Q}](q)$.*

Proof. The proof is by induction on k. The result is true for $k = 1$ trivially. If X and Y are vector fields on Q, it is a straightforward coordinate computation to show that

$$[X^{lift}, Y^{lift}] = 0.$$

If X is a vector field on Q, we compute

$$[Z_g, X^{lift}] = -Y^i \frac{\partial}{\partial q^i} + \left(\frac{\partial Y^i}{\partial q^j} v^j + \Gamma^i_{jk} Y^j v^k + \Gamma^i_{kj} v^k Y^j \right) \frac{\partial}{\partial v^i}. \tag{5}$$

Inspecting (5) shows that $[Z_g, X^{lift}](0_q) = -X(q)$. Now let $Y \in \mathfrak{X}(Q)$. We compute

$$[Y^{lift}, [Z_g, X^{lift}]] = \left(\frac{\partial Y^i}{\partial q^j} X^j + \frac{\partial X^i}{\partial q^j} Y^j + 2\Gamma^i_{jk} X^j Y^k \right) \frac{\partial}{\partial v^i}, \tag{6}$$

which is the coordinate representation of $(\nabla_X Y + \nabla_Y X)^{lift}$. This shows that the lemma is true for $k = 2$.

Now suppose that the lemma is true for $k = 1, \ldots, l$ for $l \geq 2$, and let $B \in \mathrm{Br}^{l+1}(\boldsymbol{X})$ be primitive.

(i) Suppose that $B \in \mathrm{Br}_{-1}(\boldsymbol{X})$. Without loss of generality (by Prim1) we may suppose that $B = [B_1, B_2]$ with $B_1 \in \mathrm{Br}_{-1}(\boldsymbol{X})$ and $B_2 \in \mathrm{Br}_0(\boldsymbol{X})$. Then, by the induction hypotheses, we have

$$\mathrm{Ev}(\phi)(B_1) = \alpha^i(q) \frac{\partial}{\partial v^i}, \qquad \mathrm{Ev}(\phi)(B_2) = \lambda^i(q) \frac{\partial}{\partial q^i} + \mu^i_j(q) v^j \frac{\partial}{\partial v^i}.$$

Now we compute

$$\mathrm{Ev}(\phi)([B_1, B_2]) = \left(\mu^i_j \alpha^j - \frac{\partial \alpha^i}{\partial q^j} \lambda^j \right) \frac{\partial}{\partial v^i}.$$

Note that the components in the q-direction are zero and the components in the v-direction are only functions of q. This means that this vector field is the vertical lift of a vector field on Q. This proves (i).

(ii) Suppose that $B \in \mathrm{Br}_0(\boldsymbol{X})$. Without loss of generality (by Prim2) we may suppose that either (a) $B = [X_0, B_1]$ with $B_1 \in \mathrm{Br}_{-1}(\boldsymbol{X})$ or (b) $B = [B_1, B_2]$ with $B_1, B_2 \in \mathrm{Br}_0(\boldsymbol{X})$. Let us deal with the first case. Equation (5) gives $\mathrm{Ev}(B)(\phi)(0_q) = -U_1(q)$, where U_1 is the vector field on Q so that $\mathrm{Ev}(\phi)(B_1) = U_1^{lift}$. (Such a vector field exists by (i).) For every vector field U_2 on Q we have $[U_2^{lift}, [Z_g, U_1^{lift}]] = (\nabla_{U_1} U_2 + \nabla_{U_2} U_1)^{lift}$ by (6). This proves (ii(a)).

Now suppose that we have $B_1, B_2 \in \mathrm{Br}_0(\boldsymbol{X})$. Then, by the induction hypotheses, we have

$$\mathrm{Ev}(\phi)(B_1) = \alpha^i(q) \frac{\partial}{\partial q^i} + \beta^i_j(q) v^j \frac{\partial}{\partial v^i}, \qquad \mathrm{Ev}(\phi)(B_2) = \lambda^i(q) \frac{\partial}{\partial q^i} + \mu^i_j(q) v^j \frac{\partial}{\partial v^i}.$$

We compute

$$\begin{aligned}\mathrm{Ev}(\phi)([B_1,B_2]) = &\left(\frac{\partial\lambda_i}{\partial q^j}\alpha^j - \frac{\partial\alpha^i}{\partial q^j}\lambda^j\right)\frac{\partial}{\partial q^i}\\ &+\left(\frac{\partial\mu^i_k}{\partial q^j}\alpha^j v^k + \mu^i_j\beta^j_k v^k - \frac{\partial\beta^i_k}{\partial q^j}\lambda^j v^k - \beta^i_j\mu^j_k v^k\right)\frac{\partial}{\partial v^i}.\end{aligned}$$

The components have the order in v specified by the lemma. Also, it is clear that the vector fields on Q defined by B_1 and B_2 are

$$U_{1,Q} = \alpha^i(q)\frac{\partial}{\partial q^i} \qquad \text{and} \qquad U_{2,Q} = \lambda^i(q)\frac{\partial}{\partial q^i},$$

respectively. It is easy to see that $\mathrm{Ev}(\phi)(B)(0_q) = [U_{1,Q}, U_{2,Q}](q)$. This completes the proof of the lemma. □

This lemma provides us with a positive step toward computing the value of all primitive brackets when evaluated using $\mathrm{Ev}(\phi)$. The following lemma shows that these are the *only* brackets that we need to consider.

LEMMA 5.6. (i) *Let $l \geq 1$ be an integer, and let $B \in \mathrm{Br}_l(\boldsymbol{X})$. Then $\mathrm{Ev}(\phi)(B)(0_q) = 0$ for each $q \in Q$.*

(ii) *Let $l \geq 2$ be an integer, and let $B \in \mathrm{Br}^k(\boldsymbol{X}) \cap \mathrm{Br}_{-l}(\boldsymbol{X})$ for $k \geq 2$. Then $\mathrm{Ev}(\phi)(B) = 0$.*

The proof of this lemma may be found in [8]. It goes very much like the proof of Lemma 5.5.

Let us summarize what we have done in this section. First we obtained a characterization of primitive brackets in $\boldsymbol{X}$ when we evaluate them in $\mathcal{V}$ via $\mathrm{Ev}(\phi)$. This characterization involved Lie brackets and covariant derivatives of the vector fields $Y_1, \ldots, Y_m, \mathrm{grad}V$. Then we showed in Lemma 5.6 that primitive brackets are the only ones that we need be concerned with if we are evaluating the vector fields on the zero section of TQ.

5.3. The form of the accessibility distribution restricted to $Z(TQ)$ for simple mechanical control systems. In this section we compute the accessibility distribution for (3) restricted to the zero section of TQ. By Lemma 5.4 we know that we may write the vector fields in the accessibility distribution in terms of vector fields in $\overline{\mathrm{Lie}}(\mathcal{V})$. In section 5.2 we saw some hints that we might be able to write vector fields in $\overline{\mathrm{Lie}}(\mathcal{V})$ in terms of covariant derivatives and Lie brackets of the input vector fields and grad$V$. First we resolve this issue by saying exactly what the vector fields in $\overline{\mathrm{Lie}}(\mathcal{V})$ look like when we restrict them to $Z(TQ)$. We denote by $D_{\overline{\mathrm{Lie}}(\mathcal{V})}$ the distribution defined by

$$D_{\overline{\mathrm{Lie}}(\mathcal{V})}(v) = \mathrm{span}_{\mathbb{R}}\{U(v) \mid \ U \in \overline{\mathrm{Lie}}(\mathcal{V})\}.$$

The reader will also wish to recall the ideas from symmetric algebras presented in section 3.3. We denote $\mathcal{Y} = \{Y_1, \ldots, Y_m\}$.

The following lemma describes the horizontal and vertical parts of the involutive closure of $\mathcal{V}$ restricted to $Z(TQ)$. The reader may wish to recall our remarks about the structure of the tangent bundle preceding Lemma 5.5.

LEMMA 5.7. *Let $q \in Q$. Then*

$$D_{\overline{\mathrm{Lie}}(\mathcal{V})}(0_q) \cap V_{0_q}TQ = (D_{\overline{\mathrm{Sym}}(\mathcal{Y}\cup\{\mathrm{grad}\,V\})}(q))^{lift}$$

and

$$D_{\overline{\mathrm{Lie}}(\mathcal{V})}(0_q) \cap T_qQ = D_{\overline{\mathrm{Lie}}(\overline{\mathrm{Sym}}(\mathcal{Y} \cup \{\operatorname{grad} V\}))}(q).$$

Proof. From Lemma 5.6 we know that the only brackets from $\mathrm{Br}(\boldsymbol{X})$ which we need to consider are the primitive brackets. From Lemma 5.5 we know that the brackets which are in $\mathrm{Br}_{-1}(\boldsymbol{X})$ will generate the vertical directions, and the brackets which are in $\mathrm{Br}_0(\boldsymbol{X})$ will generate the horizontal directions.

First we show that $(D_{\overline{\mathrm{Sym}}(\mathcal{Y} \cup \{\operatorname{grad} V\})}(q))^{lift} \subset D_{\overline{\mathrm{Lie}}(\mathcal{V})}(0_q)$. This may be done inductively. Define $\mathrm{Sym}^{(1)}(\mathcal{Y} \cup \{\operatorname{grad} V\}) = \mathcal{Y} \cup \{\operatorname{grad} V\}$, and inductively define

$$\mathrm{Sym}^{(k)}(\mathcal{Y} \cup \{\operatorname{grad} V\} = \{\langle U_1 : U_2 \rangle \mid U_i \in \mathrm{Sym}^{(k_i)}(\mathcal{Y} \cup \{\operatorname{grad} V\}), k_1 + k_2 = k\}.$$

Clearly

$$\overline{\mathrm{Sym}}(\mathcal{Y} \cup \{\operatorname{grad} V\}) = \bigcup_{k \in \mathbb{Z}^+} \mathrm{Sym}^{(k)}(\mathcal{Y} \cup \{\operatorname{grad} V\}).$$

It is trivially true that $(\mathrm{Sym}^{(1)}(\mathcal{Y} \cup \{\operatorname{grad} V\}))^{lift} \subset \overline{\mathrm{Lie}}(\mathcal{V})$. Now suppose that $(\mathrm{Sym}^{(k)}(\mathcal{Y} \cup \{\operatorname{grad} V\}))^{lift} \subset \overline{\mathrm{Lie}}(\mathcal{V})$ for $k = 1, \ldots, l$ for $l \geq 1$. We see that $(\mathrm{Sym}^{(l+1)}(\mathcal{Y} \cup \{\operatorname{grad} V\}))^{lift} \subset \overline{\mathrm{Lie}}(\mathcal{V})$ since we may generate all elements of $(\mathrm{Sym}^{(l+1)}(\mathcal{Y} \cup \{\operatorname{grad} V\}))^{lift}$ by considering brackets of the form $[U_1^{lift}, [Z_g, U_2^{lift}]]$, where $U_i \in \mathrm{Sym}^{(l_i)}(\mathcal{Y}, V)$ and $l_1 + l_2 = l + 1$. This follows from (6). This shows that $(D_{\overline{\mathrm{Sym}}(\mathcal{Y} \cup \{\operatorname{grad} V\})}(q))^{lift} \subset D_{\overline{\mathrm{Lie}}(\mathcal{V})}(0_q)$.

Now we show that $D_{\overline{\mathrm{Lie}}(\mathcal{V})}(0_q) \subset (D_{\overline{\mathrm{Sym}}(\mathcal{Y} \cup \{\operatorname{grad} V\})}(q))^{lift}$. To do this we must show that the image under $\mathrm{Ev}(\phi)$ of all primitive brackets in $\mathrm{Br}_{-1}(\boldsymbol{X})$ may be written as a linear combination of vector fields in $\overline{\mathrm{Sym}}(\mathcal{Y} \cup \{\operatorname{grad} V\})$. A primitive bracket in $\mathrm{Br}_{-1}(\boldsymbol{X})$ may be written as $B = [B_1, B_2]$ with $B_1 \in \mathrm{Br}_{-1}(\boldsymbol{X})$ and $B_2 \in \mathrm{Br}_0(\boldsymbol{X})$ both being primitive. Therefore, either $B_2 = [X_0, B_2']$ with B_2' primitive and in $\mathrm{Br}_{-1}(\boldsymbol{X})$ or $B_2 = [B_2', B_2'']$ with $B_2', B_2'' \in \mathrm{Br}_0(\boldsymbol{X})$ both primitive. In the first case $\mathrm{Ev}(\phi)(B) \in \mathrm{Sym}^{(k)}\mathcal{Y} \cup \{\operatorname{grad} V\})$ for some k by (6). In the second case we may use Jacobi's identity to obtain

$$B = -[B_2'', [B_1, B_2']] + [B_2', [B_1, B_2'']].$$

We may apply the above argument to the terms $[B_1, B_2']$ and $[B_1, B_2'']$, repeatedly using (6) until they are expressed in terms of covariant derivatives. When this is done, $\mathrm{Ev}(\phi)(B)$ will then be a $\mathbb{R}$-linear combination of elements in $\overline{\mathrm{Sym}}(\mathcal{Y} \cup \{\operatorname{grad} V\})$. This shows that $D_{\overline{\mathrm{Lie}}(\mathcal{V})}(0_q) \subset (D_{\overline{\mathrm{Sym}}(\mathcal{Y} \cup \{\operatorname{grad} V\})}(q))^{lift}$.

To demonstrate the proposed form of $D_{\overline{\mathrm{Lie}}(\mathcal{V})} \cap T_qQ$, by Lemma 5.5 (ii(b)) we need only show that $\overline{\mathrm{Sym}}(\mathcal{Y} \cup \{\operatorname{grad} V\})(q) \subset D_{\overline{\mathrm{Lie}}(\mathcal{V})}(0_q)$. But this is clear from Lemma 5.5 (ii(a)). This completes the proof of the lemma. ∎

Remark 5.8. Note that the constructions in the above lemma depend only upon $\{Y_1, \ldots, Y_m, \operatorname{grad} V\}$. The effects of the geodesic spray do not appear explicitly. However, its contribution is obviously important in the computations performed in section 5.2.

From Lemmas 5.4 and 5.7 we know that the vector fields which contribute to $\overline{\mathrm{Lie}}(\mathcal{V}')$ when we evaluate on $Z(TQ)$ will be $\mathbb{R}$-linear combinations of vector fields from $\overline{\mathrm{Lie}}(\overline{\mathrm{Sym}}(\mathcal{Y} \cup \{\operatorname{grad} V\}))$. Thus, to compute these vector fields, we need to figure out which vector fields need to be "removed" from $\overline{\mathrm{Lie}}(\overline{\mathrm{Sym}}(\mathcal{Y} \cup \{\operatorname{grad} V\}))$. We present an

ALGORITHM 7.1.

For $i \in \mathbb{Z}^+$ do
  For $B \in \mathrm{Br}^{(i)}(\boldsymbol{X})$ primitive do
    If $\delta_{m+1}(B) = 0$ then
      If $B \in \mathrm{Br}_{-1}(\boldsymbol{X})$ then
        $U \in \mathcal{C}_{ver}^{\frac{1}{2}(i+1)}(\mathcal{Y}, V)$ where $\mathrm{Ev}(\phi)(B) = U^{lift}$
      else
        $U \in \mathcal{C}_{hor}^{(i/2)}(\mathcal{Y}, V)$ where $U(q) = \mathrm{Ev}_{0_q}(\phi)(B)$
      end
    else
      If B has no components of the form $[X_0, X_{m+1}]$ then
        Compute $B' \in \mathrm{Br}(\boldsymbol{X})$ by replacing every occurrence of X_0 *and* X_{m+1} in B with X'_0 and by replacing every occurrence of X_a in B with X'_a for $a = 1, \ldots, m$.
        Let $B'' = 0$.
        For $\tilde{B} \in \boldsymbol{S}(B') \cap (\mathrm{Br}_{-1}(\boldsymbol{X}) \cup \mathrm{Br}_0(\boldsymbol{X}))$ do
          Write $\tilde{B}$ as a finite sum of primitive brackets in $\mathrm{Br}(\boldsymbol{X})$ by Lemma 5.2.
          $B'' = B'' + (-1)^{\delta_{m+1}(\tilde{B})}\tilde{B}$
        end
        If $B \in \mathrm{Br}_{-1}(\boldsymbol{X})$ then
          $U \in \mathcal{C}_{ver}^{\frac{1}{2}(i+1)}(\mathcal{Y}, V)$ where $\mathrm{Ev}(\phi)(B'') = U^{lift}$
        else
          $U \in \mathcal{C}_{hor}^{(i/2)}(\mathcal{Y}, V)$ where $U(q) = \mathrm{Ev}_{0_q}(\phi)(B'')$
        end
      end
    end
  end
END

FIG. 2. *Algorithm for computing* $\overline{\mathrm{Lie}}(\mathcal{V}') \mid Z(TQ)$.

algorithm which we shall prove determines exactly which $\mathbb{R}$-linear combinations from $\overline{\mathrm{Lie}}(\overline{\mathrm{Sym}}(\mathcal{Y} \cup \{\mathrm{grad}\, V\}))$ we need to compute. We define *two* sequences of families of vector fields on Q, which we shall denote by $\mathcal{C}_{ver}^{(k)}(\mathcal{Y}, V)$ and $\mathcal{C}_{hor}^{(k)}(\mathcal{Y}, V)$ where $k \in \mathbb{Z}^+$. In Figure 2 the algorithm is presented for computing these families. When we have computed these sequences we define

$$\mathcal{C}_{ver}(\mathcal{Y}, V) = \bigcup_{k \in \mathbb{Z}^+} \mathcal{C}_{ver}^{(k)}(\mathcal{Y}, V), \qquad \mathcal{C}_{hor}(\mathcal{Y}, V) = \bigcup_{k \in \mathbb{Z}^+} \mathcal{C}_{hor}^{(k)}(\mathcal{Y}, V).$$

The distributions defined by these families of vector fields shall be denoted $C_{ver}(\mathcal{Y}, V)$ and $C_{hor}(\mathcal{Y}, V)$, respectively.

We may now state the form of the accessibility distribution $\overline{\mathrm{Lie}}(\mathcal{V}')$ for (3) when restricted to the zero section of TQ.

PROPOSITION 5.9. *Let $q \in Q$. Then*

$$D_{\overline{\mathrm{Lie}}(\mathcal{V}')}(0_q) \cap V_{0_q}TQ = (C_{ver}(\mathcal{Y}, V)(q))^{lift}$$

and

$$D_{\overline{\mathrm{Lie}}(\mathcal{V}')}(0_q) \cap T_qQ = C_{hor}(\mathcal{Y}, V)(q).$$

Proof. Studying the algorithm that we have used to compute $\mathcal{C}_{ver}(\mathcal{Y}, V)$ and $\mathcal{C}_{hor}(\mathcal{Y}, V)$, the reader will note that we have exactly taken each primitive bracket $B \in \mathrm{Br}(\boldsymbol{X})$ and computed which $\mathbb{R}$-linear combinations from $\mathrm{Br}(\boldsymbol{X})$ appear along with B in the decomposition of some $B' \in \mathrm{Br}(\boldsymbol{X}')$ given by Lemma 5.4. Since it is only these primitive brackets which appear in $\overline{\mathrm{Lie}}(\mathcal{V}') \mid Z(TQ)$, this will, by construction, generate $D_{\overline{\mathrm{Lie}}(\mathcal{V}')} \mid Z(TQ)$.

We need to prove that, as stated in the first step of the algorithm, if $\delta_{m+1}(B) = 0$, then $\mathrm{Ev}_{0_q}(\phi)(B) \in D_{\overline{\mathrm{Lie}}(\mathcal{V}')}(0_q)$. To show that this is in fact the case, let $B' \in \mathrm{Br}(\boldsymbol{X}')$ be the bracket obtained by replacing X_a with X'_a for $a = 0, \ldots, m$. We claim that the only bracket in $\boldsymbol{S}(B')$ which contributes to $\mathrm{Ev}(\phi')(B')$ is B. This is true since any other brackets in $\boldsymbol{S}(B')$ are obtained by replacing X_0 in B with X_{m+1}. Such a replacement will result in a bracket which has at least one component which is in $\mathrm{Br}_{-l}(\boldsymbol{X})$ for $l \geq 2$. These brackets evaluate to zero by Lemma 5.6(ii).

We also need to show that if B has components of the form $[X_0, X_{m+1}]$, then it will not contribute to $\overline{\mathrm{Lie}}(\mathcal{V}') \mid Z(TQ)$. This is clear since, when constructing B' in the algorithm, the component $[X_0, X_{m+1}]$ will become $[X'_0, X'_0]$, which means that B' will be identically zero. □

It is perhaps useful to construct a few of the families $\mathcal{C}^{(k)}_{ver}(\mathcal{Y}, V)$ and $\mathcal{C}^{(k)}_{hor}(\mathcal{Y}, V)$ to show how the algorithm works. We shall do this for $k = 1, 2$. Our notation in these calculations follows that in the algorithm.

Let $i = 1$. The only primitive brackets in $\mathrm{Br}^{(1)}(\boldsymbol{X})$ are $X_1, \ldots, X_{m+1}$. For the brackets $B = X_a$, $a = 1, \ldots, m$, $\delta_{m+1}(B) = 0$. Note that $\mathrm{Ev}(\phi)(B) = Y_a^{lift}$ so $Y_a \in \mathcal{C}^{(1)}_{ver}(\mathcal{Y}, V)$ for $a = 1, \ldots, m$. The bracket X_{m+1} has no components of the form $[X_0, X_{m+1}]$, so it is a candidate for providing an element of $\mathcal{C}^{(1)}_{ver}(\mathcal{Y}, V)$. If $B = X_{m+1}$, we compute $B' = X'_0$. Therefore, $\boldsymbol{S}(B') = \{X_0, X_{m+1}\}$. The only element in $\boldsymbol{S}(B')$ which is in $\mathrm{Br}_{-1}(\boldsymbol{X}) \cup \mathrm{Br}_0(\boldsymbol{X})$ is X_{m+1}. Therefore, $B'' = -X_{m+1}$. We then see that $\mathrm{Ev}(\phi)(B'') = -\mathrm{grad}\, V^{lift}$, from which we conclude that $\mathrm{grad}\, V \in \mathcal{C}^{(1)}_{ver}(\mathcal{Y}, V)$. In summary,

$$\mathcal{C}^{(1)}_{ver}(\mathcal{Y}, V) = \{Y_1, \ldots, Y_m, \mathrm{grad}\, V\}.$$

Now we look at the case when $i = 2$. The primitive brackets in $\mathrm{Br}^{(2)}(\boldsymbol{X})$ are $\{[X_0, X_1], \ldots, [X_0, X_{m+1}]\}$. The brackets $B = [X_0, X_a]$, $a = 1, \ldots, m$, have the property that $\delta_{m+1}(B) = 0$. We compute $\mathrm{Ev}_{0_q}(\phi)(B) = -Y_a(q)$ and so conclude that $Y_a \in \mathcal{C}^{(1)}_{hor}(\mathcal{Y}, V)$. The bracket $[X_0, X_{m+1}]$ is not a candidate for providing an element of $\mathcal{C}^{(1)}_{hor}(\mathcal{Y}, V)$, so we have

$$\mathcal{C}^{(1)}_{hor}(\mathcal{Y}, V) = \{Y_1, \ldots, Y_m\}.$$

In a similar manner we may compute

$$\mathcal{C}^{(2)}_{ver}(\mathcal{Y}, V) = \{\langle Y_a : Y_b\rangle \mid \ a, b = 1, \ldots, m\} \cup \{\langle Y_a : \mathrm{grad}\, V\rangle \mid \ a = 1, \ldots, m\}$$

and

$$\mathcal{C}^{(2)}_{hor}(\mathcal{Y},V) = \mathcal{C}^{(2)}_{ver}(\mathcal{Y},V) \cup \{[Y_a,Y_b] \mid \ a,b=1,\dots,m\}$$
$$\cup \{2\langle Y_a : \operatorname{grad} V\rangle + [Y_a, \operatorname{grad} V] \mid \ a=1,\dots,m\}.$$

To compute the terms $2\langle Y_a : \operatorname{grad} V\rangle + [Y_a, \operatorname{grad} V]$ in $\mathcal{C}^{(2)}_{hor}(\mathcal{Y},V)$, we have used the computations of Example 5.3.

It would be interesting to be able to derive an inductive formula for computing the families $\mathcal{C}^{(k)}_{ver}(\mathcal{Y},V)$ and $\mathcal{C}^{(k)}_{hor}(\mathcal{Y},V)$. However, such an inductive formula appears to be quite complex.

There are some important statements which can easily be made regarding the distributions $C_{hor}(\mathcal{Y},V)$ and $C_{ver}(\mathcal{Y},V)$.

Remark 5.10.

(1) The generators that we have written for $\mathcal{C}^{(k)}_{ver}(\mathcal{Y},V)$ and $\mathcal{C}^{(k)}_{hor}(\mathcal{Y},V)$ are not linearly independent. Thus one should be able to generate these families with fewer calculations than are necessary to compute the generators we give. One way to do this is to choose a Philip Hall basis for $L(\boldsymbol{X}')$ and compute the image of these brackets under $\operatorname{Ev}(\phi')$. This will work for any given example. However, we are unable to give the general form for the image of a Philip Hall basis under $\operatorname{Ev}(\phi')$.

(2) We claim that $C_{hor}(\mathcal{Y},V)$ is involutive. Let $B_1', B_2' \in \operatorname{Br}(\boldsymbol{X}')$ be brackets which, when evaluated under $\operatorname{Ev}_{0_q}(\phi')$, give vector fields $U_1, U_2 \in \mathcal{C}_{hor}(\mathcal{Y},V)$. Then the decomposition of B_i given by Lemma 5.4 has the form $B_i' = B_i + \tilde{B}_i$, where $B_i \in \operatorname{Br}_0(\boldsymbol{X})$ and $\tilde{B}_i$ is a sum of brackets in $\operatorname{Br}_j(\boldsymbol{X})$ for $j \geq 2$. Therefore, $[B_1', B_2'] = [B_1, B_2] + B''$, where B'' is a sum of brackets in $\operatorname{Br}_j(\boldsymbol{X})$ for $j \geq 2$. This shows that $[U_1, U_2] \in \mathcal{C}_{hor}(\mathcal{Y},V)$. Here we have imposed the condition that brackets in $\operatorname{Br}_{-j}(\boldsymbol{X})$ are taken to be zero for $j \geq 2$ (see Lemma 5.2).

(3) An interesting special case, and one that we shall see in the examples in section 6, is that when $V = 0$. In this case we have $\mathcal{C}_{ver}(\mathcal{Y},V) = \overline{\operatorname{Sym}}(\mathcal{Y})$ and $\mathcal{C}_{hor}(\mathcal{Y},V) = \overline{\operatorname{Lie}}(\overline{\operatorname{Sym}}(\mathcal{Y}))$. This is easily seen in the algorithm by following the path when $\delta_{m+1}(B) = 0$.

(4) The calculations of this section and section 5.2 remain valid if we replace $\operatorname{grad} V$ with an arbitrary vector field on Q.

5.4. Controllability definitions for simple mechanical control systems. It is possible to simply adopt the controllability definitions from nonlinear control theory since our system may be written as a standard control system on TQ. However, since we are dealing with simple control mechanical systems, it is of more interest to us to know what is happening to the *configurations*. A good example of a question of interest in mechanics is, "What is the set of configurations which are reachable from a given configuration if we start at rest?" This is in fact exactly the question that we pose.

DEFINITION 5.11. *A* solution *of* (3) *is a pair,* (c,u)*, where* $c : [0,T] \to Q$ *is a piecewise smooth curve and* $u \in \mathcal{U}$ *such that* (c',u) *satisfies the first-order control system* (3)*.*

Note that since X_L is a second-order vector field on TQ, every solution of the control system (3) will be of the form (c',u) for some curve c on Q. We refer the reader to [1] for a discussion of second-order, and particularly Lagrangian, vector fields.

Let $q_0 \in Q$ and let U be a neighborhood of q_0. We define

$$\mathcal{R}^U_Q(q_0,T) = \{q \in Q \mid \ \text{there exists a solution } (c,u) \text{ of (3)}$$
$$\text{such that } c'(0) = 0_{q_0}, c(t) \in U \text{ for } t \in [0,T], \text{ and } c'(T) \in T_qQ\}$$

and denote $\mathcal{R}^U_Q(q_0, \leq T) = \cup_{t=0}^T \mathcal{R}^U_Q(q_0, t)$. Note that our definitions for reachable configurations do not require us to get to a point in the reachable set at *zero* velocity. They merely ask that we be able to reach that point at *some* velocity. It is, however, required that the initial velocity be zero.

We shall say that $q \in Q$ is an *equilibrium point* for L if $X_L(0_q) = 0$. Let $\mathfrak{E}(L)$ denote the set of equilibrium points for L.

We now introduce our notions of controllability.

DEFINITION 5.12. *We shall say that* (3) *is* locally configuration accessible *at $q_0 \in Q$ if there exists $T > 0$ such that $\mathcal{R}^U_Q(q_0, \leq t)$ contains a nonempty open set of Q for all neighborhoods U of q_0 and all $0 < t \leq T$. If this holds for any $q_0 \in Q$, then the system is called* locally configuration accessible.

We say that (3) *is* small-time locally configuration controllable (STLCC) *at q_0 if it is locally configuration accessible at q_0 and if there exists $T > 0$ such that q_0 is in the interior of $\mathcal{R}^U_Q(q_0, \leq t)$ for every neighborhood U of q_0 and $0 < t \leq T$. If this holds for any $q_0 \in Q$, then the system is called STLCC.*

We shall say that (3) *is* equilibrium controllable *if, for $q_1, q_2 \in \mathfrak{E}(L)$, there exists a solution (c, u) of* (3)*, where $c : [0, T] \to Q$ is such that $c(0) = q_1$, $c(T) = q_2$, and both $c'(0)$ and $c'(T)$ are zero.*

Note that these definitions may be made to apply to any control system which evolves on TQ.

5.5. Conditions for controllability of simple mechanical control systems. In [9] the authors present sufficient conditions for local configuration accessibility in the absence of potential energy. Here, since we have a complete description of $\overline{\text{Lie}}(\mathcal{Y}') \mid Z(TQ)$, we can give stronger results.

THEOREM 5.13. *The control system* (3) *is locally configuration accessible at q if $C_{hor}(\mathcal{Y}, V)(q) = T_qQ$.*

Proof. Let C denote the accessibility distribution. Since $C_{hor}(\mathcal{Y}, V)(q) \subset C(0_q)$ by Proposition 5.9 and $C_{hor}(\mathcal{Y}, V)(q) = T_qQ$ by hypothesis, $Z(TQ)$ must be an integral manifold of C. Let Λ be the maximal integral manifold which contains $Z(TQ)$. Since C is the accessibility distribution, Λ must be invariant under the system (3) and the system must be locally accessible when restricted to Λ. Thus the set $\mathcal{R}^{\tilde{U}}(0_q, \leq T)$ is open in Λ for every neighborhood $\tilde{U} \subset \Lambda$ of 0_q and for every T sufficiently small. Now let U be a neighborhood of q, and define a neighborhood of 0_q in Λ by $\tilde{U} = \tau_Q^{-1}(U) \cap \Lambda$. The set $\tau_Q(\mathcal{R}^{\tilde{U}}(0_q, \leq T))$ is open in Q for T sufficiently small since τ_Q is an open mapping. This proves the theorem. □

We also have a partial converse to Theorem 5.13 in the case when there is no potential energy.

THEOREM 5.14. *Suppose that $V = 0$ and* (3) *is locally configuration accessible. Then $C_{hor}(\mathcal{Y}, V)(q) = T_qQ$ for q in an open dense subset of Q.*

Proof. First note that if $C_{hor}(\mathcal{Y}, V)(q_0) = T_{q_0}Q$, then $C_{hor}(\mathcal{Y}, V)(q) = T_qQ$ in a neighborhood of q_0. This proves that the set of points q where $C_{hor}(\mathcal{Y}, V)(q) = T_qQ$ is open. Now suppose that $C_{hor}(\mathcal{Y}, V)(q) \subsetneq T_qQ$ in an open subset U of Q. Then there exists an open subset $\bar{U} \subset U$ so that $\text{rank}(C_{hor}(\mathcal{Y}, V)(q)) = k < n$ for all $q \in \bar{U}$. However, this contradicts local configuration accessibility. Therefore, there can be no open subset of Q on which $C_{hor}(\mathcal{Y}, V)(q) \subsetneq T_qQ$. Thus the set of points q where $C_{hor}(\mathcal{Y}, V)(q) = T_qQ$ is dense. This completes the proof. □

We may also prove an easy statement about STLCC. We need to say a few things about "good" and "bad" symmetric products. Let $\boldsymbol{Y} = \{X_1, \ldots, X_{m+1}\}$, and establish a bijection $\psi : \boldsymbol{Y} \to \mathcal{Y} \cup \{\text{grad}\, V\}$ by asking that $\psi(X_a) = Y_a$ for

$a = 1, \ldots, m$ and $\psi(X_{m+1}) = \operatorname{grad} V$. If $P \in \Pr(\boldsymbol{Y})$, we shall say that P is *bad* if $\gamma_a(P)$ is even for each $a = 1, \ldots, m$. We say that P is *good* if it is not bad. Let S_m denote the permutation group on m symbols. For $\pi \in S_m$ and $P \in \Pr(\boldsymbol{Y})$ define $\bar{\pi}(P)$ to be the bracket obtained by fixing X_{m+1} and sending X_a to $X_{\pi(a)}$ for $a = 1, \ldots, m$. Now define

$$\rho(P) = \sum_{\pi \in S_m} \bar{\pi}(P).$$

We may now state the sufficient conditions for STLCC.

THEOREM 5.15. *Suppose that* $\mathcal{Y} \cup \{\operatorname{grad} V\}$ *is such that every bad symmetric product in* $\Pr(\boldsymbol{Y})$ *has the property that*

$$\operatorname{Ev}_{0_q}(\psi)(\rho(P)) = \sum_{a=1}^{m} \xi_a \operatorname{Ev}_{0_q}(\psi)(C_a),$$

where C_a *are good symmetric products in* $\Pr(\boldsymbol{Y})$ *of lower degree than* P *and* $\xi_a \in \mathbb{R}$ *for* $a = 1, \ldots, m$. *Also, suppose that* (3) *is locally configuration accessible at* q. *Then* (3) *is STLCC at* q.

Proof. First recall from the proof of Theorem 5.13 that if (3) is locally configuration accessible at q, then $Z(TQ)$ is an integral manifold for the accessibility distribution. We let Λ be the maximal integral manifold for the accessibility distribution which contains $Z(TQ)$. Restricted to Λ, (3) is locally accessible. To show that (3) is STLCC at q, it clearly suffices to show that (3) is STLC at 0_q when restricted to Λ. We do this by showing that (3) satisfies the hypotheses of Theorem 4.1 if it satisfies the stated hypotheses on the symmetric products. To do this we shall show that there is a one-to-one correspondence between bad brackets in $\operatorname{Br}(\boldsymbol{X}')$ and bad symmetric products in $\Pr(\boldsymbol{Y})$ and good brackets in $\operatorname{Br}(\boldsymbol{X}')$ and good symmetric products in $\Pr(\boldsymbol{Y})$.

Suppose that $B' \in \operatorname{Br}(\boldsymbol{X}')$ is bad. Thus $\delta_a(B')$ is even for $a = 1, \ldots, m$ and $\delta_0(B')$ is odd. When we evaluate $\operatorname{Ev}_{0_q}(\phi')(B')$, the only terms that will remain in the decomposition of $\operatorname{Ev}(\phi')(B')$ given by Lemma 5.4 are the terms obtained from brackets in $\boldsymbol{S}(B')$ which are in $\operatorname{Br}_0(\boldsymbol{X}) \cup \operatorname{Br}_{-1}(\boldsymbol{X})$. Since B' is bad, we must have $\delta_a(B)$ even and $\delta_0(B) + \delta_{m+1}(B)$ odd for each $B \in \boldsymbol{S}(B')$. If $\delta_0(B)$ is odd, then $\delta_{m+1}(B)$ must be even. In this case we get $\sum_{a=1}^{m+1} \delta_a(B)$ as even and $\delta_0(B)$ as odd. Thus the only brackets in $\boldsymbol{S}(B')$ which contribute to $\operatorname{Ev}(\phi')(B')$ must be in $\operatorname{Br}_{-1}(\boldsymbol{X})$. This will give us a vector in $V_{0_q}TQ$ which comes from a symmetric product which is bad. Now suppose that $\delta_0(B)$ is even for $B \in \boldsymbol{S}(B')$. Then $\delta_{m+1}(B)$ must be odd. In this case $\sum_{a=1}^{m+1} \delta_a(B)$ is odd and $\delta_0(B)$ is even, and again, the only brackets in $\boldsymbol{S}(B')$ which contribute to $\operatorname{Ev}(\phi')(B')$ must be in $\operatorname{Br}_{-1}(\boldsymbol{X})$. We then conclude that $\operatorname{Ev}_{0_q}(\phi')(B')$ must be of the form $(\operatorname{Ev}_q(\psi)(P))^{lift}$, where $P \in \Pr(\boldsymbol{Y})$ is bad.

Now suppose that $B' \in \operatorname{Br}(X')$ is good. It is clear that if $\delta_a(B')$ is odd for any $a = 1, \ldots, m$, then B' cannot give rise to a bad symmetric product. Thus we may suppose that $\delta_a(B')$ is even for each $a = 0, \ldots, m$. Now let's look at what the brackets look like from $\boldsymbol{S}(B')$ which contribute to $\operatorname{Ev}(\phi')(B')$. Let B be such a bracket. We must have $\delta_a(B)$ even for $a = 1, \ldots, m$ and $\delta_0(B) + \delta_{m+1}(B)$ even. If $\delta_0(B)$ is odd, then $\delta_{m+1}(B)$ must be odd. Since B is primitive, this means that $\sum_{a=1}^{m+1} \delta_a(B)$ and $\delta_0(B)$ are odd. Therefore, B must be in $\operatorname{Br}_0(\boldsymbol{X})$. Now suppose that $\delta_0(B)$ is even. Then $\delta_{m+1}(B)$ must also be even. Thus $\sum_{a=1}^{m+1} \delta_a(B)$ and $\delta_0(B)$ are

even and so $B \in \mathrm{Br}_0(\boldsymbol{X})$. Therefore, good brackets from $\mathrm{Br}(\boldsymbol{X}')$ do not generate any bad symmetric products. □

Since the system restricted to the integral manifold Λ in the proof of the above theorem is STLC, the hypotheses of the theorem imply more than STLCC. In fact, the following corollary is easily seen to be true.

COROLLARY 5.16. *Suppose that the hypotheses of Theorem* 5.15 *hold for each* $q \in Q$. *Then the system* (3) *is equilibrium controllable.*

Remark 5.17.

(1) We have shown that it is not necessary to be able to generate *all* directions on TQ to obtain controllability in the configuration variables. Indeed, the only vertical directions that we generate are $C_{ver}(\mathcal{Y}, V)$ which need not span $V_{0_q}TQ$. This means that the notion of configuration controllability is genuinely weaker than are the standard notions of controllability if we are to simply regard the system (3) as a typical nonlinear control system.

(2) Corollary 5.16 may be made even stronger if we allow a point $q \in Q$ to be an equilibrium point if $\operatorname{grad} V(q)$ is in the span of the inputs at q.

6. Examples of mechanical control systems. In this section we present some examples. The examples are rather simple and are intended to illustrate the concepts put forward by the theory. One of the advantages of the conditions for local configuration accessibility given in Theorem 5.13 is that it lends itself to symbolic computation. Indeed, a Mathematica package was written to facilitate the computations in this section.

6.1. The robotic leg. This example, although simple, exhibits much of the subtle behavior that makes the study of mechanical systems interesting. The example is a rigid body with inertia J which is pinned to ground at its center of mass. The body has attached to it an extensible massless leg, and the leg has a point mass with mass m at its tip. The coordinate θ will describe the angle of the body, and ψ will describe the angle of the leg from an inertial reference frame. The coordinate r will describe the extension of the leg. Thus the configuration space for this problem is $Q = \mathbb{T}^2 \times \mathbb{R}^+$. See Figure 3. In the coordinates (θ, ψ, r) the Riemannian metric for the robotic leg is

$$g = Jd\theta \otimes d\theta + mr^2 d\psi \otimes d\psi + mdr \otimes dr,$$

the input one-forms are $F^1 = d\theta - d\psi$ and $F^2 = dr$, and the potential energy function is zero. We may compute the input vector fields to be

$$Y_1 = \frac{1}{J}\frac{\partial}{\partial\theta} - \frac{1}{mr^2}\frac{\partial}{\partial\psi}, \qquad Y_2 = \frac{1}{m}\frac{\partial}{\partial r}.$$

Since there is no potential energy present, the distribution $C_{hor}(\mathcal{Y}, V)$ is simply generated by the vector fields $\overline{\mathrm{Lie}}(\overline{\mathrm{Sym}}(\mathcal{Y}))$.

We will find the following computations to be sufficient:

$$\langle Y_1 : Y_1 \rangle = -\frac{2}{m^2 r^3}\frac{\partial}{\partial r}, \qquad \langle Y_1 : Y_2 \rangle = 0, \qquad \langle Y_2 : Y_2 \rangle = 0,$$

$$[Y_1, Y_2] = -\frac{2}{m^2 r^3}\frac{\partial}{\partial\psi}, \qquad [Y_1, \langle Y_1 : Y_1 \rangle] = \frac{4}{m^3 r^6}\frac{\partial}{\partial\psi}.$$

The controllability results for the robotic leg are displayed in Table 1.

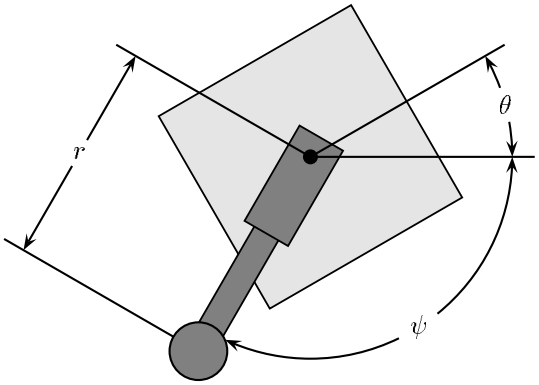


FIG. 3. *The robotic leg.*

TABLE 1

Controllability results for the robotic leg. The first column displays which inputs are present, the second column indicates whether the system is locally configuration accessible with these inputs, the third column indicates whether the system with these inputs satisfies the sufficient conditions of Theorem 5.15 *for STLCC, and the last column indicates whether the system with these inputs is actually STLCC.*

Inputs	Locally configuration accessible?	Satisfies sufficient conditions for STLCC?	STLCC?
Y_1 (torque)	yes	no	no
Y_2 (extension)	no	no	no
Y_1 and Y_2	yes	yes	yes

Remark 6.1.

(1) The linearization of this system at points of zero velocity is not controllable with any combination of inputs, so the controllability does not follow from linear results.

(2) When only the input Y_2 is present, the equations are

$$\ddot{r} - r\dot{\psi}^2 = \frac{1}{m}u_1,$$
$$\ddot{\theta} = 0,$$
$$\ddot{\psi} + \frac{2}{r}\dot{r}\dot{\psi} = 0.$$

Note that when the initial velocity is zero, the top equation decouples from the bottom two equations. Physically this means that we are simply moving the leg back and forth with no effect on the configuration of the body since the initial velocity is zero.

(3) Although the system only violates the *sufficient* conditions for STLCC with the input Y_1, one may easily see by looking at the r-component of the equations of motion that the system is, in fact, not STLCC. The reason for this is that, since $\ddot{r} \geq 0$, r will always increase no matter what happens to the other variables. Thus our initial configuration will never be in the interior of the set of reachable configurations.

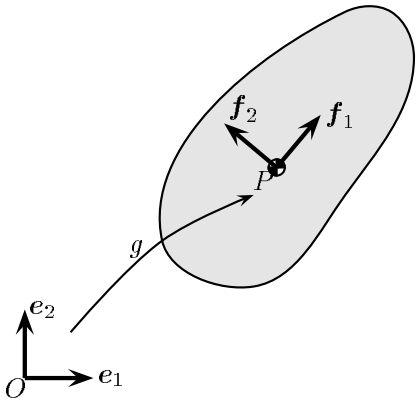


FIG. 4. *The configuration of a planar rigid body as an element of SE*(2).

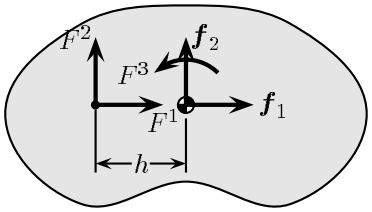


FIG. 5. *Positions for application of forces on a planar rigid body after simplifying assumptions.*

6.2. The forced planar rigid body. In this section we study the planar rigid body discussed in the introduction with various combinations of forces and torques. The configuration space for the system is the Lie group $SE(2)$. To establish the correspondence between the configuration of the body and $SE(2)$, fix a point $O \in \mathbb{R}^2$ and let $\{\boldsymbol{e}_1 = \frac{\partial}{\partial x}, \boldsymbol{e}_2 = \frac{\partial}{\partial y}\}$ be the standard orthonormal frame at that point. Let $\{\boldsymbol{f}_1, \boldsymbol{f}_2\}$ be an orthonormal frame attached to the body at its center of mass. The configuration of the body is determined by the element $g \in SE(2)$ which maps the point O with its frame $\{\boldsymbol{e}_1, \boldsymbol{e}_2\}$ to the position, P, of the center of mass of the body with its frame $\{\boldsymbol{f}_1, \boldsymbol{f}_2\}$. See Figure 4. The inputs for this problem consist of forces applied at an arbitrary point and a torque about the center of mass. Without loss of generality (by redefining our body reference frame $\{\boldsymbol{f}_1, \boldsymbol{f}_2\}$) we may suppose that the point of application of the force is a distance h along the $\boldsymbol{f}_1$ body-axis from the center of mass. The situation is illustrated in Figure 5.

With this convention fixed, we shall use coordinates (x, y, θ) for the planar rigid body, where (x, y) describe the position of the center of mass and θ describes the orientation of the frame $\{\boldsymbol{f}_1, \boldsymbol{f}_2\}$ with respect to the frame $\{\boldsymbol{e}_1, \boldsymbol{e}_2\}$. In these coordinates, the Riemannian metric for the system is

$$g = m dx \otimes dx + m dy \otimes dy + J d\theta \otimes d\theta.$$

Here m is the mass of the body and J is its moment of inertia about the center of mass. The inputs are described by the one-forms

$$F^1 = \cos\theta dx + \sin\theta dy, \qquad F^2 = -\sin\theta dx + \cos\theta dy - h d\theta, \qquad F^3 = d\theta,$$

TABLE 2

Controllability results for the planar rigid body. The first column displays which inputs are present, the second column indicates whether the system is locally configuration accessible with these inputs, the third column indicates whether the system with these inputs satisfies the sufficient conditions of Theorem 5.15 *for STLCC, and the last column indicates whether the system with these inputs is actually STLCC.*

Inputs	Locally configuration accessible?	Satisfies sufficient conditions for STLCC?	STLCC?
Y_1 (force at CM)	no	no	no
Y_2 (force $\perp$ to CM)	yes	no	no
Y_3 (torque)	no	no	no
Y_1 and Y_2	yes	yes	yes
Y_1 and Y_3	yes	yes	yes
Y_2 and Y_3	yes	no	yes

from which we compute the input vector fields as

$$Y_1 = \frac{\cos\theta}{m}\frac{\partial}{\partial x} + \frac{\sin\theta}{m}\frac{\partial}{\partial y},$$

$$Y_2 = -\frac{\sin\theta}{m}\frac{\partial}{\partial x} + \frac{\cos\theta}{m}\frac{\partial}{\partial y} - \frac{h}{J}\frac{\partial}{\partial\theta}, \qquad Y_3 = \frac{1}{J}\frac{\partial}{\partial\theta}.$$

Again, as with the robotic leg, there is no potential energy, so the distribution $C_{hor}(\mathcal{Y}, V)$ may be computed by calculating $\overline{\mathrm{Lie}}(\overline{\mathrm{Sym}}(\mathcal{Y}))$.

The following computations are sufficient to obtain the results that we desire:

$$\langle Y_1 : Y_1\rangle = 0, \qquad \langle Y_1 : Y_2\rangle = \frac{h\sin\theta}{mJ}\frac{\partial}{\partial x} - \frac{h\cos\theta}{mJ}\frac{\partial}{\partial y},$$

$$\langle Y_1 : Y_3\rangle = -\frac{\sin\theta}{mJ}\frac{\partial}{\partial x} + \frac{\cos\theta}{mJ}\frac{\partial}{\partial y}, \qquad \langle Y_2 : Y_2\rangle = \frac{2h\cos\theta}{mJ}\frac{\partial}{\partial x} + \frac{2h\sin\theta}{mJ}\frac{\partial}{\partial y},$$

$$\langle Y_2 : Y_3\rangle = -\frac{\cos\theta}{mJ}\frac{\partial}{\partial x} - \frac{\sin\theta}{mJ}\frac{\partial}{\partial y}, \qquad \langle Y_3 : Y_3\rangle = 0,$$

$$[Y_1, Y_2] = -\frac{h\sin\theta}{mJ}\frac{\partial}{\partial x} + \frac{h\cos\theta}{mJ}\frac{\partial}{\partial y}, \qquad [Y_1, Y_3] = \frac{\sin\theta}{mJ}\frac{\partial}{\partial x} - \frac{\cos\theta}{mJ}\frac{\partial}{\partial y},$$

$$[Y_2, Y_3] = \frac{\cos\theta}{mJ}\frac{\partial}{\partial x} + \frac{\sin\theta}{mJ}\frac{\partial}{\partial y}, \qquad [Y_2, \langle Y_2 : Y_2\rangle] = \frac{2h^2\sin\theta}{mJ^2}\frac{\partial}{\partial x} - \frac{2h^2\cos\theta}{mJ^2}\frac{\partial}{\partial y}.$$

With the computations done, we may proceed to determine configuration controllability for the planar rigid body with various combinations of inputs. The results are displayed in Table 2.

Remark 6.2.

(1) The linearization of this system around points of zero velocity is not controllable so the cases where the system is STLCC do not follow from the linear calculations.

(2) In this example, in the cases when the system fails to satisfy the sufficient conditions for STLCC of Theorem 5.15, we are not able to say whether the system is,

in fact, not STLCC. In fact, when the inputs Y_2 and Y_3 are present, even though the system does not satisfy the sufficient conditions of Theorem 5.15, it is easy to see that it *is* STLCC. Recent work, beyond the scope of this paper, shows that when only the input Y_2 is present, the system is not STLCC.

(3) In the case when only the input Y_1 is present, it is illustrative to represent the equations in the coordinates $(\xi, \eta, \psi) = (x\cos\theta + y\sin\theta, -x\sin\theta + y\cos\theta, \theta)$. In these coordinates the equations have the form

$$\begin{aligned}
&\ddot{\xi} + 2\left(\frac{m\eta^2}{J} - \frac{J + m\eta^2}{J}\right)\dot{\eta}\dot{\psi} + \left(\frac{m\xi\eta^2}{J} - \frac{\xi J + m\xi\eta^2}{J}\right)\dot{\psi}^2 \\
&\qquad = \left(\frac{J + m\eta^2}{J} - \frac{\eta^2}{J}\right)u_1, \\
&\ddot{\eta} + 2\left(\frac{J + m\xi^2}{J} - \frac{m\xi^2}{J}\right)\dot{\xi}\dot{\psi} + \left(\frac{m\eta\xi^2}{J} - \frac{\eta J + m\eta\xi^2}{J}\right)\dot{\psi}^2 = 0, \\
&\ddot{\psi} = 0.
\end{aligned}$$

Note that the top equation decouples from the last two equations when the initial velocity is zero. Since Y_1 is directed toward the center of mass, applying this input will cause the body to move in this direction and none of the other degrees of freedom are affected.

(4) In the case when the input Y_3 is present, the equations have the form

$$\begin{aligned}
\ddot{\theta} &= \frac{1}{J}u_3, \\
\ddot{x} &= 0, \\
\ddot{y} &= 0.
\end{aligned}$$

Again, the top equation decouples from the bottom two equations. This time the coupling is true for all initial velocities, not just zero initial velocity. In this case we see that the input simply causes a rotation of the body about its center of mass. The position of the center of mass is not affected if the initial velocity is zero.

6.3. The pendulum on a cart. To illustrate the effects of potential energy, consider the problem of a pendulum suspended from a cart. The configuration manifold for the system is $Q = \mathbb{R} \times \mathbb{S}^1$. As coordinates we shall use (x, θ) as shown in Figure 6. In this case the Riemannian metric for the system is

$$g = (M + m)dx \otimes dx + ml\cos\theta dx \otimes d\theta + ml\cos\theta d\theta \otimes dx + ml^2 d\theta \otimes d\theta.$$

Here M is the mass of the cart and m is the mass of the pendulum. The potential energy is

$$V = ma_g l(1 - \cos\theta),$$

where a_g is the acceleration due to gravity. The input is given by the one-form

$$F^1 = dx.$$

The input vector field is then readily computed to be

$$Y_1 = \frac{ml^2}{m^2l^2 + Mml^2 - m^2l^2\cos^2\theta}\frac{\partial}{\partial x} + \frac{ml\cos\theta}{m^2l^2 + Mml^2 - m^2l^2\cos^2\theta}\frac{\partial}{\partial\theta}.$$

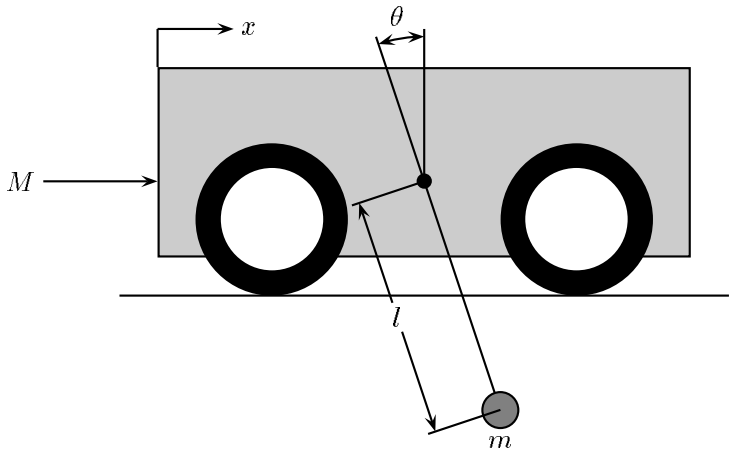


FIG. 6. *Pendulum suspended from a cart.*

To compute $C_{hor}(\mathcal{Y}, V)$ we need the following computations:

$$\langle Y_1 : Y_1 \rangle = \frac{16m\cos^2\theta\sin\theta}{l(m+2M-m\cos 2\theta)^3}\frac{\partial}{\partial x} + \frac{8(M+m)\sin\theta}{l^2(m\cos 2\theta - m - 2M)^3}\frac{\partial}{\partial\theta},$$

$$\begin{aligned}\langle Y_1 : \mathrm{grad}V \rangle = {} & \frac{4a_g m\cos\theta(m - m\cos 2\theta - 2M\cos 2\theta)}{l(m\cos 2\theta - m - 2M)^3}\frac{\partial}{\partial x} \\ & + \frac{4a_g(2M^2\cos 2\theta + 3Mm\cos 2\theta + m^2\cos 2\theta - Mm - m^2)}{l^2(m\cos 2\theta - m - 2M)^3}\frac{\partial}{\partial\theta}.\end{aligned}$$

Note that at all points $q \in Q$ except those where $\theta \in \{0, \pi\}$, the vector fields $\{Y_1, \langle Y_1 : Y_1 \rangle\}$ generate the tangent space at q. This means that the system is locally configuration accessible at these points. Also, at these points the bad symmetric product $\langle Y_1 : Y_1 \rangle$ is not a multiple of Y_1, so the system may not be STLCC at these points. At points where $\theta \in \{0, \pi\}$, the vector fields $\{Y_1, \langle Y_1 : \mathrm{grad}\, V \rangle\}$ span T_qQ, and so the system is also locally configuration accessible at these points. Most important, however, the bad symmetric product vanishes at these two points so the system is STLCC at these equilibria. This must be so as, at these two points, the linearized system is controllable.

7. Conclusions and future work. In this paper we have outlined what we regard as a *beginning* of a thorough program for analysis and synthesis for simple mechanical control systems. The first part of such a program is to determine the pertinent versions of controllability (local configuration accessibility and STLCC) and determine algebraic tests for these notions of controllability. The conditions that we present for checking our versions of controllability involve only computations on the configuration space. In determining these conditions, the symmetric product proved to play an important role. As we have presented it, the symmetric product is a useful computational tool. Our recent work provides a fairly complete description of the geometric role of the symmetric product in the control of mechanical systems. This will be the subject of an upcoming paper.

In the examples in section 6 some interesting circumstances may be observed. The most interesting of these is a comparison of the robotic leg with input Y_2 and

the planar rigid body with the inputs Y_2 and Y_3. In the former case the system does not satisfy the sufficient conditions for STLCC and may be shown to indeed not be STLCC. However, in the latter case, even though the sufficient conditions for STLCC are not met, the system *is* STLCC. It would be interesting to better understand why this happens, and perhaps arrive at a stronger condition for STLCC.

Finally we mention that, from a practical point of view, perhaps the most useful contribution is that of the notion, mentioned in section 5.4, of equilibrium controllability. If a system satisfies the hypotheses of Theorem 5.15 at each configuration, it would be interesting to determine a means of generating paths which connect points in the configuration manifold at zero velocity. Such an algorithm may involve a deeper understanding of the symmetric product.

In summary, we feel that this paper provides an effective initial understanding of mechanical control systems, and we hope that it will prove to be a useful foundation for further work in the area of mechanical control theory.

Acknowledgments. We would like to thank Jerry Marsden and Jim Ostrowski for helpful conversations. The anonymous reviewers were also very helpful in improving the presentation of the paper during the review process.

REFERENCES

[1] R. Abraham and J. E. Marsden, *Foundations of Mechanics*, 2nd ed., Addison-Wesley, Reading, MA, 1978.

[2] A. M. Bloch and P. E. Crouch, *Kinematics and dynamics of nonholonomic control systems on Riemannian manifolds*, in Proc. 32nd IEEE Conf. on Decision and Control, Tucson, AZ, Dec. 1992, IEEE, Piscataway, NJ, pp. 1–5.

[3] A. M. Bloch, M. Reyhanoglu, and N. H. McClamroch, *Control and stabilization of nonholonomic dynamic systems*, IEEE Trans. Automat. Control, 37 (1992), pp. 1746–1757.

[4] P. E. Crouch, *Geometric structures in systems theory*, Institution of Electrical Engineers. Proceedings. D. Control Theory and Applications, 128 (1981), pp. 242–252.

[5] H. Hermes, *Control systems which generate decomposable Lie algebras*, J. Differential Equations, 44 (1982), pp. 166–187.

[6] N. Jacobson, *Lie Algebras*, Interscience Tracts in Pure and Applied Mathematics 10, Interscience, New York, 1962.

[7] W. Klingenberg, *Riemannian Geometry*, Walter de Gruyter, Berlin, New York, 1982.

[8] A. D. Lewis, *Aspects of Geometric Mechanics and Control of Mechanical Systems*, Ph.D. thesis, California Institute of Technology, Technical report CIT-CDS 95-017, available electronically via http://avalon.caltech.edu/cds/.

[9] A. D. Lewis and R. M. Murray, *Configuration controllability of a class of mechanical systems*, in Proc. 34th IEEE Conf. Decision and Control, New Orleans, LA, Dec. 1995, IEEE, Piscataway, NJ, pp. 1–5.

[10] L. San Martin and P. E. Crouch, *Controllability on principal fibre bundles with compact structure group*, Systems Control Lett., 5 (1984), pp. 35–40.

[11] J.-P. Serre, *Lie Algebras and Lie Groups*, Lecture Notes in Math. 1500, Springer-Verlag, New York, Heidelberg, Berlin, 1992.

[12] H. J. Sussmann, *Lie brackets and local controllability: A sufficient condition for scalar-input systems*, SIAM J. Control Optim., 21 (1983), pp. 686–713.

[13] H. J. Sussmann, *A general theorem on local controllability*, SIAM J. Control Optim., 25 (1987), pp. 158–194.

SIAM J. CONTROL OPTIM.
Vol. 35, No. 3, pp. 791–809, May 1997

004

WEIGHTED SENSITIVITY MINIMIZATION FOR CAUSAL, LINEAR, DISCRETE TIME–VARYING SYSTEMS*

MICHEL VERHAEGEN†

Abstract. The weighted sensitivity minimization problem for discrete time–varying systems is treated in a state space framework. Given a controllable and causal (stable) state space realization of the plant to be controlled, the *first* step in the solution is the computation of an outer-inner factorization of the plant. The key algorithmic step here is the solution of a Lyapunov type of equation running backward in time. Based on the part of the realization of the inner (isometric) factor related to its output state space we then formulate and solve a Nevanlinna–Pick interpolation problem. This *second* step is also characterized by a Lyapunov equation. It is shown that the solution to the sensitivity minimization problem exists when the solution to this Lyapunov equation is positive definite for all time instances. Finally, we pay special attention to the minimal disturbance attenuation level when the latter is assumed to be equal to a constant scalar for all time instances as well as to a square root implementation of the recursive equations.

Key words. discrete time–varying systems, Nevanlinna–Pick interpolation, outer-inner factorization, Lyapunov equations

AMS subject classifications. Primary, 93B36; Secondary, 93C50, 93C55, 93D15

PII. S036301299325339X

1. Introduction.

The sensitivity minimization problem was introduced by Zames in [1]. This type of problem constitutes a fundamental problem in the area of robust (H_∞) control.

So far, only a limited number of papers have appeared addressing the problem in a time-varying context. For discrete periodic time-varying systems, we mention the work of [2] and [3], which presents an algorithm for the operator theoretic solution outlined in [2]. Basically, this solution is characterized by "lifting up" the periodic time-variant system into a time-invariant system of big state dimension and then solving the problem using standard tools developed for time-invariant systems.

This approach fails when the system varies arbitrarily in time. Recently, based on [7], the sensitivity minimization problem for discrete time–varying systems was formulated and solved as a Nevanlinna–Pick interpolation (NPI) problem in [8]. In this way, this approach follows the original strategy of the solution presented in [10].

To formulate the NPI problem in [8], it was assumed that the inverse of the plant to be controlled is given into a simple partial fraction expansion format. Furthermore, that paper is restricted to systems of constant state, input, and output dimension. Both conditions will hamper the application of the outlined solution to practical circumstances. In a more realistic environment, the starting point in the controller design is a state space description of the plant P to be controlled. Therefore, the computation of the specific partial fraction expansion still needs to be performed. In [11], a procedure has been developed to calculate a state space realization of the required anticausal part of P^{-1}. This solution requires solving a Riccati equation, and also it is restricted to the case of constant state dimensions.

*Received by the editors August 6, 1993; accepted for publication (in revised form) March 15, 1996. This research was supported by a senior research fellowship from the Royal Dutch Academy of Arts and Sciences.

http://www.siam.org/journals/sicon/35-3/25339.html

†Department of Electrical Engineering, Systems and Control Laboratory, Delft University of Technology, P.O. Box 5031, NL-2600 GA Delft, The Netherlands (M.Verhaegen@et.tudelft.nl).

It has been observed that in realistic circumstances as reported in [13], the state dimension can change. Other examples include mechanical systems, such as a robot arm, where the degrees of freedom may change during operation. This in combination with the important class of problems related to multirate sampled data systems requires the treatment of time-varying systems exhibiting nonconstant state, input, and output dimensions.

To overcome the above drawbacks of existing methods, we present a solution to the weighted sensitivity minimization problem for discrete time–varying systems based on the inner-outer factorization technique developed in [15]. This solution only requires a state space realization of the plant P and allows varying state, input, and output dimensions. The key algorithmic steps are two recursive Lyapunov equations, both running backward in time.

An alternative to the interpolation approach in tackling robust control problems is the so-called "Riccati" approach, outlined in [9] for the continuous time–invariant case. Recently, this "standard" robust control problem has been extended to the continuous time–variant case with constant system dimensions in [4] and to the discrete time–variant case with varying system dimensions in [5].

Reformulating the sensitivity minimization problem as a standard robust control problem and solving the latter problem has two major drawbacks:

(1) The reformulation of the sensitivity minimization problem, precisely defined later on in section 3 (Problem 1), to a standard robust control problem is pictured in Figure 1. Here we adopt the standard notation such as used in [9]. This "standard" formulation does, however, violate one of the standard assumptions, namely that (one of) the diagonal operators of the feedthrough operator of the generalized plant G is nonzero. Making this formulation of the sensitivity minimization problem conformal to the standard assumptions substantially complicates the formulas which arise in the solution to the standard problem under standard assumptions [9].

(2) Even when one is willing to invest the energy in reformulating the sensitivity minimization problem as a standard robust control problem, the solution remains characterized by recursive Riccati equations. In [14], it is shown that the numerical solution of these recursive Riccati equations may diverge due to the accumulation of round-off errors. Coping with this phenomenon again requires additional precautions which complicate the overall solution.

In this paper, we overcome both drawbacks by tackling the sensitivity minimization problem for discrete time–varying systems with varying system dimensions in a more direct way. This direct route shows that the solution to this problem only requires the solution of two (recursive) Lyapunov equations instead of two (recursive) Riccati equations. These Lyapunov equations do not suffer from the divergence problems related to the Riccati equations. Both the direct way of addressing the problem and the solution of the problem via Lyapunov equations make the solution derived in this paper more robust and elegant compared with existing solutions in a time-variant context.

The outline of the paper is as follows. After reviewing the nomenclature of linear time-varying (LTV) systems in section 2, we formulate the sensitivity minimization problem in section 3. The reformulation of this problem as a NPI problem is done in section 4. In section 5, we briefly review its solution and mainly focus on the different viewpoint taken in this paper compared with that in [7]. These include (a) the formulation of the sensitivity minimization problem without making use of the generalized point evaluation map as done in [7, 8] and (b) the treatment of varying

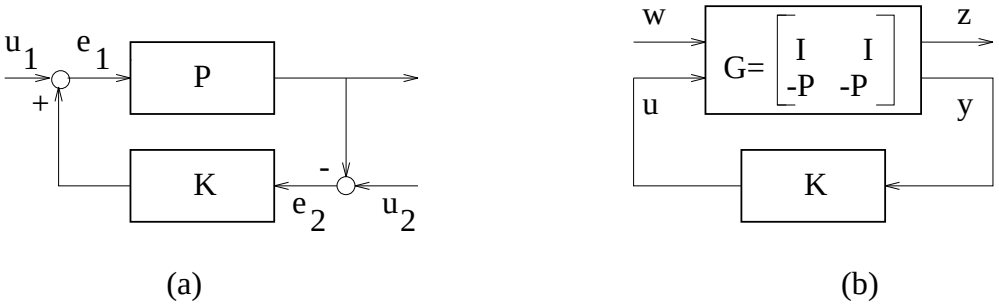


FIG. 1. (a) *The closed-loop configuration in the sensitivity minimization problem and* (b) *the closed-loop configuration in the formulation of the sensitivity minimization problem as a standard robust control problem.*

system dimensions. A summary of the numerical procedure and a brief highlight of some computational aspects are given in section 6. Finally section 7 concludes with some remarks.

2. Preliminaries. In this section, we introduce the notation used in representing LTV systems and collect the basic results from [15] about outer-inner factorization.

A state space realization of the LTV system P to be controlled is denoted on a local time scale as

$$\begin{aligned} x_{k+1} &= x_k A_k + u_k B_k, \\ y_k &= x_k C_k + u_k D_k, \end{aligned} \tag{1}$$

where x_k, u_k, and y_k are (finite-dimensional) row vectors in, respectively, $\mathbb{C}^{N_k}, \mathbb{C}^{M_k}$, and $\mathbb{C}^{L_k}$ and the matrices $\{A_k, B_k, C_k, D_k\}$ are bounded matrices of appropriate dimensions. Remark that this notation is compatible with the earlier work on LTV systems as reported in [16, 17, 18, 19].

To denote the state space representation more compactly, we introduce, as in [16, 17, 18, 19], the dimension space sequences $\mathcal{B}$:

$$\mathcal{B} = \cdots \times \boxed{\mathcal{B}_0} \times \mathcal{B}_1 \times \cdots,$$

where $\mathcal{B}_k = \mathbb{C}^{N_k}$ and the square box identifies the space of the zeroth entry. In a similar way, we introduce the dimension space sequence $\mathcal{M}$ and $\mathcal{N}$ from the integer sequences $\{M_k\}$ and $\{L_k\}$. It is allowed that some integers in these sequences are zero. The space of sequences in $\mathcal{B}$ with finite 2-norm will be denoted by $\ell_2^{\mathcal{B}}$. Next we stack the sequence of state vectors x_k, input vectors u_k, and output vectors y_k into ∞-dimensional row vectors x, u, and y, denoted explicitly for the state vector sequence as

$$x = \begin{bmatrix} \cdots & x_{-1} & \boxed{x_0} & x_1 & \cdots \end{bmatrix},$$

where the square identifies the position of the zeroth entry. Let $\mathcal{B}^{(-1)}$ denote the shifted dimension space sequence of $\mathcal{B}$, i.e.,

$$\mathcal{B}^{(-1)} = \cdots \times \boxed{\mathcal{B}_1} \times \mathcal{B}_2 \times \cdots,$$

and let $\mathcal{D}(\mathcal{M},\mathcal{N})$ denote the Hilbert space of bounded diagonal operators $\ell_2^{\mathcal{M}} \to \ell_2^{\mathcal{N}}$. Then we can stack the system operators A_k, B_k, C_k, and D_k into the diagonal

operators A, B, C, and D as (denoted only explicitly for A)

$$A = \operatorname{diag}\left[\begin{array}{ccccc} \cdots & A_{-1} & \boxed{A_0} & A_1 & \cdots \end{array}\right] \in \mathcal{D}(\mathcal{B}, \mathcal{B}^{(-1)}), \qquad \begin{array}{rcl} C & \in & \mathcal{D}(\mathcal{B}, \mathcal{N}), \\ B & \in & \mathcal{D}(\mathcal{M}, \mathcal{B}^{(-1)}), \\ D & \in & \mathcal{D}(\mathcal{M}, \mathcal{N}). \end{array}$$

Let the causal bilateral shift operator on sequences be denoted by Z such that

$$\left[\begin{array}{ccccc} \cdots & x_{-1} & \boxed{x_0} & x_1 & \cdots \end{array}\right] Z = \left[\begin{array}{ccccc} \cdots & x_{-2} & \boxed{x_{-1}} & x_0 & \cdots \end{array}\right];$$

then a compact notation on a global time scale of the state space representation (1) is

$$\begin{array}{rcl} xZ^{-1} & = & xA + uB, \\ y & = & xC + uD, \end{array} \quad \text{also denoted as} \quad \mathbf{P} = \left[\begin{array}{cc} A & C \\ B & D \end{array}\right]. \tag{2}$$

With this notation it is possible to represent an LTV system as an operator. Let the transition operator $\Phi(j,k)$ of the system with state space representation (2) be defined as

$$\Phi(j,k) = \begin{cases} A_k A_{k+1} \cdots A_{j-1}, & j > k, \\ I, & j = k, \\ \text{undefined}, & j < k, \end{cases}$$

and let $\lim_{j\to\infty} \Phi(j,k) = 0 \ \forall k < \infty$; then the inverse of the operator $(I - AZ)$ exists and is in $\mathcal{U}$, and the operator representation of the (*asymptotically stable*) LTV system P becomes

$$P = D + BZ(I - AZ)^{-1}C. \tag{3}$$

This transfer operator is *upper* triangular and in general the Hilbert space of bounded upper operators acting from $\ell_2^{\mathcal{M}}$ to $\ell_2^{\mathcal{N}}$ is denoted by $\mathcal{U}(\mathcal{M}, \mathcal{N})$ or denoted in short by $\mathcal{U}$. When the dimension N_k of the state vector is finite $\forall\ k$, then the operator represented as in (3) is *locally finite*. In the same way as $\mathcal{U}$, we denote the space of bounded operators by $\mathcal{X}(\mathcal{M}, \mathcal{N})$ and the space of bounded *lower* triangular operators by $\mathcal{L}(\mathcal{M}, \mathcal{N})$. A specific operator in $\mathcal{D}(\mathcal{M}, \mathcal{M})$ is the identity operator denoted by $I_{\mathcal{M}}$.

We also need the Hilbert–Schmidt space $\mathcal{U}_2$ and $\mathcal{D}_2$, which consists of those elements of $\mathcal{U}$ (respectively, $\mathcal{D}$) with square summable norms of their entries.

With the operator representation of P in (3), the closure of the output state space [20], denoted by $\overline{\mathcal{H}}_o(P)$, equals

$$\overline{\mathcal{H}}_o(P) = \mathcal{D}_2(\mathcal{B}, \mathcal{B})(I - AZ)^{-1}C \subset \mathcal{U}.$$

An operator $P_O \in \mathcal{U}$ is *left outer* if

$$\overline{\mathcal{U}_2 P_O} = \mathcal{U}_2,$$

and an operator $P_I \in \mathcal{U}$ is a *right isometry* if

$$P_I P_I^* = I,$$

where P_I^* denotes the adjoint of P_I.

Given a *causal* or upper operator P, the existence of an *outer-inner* factorization was first established by Arveson in [21]. Later, van der Veen [15], continuing in this work, devised a numerical procedure to calculate a realization of the inner and outer factor given a state realization of P.

Let us recall the existence theorem of the outer-inner factorization, given by Theorem 5 in [15].

THEOREM 2.1 (see [15, Theorem 5]). *Let $P \in \mathcal{U}(\mathcal{M},\mathcal{N})$. Then P has a factorization*

$$P = P_O P_I,$$

where $P_I \in \mathcal{U}(\mathcal{M}_I,\mathcal{N})$ is a right isometry, $P_O \in \mathcal{U}(\mathcal{M},\mathcal{M}_I)$ is outer, and $\mathcal{M}_I \subset \mathcal{M}$ (entrywise).

The key part of the algorithm to compute the inner-outer factorization is summarized in Proposition 6 of [15]. In the present paper, we state its dual variant, as this is required in solving the weighted sensitivity minimization problem and considers the outer-inner factorization. Due to their close relationship, the latter variant is stated without proof.

PROPOSITION 2.2 (Dual of Proposition 6 in [15]). *Let $P \in \mathcal{U}$ be a locally finite transfer operator, let $P = \{A, B, C, D\}$ be a controllable realization of P, and assume that the realization is asymptotically stable. Let P_I be a right isometric factor of P so that $P_O = PP_I^*$ is left outer. Then the pair (A_I, C_I) that defines an orthonormal basis for $\overline{\mathcal{H}_o}(P_I)$ satisfies*

$$\begin{array}{ll} \text{(i)} & Y = AY^{(-1)}A_I^* + CC_I^*, \\ \text{(ii)} & 0 = BY^{(-1)}A_I^* + DC_I^*, \\ \text{(iii)} & I = A_I A_I^* + C_I C_I^*, \\ \text{(iv)} & \ker(Y.) = \emptyset, \end{array}$$

where Y is some bounded diagonal operator and $Y^{(-1)} = ZYZ^{-1}$. Conversely, all solutions (A_I, C_I) of these equations give a basis representation of $\overline{\mathcal{H}_o}(P_I)$.

The computational scheme that can be derived from Proposition 2.2 (see section 6) returns a *uniformly observable* [23] and asymptotically stable pair (A_I, C_I).

Finally we end this section with the definition of the notion of uniform positivity of an Hermitian operator.

An Hermitian operator X is uniformly positive, denoted by $X \gg 0$, if

$$\exists\, \epsilon > 0 : \quad \| uXu^* \| > \epsilon \| uu^* \| \quad \forall\, u \in \ell_2\,.$$

If X is uniformly positive, then it is boundedly invertible.

3. The sensitivity minimization problem. Assume the feedback configuration in Figure 1(a). The plant P is assumed to be in $\mathcal{U}(\mathcal{M},\mathcal{N})$ and to be given by the following finite-dimensional LTV state space representation:

$$\mathbf{P} = \begin{bmatrix} A & C \\ B & D \end{bmatrix}, \qquad \begin{array}{ll} A \in \mathcal{D}(\mathcal{B},\mathcal{B}^{(-1)}), & C \in \mathcal{D}(\mathcal{B},\mathcal{N}), \\ B \in \mathcal{D}(\mathcal{M},\mathcal{B}^{(-1)}), & D \in \mathcal{D}(\mathcal{M},\mathcal{N}). \end{array}$$

The closed-loop system in Figure 1(a) is described by the systems of equations

$$\begin{cases} e_1 &= u_1 + e_2 K, \\ e_2 &= u_2 - e_1 P. \end{cases}$$

The overall closed-loop system is assumed to be *well posed*; that is, we can solve the above system of equations for the internal signals $[e_1,\ e_2]$ in terms of the input, respectively, disturbance signals $[u_1,\ u_2]$. This is indeed the case in which the operator $(I + PK)$ has a bounded inverse and then the map H from $[u_1,\ u_2]$ to $[e_1,\ e_2]$ is in $\mathcal{X}(\mathcal{M} \oplus \mathcal{N}, \mathcal{M} \oplus \mathcal{N})$ and given by

$$H = \begin{bmatrix} (I+PK)^{-1} & -P(I+KP)^{-1} \\ K(I+PK)^{-1} & (I+KP)^{-1} \end{bmatrix}.$$

An important notion of the closed-loop system is *internal stability*. This is defined next.

DEFINITION 3.1. *The system P in Figure* 1(a) *is internally stabilized by the controller K if and only if H is causal; that is, $H \in \mathcal{U}(\mathcal{M} \oplus \mathcal{N}, \mathcal{M} \oplus \mathcal{N})$.* ∎

The sensitivity map S is defined as the map from u_2 to e_2; that is,

$$S = (I + KP)^{-1}. \tag{4}$$

We are now in a position to state the sensitivity minimization problem.

Problem 1. For a given disturbance attenuation level $\Gamma \in \mathcal{D}$ with $\Gamma^*\Gamma > 0$ and an outer weighting map $W \in \mathcal{U}(\mathcal{N},\mathcal{N})$ $(W^{-1} \in \mathcal{U}(\mathcal{N},\mathcal{N}))$ with given state space realization

$$W = D_w + B_w Z(I - A_w Z)^{-1} C_w, \tag{5}$$

find a compensator K (if any exist) such that

$$\begin{aligned} &\text{(i)} \quad \text{the closed-loop system } H \text{ is internally stable,} \\ &\text{(ii)} \quad \|W S \Gamma^{-1}\| < 1, \end{aligned} \tag{6}$$

where $\|W S \Gamma^{-1}\|$ is the norm of $W S \Gamma^{-1}$ as a bounded operator on $\mathcal{N}$.

Remark 1. An extension of the above problem is treated in section 6. Here we do not assume the disturbance attenuation level to be specified but additionally consider the problem of determining the optimal (minimal) Γ for the special case $\Gamma = \gamma I$, where γ is a positive real constant.

4. An NPI problem. In this section, we formulate the sensitivity minimization Problem 1 as an NPI problem. For that purpose, we need the Youla parametrization of all internally stabilizing controllers in the present time-varying context. That this parametrization is valid in this context is stated in our first theorem. Since it is a special case of Theorem 11 of [6], we state it without proof.

THEOREM 4.1. *Let the plant $P \in \mathcal{U}(\mathcal{M},\mathcal{N})$; then a controller K internally stabilizes H if and only if there exists a $Q \in \mathcal{U}(\mathcal{N},\mathcal{M})$ such that*

$$K = (I - QP)^{-1} Q. \tag{7}$$

With the parametrization of K in (7), the sensitivity map S becomes

$$S = I - QP. \tag{8}$$

Suppose that in addition to the stability of P, P has an outer-inner factorization

$$P = P_O P_I. \tag{9}$$

Then we have

$$WS = W(I - QP)$$

from (8), and hence

$$W(I - S) = WQP \quad \in \mathcal{U}(\mathcal{N}, \mathcal{N}). \tag{10}$$

Using the expression for P in (9), we can state that when

$$W(I - S)P_I^* = WQP_O \quad \in \mathcal{U}(\mathcal{N}, \mathcal{M}_I), \tag{11}$$

we have found, since $W^{-1} \in \mathcal{U}$, a $Q \in \mathcal{U}$ that parametrizes an internally stabilizing controller K.

Now we first decompose the product WP_I^* into a component in $\mathcal{U}$ and one in $\mathcal{L}Z^{-1}$. This is done in our first lemma.

LEMMA 4.2. *Let the input-output map $W \in \mathcal{U}$ be given as in* (5)*, and let the input-output map $P_I \in \mathcal{U}(\mathcal{M}_I, \mathcal{N})$ be given as*

$$P_I = B_I Z(I - A_I Z)^{-1} C_I,$$

with A_I and A_w both asymptotically stable. Then

$$\exists! X_w, Y_w \in \mathcal{D} : WP_I^* = \underbrace{B_w Y_w^{(-1)} B_I^* + B_w Z(I - A_w Z)^{-1} X_w B_I^*}_{\in \mathcal{U}} + \underbrace{(D_w C_I^* + B_w Y_w^{(-1)} A_I^*)(I - Z^* A_I^*)^{-1} Z^* B_I^*}_{\in \mathcal{L}Z^{-1}},$$

with X_w and Y_w satisfying

$$C_w C_I^* = Y_w - A_w Y_w^{(-1)} A_I^*,$$
$$X_w = A_w Y_w^{(-1)}.$$

Proof. With the expressions for W and P_I in Lemma 4.2, WP_I^* becomes

$$WP_I^* = D_w C_I^* (I - Z^* A_I^*)^{-1} Z^* B_I^* + B_w Z \left[\; (I - A_w Z)^{-1} C_w C_I^* (I - Z^* A_I^*)^{-1} Z^* \; \right] B_I^*. \tag{12}$$

It is shown next that there exist operators $X_w, Y_w \in \mathcal{D}$ such that the term between square brackets can be decomposed as

$$\left[(I - A_w Z)^{-1} X_w + Y_w (I - Z^* A_I^*)^{-1} Z^* \right]. \tag{13}$$

To find these operators X_w and Y_w, multiply the left- and right-hand sides of both terms between square brackets by $(I - A_w Z)$ and $Z(I - Z^* A_I^*)$, respectively. This yields the following sequence of results:

$$\begin{aligned} C_w C_I^* &= X_w Z(I - Z^* A_I^*) + (I - A_w Z) Y_w \\ &= X_w Z - X_w A_I^* + Y_w - A_w Y_w^{(-1)} Z \\ &= (X_w - A_w Y_w^{(-1)}) Z + (Y_w - X_w A_I^*). \end{aligned}$$

Since X_w and Y_w are diagonal operators, we have that

$$X_w = A_w Y_w^{(-1)},$$
$$C_w C_I^* = Y_w - X_w A_I^*.$$

When we combine these expressions, we conclude that the Y_w operator satisfies

$$C_w C_I^* = Y_w - A_w Y_w^{(-1)} A_I^*.$$

This Stein equation has a unique and bounded solution since A_I (and A_w) are asymptotically stable. Therefore both X_w and Y_w are in $\mathcal{D}$. Replacing the term between square brackets in (12) with the one obtained in (13) yields

$$WP_I^* = D_w C_I^*(I - Z^* A_I^*)^{-1} Z^* B_I^* + B_w Z(I - A_w Z)^{-1} X B_I^* + B_w Z Y_w (I - Z^* A_I^*)^{-1} Z^* B_I^*.$$

Since the last term in this expression is equal to

$$B_w Y_w^{(-1)} B_I^* + B_w Y_w^{(-1)} A_I^* (I - Z^* A_I^*)^{-1} Z^* B_I^*,$$

the result of Lemma 4.2 is proved. □

With the result of Lemma 4.2, (11) can be written as

$$\begin{aligned} &\left(B_w Y_w^{(-1)} B_I^* + B_w Z(I - A_w Z)^{-1} X_w B_I^* \right) \\ &\quad + \left(D_w C_I^* + B_w Y_w^{(-1)} A_I^*) - (WS\Gamma^{-1})\Gamma C_I^* \right) (I - Z^* A_I^*)^{-1} Z^* B_I^* \in \mathcal{U}. \end{aligned}$$

Therefore, since the first term of the summation is in $\mathcal{U}$, (11) reduces to

$$\left((D_w C_I^* + B_w Y_w^{(-1)} A_I^*) - (WS\Gamma^{-1})\Gamma C_I^* \right) (I - Z^* A_I^*)^{-1} Z^* B_I^* \in \mathcal{U}.$$

Introduce the operator $J \in \mathcal{D}^2$ equal to $\begin{pmatrix} I_{\mathcal{N}} & \\ & -I_{\mathcal{N}} \end{pmatrix}$; then the Hermitian transpose of this last relationship can also be denoted as

$$B_I Z(I - A_I Z)^{-1} \begin{bmatrix} C_I \Gamma^* & C_I D_w^* + A_I Y_w^{*(-1)} B_w^* \end{bmatrix} J \begin{bmatrix} \Gamma^{-*} S^* W^* \\ I \end{bmatrix} \in \mathcal{L}.$$

The above exposure shows that the sensitivity minimization problem can be formulated as an NPI problem.

Problem 2. Find the weighted sensitivity map $\overline{S} = (WS\Gamma^{-1}) \in \mathcal{U}(\mathcal{N}, \mathcal{N})$ such that

$$\text{(14) (i)} \quad B_I Z(I - A_I Z)^{-1} \begin{bmatrix} C_I \Gamma^* & C_I D_w^* + A_I Y_w^{*(-1)} B_w^* \end{bmatrix} J \begin{bmatrix} \overline{S}^* \\ I \end{bmatrix} \in \mathcal{L}(\mathcal{M}_I, \mathcal{N}),$$

$$\text{(15) (ii)} \quad \|\overline{S}\| < 1.$$

Once the weighted sensitivity map $\overline{S}$ has been determined, the controller K directly follows from (4) and is given by

$$K = (\Gamma^{-1} \overline{S}^{-1} W - I) P^{-1}. \tag{16}$$

It should be remarked that this controller need not be causal.

5. Solving the NPI problem. NPI problems in the context of discrete time–varying systems have been studied and solved in [7]. That paper formulates and treats the NPI problem based on the so-called generalized "point evaluation map." In the present paper we take a different viewpoint. Instead we formulate the interpolation condition as in (14). As such we do not require the calculation of generalized points

as in [7], but we directly operate on the state space matrices of the given plant, its inner factor of an outer-inner factorization, and the given weighting. The calculation of these generalized points from the state space representation of the given plant P is still an unresolved problem. This alternative viewpoint leads to an immediate identification of the output state space of the J-isometric operator to be constructed in solving the NPI problem. This is highlighted in Lemma 5.2. A further generalization of the treatment given in [7] is that we allow the system dimensions to vary in time. Despite this more general context, a number of the key results derived in [7] simply carry over with minor modifications. As a consequence, we do not include proofs of the results requiring minor modifications and prove only the lemma's addressing the different viewpoint.

As outlined in [7], the solution of NPI problems requires the construction of a J-inner operator. In the present context of varying system dimension we focus on the construction of an operator $\Theta \in \mathcal{U}(\mathcal{M}_1 \oplus \mathcal{N}, \mathcal{N} \oplus \mathcal{N})$ which is J-isometric in the following generalized sense [19]:

$$\Theta J \Theta^* = J\prime = \begin{bmatrix} I_{\mathcal{M}_1} & \\ & -I_{\mathcal{N}} \end{bmatrix}, \tag{17}$$

where the dimension space sequence $\mathcal{M}_1$ still needs to be determined and which satisfies the interpolation condition

$$B_I Z (I - A_I Z)^{-1} \begin{bmatrix} C_I \Gamma^* & C_I D_w^* + A_I Y^{*(-1)} B_w^* \end{bmatrix} J\Theta^* \in \mathcal{L}(\mathcal{M}_I, \mathcal{M}_1 \oplus \mathcal{N}) \tag{18}$$

The following series of lemmas list a number of properties of J-isometric operators which are important in solving the NPI problem 2. In these lemmas we assume the operator Θ to be partitioned as

$$\Theta = \begin{bmatrix} \Theta_{11} & \Theta_{12} \\ \Theta_{21} & \Theta_{22} \end{bmatrix}, \qquad \begin{matrix} \Theta_{11} \in \mathcal{U}(\mathcal{M}_1, \mathcal{N}), & \Theta_{12} \in \mathcal{U}(\mathcal{M}_1, \mathcal{N}), \\ \Theta_{21} \in \mathcal{U}(\mathcal{N}, \mathcal{N}), & \Theta_{22} \in \mathcal{U}(\mathcal{N}, \mathcal{N}), \end{matrix}$$

and have the following state space realization:

$$\boldsymbol{\Theta} = \begin{bmatrix} A_\Theta & C_\Theta \\ B_\Theta & D_\Theta \end{bmatrix} = \begin{bmatrix} A_\Theta & C_1 & C_2 \\ B_1 & D_{11} & D_{12} \\ B_2 & D_{21} & D_{22} \end{bmatrix}. \tag{19}$$

As in [7], the state space realization $\boldsymbol{\Theta}$ of a J-isometric operator Θ satisfies the following condition.

LEMMA 5.1. *Let Θ be an operator in $\mathcal{U}^2$ with state space realization as given in* (19)*, and let the following relationship hold for some Hermitian operator $Q \in \mathcal{D}$:*

$$\boldsymbol{\Theta} \begin{bmatrix} Q^{(-1)} & \\ & J \end{bmatrix} \boldsymbol{\Theta}^* = \begin{bmatrix} Q & \\ & J\prime \end{bmatrix}; \tag{20}$$

then Θ is J-isometric in the generalized sense given in (17).

The identification of the output state space of the operator Θ by the interpolation condition (18) is stated in the following lemma.

LEMMA 5.2. *Let Θ be an operator in $\mathcal{U}^2$ with state space realization* (19)*, and let* (20) *hold for some Hermitian operator $Q \in \mathcal{D}$; then*

$$\mathcal{D}(I - A_\Theta Z)^{-1} C_\Theta J \Theta^* \in \mathcal{L}^2 Z^{-1}.$$

Proof. Since $\Theta \in \mathcal{U}^2$, its state space realization given in (19) allows us to express Θ as

$$\Theta = D_\Theta + B_\Theta Z(I - A_\Theta Z)^{-1} C_\Theta.$$

Hence,

$$(I - A_\Theta Z)^{-1} C_\Theta J \Theta^* = (I - A_\Theta Z)^{-1} C_\Theta J (D_\Theta^* + C_\Theta^* (I - Z^* A_\Theta^*)^{-1} Z^* B_\Theta^*).$$

Using (20), the right-hand side equals

$$\begin{aligned}
&= (I - A_\Theta Z)^{-1}(-A_\Theta Q^{(-1)} + (Q - A_\Theta Q^{(-1)} A_\Theta^*)(I - Z^* A_\Theta^*)^{-1} Z^*) B_\Theta^* \\
&= (I - A_\Theta Z)^{-1}(-A_\Theta Q^{(-1)} Z(I - Z^* A_\Theta^*) + Q - A_\Theta Q^{(-1)} A_\Theta^*)(I - Z^* A_\Theta^*)^{-1} Z^*) B_\Theta^* \\
&= (I - A_\Theta Z)^{-1}(-A_\Theta Q^{(-1)} Z + A_\Theta Q^{(-1)} A_\Theta^* + Q - A_\Theta Q^{(-1)} A_\Theta^*)(I - Z^* A_\Theta^*)^{-1} Z^* B_\Theta^* \\
&= (I - A_\Theta Z)^{-1}(-A_\Theta Z + I) Q (I - Z^* A_\Theta^*)^{-1} Z^* B_\Theta^* \\
&= Q(I - Z^* A_\Theta^*)^{-1} Z^* B_\Theta^* \quad \in \mathcal{L}^2 Z^{-1}. \qquad \square
\end{aligned}$$

The key condition on the J-isometric operator solving the NPI Problem 2 is stipulated in Lemma 5.5. Prior to stating that lemma, we have the following definition and lemma.

DEFINITION 5.3. *The pair* (A, C) *is uniformly detectable if and only if there exists a bounded operator* $K \in \mathcal{D}(\mathcal{N}, \mathcal{B}^{(-1)})$ *such that the state space model* $xZ^{-1} = x(A + CK)$ *is asymptotically stable.*

LEMMA 5.4. *Suppose the pair* (A, C) *is uniformly detectable. Then if there exists a solution* $X \in \mathcal{D}(\mathcal{B}, \mathcal{B})$ *and* $X \geq 0$ *of*

$$AX^{(-1)}A^* + CC^* = X, \tag{21}$$

then the state space model $xZ^{-1} = xA$ *is asymptotically stable. Conversely, if the state space model* $xZ^{-1} = xA$ *is asymptotically stable, then there exists a unique bounded solution* $X \geq 0$ *of* (21).

LEMMA 5.5. *Let* Θ *be an operator in* $\mathcal{U}^2$ *with state space realization* (19), *let* (20) *hold for a uniformly positive operator* $Q \in \mathcal{D}$, *and let the pair* (A_Θ, C_2) *be uniformly detectable; then*

(i) $\Theta_{22}^{-1} \in \mathcal{U}(\mathcal{N}, \mathcal{N})$,
(ii) $\Theta_{22}^{-1}\Theta_{21} \in \mathcal{U}(\mathcal{N}, \mathcal{N})$ *and* $\|\Theta_{22}^{-1}\Theta_{21}\| < 1$.

Proof. (i) From the state space realization of Θ given in the right-hand side in (19), the state space realization of Θ_{22} is

$$\Theta_{22} = \begin{bmatrix} A_\Theta & C_2 \\ B_2 & D_{22} \end{bmatrix}.$$

When the conditions

(a) D_{22} is invertible,

(b) the operator $(A_\Theta - C_2 D_{22}^{-1} B_2)$ is asymptotically stable

hold, then since the state space realization of Θ_{22}^{-1} equals

$$\begin{bmatrix} A_\Theta - C_2 D_{22}^{-1} B_2 & -C_2 D_{22}^{-1} \\ D_{22}^{-1} B_2 & D_{22}^{-1} \end{bmatrix},$$

we have that item (i) of the lemma is proved. We now prove both conditions (a) and (b), making use of the following relationships derived from (20):

$$A_\Theta Q^{(-1)} A_\Theta^* + C_1 C_1^* - C_2 C_2^* = Q, \tag{22}$$

$$B_2 Q^{(-1)} A_\Theta^* + D_{21} C_1^* - D_{22} C_2^* = 0, \tag{23}$$

$$B_2 Q^{(-1)} B_2^* + D_{21} D_{21}^* - D_{22} D_{22}^* = -I. \tag{24}$$

For condition (a), (24) shows that

$$D_{22} D_{22}^* = I + D_{21} D_{21}^* + B_2 Q^{(-1)} B_2^*.$$

Since $Q \gg 0$ it follows that $D_{22} D_{22}^* \gg 0$ and D_{22} is invertible.

For condition (b), consider the product of operators

$$(A_\Theta - C_2 D_{22}^{-1} B_2) Q^{(-1)} (A_\Theta - C_2 D_{22}^{-1} B_2)^*.$$

Using (23)–(24), this product can be written as

$$\begin{aligned} &= A_\Theta Q^{(-1)} A_\Theta^* - C_2 C_2^* \\ &\quad - \left[C_2 D_{22}^{-1} D_{22}^{-*} C_2^* + C_2 D_{22}^{-1} D_{21} D_{21}^* D_{22}^{-*} C_2^* - C_2 D_{22}^{-1} D_{21} C_1^* - C_1 D_{21}^* D_{22}^{-*} C_2^* \right]. \end{aligned}$$

And, therefore, using (22) we obtain the following Lyapunov equation:

$$\begin{aligned} &(A_\Theta - C_2 D_{22}^{-1} B_2) Q^{(-1)} (A_\Theta - C_2 D_{22}^{-1} B_2)^* + (C_1 - C_2 D_{22}^{-1} D_{21})(C_1 - C_2 D_{22}^{-1} D_{21})^* \\ &\qquad + C_2 D_{22}^{-1} D_{22}^{-*} C_2^* = Q. \end{aligned} \tag{25}$$

Since the pair $(A_\Theta,\ C_2)$ is uniformly detectable, also the pair $(A_\Theta - C_2 D_{22}^{-1} B_2,\ \left[C_1 - C_2 D_{22}^{-1} D_{21} \quad C_2 D_{22}^{-1}\right])$ is uniformly detectable. Therefore, since $Q \gg 0$, equation (25) and Lemma 5.4 show that condition (b) holds and therefore item (i) is proved.

(ii) Lemma 5.1 shows that

$$\Theta J \Theta^* = J\prime$$

and, therefore,

$$\Theta_{21} \Theta_{21}^* - \Theta_{22} \Theta_{22}^* = -I.$$

This shows that

$$\Theta_{22} \Theta_{22}^* = I + \Theta_{21} \Theta_{21}^* \gg 0;$$

since Θ_{22}^{-1} exists this last equation shows that

$$I - \Theta_{22}^{-1} \Theta_{22}^{-*} = \Theta_{22}^{-1} \Theta_{21} \Theta_{21}^* \Theta_{22}^{-*}$$

and therefore $\|\Theta_{22}^{-1} \Theta_{21}\| < 1$. Since $\Theta_{22}^{-1} \in \mathcal{U}$ and $\Theta_{21} \in \mathcal{U}$, it follows that $\Theta_{22}^{-1} \Theta_{21} \in \mathcal{U}$. □

Lemmas 5.1, 5.2, and 5.5 are used in the next section to calculate a state space realization of Θ that solves the interpolation condition (18).

In the remaining part of this section, we assume that such a Θ has been calculated and demonstrate how this Θ allows us to parametrize, as in [7], all solutions to the NPI Problem 2.

The first step in the parametrization is to observe that for $G_1 \in \mathcal{U}(\mathcal{N}, \mathcal{M}_1)$ and $G_2 \in \mathcal{U}(\mathcal{N}, \mathcal{N})$, equation (18) shows that

$$B_I Z(I - A_I Z)^{-1} \begin{bmatrix} C_I \Gamma^* & C_I D_w^* + A_I Y^{*(-1)} B_w^* \end{bmatrix} J \begin{bmatrix} \Theta_{11}^* & \Theta_{21}^* \\ \Theta_{12}^* & \Theta_{22}^* \end{bmatrix} \begin{bmatrix} G_1^* \\ G_2^* \end{bmatrix} \in \mathcal{L},$$

and hence the solution to the interpolation condition (14) requires G_1 and G_2 to be selected such that

$$\begin{bmatrix} \overline{S} & I \end{bmatrix} = \begin{bmatrix} G_1 & G_2 \end{bmatrix} \begin{bmatrix} \Theta_{11} & \Theta_{12} \\ \Theta_{21} & \Theta_{22} \end{bmatrix}. \tag{26}$$

From this relationship we parametrize all solutions $\overline{S}$ to Problem 2 as shown in the following two theorems. These theorems correspond to Theorems 3.2 and 3.3 of [7], adapted to the present context of nonconstant state, input and output dimensions. Since only minor adaptations are required, we state both modifications without proof.

THEOREM 5.6. *Let Θ be an operator in $\mathcal{U}^2$ satisfying the interpolation condition* (18). *Then $\overline{S}$ satisfies the interpolation condition* (14) *if and only if $\overline{S}$ has a representation of the form*

$$\overline{S} = (G_1\Theta_{12} + G_2\Theta_{22})^{-1}(G_1\Theta_{11} + G_2\Theta_{21})$$

for some pair of upper triangular operators $G_1 \in \mathcal{U}$ and $G_2 \in \mathcal{U}$ such that $(G_1\Theta_{12} + G_2\Theta_{22})$ is invertible with inverse again in $\mathcal{U}$.

THEOREM 5.7. *Let Θ be an operator in $\mathcal{U}^2$ satisfying the interpolation condition* (18). *Let Θ be partitioned as*

$$\Theta = \begin{bmatrix} \Theta_{11} & \Theta_{12} \\ \Theta_{21} & \Theta_{22} \end{bmatrix}$$

and have a state space realization as in (19) *such that* (20) *is satisfied for $Q \gg 0$. Then*

(1) *there exist solutions $\overline{S} \in \mathcal{U}$ to the NPI Problem* 2*;*

(2) *any solution $\overline{S}$ is given by*

$$\overline{S} = (S_L\Theta_{12} + \Theta_{22})^{-1}(S_L\Theta_{11} + \Theta_{21})$$

with $S_L \in \mathcal{U}(\mathcal{N}, \mathcal{M}_1)$ satisfying $\|S_L\| < 1$.

6. Calculation of a state space realization of Θ and a particular $\overline{S}$. Lemma 5.2 shows that the interpolation condition (18) fixes the pair $(A_\Theta,\ C_\Theta)$ of the state space realization of Θ, namely,

$$\begin{bmatrix} A_\Theta & C_\Theta \end{bmatrix} = \begin{bmatrix} A_I & C_I\Gamma^* & C_I D_w^* + A_I Y^{*(-1)} B_w^* \end{bmatrix}. \tag{27}$$

Computing the remaining part $(B_\Theta,\ D_\Theta)$ of the state space representation of Θ is done in such a way that (20) is satisfied for a uniformly positive diagonal operator Q. With the specification of $(A_\Theta,\ C_\Theta)$ as in (27), the relationships to be satisfied then are

$$\begin{bmatrix} A_I & C_I\Gamma^* & G \\ B_1 & D_{11} & D_{12} \\ B_2 & D_{21} & D_{22} \end{bmatrix} \begin{bmatrix} Q^{(-1)} & & \\ & I_{\mathcal{N}} & \\ & & -I_{\mathcal{N}} \end{bmatrix} \begin{bmatrix} A_I^* & B_1^* & B_2^* \\ \Gamma C_I^* & D_{11}^* & D_{21}^* \\ G^* & D_{12}^* & D_{22}^* \end{bmatrix} = \begin{bmatrix} Q & & \\ & J_{\mathcal{I}} & \end{bmatrix}, \tag{28}$$

where G equals $C_I D_w^* + A_I Y^{*(-1)} B_w^*$.

The relationships to compute $\boldsymbol{\Theta}$, which can directly be derived from (28), can be divided in two parts. One part is the following Lyapunov equation:

$$A_I Q^{(-1)} A_I^* + C_I \Gamma^* \Gamma C_I^* - GG^* = Q. \tag{29}$$

From this equation the operator Q can be calculated. The other part is the set of equations

$$\begin{bmatrix} B_1 & D_{11} & D_{12} \\ B_2 & D_{21} & D_{22} \end{bmatrix} \begin{bmatrix} Q^{(-1)} & & \\ & I_{\mathcal{N}} & \\ & & -I_{\mathcal{N}} \end{bmatrix} \begin{bmatrix} A_I^* \\ \Gamma C_I^* \\ G^* \end{bmatrix} = 0, \tag{30}$$

$$\begin{bmatrix} B_1 & D_{11} & D_{12} \\ B_2 & D_{21} & D_{22} \end{bmatrix} \begin{bmatrix} Q^{(-1)} & & \\ & I_{\mathcal{N}} & \\ & & -I_{\mathcal{N}} \end{bmatrix} \begin{bmatrix} B_1^* & B_2^* \\ D_{11}^* & D_{21}^* \\ D_{12}^* & D_{22}^* \end{bmatrix} = J\imath, \tag{31}$$

from which we determine the system matrices $\begin{pmatrix} B_1 & D_{11} & D_{12} \\ B_2 & D_{21} & D_{22} \end{pmatrix}$, given the solution Q to (29).

6.1. Selecting the minimal scalar γ when $\boldsymbol{\Gamma = \gamma I}$ in solving for the operator $\boldsymbol{Q}$. Consider the following two Lyapunov equations:

$$A_I Q_1^{(-1)} A_I^* + \gamma^2 C_I C_I^* = Q_1, \tag{32}$$

$$A_I Q_2^{(-1)} A_I^* - GG^* = Q_2. \tag{33}$$

Since A_I is asymptotically stable both equations have a unique solution. In addition $Q_1 \gg 0$, since the pair $(A_I \;\; C_I)$ is uniformly detectable. We now consider the scaled Lyapunov equations

$$A_I \overline{Q}_1^{(-1)} A_I^* + C_I C_I^* = \overline{Q}_1,$$

$$A_I \overline{Q}_2^{(-1)} A_I^* + GG^* = \overline{Q}_2.$$

And the relation between the solution of both pairs of Lyapunov equations is

$$Q_1 = \gamma^2 \overline{Q}_1, \quad Q_2 = -\overline{Q}_2.$$

In this way $\overline{Q}_2$ is semipositive definite. Furthermore, let $\overline{Q}_1 = S_1 S_1^*$ and $\overline{Q}_2 = S_2 S_2^*$; then we can easily construct the following square root update mechanisms for the above Lyapunov equations:

$$\begin{bmatrix} A_I S_1^{(-1)} & C_I \end{bmatrix} T_1 = \begin{bmatrix} S_1 & 0 \end{bmatrix},$$

$$\begin{bmatrix} A_I S_2^{(-1)} & G \end{bmatrix} T_2 = \begin{bmatrix} S_2 & 0 \end{bmatrix},$$

with both T_1 and T_2 diagonal unitary operators satisfying $T_1 T_1^* = T_2 T_2^* = I$. The actual solution to the original Lyapunov equation (29) can be derived from these two square roots and reads

$$Q = \left(\gamma^2 S_1 S_1^* - S_2 S_2^*\right).$$

Since $Q_1 \gg 0$, S_1 is invertible and we have that

$$Q = S_1 \left(\gamma^2 I - S_1^{-1} S_2 S_2^* S_1^{-*}\right) S_1^*. \tag{34}$$

Therefore, let $\mathbf{sv}(M)$ denote the singular values of the operator M; then

$$Q \gg 0 \quad \Leftrightarrow \gamma > \min \ \mathbf{sv}(S_1^{-1} S_2),$$

and we see that we can compute the minimal γ directly from the square roots S_1 and S_2. We also observe that as in the time-invariant case approaching the optimal γ leads to a singular solution of the Lyapunov equation. A close approximation of the minimal scalar disturbance attenuation level γ may be taken as

$$\widetilde{\gamma_{opt}} = \min \ \mathbf{sv}(S_1^{-1} S_2) + \epsilon,$$

where ϵ is a small positive real number.

Remark 2. In the present time-varying context, there is no need to fix Γ to γI. The diagonal operator Γ may have varying entries along its diagonal. However, determining such an optimal time-varying disturbance attenuation level is a very complex problem even when we assume Γ_k to be equal to $\gamma_k I$, for γ_k scalar. This is because a particular choice of such a γ_k at a local time instance k has a global effect on the solution (32), with γ replaced by γ_k. In addition this global effect of γ_k on $Q_{1,j}$ for $j \leq k$ is implicit rather than explicit. □

6.2. Completing the state space realization of Θ. We demonstrate in this section that the solution presented in [7] to complete the state space realization of $\boldsymbol{\Theta}$ using the solution of the Lyapunov equation (29) still holds when the system dimensions are time variant. In addition we present a square root solution. The solution is discussed on a local time scale.

Let $A_{I,k} \in \mathbb{R}^{N_k \times N_{k+1}}$ and $C_{I,k} \in \mathbb{R}^{N_k \times L_k}$; then the Lyapunov equation (29) with $Q_k > 0$ shows that [12]

$$N_k \leq N_{k+1} + L_k.$$

Furthermore, equation (29) shows that the rows of the matrix $\begin{bmatrix} A_{I,k} & \gamma_k C_{I,k} & G_k \end{bmatrix} \in \mathbb{R}^{N_k \times (N_{k+1}+2L_k)}$ are independent, and so are the columns of the matrix

$$\begin{bmatrix} Q_{k+1} & & \\ & I_{L_k} & \\ & & -I_{L_k} \end{bmatrix} \begin{bmatrix} A_{I,k}^* \\ \gamma_k C_{I,k}^* \\ G_k^* \end{bmatrix}.$$

Therefore, we can always find a matrix $V_k \in \mathbb{R}^{(N_{k+1}-N_k+2L_k)\times(N_{k+1}+2L_k)}$ with independent rows such that

$$V_k \begin{bmatrix} Q_{k+1} & & \\ & I_{L_k} & \\ & & -I_{L_k} \end{bmatrix} \begin{bmatrix} A_{I,k}^* \\ \gamma_k C_{I,k}^* \\ G_k^* \end{bmatrix} = 0.$$

The compound matrix $\begin{bmatrix} A_{I,k} & \gamma_k C_{I,k} & G_k \\ & V_k & \end{bmatrix}$ is square and of full rank, as is shown in the following theorem.

THEOREM 6.1. *Let* $V_k \in \mathbb{R}^{(N_{k+1}-N_k+2L_k)\times(N_{k+1}+2L_k)}$ *have independent rows, and let the following relationship hold:*

$$\underbrace{\begin{bmatrix} A_{I,k} & \gamma_k C_{I,k} & G_k \\ & V_k & \end{bmatrix}}\begin{bmatrix} Q_{k+1} & & \\ & I_{L_k} & \\ & & -I_{L_k} \end{bmatrix}\begin{bmatrix} A_{I,k}^* & & \\ \gamma_k C_{I,k}^* & & V_k^* \\ G_k^* & & \end{bmatrix} = \begin{bmatrix} Q_k & 0 \\ 0 & J'' \end{bmatrix}, \tag{35}$$

with both Q_k *and* Q_{k+1} *positive definite, then the underbraced matrix has full rank.*

Proof. The underbraced matrix is square since $V_k \in \mathbb{R}^{(N_{k+1}-N_k+2L_k)\times(N_{k+1}+2L_k)}$ and the matrix $\begin{bmatrix} A_{I,k} & \gamma_k C_{I,k} & G_k \end{bmatrix} \in \mathbb{R}^{N_k\times(N_{k+1}+2L_k)}$. Now suppose that this matrix is not invertible; i.e.,

$$\exists \begin{bmatrix} x_1 & x_2 \end{bmatrix} \neq 0 : \begin{bmatrix} x_1 & x_2 \end{bmatrix}\begin{bmatrix} A_{I,k} & \gamma_k C_{I,k} & G_k \\ & V_k & \end{bmatrix} = 0.$$

However, from (35) we derive that

$$\begin{bmatrix} x_1 & x_2 \end{bmatrix}\begin{bmatrix} A_{I,k} & \gamma_k C_{I,k} & G_k \\ & V_k & \end{bmatrix}\begin{bmatrix} Q_{k+1} & & \\ & I_{L_k} & \\ & & -I_{L_k} \end{bmatrix}\begin{bmatrix} A_{I,k}^* & & \\ \gamma_k C_{I,k}^* & & \\ G_k^* & & \end{bmatrix} = x_1 Q_k.$$

Since Q_k is nonsingular, $x_1 = 0$ and, as a consequence since V_k has independent rows, $x_2 = 0$. This is a contradiction and the theorem is proved. □

This theorem allows us to determine the inertia of the matrix J''. Let $\#_+(M_k)$ denote the number of $+$'s in the inertia of M_k, and similarly let $\#_-(M_k)$ denote the number of $-$'s; then (35) and Theorem 6.1 show that

$$\begin{aligned} N_{k+1} + L_k &= N_k + \#_+(J''), \\ L_k &= \#_-(J''). \end{aligned} \tag{36}$$

Let $\#_+(J'') = \alpha_k$; then we complete the calculation of the state space realization of Θ by computing the following factorization:

$$V_k \begin{bmatrix} Q_{k+1} & & \\ & I_{L_k} & \\ & & -I_{L_k} \end{bmatrix} V_k^* = J'' = U_k \begin{bmatrix} I_{\alpha_k} & \\ & -I_{L_k} \end{bmatrix} U_k^*. \tag{37}$$

Such a factorization with U_k invertible always exists. As a consequence the pair $\begin{bmatrix} B_\Theta & D_\Theta \end{bmatrix}$ now follows as

$$\begin{bmatrix} B_\Theta & D_\Theta \end{bmatrix} = U_k^{-1} V_k. \tag{38}$$

We end this section by presenting a square root variant to compute U_k directly from the square roots $S_{1,k}$ and $S_{2,k}$.

First note that with an additional singular value decomposition (SVD) of the matrix $S_{1,k+1}^{-1} S_{2,k+1}$ given as

$$S_{1,k+1}^{-1} S_{2,k+1} = U_{S,k+1} \Sigma_{S,k+1} V_{S,k+1}^*,$$

we can write (34) on the time instance $k+1$ as

$$Q_{k+1} = S_{1,k+1} U_{S,k+1} (\gamma^2 I - \Sigma_{S,k+1}^2)^{\frac{1}{2}} (\gamma^2 I - \Sigma_{S,k+1}^2)^{\frac{*}{2}} U_{S,k+1}^* S_{1,k+1}^*.$$

Let us denote the square root of Q_{k+1} by $S_{Q,k+1}$; then we can read off this quantity from the above expression.

We partition the matrix V_k as $\begin{bmatrix} V_{1,k} & V_{2,k} \end{bmatrix}$ with $V_{1,k} \in \mathbb{R}^{(N_{k+1}-N_k+2L_k)\times(N_{k+1}+L_k)}$ and the matrix U_k as $\begin{bmatrix} U_{1,k} & U_{2,k} \end{bmatrix}$ with $U_{1,k} \in \mathbb{R}^{(N_{k+1}-N_k+2L_k)\times\alpha_k}$ in addition to defining $V'_{1,k}$ as $V_{1,K}\begin{bmatrix} S_{Q,k+1} & \\ & I_{L_k} \end{bmatrix}$; then we can alternatively denote (37) as

$$V'_{1,k}V'^{*}_{1,k} - V_{2,k}V^{*}_{2,k} = U_{1,k}U^{*}_{1,k} - U_{2,k}U^{*}_{2,k}. \tag{39}$$

Finally, consider the generalized SVD [22] of the matrix pair $\left(V'_{1,k}, V_{2,k}\right)$, and denote this decomposition as

$$V'_{1,k} = X_k\Sigma_{V_1,k}U^{*}_{V_1,k}, \qquad V_{2,k} = X_k\Sigma_{V_2,k}U^{*}_{V_2,k},$$

where X_k is a square nonsingular $(L_k+\alpha_k)\times(L_k+\alpha_k)$ matrix, $\Sigma_{V_j,k}$ for $j=1,2$ is a diagonal matrix with positive or zero entries, and $U_{V_j,k}$ for $j=1,2$ are orthogonal square matrices. Based on this decomposition the left-hand side of (39) can be written as

$$V'_{1,k}V'^{*}_{1,k} - V_{2,k}V^{*}_{2,k} = X_k\left(\Sigma^2_{V_1,k} - \Sigma^2_{V_2,k}\right)X^{*}_k.$$

From this equation we can read off the required factors $U_{1,k}$ and $U_{2,k}$.

6.3. The state space realization for a particular $\overline{S}$. From Theorem 5.7 we recall that a particular solution $\overline{S}$ to the NPI Problem 2 is

$$\overline{S} = \Theta_{22}^{-1}\Theta_{21}.$$

With the state space representation for Θ calculated in the previous subsection, the latter two operators are determined as follows:

$$\begin{aligned}\Theta_{21} &= D_{21} + B_2Z(I - A_IZ)^{-1}C_I\Gamma^{*},\\ \Theta_{22} &= D_{22} + B_2Z(I - A_IZ)^{-1}G.\end{aligned}$$

Hence, because of Lemma 5.5,

$$\Theta_{22}^{-1} = D_{22}^{-1} - D_{22}^{-1}B_2Z(I - (A_I - GD_{22}^{-1}B_2)Z)^{-1}GD_{22}^{-1} \quad \text{exists and is} \quad \in \mathcal{U},$$

and the product $\Theta_{22}^{-1}\Theta_{21}(\in\mathcal{U})$ has the state following state space representation:

$$\begin{bmatrix} (A_I - GD_{22}^{-1}B_2) & -GD_{22}^{-1}B_2 & -GD_{22}^{-1}D_{21} \\ 0 & A_I & C_I\Gamma^{*} \\ D_{22}^{-1} & D_{22}^{-1}B_2 & D_{22}^{-1}D_{21} \end{bmatrix}.$$

Consider the following constant similarity transformation:

$$\begin{bmatrix} I & I & 0 \\ 0 & I & 0 \\ 0 & 0 & I \end{bmatrix}\begin{bmatrix} (A_I - GD_{22}^{-1}B_2) & -GD_{22}^{-1}B_2 & -GD_{22}^{-1}D_{21} \\ 0 & A_I & C_I\Gamma^{*} \\ D_{22}^{-1}B_2 & D_{22}^{-1}B_2 & D_{22}^{-1}D_{21} \end{bmatrix}\begin{bmatrix} I & -I & 0 \\ 0 & I & 0 \\ 0 & 0 & I \end{bmatrix}$$

$$= \begin{bmatrix} (A_I - GD_{22}^{-1}B_2) & 0 & C_I\Gamma^{*} - GD_{22}^{-1}D_{21} \\ 0 & A_I & C_I\Gamma^{*} \\ D_{22}^{-1}B_2 & 0 & D_{22}^{-1}D_{21} \end{bmatrix};$$

then $\Theta_{22}^{-1}\Theta_{21}$ is given as

$$\Theta_{22}^{-1}\Theta_{21} = D_{22}^{-1}D_{21} + D_{22}^{-1}B_2(I - (A_I - GD_{22}^{-1}B_2)Z)^{-1}(C_I\Gamma^{*} - GD_{22}^{-1}D_{21}). \tag{40}$$

6.4. Summary of the computational scheme. The analysis made in the previous sections can readily be summarized in a computational scheme to solve the weighted sensitivity minimization problem. In this section, we restrict ourselves to summarizing the key numerical equations. These are the following recursive equations, which are all of Lyapunov type running backwards in time:

(a) The computation of the part of the outer-inner factorization of the given plant P necessary in solving the sensitivity minimization problem is highlighted in Proposition 2.2. The equations in this proposition are reordered such that the sequential calculations occur from top to bottom. This results on a local time scale in the following equations:

$$\begin{array}{lcl} \left[\begin{array}{cc} B_k Y_{k+1} & D_k \end{array}\right] \left[\begin{array}{c} A^*_{I,k} \\ C^*_{I,k} \end{array}\right] & = & 0, \\ A_{I,k}A^*_{I,k} + C_{I,k}C^*_{I,k} & = & I, \\ A_k Y_{k+1} A^*_{I,k} + C_k C^*_{I,k} & = & Y_k. \end{array}$$

(b) In formulating the interpolation condition (14) for the weighted sensitivity minimization problem, the quantity $Y_{w,k}$ is required. According to Lemma 4.2, $Y_{w,k}$ satisfies

$$C_{w,k}C^*_{I,k} + A_{w,k}Y_{w,k+1}A^*_{I,k} \quad = \quad Y_{w,k}.$$

(c) Finally, the key equation in solving the NPI problem is the Lyapunov equation (29). This equation on a local time scale reads

$$A_{I,k}Q_{k+1}A^*_{I,k} + C_{I,k}\Gamma^*_k\Gamma_k C^*_{I,k} - G_k G^*_k \quad = \quad Q_k.$$

All other quantities that are computed in the course of solving the sensitivity minimization problem satisfy local expressions, making use of the solutions of the above recursive equations. For that reason, these three groups of recursive equations constitute the heart of the computational procedure. The actual solution of these equations requires the specification of end conditions. These end conditions can be calculated with a time-invariant solution to the sensitivity minimization problem for the case the plant becomes time-invariant from a particular time-instant on.

This situation occurs, e.g., in practical engineering problems when changing the stationary operation condition of industrial plants. Generally, in these operation conditions the system accurately behaves as a linear time-invariant system. Furthermore, the transition between operation conditions can often be approximated by a linear time-varying system.

7. Concluding remarks. The solution presented in this paper to the weighted sensitivity minimization problem is characterized by recursive calculations involving the system matrices of the given state space model of the plant to be controlled. The latter is assumed to be controllable and stable.

The key recursive equations are Lyapunov equations which run backward in time. For the special case the system becomes time-invariant from a specific point in time; it is possible to find initial conditions for both Lyapunov recursions making use of standard solutions for time-invariant systems.

The solution presented in this paper is computationally more efficient and elegant compared with both the lifting approach presented in [3] when the linear system is periodic as well as the time-varying solution in [4], [5]. Furthermore, the approach

outlined in this paper is capable of treating varying input and output dimensions. As a consequence the important class of problems of multirate sampled data systems also fits within the constructed framework.

Continued research on the extension of the present approach to treat unstable plants is currently under way. Here we will also consider the problem of making the disturbance attenuation level time-varying.

Acknowledgment. The author acknowledges Prof. H. Dym of the Weizmann Institute for fruitful discussions related to Theorem 6.1.

REFERENCES

[1] G. Zames, *Feedback and optimal sensitivity*, IEEE Trans. Automat. Control, 26 (1981), pp. 310–320.

[2] A. Feintuch, P. P. Khargonekar, and A. Tannenbaum, *On the sensitivity minimization problem for linear time-varying systems*, SIAM J. Control Optim., 24 (1986), pp. 1076–1085.

[3] T. T. Georgiou and P. P. Khargonekar, *A constructive algorithm for sensitivity optimization of periodic systems*, SIAM J. Control Optim., 25 (1987), pp. 334–340.

[4] R. Ravi, K. M. Nagpal, and P. P. Khargonekar, *H_∞ control of linear time-varying systems: A state space approach*, SIAM J. Control Optim., 29 (1991), pp. 1394–1413.

[5] M. Verhaegen and A. J. van der Veen, *The bounded real lemma for discrete time–varying systems with application to robust output feedback*, in Proc. 32nd Conference on Decision and Control, San Antonio, TX, 1993, pp. 45–50.

[6] W. N. Dale and M. C. Smith, *Stabilizability and existence of system representations for discrete time–varying systems*, in Proc. American Control Conference, 1991, pp. 2737–2742.

[7] J. A. Ball, I. Gohberg, and M. A. Kaashoek, *Nevanlinna Pick interpolation for linear time-varying input-output maps: The discrete case*, in Operator Theory: Advances and Applications, vol. OT 56, Birkhäuser-Verlag, Basel, Switzerland, 1992, pp. 1–51.

[8] J. A. Ball, I. Gohberg, and M. A. Kaashoek, *Time-varying systems: Nevanlinna-Pick interpolation and sensitivity minimization*, in Recent Advances in Mathematical Theory of Systems, Control, Networks and Signal Processing, vol. I, Mita Press, Tokyo, 1992, pp. 53–58.

[9] J. C. Doyle, K. Glover, P. P. Khargonekar, and B. A. Francis, *State-space solutions to standard H_2 and H_∞ control problems*, IEEE Trans. Automat. Control, 34 (1989), pp. 831–847.

[10] B. A. Francis, J. W. Helton, and G. Zames, *H^∞-optimal feedback controllers for linear multivariable systems*, IEEE Trans. Automat. Control, 29 (1984), pp. 888–900.

[11] M. Verhaegen and P. M. Dewilde, *Calculating the anti-causal part of the inverse of a causal, time-varying discrete system in the framework of sensitivity minimization*, in Challenges of a Generalized System Theory, Essays of the Royal Dutch Academy of Sciences, P. Dewilde, M. A. Kaashoek, and M. Verhaegen., eds., North-Holland, Amsterdam, The Netherlands, 1993, pp. 119–136.

[12] J. Kos, *Higher order time-varying Nevanlinna-Pick interpolation*, in Challenges of a Generalized System Theory, Essays of the Royal Dutch Academy of Sciences, P. Dewilde, M. A. Kaashoek, and M. Verhaegen., eds., North-Holland, Amsterdam, The Netherlands, 1993.

[13] Y. Xiaode and M. Verhaegen, *Application of a time-varying subspace model identification scheme to the identification of the human joint dynamics*, in European Control Conference 1993, Groningen, The Netherlands, 1993, pp. 603–608.

[14] M. Verhaegen and P. Van Dooren, *Numerical aspects of different Kalman filter implementations*, IEEE Trans. on Autom. Control, 31 (1986), pp. 907–917.

[15] A. J. van der Veen, *Computation of the inner-outer factorization for time-varying systems*, in Challenges of a Generalized System Theory, Essays of the Royal Dutch Academy of Sciences, P. Dewilde, M. A. Kaashoek, and M. Verhaegen, eds., North-Holland, Amsterdam, The Netherlands, 1993.

[16] P. Dewilde and H. Dym, *Interpolation for upper triangular operators*, in Operator Theory: Advances and Applications, vol. OT 56, I. Gohberg, ed., Birkhaüser-Verlag, Basel, Switzerland, 1992, pp. 153–260.

[17] A. J. van der Veen and P. M. Dewilde, *Time-varying system theory for computational networks*, in Algorithms and Parallel VLSI Architectures, Vol. II, P. Quinton and Y. Robert, eds., Elsevier, New York, 1991, pp. 103–127.
[18] A. J. van der Veen and P. M. Dewilde, *Embedding of time-varying contractive systems in lossless realizations*, Math. Control Signals Systems, 7 (1994), pp. 306–331.
[19] P. M. Dewilde and A. J. van der Veen, *On the Hankel-Norm approximation of upper-triangular operators and matrices*, Integral Equations Operator Theory, 17 (1993), pp. 1–45.
[20] A. J. van der Veen, *Time-Varying System Theory and Computational Modeling: Realization, Approximation and Factorization*, Ph.D. thesis, Delft University of Technology, Delft, The Netherlands, 1993.
[21] W. Arveson, *Interpolation problems in nest algebras*, J. Funct. Anal., 20 (1975), pp. 208–233.
[22] G. Golub and C. Van Loan, *Matrix Computations*, The Johns Hopkins Univ. Press, Baltimore, 1989.
[23] A. H. Jazwinski, *Stochastic Processes and Filtering Theory*, Academic Press, New York, 1970.

SIAM J. CONTROL OPTIM.
Vol. 35, No. 3, pp. 810–859, May 1997

005

OUTPUT-INDUCED SUBSPACES, INVARIANT DIRECTIONS, AND INTERPOLATION IN LINEAR DISCRETE-TIME STOCHASTIC SYSTEMS*

ANDERS LINDQUIST† AND GYÖRGY MICHALETZKY‡

Abstract. In this paper we analyze the structure of the class of discrete-time linear stochastic systems in terms of the geometric theory of stochastic realization. We discuss the role of invariant directions, zeros of spectral factors, and output-induced subspaces in determining the systems-theoretical properties of the stochastic systems. A prototype interpolation problem for recovering lost state information is discussed, and it is shown how it can be solved via Kalman filtering recursions tying together the state processes of a family of totally ordered splitting subspaces.

Key words. invariant directions, zero dynamics, discrete-time stochastic systems, splitting subspaces

AMS subject classifications. 93E03, 93B27, 60G10

PII. S0363012994268679

1. Introduction.

It is a somewhat surprising fact that, in the discrete-time case, the family of minimal state-space representations

$$\begin{cases} x(t+1) = Ax(t) + Bu(t), \\ y(t) = Cx(t) + Du(t) \end{cases} \tag{1.1}$$

of a stationary stochastic vector process $\{y(t); t \in \mathbb{Z}\}$ with a rational spectral density exhibits a remarkably rich structure, affecting the implementation of most estimation algorithms, and that much of this structure is not present in the corresponding continuous-time setting. This diversity is also reflected in the structure of the corresponding family of matrix Riccati equations, studied in detail in [22] in the context of invariant directions of matrix Riccati equations [5, 25, 26], a phenomenon that is not present in the continuous-time case.

As usual, $\{u(t); t \in \mathbb{Z}\}$ is a vector-valued white noise process which, passed through a stable filter with transfer function

$$W(z) = C(zI - A)^{-1}B + D \tag{1.2}$$

beginning at the remote past, produces the output process $\{y(t); t \in \mathbb{Z}\}$, say, of dimension m and with an $m \times m$ spectral density $\Phi(z) = W(z)W(1/z)'$. (Here $'$ denotes transpose, and the white noise u is a zero-mean process such that $\mathrm{E}\{u(t)u(s)'\} = I\delta_{ts}$.)

The output process y is of course purely nondeterministic, and we assume that its spectral density Φ is full rank. The representation (1.1) is a minimal realization in the sense that the *state process* x has as few components as possible.

*Received by the editors June 1, 1994; accepted for publication (in revised form) March 26, 1996. This research was supported in part by grants from the Swedish Research Council for Engineering Sciences and the Göran Gustafsson Foundation and Hungarian National Science Foundation grants 2042 and T15668.

http://www.siam.org/journals/sicon/35-3/26867.html

†Division of Optimization and Systems Theory, Royal Institute of Technology, 100 44 Stockholm, Sweden (alq@math.kth.se).

‡Department of Probability Theory and Statistics, Eötvös Loránd University, and Laboratory of Operations Research and Decision Systems, Computer and Automation Institute, Hungarian Academy of Sciences, Budapest, Hungary (micha@math.elte.hu).

Obviously W is a rational spectral factor having all its poles strictly inside the unit circle, implying that the same holds for the eigenvalues of A. Note that we are not confining ourselves to square spectral factors W, since the dimension p of the input noise process could be larger than m. In particular, this implies that the state vector may not be expressible in terms of the output process $y(t), t \in \mathbb{Z}$, alone but that it depends on some unobserved exogenous noise also. Another consequence is that the number of zeros of W may be fewer than the number of poles of W, even if D is full rank.

Part of this paper is devoted to the following prototype *interpolation problem*, which is of a somewhat different type from the interpolation problem considered in [23, 24]. Suppose that we observe the state x as well as the output y on some finite or infinite interval except that there is a blackout of state information on some finite subinterval (t_0, t_1). Then the problem is to reconstruct the lost state information in the least squares sense, given the noisy output and the remaining states information. This problem provides a framework for studying many important questions concerning the structure of discrete-time linear stochastic systems.

This interpolation problem is preferably studied in the context of the geometric theory of stochastic realization (see [17, 18, 14, 15] and references therein), in which the properties of the state representation (1.1) are expressed in terms of the *minimal Markovian splitting subspace*

$$X = \{a'x(0) \mid a \in \mathbb{R}^n\}, \tag{1.3}$$

where $n := \dim X$ is the number of components of $x(0)$ so that $x(0)$ forms a basis in X. Due to stationarity, it is sufficient to study $\{x(t); t \in \mathbb{Z}\}$ at time $t = 0$. The family of such X corresponding to a given y will be denoted $\mathcal{X}$. It is known that $\mathcal{X}$ is endowed with a certain partial ordering. This ordering, reviewed in section 2, will play an important role in this paper.

A basic tool in the analysis of the interpolation problem and, more generally, the structural properties of the family $\mathcal{X}$ of state-space representations is a pair $(\sigma, \bar{\sigma})$ of shift operators on $\mathcal{X}$, which given any $X \in \mathcal{X}$ produces a family $\{X^{(k)} \mid k \in \mathbb{Z}\}$ of totally ordered splitting subspaces. We show that these splitting subspaces are tied together by Kalman filtering recursions in the sense that we can pass from one state process $x^{(k)}$ to the next by (forward or backward) Kalman filtering, a remarkable fact that enables us actually to compute these spaces.

These sequences of splitting subspaces provide a deeper insight into the structure of the related discrete-time matrix Riccati difference equation. In fact, the corresponding sequence of state covariance matrices constitutes a solution of this Riccati equation. It is well known that the limits at $-\infty$ and at ∞ are solutions of the steady-state (algebraic) Riccati equation, but our procedure also enables us to study the transient behavior of these equations. This should be compared with the corresponding continuous-time results in [13].

The interpolation estimate of $x(t)$ on the interval (t_0, t_1) turns out to be a linear combination of $x^{(t_0-t)}(t)$ and $x^{(t_1-t)}(t)$, the state processes of $X^{(t_0-t)}$ and $X^{(t_1-t)}$, respectively, in a certain uniform choice of coordinates, enabling us to use these Kalman recursions to determine the estimate. As $t_0 \to -\infty$ and $t_1 \to \infty$, we obtain the corresponding prototype *smoothing problem*, and the structure of the solution is similar to those presented in [3] and in [18].

We show that the computational burden of determining the interpolation estimates depends on the dimension of the *internal subspace* $X \cap H_0$ of the splitting

subspace, where H_0 is the closure, in the inner product $\langle \xi, \eta \rangle := \mathrm{E}\{\xi, \eta\}$, of all random variables

$$\{y_i(t) \mid i = 1, 2, \dots, m;\ t \in \mathbb{Z}\}$$

of the output process. In fact, we show that if $\dim X \cap H_0 = n - \nu$, we need only solve matrix Riccati equations of dimension at most $\nu \times \nu$ rather than $n \times n$ to compute the appropriate filter estimates. Sometimes, however, we need an initial number of time steps to achieve this reduction and to understand this better we need to study the structure of the internal subspace $X \cap H_0$.

In this paper we show among other things that the internal subspace has the direct sum decomposition

$$X \cap H_0 = X \cap \{y(-1), \dots, y(-n)\} + Y^* + X \cap \{y(0), \dots, y(n-1)\},$$

where the subspace Y^* can be determined by algorithms akin to the one used to compute the maximal output-nulling subspace in geometric control theory [31]. This decomposition and the theoretical framework in which it is developed give a considerable amount of information about the structure of the discrete-time linear stochastic system (1.1).

First, if the *predictable subspace* $X \cap \{y(-1), \dots, y(-n)\}$ is nontrivial, there is an $a \in \mathbb{R}^n$ such that

$$a'x(t) \in \{y(t-1), y(t-2), \dots, y(t-n)\},$$

and consequently the *usual* Kalman filtering problem of estimating $x(t)$ given the data $y(t-1), y(t-2), \dots, y(0)$ reaches steady state in a finite number of steps in the direction a. An analogous statement holds for the initial point smoothing problem and the *smoothable subspace* $X \cap \{y(0), \dots, y(n-1)\}$. Nontrivial such directions a are known as *invariant directions* and have been studied extensively in the literature [5, 25, 26, 22], but the connections to the geometric theory of Markovian splitting subspaces are presented here for the first time.

Second, the basic reason why discrete-time models (1.1) are more complicated, and the study of them is more challenging, than in the continuous-time case is that DD' varies over $\mathcal{X}$. If $DD' > 0$ for all $X \in \mathcal{X}$, the results and the analysis of the (coercive) continuous-time case generally carry over verbatim. This is known as the *regular case*. In the regular case there are no invariant directions and $Y^* = X \cap H_0$. In this paper we give several geometric characterizations of regularity and investigate the fine structure of the nonregular case.

Third, the zero structure of the transfer function (1.2) plays an important role in the analysis of the interpolation problem, and it can be studied in terms of *output-induced subspaces*, i.e., subspaces of $X \cap H_0$ with certain invariance properties to be specified below. The output-induced subspaces also provide a link between stochastic realization theory and geometric control theory [31, 4] (see Remark 7.3). This program was initiated in [18, 19] and was continued in [13] and [29], where, in particular, the connections to geometric control theory are discussed in great detail in continuous and discrete time, respectively. In this paper we introduce the concept of *strictly output-induced subspaces*, a refinement needed to study the discrete-time case. In particular, Y^* is the maximal strictly output-induced subspace, which plays the role of $X \cap H_0$ in the nonregular case. The zero structure also provides information about the possible reduction of the Riccati recursions in the interpolation problem.

The paper is organized as follows. Section 2 is devoted to preliminaries on the geometric theory of stochastic realization theory and to notations. In section 3 we introduce the operators σ and $\bar{\sigma}$, characterize regularity in terms of them, and establish the properties of the family $\{X^{(k)}|k \in \mathbb{Z}\}$, and in section 4 we introduce the interpolation problem and relate it to the results in section 3. Section 5 is about the zero structure of $\{X^{(k)}|k \in \mathbb{Z}\}$ in the regular case. In section 6 we discuss output-induced subspaces, and in section 7 the role of invariant directions is investigated and the algorithm for determining Y^* is given. Finally, in section 8, the change in zero structure when applying σ and $\bar{\sigma}$ is discussed, and the connections to the zero dynamics operators and the reduction of the Riccati equations in the interpolation problem are explained.

2. Preliminaries and notations. Given a stationary purely nondeterministic m-dimensional stochastic process $\{y(t); t \in \mathbb{Z}\}$, any stochastic realization (1.1) of y may be represented in a coordinate-free manner by a triplet (X, H, U), where X is given by (1.3), underscoring the fact that two representations (1.1) are considered identical if they differ only by the choice of coordinates in X. Here H is the Hilbert space generated by the random variables

$$\{u_i(t) \mid i = 1, 2, \dots, p;\ t \in \mathbb{Z}\}$$

with inner product

$$\langle \xi, \eta \rangle = \mathrm{E}\{\xi, \eta\},$$

and the unitary operator $U : H \to H$ is the shift determined by

$$Uu_i(t) = u_i(t+1).$$

Then U acts as the shift for all processes in the system, i.e., $Uy_i(t) = y_i(t+1)$ and $Ux_i(t) = x_i(t+1)$. We always assume that the matrix $\begin{bmatrix} B \\ D \end{bmatrix}$ has linearly independent columns so that H is generated also by

$$\{y_i(t), x_j(t) \mid i = 1, \dots, m;\ j = 1, \dots, n;\ t \in \mathbb{Z}\}.$$

The Hilbert space H so defined is called the *ambient space* of X.

For any subspace $Y \subset H$ we shall write $\mathrm{E}^Y \lambda$ to denote the orthogonal projection of $\lambda \in H$ onto Y. Occasionally we shall misuse notations somewhat by writing $\mathrm{E}^Y z$ when z is a random vector to denote the vector with components $\{\mathrm{E}^Y z_i\}$. By $\mathrm{E}^Y Z$ we shall mean the closure of $\{\mathrm{E}^Y \zeta \mid \zeta \in Z\}$. For any pair of subspaces Y and Z we write $Y + Z$ to denote direct sum (implying that $Y \cap Z = 0$), $Y \oplus Z$ for orthogonal direct sum, and $Y \vee Z$ for the vector sum in the general case, i.e., for closure$\{\eta + \zeta \mid \eta \in Y, \zeta \in Z\}$. Moreover, we write $Z^\perp$ to denote the orthogonal complement $H \ominus Z$ of Z in the ambient space H. Finally, we write $Z \perp Y \mid X$ to denote that Z and Y are *conditionally orthogonal* given X, i.e., that

$$\langle \eta - E^X \eta, \zeta - E^X \zeta \rangle = 0 \quad \text{for all } \eta \in Y,\ \zeta \in Z.$$

There are some important subspaces related to the given process y, which are subspaces of H for each representation (X, H, U) and which are considered fixed in this analysis. Define the *past space* H^- as the subspace generated by the random variables

$$\{y_i(t) \mid i = 1, 2, \dots, m;\ t = -1, -2, -3, \dots\}$$

and the *future space* H^+ as the subspace generated by

$$\{y_i(t) \mid i = 1, 2, \dots, m;\ t = 0, 1, 2, \dots\},$$

and let

$$H_0 := H^- \vee H^+ \subset H \tag{2.1}$$

be the space generated by all random variables in y. We shall also consider finite-dimensional subspaces $\{y(j), \dots, y(k)\}$ spanned by the components of the random vectors depicted inside the curly brackets. We shall also use the shorthand notation H_{t-1}^- and H_t^+ for $U^t H^-$ and $U^t H^+$, respectively, the past and future spaces shifted to time t. Then $H_{-1}^- = H^-$ and $H_0^+ = H^+$, which is consistent with the asymmetric definition of past and future.

It is well known that X is a minimal Markovian splitting subspace [17, 18] and that it can be represented *uniquely* in terms of a pair $(S, \bar{S})$ of subspaces such that

$$S \supset H^- \quad \text{and} \quad \bar{S} \supset H^+, \tag{2.2}$$

$$U^{-1}S \subset S \quad \text{and} \quad U\bar{S} \subset \bar{S}, \tag{2.3}$$

and

$$H = \bar{S}^\perp \oplus X \oplus S^\perp. \tag{2.4}$$

Consequently, S and $\bar{S}$ may be regarded as extensions of the past space H^- and future space H^+, respectively, inheriting their invariance properties, and they intersect perpendicularly so that

$$X = S \cap \bar{S} = \mathrm{E}^S \bar{S} = \mathrm{E}^{\bar{S}} S. \tag{2.5}$$

Conversely, S and $\bar{S}$ can be recovered from X in terms of

$$\begin{cases} S &= H^- \vee X^-, \\ \bar{S} &= H^+ \vee X^+, \end{cases} \tag{2.6}$$

where $X^- := \bigvee_{t=-\infty}^{0} U^t X$ and $X^+ := \bigvee_{t=0}^{\infty} U^t X$. We shall write $X \sim (S, \bar{S})$ to exhibit the one-to-one correspondence between X and $(S, \bar{S})$.

Clearly, the ambient space has the representation

$$H = S \vee \bar{S}, \tag{2.7}$$

and $S \perp \bar{S} \mid X$, which is equivalent to

$$\mathrm{E}^S \lambda = \mathrm{E}^X \lambda \quad \text{for } \lambda \in \bar{S} \tag{2.8}$$

and to

$$\mathrm{E}^{\bar{S}} \lambda = \mathrm{E}^X \lambda \quad \text{for } \lambda \in S. \tag{2.9}$$

In particular, $H^- \perp H^+ \mid X$; i.e., X is a *splitting subspace*.

We recall that $X \sim (S, \bar{S})$ is *minimal* both in the sense of subspace inclusion and in the sense of dimension, two concepts of minimality which can be shown to be equivalent, if and only if

$$\bar{S} = H^+ \vee S^\perp \tag{2.10}$$

and

$$S = H^- \vee \bar{S}^\perp \tag{2.11}$$

[17, 18]. Condition (2.10) is equivalent to $X \cap (H^+)^\perp = 0$, i.e., to X being *observable*, and (2.11) to $X \cap (H^-)^\perp = 0$, i.e., to X being *constructible*. Therefore, in view of (2.5), we have

$$X = \mathrm{E}^S H^+ = \mathrm{E}^{\bar{S}} H^- \tag{2.12}$$

whenever X is minimal.

The space S is actually identical to the subspace generated by the past of the driving white noise u in (1.1), so u can be constructed from S by Wold decomposition [14, 15]. Analogously, $\bar{S}$ corresponds to another white noise process $\{\bar{u}(t); t \in \mathbb{Z}\}$, the future space of which coincides with $\bar{S}$, and, passed through an antistable filter with transfer function

$$\bar{W}(z) = z\bar{C}(I - zA')^{-1}\bar{B} + \bar{D} \tag{2.13}$$

from the remote future, $\bar{u}$ produces a backward realization of y, namely,

$$\begin{cases} \bar{x}(t-1) = A'\bar{x}(t) + \bar{B}\bar{u}(t-1), \\ y(t-1) = \bar{C}\bar{x}(t) + \bar{D}\bar{u}(t-1). \end{cases} \tag{2.14}$$

Here $\bar{x}(0)$ is just another basis in X such that

$$\bar{x}(t) = P^{-1}x(t), \tag{2.15}$$

where P is the state covariance

$$P = \mathrm{E}\{x(0)x(0)'\}. \tag{2.16}$$

The ambient space H will of course vary over the family $\mathfrak{X}$ of minimal Markovian splitting subspaces. If $X \subset H_0$, then $H = H_0$ and we say that X is *internal*. We write $\mathfrak{X}_0$ to denote the subclass of internal $X \in \mathfrak{X}$.

The family $\mathfrak{X}$ is endowed with a natural partial ordering [18]. We say that $X_1 \leq X_2$ if

$$\|\mathrm{E}^{X_1}\lambda\| \leq \|\mathrm{E}^{X_2}\lambda\| \quad \text{for all } \lambda \in H^+$$

or, equivalently,

$$\|\mathrm{E}^{X_2}\lambda\| \leq \|\mathrm{E}^{X_1}\lambda\| \quad \text{for all } \lambda \in H^-.$$

In this ordering the *predictor space* $X_- := \mathrm{E}^{H^-}H^+$ is the minimal element in $\mathfrak{X}$ and $X_+ := \mathrm{E}^{H^+}H^-$ is the maximal element, i.e.,

$$X_- \leq X \leq X_+ \quad \text{for all } X \in \mathfrak{X}. \tag{2.17}$$

Obviously, both $X_- \sim (S_-, \bar{S}_-)$ and $X_+ \sim (S_+, \bar{S}_+)$ are internal.

This ordering can be used to introduce a *uniform choice of bases* (or coordinates) in all $X \in \mathcal{X}$. In fact, let $x_+(0)$ be an arbitrary choice of basis in X_+, and define

$$x(0) = \mathrm{E}^X x_+(0) \tag{2.18}$$

for all $X \in \mathcal{X}$. This will ensure the invariance of the matrices A and C over the class of forward minimal realizations (1.1). In the same way, we define

$$\bar{x}(0) = \mathrm{E}^X \bar{x}_-(0), \tag{2.19}$$

where $\bar{x}_-(0) = P_-^{-1} x_-(0)$, $x_-(0)$ being formed via (2.18) for $X = X_-$ and P_- being the corresponding state covariance (2.16). Then $\bar{C}$ and A' will be invariant over the set of backward realizations (2.14).

Introducing coordinates in this uniform fashion, we can also parameterize the family $\mathcal{X}$ in terms of the corresponding class $\mathcal{P}$ of state covariances (2.16). The usual partial ordering of these positive definite matrices reflects the partial ordering of splitting subspaces in $\mathcal{X}$ introduced above. Consequently, in this parameterization (2.17) becomes

$$P_- \leq P \leq P_+ \quad \text{for all } P \in \mathcal{P} \tag{2.20}$$

(cf. [8]). In the same way, we can parameterize $\mathcal{X}$ in terms of the family $\bar{\mathcal{P}}$ of covariance matrices

$$\bar{P} := \mathrm{E}\{\bar{x}(0)\bar{x}(0)'\} = P^{-1} \tag{2.21}$$

of the backward realizations (2.14). Then (2.17) becomes

$$\bar{P}_+ \leq \bar{P} \leq \bar{P}_- \quad \text{for all } \bar{P} \in \bar{\mathcal{P}}. \tag{2.22}$$

3. An ordered family of splitting subspaces. A fact of central importance in this paper is that each splitting subspace $X \in \mathcal{X}$ can be naturally imbedded in a doubly infinite sequence of elements in $\mathcal{X}$, which contains finitely many different splitting subspaces if and only if $X \in \mathcal{X}_0$; i.e., X is internal. To see this, define operators σ and $\bar{\sigma}$ on $\mathcal{X}$ so that, for $X \sim (S, \bar{S})$,

$$\sigma X = \mathrm{E}^{H^- \vee (U^{-1}S)} X, \tag{3.1}$$

$$\bar{\sigma} X = \mathrm{E}^{H^+ \vee (U\bar{S})} X. \tag{3.2}$$

Observe that the operator σ is the geometric counterpart of a one-step-ahead state predictor given past output and state information. Our first result states, among other things, that σX is itself a splitting subspace so that $\sigma X \in \mathcal{X}$. Remarkably, as we shall see in section 4, the states corresponding to $\{\sigma^k X\}$ are actually generated by a Kalman filter. Analogous statements hold for $\bar{\sigma}$ with respect to the backward setting.

THEOREM 3.1. *Let $X \sim (S, \bar{S})$ be a minimal Markovian splitting subspace. Then*

(i) *σX and $\bar{\sigma} X$ are minimal Markovian splitting subspaces and*

$$\sigma X \leq X \leq \bar{\sigma} X. \tag{3.3}$$

Moreover, they have the same ambient spaces, namely, $S \vee \bar{S}$.

(ii) $\sigma X = X$ *if and only if*

$$UX \subset X \vee \{y(0)\}, \tag{3.4}$$

and $\bar{\sigma} X = X$ *if and only if*

$$U^{-1}X \subset X \vee \{y(-1)\}. \tag{3.5}$$

(iii) *The fixed points of* σ *and* $\bar{\sigma}$ *are internal minimal Markovian splitting subspaces.*

Proof. We prove all statements involving σ. Then those involving $\bar{\sigma}$ follow by symmetry, replacing H^-, S, and U^{-1} with H^+, $\bar{S}$, and U. Since $X \sim (S, \bar{S})$ is a minimal Markovian splitting subspace, S is U^{-1}-invariant and $X = \mathrm{E}^S H^+$. Obviously $S^{(-1)} := H^- \vee (U^{-1}S)$ is also U^{-1}-invariant and $S^{(-1)} \subset S$. Therefore

$$\sigma X = \mathrm{E}^{S^{(-1)}} \mathrm{E}^S H^+ = \mathrm{E}^{S^{(-1)}} H^+$$

is an observable Markovian splitting subspace. Since, in addition, $S^{(-1)} \subset S \perp H^+ \cap (H^-)^\perp$, σX is minimal [18, Theorem 4.10]. Since

$$\mathrm{E}^{S^{(-1)}} \mathrm{E}^S \lambda = \mathrm{E}^{S^{(-1)}} \lambda \quad \text{for each } \lambda \in H^+,$$

the splitting property (2.8) and the fact that $\|\mathrm{E}^{S^{(-1)}} \xi\| \leq \|\xi\|$ imply that $\sigma X \leq X$. Since X and σX are finite dimensional and hence proper [17, 18], their ambient spaces are $\bigvee_{t=0}^{\infty} U^t S$ and $\bigvee_{t=0}^{\infty} U^t S^{(-1)}$, which must coincide in view of the fact that

$$U^{-1}S \subset S^{(-1)} \subset S.$$

In the same way we show that X and $\bar{\sigma} X$ have the same ambient space. This proves (i).

Next, we show that if $\sigma X = X$, then $X \subset H_0$. Now, $\sigma X = X$ is equivalent to $X \subset H^- \vee U^{-1}S$ and hence to $UX \subset S \vee \{y(0)\}$. However, $S = X \oplus \bar{S}^\perp$ and $\{y(0)\} \subset H^+ \subset \bar{S} \perp \bar{S}^\perp$, so

$$UX \subset [X \vee \{y(0)\}] \oplus \bar{S}^\perp.$$

Since $UX \perp U\bar{S}^\perp \supset \bar{S}^\perp$ we have thus established that $\sigma X = X$ if and only if

$$UX \subset X \vee \{y(0)\}, \tag{3.6}$$

which is the first part of (ii). A symmetric argument yields the second part.

Finally, to prove (iii), we note that (3.6) implies that

$$U\mathrm{E}^{H_0^\perp} X \subset \mathrm{E}^{H_0^\perp} X.$$

But the subspace $\mathrm{E}^{H_0^\perp} X$ is finite dimensional. Since U is a bilateral shift, it has no eigenvalues [30] and hence cannot have a nontrivial finite-dimensional invariant subspace. Consequently we must have $X \subset H_0$. ∎

COROLLARY 3.2. *Let* $X \in \mathcal{X}$ *and* $X \sim (S, \bar{S})$. *Then*

$$X^{(k)} = \begin{cases} \sigma^{-k} X & \text{for } k = 0, -1, -2, \dots, \\ \bar{\sigma}^k X & \text{for } k = 0, 1, 2, \dots \end{cases} \tag{3.7}$$

defines a sequence $\{X^{(k)} \mid k \in \mathbb{Z}\}$ *of elements in* $\mathfrak{X}$ *which have the same ambient space and which are totally ordered with* $X^{(0)} = X$. *More precisely,*

$$\cdots \leq X^{(-2)} \leq X^{(-1)} \leq X \leq X^{(1)} \leq X^{(2)} \leq \cdots .$$

Moreover, for each $k \in \mathbb{Z}$, $X^{(k)} \sim (S^{(k)}, \bar{S}^{(k)})$, *where*

$$S^{(k)} = H^- \vee U^k S, \quad \bar{S}^{(k)} = H^+ \vee \left[S^{(k)}\right]^\perp \quad \text{for } k \leq 0$$

and

$$\bar{S}^{(k)} = H^+ \vee U^k \bar{S}, \quad S^{(k)} = H^- \vee \left[\bar{S}^{(k)}\right]^\perp \quad \text{for } k \geq 0.$$

Here the orthogonal complement $^\perp$ *is taken with respect to the common ambient space* $S \vee \bar{S}$.

Proof. This corollary follows from the proof of Theorem 3.1, (2.10), and (2.11). In fact, it follows by induction that $\sigma^k X$ is also a minimal Markovian splitting subspace and that

$$\sigma^k X = \mathrm{E}^{S^{(-k)}} X \quad \text{for } k = 0, 1, 2, \ldots,$$

where $S^{(-k)} := H^- \vee U^{-k} S$. The statement about $\bar{\sigma}$ follows by symmetry. □

Next we show that the sequence $\{X^{(k)} \mid k \in \mathbb{Z}\}$ can be extended to include limits $X^{(-\infty)}$ and $X^{(\infty)}$.

THEOREM 3.3. *The limits* $\lim_{k\to-\infty} E^{S^{(k)}}\xi$ *and* $\lim_{k\to\infty} E^{\bar{S}^{(k)}}\xi$ *exist for all* $\xi \in X$, *and the spaces*

$$X^{(-\infty)} := \left\{ \lim_{k\to-\infty} E^{S^{(k)}}\xi \mid \xi \in X \right\}, \tag{3.8}$$

$$X^{(\infty)} := \left\{ \lim_{k\to\infty} E^{\bar{S}^{(k)}}\xi \mid \xi \in X \right\} \tag{3.9}$$

are internal minimal Markovian splitting subspaces. Moreover, the sequences of splitting subspaces $\{X^{(-k)} \mid k = 0, 1, 2, \ldots\}$ *and* $\{X^{(k)} \mid k = 0, 1, 2, \ldots\}$ *converge in a finite number of steps if and only if* X *is internal. In that case the number of steps is no greater than* $\dim X$.

Proof. Since $\{S^{(-k)} \mid k = 0, 1, 2, \ldots\}$ is a nonincreasing sequence of subspaces, i.e.,

$$S \supset S^{(-1)} \supset S^{(-2)} \supset S^{(-3)} \supset \cdots , \tag{3.10}$$

it is well known [6, p. 24] that $\xi_{-\infty} = \lim_{k\to\infty} \mathrm{E}^{S^{(-k)}}\xi$ exists for all $\xi \in X$ and that $\xi_{-\infty} = \mathrm{E}^{S^{(-\infty)}}\xi$, where $S^{(-\infty)} = \bigcap_{k=0}^{\infty} S^{(-k)}$. Thus $X^{(-\infty)}$ is well defined, and since $X = \mathrm{E}^S H^+$,

$$X^{(-\infty)} = \mathrm{E}^{S^{(-\infty)}} H^+.$$

Therefore, since $S^{(-\infty)}$ is U^{-1}-invariant, $X^{(-\infty)}$ is an observable Markovian splitting subspace. But $S^{(-\infty)} \subset S \perp H^+ \cap (H^-)^\perp$, and hence $X^{(-\infty)}$ is minimal. It remains to show that $X^{(-\infty)}$ is internal. In view of Theorem 3.1, this would follow if $X^{(-\infty)}$ were a fixed point for σ. Next, we prove that this is in fact the case.

Consequently we want to prove that $\sigma X^{(-\infty)} = X^{(-\infty)}$, which follows from

$$H^- \vee U^{-1}S^{(-\infty)} = S^{(-\infty)}. \tag{3.11}$$

Let us prove (3.11). Since $S^{(-\infty)} = \bigcap_{k=0}^{\infty} S^{(-k)}$, $U^{-1}S^{(-k)} \subset S^{(-k)}$, and $H^- \subset S^{(-k)}$, it is trivial that

$$H^- \vee U^{-1}S^{(-\infty)} \subset S^{(-\infty)}.$$

It remains to prove the converse. To this end, note that

$$H^- \vee U^{-1}S^{(-k)} = \{y(-1)\} \vee U^{-1}S^{(-k)}.$$

This sum is in general not direct, so we want to reformulate it into such a sum. Therefore, observe that $\{y(-1)\} \cap U^{-1}S^{(k)}$ is nonincreasing in k and finite dimensional and thus there is a k_0 such that

$$\{y(-1)\} \cap U^{-1}S^{(k)} = \{y(-1)\} \cap U^{-1}S^{(k_0)} \quad \text{for } k \geq k_0.$$

Let V be a complement of $\{y(-1)\} \cap U^{-1}S^{(k_0)}$ in $\{y(-1)\}$. Then, for $k \geq k_0$,

$$H^- \vee U^{-1}S^{(-k)} = V + U^{-1}S^{(-k)} \tag{3.12}$$

is a direct sum. Now, if $\xi \in S^{(-\infty)} = \bigcap_{k=0}^{\infty}(H^- \vee U^{-1}S^{(-k)})$, then

$$\xi = v_k + \eta_k$$

with $v_k \in V$ and $\eta_k \in U^{-1}S^{(-k)}$ is a unique representation for each $k \geq k_0$. Therefore, since (3.12) is nonincreasing, $v_k = v \in V$ and $\eta_k = \eta \in \bigcap_{k=0}^{\infty} U^{-1}S^{(-k)} = U^{-1}S^{(-\infty)}$ for $k \geq k_0$. Hence

$$\xi = v + \eta \in V \vee U^{-1}S^{(-\infty)} \subset H^- \vee U^{-1}S^{(-\infty)},$$

proving that $X^{(-\infty)}$ is a fixed point.

Next, we assume that X is noninternal and prove that, in this case, all elements of the sequences $\{X^{(-k)} \mid k = 0, 1, 2, \dots\}$ and $\{X^{(k)} \mid k = 0, 1, 2, \dots\}$ are noninternal and that consequently these sequences cannot converge in a finite number of steps, the limits being internal. To see this, take a $\xi \in S$ such that $\xi \neq H_0$. Then $U^{-k}\xi \in S^{(-k)}$ but $U^{-k}\xi \notin H_0$, showing that $X^{(-k)}$ is noninternal for $k = 0, 1, 2, \dots$. A symmetric argument involving $\bar{S}$ shows that $X^{(k)}$ is also noninternal for $k = 0, 1, 2, \dots$.

To prove the converse, first recall that, for any internal $X \sim (S, \bar{S})$,

$$X = (X \cap X_-) \vee (X \cap X_+) \tag{3.13}$$

and that $S = H^- \vee X$. Relation (3.13) is proven in the same way as Lemma 2.9 in [13] and can also be found in [21]. Hence,

$$S = H^- \vee (X \cap X_+). \tag{3.14}$$

Since $X = \mathrm{E}^S H^+$, the subspace $X \cap X_+$ thus uniquely determines the internal splitting subspace X. In view of (3.14), (3.10) implies that the sequence $\{X^{(-k)} \cap X_+ \mid k = 0, 1, 2, \dots\}$ of finite-dimensional subspaces is nonincreasing. Therefore, it must converge in a finite number of steps which cannot be larger than $\dim X$, implying via (3.14) that the same holds for the sequence $\{X^{(-k)} \mid k = 0, 1, 2, \dots\}$. ∎

We remark that this proof also shows that, if X is internal, the whole sequence $\{X^{(k)} \mid k \in \mathbb{Z}\}$ cannot have more than $\dim X + 1$ different elements.

As pointed out in the end of the proof of Theorem 3.3, an internal X is completely characterized by its intersection $X \cap X_+ = X \cap H^+$ with the future via (3.14). In the same way,

$$\bar{S} = H^+ \vee (X \cap X_-), \tag{3.15}$$

so $X \in \mathfrak{X}_0$ is also characterized by its intersection $X \cap X_- = X \cap H^-$ with the past. In particular, we have the following characterizations of σX_+ and $\bar{\sigma} X_-$. Similar characterizations for arbitrary X are given in section 8.

PROPOSITION 3.4. *The intersection of σX_+ with the past H^- is described by*

$$(\sigma X_+) \cap X_- = (X_+ \cap X_-) \vee (\{y(-1)\} \cap X_-), \tag{3.16}$$

and the intersection of $\bar{\sigma} X_-$ with the future H^+ is described by

$$(\bar{\sigma} X_-) \cap X_+ = (X_- \cap X_+) \vee (\{y(0)\} \cap X_+). \tag{3.17}$$

Proof. First observe that (3.16) is equivalent to

$$\bar{S}_+^{(-1)} = H^+ \vee (\{y(-1)\} \cap X_-). \tag{3.18}$$

In fact, that (3.18) implies (3.16) follows from the facts that

$$\bar{S}_+^{(-1)} \cap X_- = (\sigma X_+) \cap X_-$$

and

$$\left[H^+ \vee (\{y(-1)\} \cap X_-)\right] \cap X_- = (X_+ \cap X_-) \vee (\{y(-1)\} \cap X_-),$$

while the opposite implication follows from the fact that

$$\bar{S}_+^{(-1)} = H^+ \vee \left[(\sigma X_+) \cap X_-\right].$$

Next let us prove (3.18). We have

$$\begin{aligned}
\bar{S}_+^{(-1)} &= H^+ \vee (S_+^{(-1)})^\perp = H^+ \vee (H^- \vee U^{-1} S_+)^\perp \\
&= H^+ \vee \left[(H^-)^\perp \cap (U^{-1} H^+) \cap (U^{-1} H^-)^\perp\right] \\
&= H^+ \vee \left[(H^-)^\perp \cap (U^{-1} H^+)\right] \\
&= H^+ \vee \left[(H^-)^\perp \cap (H^+ \vee \{y(-1)\})\right],
\end{aligned}$$

where in the third step we have used the fact that $S_+^\perp = H^+ \cap (H^-)^\perp$. (See, for example, [18, Example 4.4].) Now suppose that $\xi \in (H^-)^\perp \cap (H^+ \vee \{y(-1)\})$. Then $\xi = \alpha + \beta$, where $\alpha \in H^+$ and $\beta \in \{y(-1)\} \subset H^-$. But $\alpha + \beta \perp H^-$, and consequently

$$\beta = -\mathrm{E}^{H^-} \alpha \in \mathrm{E}^{H^-} H^+ = X_-$$

so that $\beta \in \{y(-1)\} \cap X_-$. Hence $\bar{S}_+^{(-1)} \subset H^+ \vee (\{y(-1)\} \cap X_-)$.

Conversely, if $\beta \in \{y(-1)\} \cap X_-$, there is an $\alpha \in H^+$ such that $\beta = -\mathrm{E}^{H^-} \alpha$, which implies that $\alpha + \beta \perp H^-$ and that $\alpha + \beta \in (H^+ \vee \{y(-1)\})$. Hence

$$\{y(-1)\} \cap X_- \subset H^+ \vee \left[(H^-)^\perp \cap (H^+ \vee \{y(-1)\})\right],$$

which concludes the proof of (3.16) and (3.18).

The proof of the dual statement is completely symmetric, using the fact that

$$S_-^{(1)} = H^- \vee (\{y(0)\} \cap X_+) \tag{3.19}$$

is equivalent to (3.17). □

Since a minimal internal splitting subspace is completely characterized by its intersection with X_- via (3.15) and by its intersection with X_+ via (3.14), Proposition 3.4 has the following corollary, which we shall need later.

COROLLARY 3.5. *The splitting subspace X_+ is a fixed point of the operator σ if and only if $X_- \cap \{y(-1)\} = 0$. Likewise, X_- is a fixed point of the operator $\bar{\sigma}$ if and only if $X_+ \cap \{y(0)\} = 0$.*

The proof of Proposition 3.4 is easily modified to yield the following amplification, describing the chains of splitting subspaces $\{X_+^{(k)}\}$ and $\{X_-^{(k)}\}$.

PROPOSITION 3.6. *For $k = 1, 2, 3, \dots$, we have*

$$(\sigma^k X_+) \cap X_- = (X_+ \cap X_-) \vee (\{y(-1), \dots, y(-k)\} \cap X_-) \tag{3.20}$$

or, equivalently,

$$\bar{S}_+^{(-k)} = H^+ \vee (\{y(-1), \dots, y(-k)\} \cap X_-); \tag{3.21}$$

and

$$(\bar{\sigma}^k X_-) \cap X_+ = (X_- \cap X_+) \vee (\{y(0), \dots, y(k-1)\} \cap X_+) \tag{3.22}$$

or, equivalently,

$$S_-^{(k)} = H^- \vee (\{y(0), \dots, y(k-1)\} \cap X_+). \tag{3.23}$$

We shall now characterize the fixed points of the operators σ and $\bar{\sigma}$ in terms of the matrices D and $\bar{D}$ in equations (1.1) and (2.14), respectively.

COROLLARY 3.7. *Let $X \in \mathfrak{X}$ and let D and $\bar{D}$ be the corresponding matrices in the models* (1.1) *and* (2.14). *Then $\sigma X = X$ if and only if $\ker D' = 0$ and $\bar{\sigma} X = X$ if and only if $\ker \bar{D}' = 0$.*

Proof. Given (1.1) an elementary calculation yields

$$x(1) = Ax(0) + BD'(DD')^\sharp [y(0) - Cx(0)] + B_2 u(0),$$

where $B_2 := B - BD'(DD')^\sharp D$ and $(DD')^\sharp$ is a pseudoinverse of DD'. In particular, this implies that

$$\mathrm{E}\{B_2 u(0) y(0)'\} = BD' - BD'(DD')^\sharp DD' = 0.$$

Since therefore the components of $B_2 u(0)$ are orthogonal to those of both $x(0)$ and $y(0)$, (3.4) is equivalent to $B_2 = 0$, which in turn is equivalent to

$$\begin{bmatrix} B \\ D \end{bmatrix} [I - D'(DD')^\sharp D] = 0.$$

But the columns of $\begin{bmatrix} B \\ D \end{bmatrix}$ are—according to our assumption—linearly independent so (3.4) is equivalent to $D'(DD')^\sharp D = I$, which holds if and only if DD' is full rank, i.e.,

if $(DD')^{-1}$ exists. Then the first statement follows from Theorem 3.1(ii). The second statement follows by symmetry. □

Remark 3.8. In view of Corollary 3.7 we have another proof of the fact that any fixed point of σ is internal. In fact, we established in the proof above that (3.4) is equivalent to $B_2 = 0$, which in the case when DD' is full rank implies that the transfer function (1.2) of (1.1) must be a square spectral factor and thus correspond to an internal realization [16, Theorem 5.2].

Theorem 3.1 and Corollary 3.7 give characterizations of precisely which internal X are fixed points of σ and $\bar{\sigma}$. It follows trivially from the definitions (3.1) and (3.2) that

$$\sigma X_- = X_- \quad \text{and} \quad \bar{\sigma} X_+ = X_+,$$

which, by Corollary 3.7, implies that D_- and $\bar{D}_+$ are always full rank, a well-known property of the innovations models. The following proposition together with Corollary 3.7 gives a more global picture on this question. (Also see [14].)

PROPOSITION 3.9. *Let $X \in \mathcal{X}$, and let D and $\bar{D}$ be the corresponding matrices in the models* (1.1) *and* (2.14). *Then*

$$\dim \ker D' = \dim(X \cap \{y(0)\}) \leq \dim(X_+ \cap \{y(0)\}) = \dim \ker D'_+ \tag{3.24}$$

and

$$\dim \ker \bar{D}' = \dim(X \cap \{y(-1)\}) \leq \dim(X_- \cap \{y(-1)\}) = \dim \ker \bar{D}'_-. \tag{3.25}$$

Proof. Let $a \in \ker D'$. Then $a'D = 0$ so that $a'y(0) \in X$. Conversely, suppose that $a'y(0) \in X$. Then, since $a'Cx(0) \in X$, we must have $a'Du(0) \in X \perp \{u(0)\}$, implying that $a'D = 0$. This proves the equalities in (3.24). To prove the inequality, note that

$$X \cap X_+ \cap \{y(0)\} = X \cap H^+ \cap \{y(0)\} = X \cap \{y(0)\}.$$

A symmetric argument yields (3.25). □

COROLLARY 3.10. *The splitting subspace X_+ is a fixed point of the operator σ if and only if $X_+ \cap \{y(0)\} = 0$. Likewise, X_- is a fixed point of the operator $\bar{\sigma}$ if and only if $X_- \cap \{y(-1)\} = 0$.*

Proof. In view of Corollary 3.7, this follows from the last equalities in (3.24) and (3.25), respectively. □

Comparing Corollaries 3.5 and 3.10, we can now see that the two conditions $X_+ \cap \{y(0)\} = 0$ and $X_- \cap \{y(-1)\} = 0$ are actually equivalent. We shall refer to the situation when they are satisfied as the *regular case*. From Proposition 3.9 and Corollary 3.7 it readily follows that, in the regular case and only in the regular case, *all* $X \in \mathcal{X}_0$ are fixed points of *both* σ and $\bar{\sigma}$. All this could also have been shown without using Corollary 3.5 by instead invoking the fact, proven in [14, Theorem 10.2], that D_+ has full rank if and only if $\bar{D}_-$ has.

The fact that $\sigma X_+ = X_+$ and $\bar{\sigma} X_- = X_-$ are the critical conditions in this analysis is also reflected in the ordering of covariances. In fact,

$$DD' = \Lambda_0 - CPC' \geq \Lambda_0 - CP_+C' = D_+D'_+$$

for all $P \leq P_+$ so that regularity is equivalent to

$$\Lambda_0 - CPC' > 0 \quad \text{for all } P \in \mathcal{P}$$

and analogous in the backward setting. This is also equivalent to all minimal spectral factors having zeros neither at zero nor at infinity.

We collect the regularity conditions in the following proposition. Some other characterizations can be found in [22, Theorem 3.2].

PROPOSITION 3.11. *The following regularity conditions are equivalent.*

(i) $\Lambda_0 - CP_+C' > 0$.
(ii) $\Lambda_0 - \bar{C}\bar{P}_-\bar{C}' > 0$.
(iii) $X_+ \cap \{y(0)\} = 0$.
(iv) $X_- \cap \{y(-1)\} = 0$.
(v) $\sigma X_+ = X_+$.
(vi) $\bar{\sigma} X_- = X_-$.

Clearly regularity is a property of the output process y. Therefore, we introduce the following definition.

DEFINITION 3.12. *The process y is* regular *if the conditions of Proposition* 3.11 *are satisfied.*

The regularity conditions can also be stated in terms of the whole family of minimal realizations.

PROPOSITION 3.11′. *Each of the following regularity conditions is equivalent to those in Proposition* 3.11.

(i)′ $\Lambda_0 - CPC' > 0$ *for all* $P \in \mathcal{P}$.
(ii)′ $\Lambda_0 - C\bar{P}C' > 0$ *for all* $\bar{P} \in \bar{\mathcal{P}}$.
(iii)′ $X \cap \{y(0)\} = 0$ *for all* $X \in \mathcal{X}$.
(iv)′ $X \cap \{y(-1)\} = 0$ *for all* $X \in \mathcal{X}$.
(v)′ $\sigma X = X$ *for all* $X \in \mathcal{X}_0$.
(vi)′ $\bar{\sigma} X = X$ *for all* $X \in \mathcal{X}_0$.

We shall next prove that the operators σ and $\bar{\sigma}$ are invertible in the regular case and that $\bar{\sigma} = \sigma^{-1}$. In fact, as we shall see in Theorem 3.13 and Corollary 3.14 below, this property characterizes the regularity of the process y. In section 6 we study the nonregular case and give a more complete description of the subspaces σX, $\bar{\sigma} X$ for any $X \in \mathcal{X}$.

In view of Proposition 3.4, a straightforward calculation based on (3.18) shows that

$$\bar{\sigma}\sigma X_+ = X_+ \quad \text{and} \quad \sigma\bar{\sigma} X_- = X_-. \tag{3.26}$$

A natural question is under what conditions these fixed-point properties can be generalized to arbitrary $X \in \mathcal{X}$.

THEOREM 3.13. *Let $X \in \mathcal{X}$. Then*

$$\sigma\bar{\sigma} X \leq X \leq \bar{\sigma}\sigma X \tag{3.27}$$

and

$$\bar{\sigma}\sigma X = X \Longleftrightarrow \{y(0)\} \cap X = \{y(0)\} \cap X_+. \tag{3.28}$$

Symmetrically,

$$\sigma\bar{\sigma} X = X \Longleftrightarrow \{y(-1)\} \cap X = \{y(-1)\} \cap X_-. \tag{3.29}$$

Proof. We prove (3.29) and the first inequality in (3.27). Then, the rest follows by symmetry. First observe that, since $X \cap H^- = X \cap X_- \subset X_-$ and $\{y(-1)\} \subset H^-$, it always holds that

$$\{y(-1)\} \cap X \subset \{y(-1)\} \cap X_-.$$

In view of Corollary 3.2, $\bar{\sigma}X \sim (S^{(1)}, \bar{S}^{(1)})$, where

$$\begin{aligned}\bar{S}^{(1)} &= H^+ \vee U\bar{S},\\ S^{(1)} &= H^- \vee (\bar{S}^{(1)})^\perp = H^- \vee \left[(H^+)^\perp \cap (U\bar{S}^\perp)\right].\end{aligned}$$

Then apply σ to $\bar{\sigma}X$ to obtain

$$H^- \vee U^{-1}S^{(1)} = H^- \vee U^{-1}H^- \vee \left[(U^{-1}H^+)^\perp \cap \bar{S}^\perp\right].$$

Since $U^{-1}H^- \subset H^-$ and $U^{-1}H^+ = \{y(-1)\} \vee H^+$, we have

$$H^- \vee U^{-1}S^{(1)} = H^- \vee \left[\bar{S}^\perp \cap \{y(-1)\}^\perp\right] \subset S, \tag{3.30}$$

the last of which is a consequence of the condition $S = H^- \vee \bar{S}^\perp$. Hence $\sigma\bar{\sigma}X \leq X$. To find a condition under which $\sigma\bar{\sigma}X = X$, we need to characterize the converse inequality. To this end, we consider the converse inclusion of (3.30) and take orthogonal complements in it to obtain

$$(H^-)^\perp \cap (\bar{S} \vee \{y(-1)\}) \subset (H^-)^\perp \cap \bar{S}. \tag{3.31}$$

Now, let ξ be an element in the subspace on the left side of (3.31). Then, $\xi = \alpha + \beta$, where $\alpha \in \bar{S}$ and $\beta \in \{y(-1)\} \subset H^-$, and $\xi \perp H^-$. Consequently,

$$\beta = -\mathrm{E}^{H^-}\alpha \in \mathrm{E}^{H^-}\bar{S} = X_-,$$

and hence $\beta \in \{y(-1)\} \cap X_-$. So if $\{y(-1)\} \cap X = \{y(-1)\} \cap X_-$ holds, $\beta \in \{y(-1)\} \cap X \subset \bar{S}$. Therefore, since $\alpha \in \bar{S}$, we have $\beta \in \bar{S}$, and hence (3.31) holds.

Conversely, suppose that (3.31) holds. Consider a $\beta \in \{y(-1)\} \cap X_-$. Then, there is an $\alpha \in \bar{S}$ such that $\beta = -\mathrm{E}^{H^-}\alpha$ so that

$$\alpha + \beta \in (H^-)^\perp \cap (\bar{S} \vee \{y(-1)\}).$$

Using condition (3.31), we obtain $\alpha + \beta \in \bar{S}$. But $\alpha \in \bar{S}$, and hence $\beta \in \bar{S}$. In other words, (3.31) implies that $\{y(-1)\} \cap X_- \subset \bar{S}$, and consequently

$$\{y(-1)\} \cap X_- = \{y(-1)\} \cap X_- \cap \bar{S} \subset \{y(-1)\} \cap X,$$

since $X_- \cap \bar{S} = X_- \cap S \cap \bar{S} = X_- \cap X$. But the converse inclusion has already been proven above. Hence we have established (3.29). □

COROLLARY 3.14. *In the regular case the operators σ and $\bar{\sigma}$ are invertible and*

$$\bar{\sigma} = \sigma^{-1}. \tag{3.32}$$

Proof. This follows from regularity condition (iii) in Proposition 3.11 and (3.28) in Theorem 3.13. □

In particular Corollary 3.14 implies that relations (3.7) can be extended so that

$$X^{(j)} = \sigma^{j-k}X^{(k)} = \bar{\sigma}^{k-j}X^{(k)} \quad \text{for all } j, k \in \mathbb{Z}. \tag{3.33}$$

4. An interpolation problem. The ordered family of splitting subspaces introduced in section 3 is intimately connected to the following estimation problem. Given a minimal stochastic system

$$\begin{cases} x(t+1) = Ax(t) + Bu(t), \\ y(t) = Cx(t) + Du(t) \end{cases} \tag{4.1}$$

of the type defined in section 2 and integers t_0, t_1 such that $t_0 < t_1$, find, for each time t between t_0 and t_1, the linear least squares estimate[1]

$$\hat{x}(t \mid t_0, t_1) = \mathrm{E}\{x(t) \mid y(s), s \in \mathbb{Z}; x(\tau), \tau \in \mathbb{Z} \setminus \{t_0 + 1, \dots, t_1 - 1\}\} \tag{4.2}$$

of the state $x(t)$ given the whole output process y and the whole state process x except for times τ such that $t_0 < \tau < t_1$.

This interpolation problem is a prototype of an estimation problem of the following type. The state of a linear stochastic system is being observed both directly and through a noisy channel. During an interval of time (t_0, t_1) the direct state information is lost, and the problem is to estimate the lost states from the noisy observations and the remaining state information. Letting $t_0 \to -\infty$ and $t_1 \to \infty$, we obtain a *smoothing problem.* In a practical situation one would of course expect the information to be given on a finite interval containing $[t_0, t_1]$ and not on all of $\mathbb{Z}$ as here. However, as will be seen in Theorem 4.6 below, our solution will depend only on data from the interval $[t_0, t_1]$ and hence applies also to this situation, a remarkable fact that derives from the Markov property and allows us to use Kalman filtering. Nevertheless, it is convenient to formulate the problem in terms of infinite data.

In the more compact notation of section 2, the interpolation estimate may be written

$$\hat{x}(t \mid t_0, t_1) = \mathrm{E}^{H_0 \vee (U^{t_0} X^-) \vee (U^{t_1} X^+)} x(t), \tag{4.3}$$

where X is the splitting subspace corresponding to (4.3) and X^- and X^+ are the past and future of X as defined after (2.6). Now, one of the main results of this section is that the estimate (4.3) can be represented as a linear combination of the two estimates

$$x^{(t_0 - t)}(t) = \mathrm{E}^{H^-_{t-1} \vee (U^{t_0} X^-)} x(t), \tag{4.4}$$

based on the past information, and

$$x^{(t_1 - t)}(t) = \mathrm{E}^{H^+_t \vee (U^{t_1} X^+)} x(t), \tag{4.5}$$

based on the future information. As we demonstrate below, this is due to the fact that $x^{(t_0 - t)}$ and $x^{(t_1 - t)}$ are state processes of minimal realizations of y, the splitting subspaces of which bound X from below and from above in the ordering defined in section 2. In fact, $x^{(t_0 - t)}(t) = U^t x^{(t_0 - t)}(0)$ and $x^{(t_1 - t)}(t) = U^t x^{(t_1 - t)}(0)$, where

$$x^{(-k)}(0) = \mathrm{E}^{H^- \vee U^{-k} X^-} x(0) \tag{4.6}$$

[1] Clearly $\mathrm{E}\{\cdot \mid \cdot\}$ denotes *wide sense* conditional expectation unless the system is assumed to be Gaussian.

and

$$x^{(k)}(0) = \mathrm{E}^{H^+ \vee U^k X^+} x(0) \tag{4.7}$$

are defined for $k = 0, 1, 2, \dots$. Obviously $x^{(0)} = x$ by both formulas. This relates the estimates (4.4) and (4.5) to the operators σ and $\bar{\sigma}$ defined in section 3.

PROPOSITION 4.1. *The family of subspaces* $\{X^{(k)} \mid k \in \mathbb{Z}\}$, *defined in terms of* (4.6) *and* (4.7) *by*

$$X^{(k)} = \{a' x^{(k)}(0) \mid a \in \mathbb{R}^n\}, \tag{4.8}$$

is a family of minimal Markovian splitting subspaces such that

$$X^{(-k)} = \sigma^k X \quad \textit{and} \quad X^{(k)} = \bar{\sigma}^k X \tag{4.9}$$

for $k = 0, 1, 2, \dots$, *where* $X = \{a' x(0) \mid a \in \mathbb{R}^n\}$ *and* $\sigma, \bar{\sigma}$ *are the operators defined by* (3.1) *and* (3.2). *Moreover, for each* $k \in \mathbb{Z}$, $x^{(k)}(0)$ *is the basis in* $X^{(k)}$ *in the same uniform choice of coordinates as* $x(0)$.

Proof. Let $X \sim (S, \bar{S})$. Then $S = H^- \vee X^-$ and $\bar{S} = H^+ \vee X^+$ and so, since $U^{-1}S \subset S$ and $U\bar{S} \subset \bar{S}$,

$$x^{(k)}(0) = \begin{cases} \mathrm{E}^{H^- \vee U^k S} x(0) = \mathrm{E}^{S^{(k)}} x(0) & \text{for } k \le 0, \\ \mathrm{E}^{H^+ \vee U^k \bar{S}} x(0) = \mathrm{E}^{\bar{S}^{(k)}} x(0) & \text{for } k \ge 0, \end{cases} \tag{4.10}$$

where $S^{(k)}$ and $\bar{S}^{(k)}$ are defined as in Corollary 3.2. Then the first statement is an immediate consequence of (4.10) and Corollary 3.2. Moreover, since $S^{(k)} \subset S$ for $k \le 0$,

$$x^{(k)}(0) = \mathrm{E}^{S^{(k)}} x(0) = \mathrm{E}^{S^{(k)}} \mathrm{E}^{S} x_+(0) = \mathrm{E}^{S^{(k)}} x_+(0)$$

for the appropriate choice of basis in X_+. Similarly, for $k \ge 0$, $\bar{S}^{(k)} \subset \bar{S}$ so that

$$\bar{x}^{(k)}(0) = \mathrm{E}^{\bar{S}^{(k)}} \bar{x}(0) = \mathrm{E}^{\bar{S}^{(k)}} \mathrm{E}^{\bar{S}} \bar{x}_-(0) = \mathrm{E}^{\bar{S}^{(k)}} \bar{x}_-(0),$$

where $\bar{x}_-(0) = P_-^{-1} x_-(0)$ and $x_-(0) = \mathrm{E}^{X_-} x_+(0)$, and this proves the second statement. □

Consequently, we have established that

$$\hat{x}(0 \mid t_0, t_1) = \mathrm{E}^{S^{(t_0 - t)} \vee \bar{S}^{(t_1 - t)}} x(0), \tag{4.11}$$

where $X^{(t_0 - t)} \sim (S^{(t_0 - t)}, \bar{S}^{(t_0 - t)})$ and $X^{(t_1 - t)} \sim (S^{(t_1 - t)}, \bar{S}^{(t_0 - t)})$ are elements in $\mathfrak{X}$ such that

$$X^{(t_0 - t)} \le X \le X^{(t_1 - t)} \tag{4.12}$$

and such that $S^{(t_0 - t)} \subset S$ and $\bar{S}^{(t_1 - t)} \subset \bar{S}$. The following chain of lemmas deals with this setup and leads to the first main result of this section.

LEMMA 4.2. *Let* $X \sim (S, \bar{S})$, $X_1 \sim (S_1, \bar{S}_1)$, *and* $X_2 \sim (S_2, \bar{S}_2)$ *be minimal Markovian splitting subspaces such that* $S_1 \subset S$ *and* $\bar{S}_2 \subset \bar{S}$. *Then* $X_1 \le X \le X_2$ *and*

$$x_1(0) = E^{X_1} x_2(0) \tag{4.13}$$

for any uniform choice $x_1(0), x_2(0)$ *of bases in* X_1 *and* X_2.

Proof. Since $\bar{S}_2 \subset \bar{S}$, we have

$$\bar{S}_2 \cap (H^-)^\perp \subset \bar{S} \cap (H^-)^\perp.$$

Since X and X_2 are minimal, this is equivalent to $S_2^\perp \subset S^\perp$ [18, Corollary 4.5] or, equivalently,

$$(\bar{S}_2 \vee S_2) \ominus S_2 \perp S. \tag{4.14}$$

In fact, $\bar{S}_2 \vee S_2$ is the ambient space of X_2. But (4.14) is equivalent to $\bar{S}_2 \perp S \mid S_2$ [18, Proposition 2.1] and also to

$$\mathrm{E}^S\lambda = \mathrm{E}^S\mathrm{E}^{S_2}\lambda \quad \text{for all } \lambda \in \bar{S}_2. \tag{4.15}$$

Apply E^{X_1} to this. Since $X_1 \subset S_1 \subset S$ and $H^+ \subset \bar{S}_2$, we obtain in particular

$$\mathrm{E}^{X_1}\lambda = \mathrm{E}^{X_1}\mathrm{E}^{S_2}\lambda \quad \text{for all } \lambda \in H^+.$$

But X_2 is a splitting subspace, so $\mathrm{E}^{S_2}\lambda = \mathrm{E}^{X_2}\lambda$ for all $\lambda \in H^+$ and $X_+ \subset H^+$. Consequently

$$\mathrm{E}^{X_1}\lambda = \mathrm{E}^{X_1}\mathrm{E}^{X_2}\lambda \quad \text{for all } \lambda \in X^+.$$

Hence, for an arbitrary choice of basis $x_+(0)$ in X_+, we have

$$\mathrm{E}^{X_1}x_+(0) = \mathrm{E}^{X_1}\mathrm{E}^{X_2}x_+(0),$$

which is equivalent to (4.13) with $x_1(0)$ and $x_2(0)$ being the corresponding bases in X_1 and X_2. The fact that $X_1 \le X$ follows immediately from $\mathrm{E}^{S_1} = \mathrm{E}^{S_1}\mathrm{E}^S$ and [18, Lemma 6.7]. The relation $X \le X_2$ follows analogously from $\bar{S}_2 \subset \bar{S}$. □

LEMMA 4.3. *Let $X \sim (S, \bar{S})$, $X_1 \sim (S_1, \bar{S}_1)$, and $X_2 \sim (S_2, \bar{S}_2)$ be minimal Markovian splitting subspaces such that $S_1 \subset S$, and $\bar{S}_2 \subset \bar{S}$. Then*

$$E^{S_1 \vee \bar{S}_2}X \subset X_1 \vee X_2. \tag{4.16}$$

Proof. Applying the projection operator E^S to $H^+ \subset \bar{S}_2 \subset \bar{S}$ we obtain

$$\mathrm{E}^S H^+ \subset \mathrm{E}^S\bar{S}_2 \subset \mathrm{E}^S\bar{S} = X.$$

But since X is observable, $X = \mathrm{E}^S H^+$, and therefore

$$X = \mathrm{E}^S\bar{S}_2. \tag{4.17}$$

Moreover, since $S_1 \subset S$, we have $\mathrm{E}^{S_1}H^+ = \mathrm{E}^{S_1}\mathrm{E}^S H^+$, and therefore, since X_1 and X are both observable,

$$X_1 = \mathrm{E}^{S_1}X, \tag{4.18}$$

which together with (4.17) yields

$$X_1 = \mathrm{E}^{S_1}\bar{S}_2. \tag{4.19}$$

Now, it is well known and easy to check that the orthogonal decomposition

$$A = (\mathrm{E}^A B) \oplus (A \cap B^\perp) \tag{4.20}$$

holds for all pairs of subspaces A, B. Therefore, in view of (4.19), we have

$$S_1 = X_1 \oplus (S_1 \cap \bar{S}_2^{\perp}). \tag{4.21}$$

A completely symmetric argument yields

$$\bar{S}_2 = X_2 \oplus (\bar{S}_2 \cap S_1^{\perp}). \tag{4.22}$$

Therefore, since $X_1 \perp S_1^{\perp}$ and $X_2 \perp \bar{S}_2^{\perp}$, we have

$$S_1 \vee \bar{S}_2 = (S_1 \cap \bar{S}_2^{\perp}) \oplus (X_1 \vee X_2) \oplus (\bar{S}_2 \cap S_1^{\perp}). \tag{4.23}$$

To prove (4.16), take any $\xi \in X$. Then

$$\xi - \mathrm{E}^{S_1}\xi \perp S_1 \supset S_1 \cap \bar{S}_2^{\perp}$$

and, by (4.18) and (4.21),

$$\mathrm{E}^{S_1}\xi \in X_1 \perp S_1 \cap \bar{S}_2^{\perp}.$$

Consequently, $\xi \perp S_1 \cap \bar{S}_2^{\perp}$. In the same way we show that $\xi \perp \bar{S}_2 \cap S_1^{\perp}$, and therefore it follows from (4.23) that

$$\mathrm{E}^{S_1 \vee \bar{S}_2}\xi \in X_1 \vee X_2,$$

establishing (4.16). □

LEMMA 4.4. *Let $X \sim (S, \bar{S})$, $X_1 \sim (S_1, \bar{S}_1)$, and $X_2 \sim (S_2, \bar{S}_2)$ be minimal Markovian splitting subspaces such that $S_1 \subset S$ and $\bar{S}_2 \subset \bar{S}$, and let $x(0)$, $x_1(0)$, and $x_2(0)$ be a uniform choice of bases in X, X_1, and X_2 with covariances P, P_1, and P_2, respectively. Then*

$$E^{X_1 \vee X_2}x(0) = (I - L)x_1(0) + Lx_2(0) \tag{4.24}$$

for any $n \times n$ matrix solution L of the linear system of equations

$$P - P_1 = L(P_2 - P_1). \tag{4.25}$$

Proof. Setting $\hat{x}(0) := \mathrm{E}^{X_1 \vee X_2}x(0)$, we have

$$\hat{x}(0) = Kx_1(0) + Lx_2(0) \tag{4.26}$$

for some $n \times n$ matrices K and L. By construction, $a'[x(0) - \hat{x}(0)] \perp X_1 \vee X_2$ for all $a \in \mathbb{R}^n$, which in particular implies that

(i) $a'[x(0) - \hat{x}(0)] \perp X_1$ for all $a \in \mathbb{R}^n$,
(ii) $a'[x(0) - \hat{x}(0)] \perp X_2$ for all $a \in \mathbb{R}^n$.

Condition (i) together with (4.26) yields

$$\mathrm{E}\{x(0)x_1(0)'\} - KP_1 - L\mathrm{E}\{x_2(0)x_1(0)'\} = 0.$$

But from Lemma 4.2 it follows that

$$\mathrm{E}\{x_2(0)x_1(0)'\} = P_1 \quad \text{and} \quad \mathrm{E}\{x(0)x_1(0)'\} = P_1,$$

and therefore, since P_1 is nonsingular,

$$K = I - L. \tag{4.27}$$

In the same way, condition (ii) implies that

$$P = KP_1 + LP_2, \tag{4.28}$$

where again we have used Lemma 4.2 to see that

$$\mathrm{E}\{x(0)x_2(0)'\} = P \quad \text{and} \quad \mathrm{E}\{x_1(0)x_2(0)'\} = P_1.$$

Then (4.24) and (4.25) follow from (4.26)–(4.28).

To show that any solution L of (4.25) yields the same estimate $\hat{x}(0)$, let L_1 and L_2 be any two such solutions and let $\hat{x}_1(0)$ and $\hat{x}_2(0)$ be the corresponding estimates (4.26). Then

$$(L_1 - L_2)(P_2 - P_1) = 0 \tag{4.29}$$

and

$$\hat{x}_1(0) - \hat{x}_2(0) = (L_1 - L_2)\,[x_2(0) - x_1(0)]\,. \tag{4.30}$$

Equating the covariances of each side in (4.30), equation (4.29) implies that $\hat{x}_1(0) = \hat{x}_2(0)$, as claimed. □

This immediately yields the following representation formula for the interpolation estimate.

THEOREM 4.5. *Given the stochastic system* (4.1) *and* $t_0, t_1 \in \mathbb{Z}$ *such that* $t_0 < t_1$, *the state estimate*

$$\hat{x}(t \mid t_0, t_1) = E\{x(t) \mid y(s), s \in \mathbb{Z}; x(\tau), \tau \in (-\infty, t_0] \vee [t_1, \infty)\} \tag{4.31}$$

is given by

$$\hat{x}(t \mid t_0, t_1) = [I - L(t_0 - t, t_1 - t)]\, x^{(t_0 - t)}(t) + L(t_0 - t, t_1 - t) x^{(t_1 - t)}(t), \tag{4.32}$$

where $\{x^{(k)} \mid k \in \mathbb{Z}\}$ *is the estimation sequence* (4.6)–(4.7) *corresponding to* x *with covariances* $\{P^{(k)} \mid k \in \mathbb{Z}\}$ *and* $L(\tau, s)$ *is an arbitrary solution of*

$$P - P^{(\tau)} = L(\tau, s)\left[P^{(s)} - P^{(\tau)}\right]. \tag{4.33}$$

It remains to design a procedure for determining the estimation sequence $\{x^{(k)} \mid k \in \mathbb{Z}\}$. We shall address this question next. For this we need the following important consequence of the Markov property.

THEOREM 4.6. *The state estimate* (4.31) *depends only on the data from the interval* $[t_0, t_1]$ *or, more precisely, on* $x(t_0)$, $x(t_1)$, *and* $y(t)$, $t = t_0, t_0 + 1, \ldots, t_1$. *In particular,*

$$x^{(t_0 - t)}(t) := E\{x(t) \mid x(t_0), y(t_0), \ldots, y(t-1)\} \quad \textit{for } t > t_0, \tag{4.34}$$

$$x^{(t_1 - t)}(t) := E\{x(t) \mid y(t), \ldots, y(t_1), x(t_1)\} \quad \textit{for } t \le t_1, \tag{4.35}$$

where $\{x^{(k)}; k \in \mathbb{Z}\}$ *is the sequence of estimation processes defined by* (4.6) *and* (4.7).

Proof. Let $X \sim (S, \bar{S})$ be the splitting subspace corresponding to the state process x. In view of the definition (4.4), the first statement (4.34) is equivalent to

$$\mathrm{E}^{H^-_{k-1}(y) \vee X^-} \xi = \eta \quad \text{for all } \xi \in U^k X \text{ and } k \ge 0, \tag{4.36}$$

where

$$\eta := \mathrm{E}^{\{y(0),\dots,y(k-1)\}\vee X}\xi.$$

The original statement is obtained from (4.36) by merely applying the shift U^{t_0}. To prove (4.36) first note that, since $S = H^- \vee X^-$,

$$\begin{aligned} H_{k-1}^- \vee X^- &= \{y(0),\dots,y(k-1)\} \vee S \\ &= [\{y(0),\dots,y(k-1)\} \vee X] \oplus [S \ominus X]\,. \end{aligned}$$

To see this, note that $\{y(0),\dots,y(k-1)\} \subset H^+ \subset \bar{S} \perp S \ominus X$. Moreover, $\xi \in U^k X \subset U^k \bar{S} \subset \bar{S}$, and hence $\xi \perp S \ominus X$, which implies (4.36). A completely symmetric argument yields (4.35). □

Note that (4.34) and (4.35) are really forward and backward Kalman estimates initiated at $x(t_0)$ and $x(t_1)$, respectively, enabling us to use Kalman filtering techniques to generate them. Due to the fact that the initial conditions are states, these Kalman filters will have some remarkable properties, especially in the regular case when the reversibility condition (3.32) holds. This will be further discussed below.

The estimate (4.34) is generated by the recursion

$$\text{(4.37)}\quad \begin{cases} x^{(t_0-t)}(t) = Ax^{(t_0-t+1)}(t-1) + K^{(t_0-t)}\left[y(t-1) - Cx^{(t_0-t+1)}(t-1)\right], \\ x^{(0)}(t_0) = x(t_0), \end{cases}$$

where

$$K^{(-k)} = (\bar{C}' - AP^{(-k)}C')(\Lambda_0 - CP^{(-k)}C')^\sharp.$$

Here $^\sharp$ denotes pseudoinverse, and the state covariance

$$P^{(-k)} = \mathrm{E}\{x^{(-k)}(0)x^{(-k)}(0)'\}$$

is given by the matrix Riccati equation

$$\text{(4.38)}\quad \begin{cases} P^{(-k-1)} = AP^{(-k)}A' + (\bar{C}' - AP^{(-k)}C')(\Lambda_0 - CP^{(-k)}C')^\sharp(\bar{C}' - AP^{(-k)}C')', \\ P^{(0)} = P. \end{cases}$$

Note that this is the (invariant) formulation of the Kalman filter used in stochastic realization theory [1, 8, 16].

In the same way, the estimate (4.35) can be generated by a backward Kalman filter applied to the backward model

$$\text{(4.39)}\quad \begin{cases} \bar{x}(t-1) = A'\bar{x}(t) + \bar{B}\bar{u}(t-1), \\ y(t-1) = \bar{C}\bar{x}(t) + \bar{D}\bar{u}(t-1) \end{cases}$$

of X. Using a similar calculation as that in the forward direction, it is not hard to see that the process $\bar{x}^{(k)}(t) = [P^{(k)}]^{-1}x^{(k)}(t)$ is the solution of the backward Kalman filter

$$\text{(4.40)}\quad \begin{cases} \bar{x}^{(t_1-t)}(t) = A'\bar{x}^{(t_1-t-1)}(t+1) + \bar{K}^{(t_1-t-1)}\left[y(t) - \bar{C}\bar{x}^{(t_1-t-1)}(t)\right], \\ \bar{x}^{(0)}(t_1) = x(t_1), \end{cases}$$

where

$$\bar{K}^{(k)} = (C' - A'\bar{P}^{(k)}\bar{C}')(\Lambda_0 - \bar{C}\bar{P}^{(k)}\bar{C}')^\sharp$$

and the backward covariance matrix $\bar{P}^{(k)} = (P^{(k)})^{-1}$ is given by the matrix Riccati equation

$$\begin{cases} \bar{P}^{(k+1)} = A'\bar{P}^{(k)}A + (C' - A'\bar{P}^{(k)}\bar{C}')(\Lambda_0 - \bar{C}\bar{P}^{(k)}\bar{C}')^\sharp(C' - A'\bar{P}^{(k)}\bar{C}')', \\ \bar{P}^{(0)} = P^{-1}. \end{cases} \tag{4.41}$$

Then the process $x^{(t_1-t)}$ is given by

$$x^{(t_1-t)}(t) = (\bar{P}^{(t_1-t)})^{-1}\bar{x}^{(t_1-t)}(t),$$

defining $x^{(k)}$ for $k \geq 0$.

At least in the regular case, the inverse $(\Lambda_0 - CP^{(-k)}C')^{-1}$ will exist for all $k \in \mathbb{Z}$ (Proposition 3.11) and the pseudoinverses can be replaced with inverses. In the regular case we also have the reversibility property $\bar{\sigma} = \sigma^{-1}$ (Corollary 3.14) leading to (3.33). This useful property can be expressed in terms of estimation processes as

$$\mathrm{E}\{x(t) \mid H_{[t,t_1]}(y), x^{(t_0-t_1)}(t_1)\} = \mathrm{E}\{x(t) \mid H_{[t_0,t-1]}(y), x(t_0)\} = x^{(t_0-t)}(t), \tag{4.42}$$

i.e., tying together forward and backward estimation. This relation illustrates an important property of the Kalman recursions (4.37) and (4.40), namely, that a consecutive application of forward and backward Kalman filtering brings us back through the same sequence of state processes of totally ordered stochastic realizations. This remarkable fact, which is due to the invertibility of the operator σ, can also be justified by elementary calculations expressing $x^{(t_0-t+1)}(t-1)$ in terms of $x^{(t_0-t)}(t)$ and $y(t-1)$ in (4.37), leading to a backward Kalman filter which is an extension of (4.40) for negative $k = t_0 - t$. Similarly (4.40) can be reversed to give a forward Kalman filter identical to (4.37) for positive $k = t_1 - t$.

Given a stochastic realization (4.1) of y and a corresponding splitting subspace $X \sim (S, \bar{S})$, we have thus constructed a sequence of splitting subspaces $\{X^{(k)}; k \in \mathbb{Z}\}$ with bases

$$x^{(k)}(0) = \begin{cases} \mathrm{E}^{H^- \vee (U^k S)}x(0), & k \leq 0, \\ P^{(k)}P^{-1}\mathrm{E}^{H^+ \vee (U^k \bar{S})}x(0), & k \geq 0, \end{cases} \tag{4.43}$$

which are tied together by the Kalman filtering recursions (4.37) and (4.40). Each such basis vector defines a vector process

$$x^{(k)}(t) = U^k x^{(k)}(0),$$

which is the state process of a (forward) realization

$$\begin{cases} x^{(k)}(t+1) = Ax^{(k)}(t) + B^{(k)}u^{(k)}(t), \\ y(t) = Cx^{(k)}(t) + D^{(k)}u^{(k)}(t) \end{cases} \tag{4.44}$$

connected with a spectral factor

$$W^{(k)}(z) = C(zI - A)^{-1}B^{(k)} + D^{(k)}. \tag{4.45}$$

It is a manifestation of the fact that (4.43) is a uniform choice of bases for the splitting subspaces $\{X^{(k)} \mid k \in \mathbb{Z}\}$ in the sense defined in section 2, and it is also easy to check that the system matrices A and C remain constant for all $k \in \mathbb{Z}$, while $B^{(k)}$, $D^{(k)}$, and $P^{(k)}$ will vary. We shall not need to determine $\{B^{(k)}\}$ and $\{D^{(k)}\}$, but we note that this is easy to do either from the Riccati equation (4.37) or by means of a "fast algorithm" formulated directly in terms of $\{B^{(k)}, D^{(k)}\}$ as reported in Badawi [2].

Remark 4.7. Let us point out that the Riccati equation (4.38) can be written in the following form:

$$P^{(-k-1)} = P^{(-k)} - B^{(-k)}\{I - (D^{(-k)})'[D^{(-k)}(D^{(-k)})']^{\sharp}D^{(-k)}\}(B^{(-k)})'.$$

The last term is nothing other than the covariance matrix of that part of the noise in the state-space equation, i.e., of $B^{(-k)}u^{(-k)}$, which cannot be explained using the noise in the corresponding observation equation, i.e., via $D^{(-k)}u^{(-k)}$.

A similar statement can be formulated for the Riccati equation (4.41).

In the next section we show that, in the regular case, all spectral factors $\{W^{(k)} \mid k \in \mathbb{Z}\}$ have the same zeros, and in section 8 we demonstrate that this is no longer the case in the nonregular case.

5. The zero structure of the estimation sequence in the regular case. Let us recall that $\lambda \in \mathbb{C}$ is an *(invariant) zero* of a spectral factor

$$W(z) = C(zI - A)^{-1}B + D$$

if there are row vectors a and b so that

$$\begin{bmatrix} a & b \end{bmatrix} \begin{bmatrix} A - \lambda I & B \\ C & D \end{bmatrix} = 0$$

or, in other words,

$$\begin{bmatrix} a & b \end{bmatrix} \begin{bmatrix} A & B \\ C & D \end{bmatrix} = \begin{bmatrix} \lambda a & 0 \end{bmatrix}. \tag{5.1}$$

Here a is called a *zero direction* (of order one) of W. In the regular case, when $DD' > 0$, we may eliminate b in these equations to obtain

$$\begin{cases} a\Gamma = \lambda a, \\ aB_2 = 0, \end{cases}$$

where

$$\begin{aligned} \Gamma &:= A - BD'(DD')^{-1}C, \\ B_2 &:= B - BD'(DD')^{-1}D, \end{aligned}$$

showing that a is perpendicular to the reachability space

$$\langle \Gamma \mid B_2 \rangle = \operatorname{Im}(B_2, \Gamma B_2, \Gamma^2 B_2, \dots).$$

More generally, the zero directions (of any order) of W are defined using the Jordan structure of Γ. Then it can be proven that the orthogonal complement $\langle \Gamma \mid B_2 \rangle^{\perp}$ of this space in $\mathbb{R}^n$ is spanned by the *zero directions* of W. Hence, if Π is a matrix whose rows form a basis in $\langle \Gamma \mid B_2 \rangle^{\perp}$, i.e.,

$$\ker \Pi = \langle \Gamma \mid B_2 \rangle, \tag{5.2}$$

then there is a matrix Λ such that

$$\begin{cases} \Pi\Gamma = \Lambda\Pi, \\ \Pi B_2 = 0. \end{cases} \tag{5.3}$$

Conversely, if Π is a matrix satisfying (5.3), then

$$\ker \Pi \supset \langle \Gamma \mid B_2 \rangle. \tag{5.4}$$

This fact can also be expressed in terms of a generalization of (5.1): Relation (5.4) is equivalent to the existence of matrices Λ and M so that

$$\begin{bmatrix} \Pi & -M \end{bmatrix} \begin{bmatrix} A & B \\ C & D \end{bmatrix} = \begin{bmatrix} \Lambda\Pi & 0 \end{bmatrix}. \tag{5.5}$$

The row vectors of the maximal solution Π satisfying (5.2) (in the sense of having maximal rank) are the generalized zero directions, and the eigenvalues of the corresponding matrix Λ are of course precisely the finite zeros of W.

Remark 5.1. The matrix equation (5.5) is the appropriate generalization of (5.6) also in the nonregular case to be discussed in sections 7 and 8; see [20]. Note, however, that W may have zeros at infinity in the nonregular case, so the eigenvalues of Λ corresponding to the maximal solution of (5.5) are here the *finite* zeros of W.

The following lemma enables us to characterize the zero directions in terms of a connection between the state x and the output y.

LEMMA 5.2. *A matrix* Π *satisfies* (5.3) *if and only if there are matrices* Λ *and* M *such that*

$$\Pi x(t+1) = \Lambda\Pi x(t) + My(t). \tag{5.6}$$

We shall here give a proof which exhibits the connection between the Γ-matrix and the zero directions and which works in the present regular case. In section 8, we shall provide an alternative proof which also works in the nonregular case—in fact, even when the Γ-matrix cannot be defined.

Proof. As mentioned in the proof of Corollary 3.7, the state equation can be reformulated in the form

$$x(t+1) = \Gamma x(t) + BD'(DD')^{-1}y(t) + B_2u(t), \tag{5.7}$$

which, in the present regular case, is a unique decomposition of $x(t+1)$ in terms of $x(t)$, $y(t)$ and $B_2u(t)$. Hence

$$\Pi x(t+1) = \Pi\Gamma x(t) + \Pi BD'(DD')^{-1}y(t) + \Pi B_2u(t)$$

so that if Π satisfies (5.3), then (5.6) is also satisfied with $M = \Pi BD'(DD')^{-1}$. Conversely, if there are Λ and M so that (5.6) holds, then the uniqueness of decomposition (5.7) implies that (5.3) holds. ∎

Remark 5.3. Since Λ has no zero eigenvalues in the regular case, (5.6) may be written

$$\Pi P\bar{x}(t-1) = \Lambda^{-1}\Pi P\bar{x}(t) - \Lambda^{-1}My(t-1),$$

showing that the zeros of $\bar{W}(z^{-1})$ are precisely the eigenvalues of Λ^{-1}. Consequently, the forward and the backward models have the same zeros although the zero directions are transformed by the covariance matrix P. In fact, introducing the matrix

$$\bar{\Gamma} = A' - \bar{B}\bar{D}'(\bar{D}\bar{D}')^{-1}\bar{C},$$

the zeros of $\bar{W}$ are connected to the reciprocals of the eigenvalues of $\bar{\Gamma}$ in a manner analogous to (5.3).

Let us note the similarity between (5.6) and (3.4). In Theorem 3.1 we proved that (3.4) implies that $X \subset H_0$. In view of this, it is not surprising that (5.6) characterizes the subspace $X \cap H_0$. Recall that

$$X \cap H_0 = (X \cap X_-) \vee (X \cap X_+), \tag{5.8}$$

where the sum is direct if and only if $H^- \cap H^+ = 0$, i.e., if and only if

$$P_+ - P_- > 0 \tag{5.9}$$

[13, Lemma 2.9]. We note that $X \cap X_-$ is connected to the stable zeros of W (including the zeros on the unit circle) and that $X \cap X_+$ is connected to the antistable zeros (again including the zeros on the unit circle). If (5.9) holds, these sets of zeros are disjoint, there being no zeros on the unit circle.

As explained in [13], the subspaces $\ker(P - P_-)$ and $\ker(P_+ - P)$ are isomorphic to the subspaces $X \cap X_-$ and $X \cap X_+$, respectively, under the bijection $a \mapsto a'x(0)$. Based on these observations it can be proven that the zeros of W form a subset of those of W_- and $\bar{W}_+$. Let us collect the statements about the zeros of W in the following theorem. Proofs can be found in [13, 19, 29].

THEOREM 5.4. *The subspace* $\ker(P-P_-)$ *is invariant under* Γ'_- *and* Γ'. *Moreover,*

$$\Gamma'|_{\ker(P-P_-)} = \Gamma'_-|_{\ker(P-P_-)}. \tag{5.10}$$

The stable zeros of W *and* $\bar{W}$ *(including the ones on the unit circle) are the eigenvalues of* (5.10)*, and the corresponding zero directions of* W *span the subspace* $\ker(P - P_-)$. *Similarly,* $\ker(\bar{P} - \bar{P}_+)$ *is invariant under* $\bar{\Gamma}'_+$ *and* $\bar{\Gamma}'$. *Moreover,*

$$\bar{\Gamma}'|_{\ker(\bar{P}-\bar{P}_+)} = \bar{\Gamma}'_+|_{\ker(\bar{P}-\bar{P}_+)}. \tag{5.11}$$

The antistable zeros of W *and* $\bar{W}$ *(including the ones on the unit circle) are the reciprocals of the eigenvalues of* (5.11)*, and the corresponding zero directions of* $\bar{W}$ *span the subspace* $\ker(\bar{P} - \bar{P}_+)$.

Note that in the nonregular case, to be considered in sections 7 and 8, the matrix Γ may not be well defined for all X. Nevertheless all other statements of the theorem remain true.

To obtain coordinate-free versions of Γ' and $\bar{\Gamma}'$ we first observe that, in the regular case and with Π maximal so that $\ker \Pi = \langle \Gamma \mid B_2 \rangle$, (5.6) is equivalent to

$$U(X \cap H_0) \subset X \cap H_0 + \{y(0)\}, \tag{5.12}$$

where the sum is direct because of the regularity condition (iii)$'$ of Proposition 3.11. Similarly,

$$U^{-1}(X \cap H_0) \subset X \cap H_0 + \{y(-1)\}. \tag{5.13}$$

Now, following [13], let us introduce the *zero dynamics operators* in the regular case.

DEFINITION 5.5 (regular case). *Let the operators* $G : X \cap H_0 \to X \cap H_0$ *and* $\bar{G} : X \cap H_0 \to X \cap H_0$ *be defined as*

$$G = \pi U|_{X \cap H_0} \tag{5.14}$$

and

$$\bar{G} = \bar{\pi} U^{-1}|_{X \cap H_0}, \tag{5.15}$$

where $\pi : (X \cap H_0) + \{y(0)\} \to X \cap H_0$ *and* $\bar{\pi} : (X \cap H_0) + \{y(-1)\} \to X \cap H_0$ *are the oblique projectors projecting parallel to* $\{y(0)\}$ *and* $\{y(-1)\}$*, respectively.*

In view of Definition 5.5, (5.10) and (5.11) may be written

$$G|_{X \cap X_-} = G_-|_{X \cap X_-}$$

and

$$\bar{G}|_{X \cap X_+} = \bar{G}_+|_{X \cap X_+},$$

respectively. Moreover, $X \cap X_-$ is invariant under both G and G_-, and $X \cap X_+$ under both $\bar{G}$ and $\bar{G}_+$. In the nonregular case, the operators G and $\bar{G}$ may not be defined on all of $X \cap H_0$ but only on a subset of it, a circumstance manifest in the fact that Γ and $\bar{\Gamma}$ cannot be defined as above. However, G_- and $\bar{G}_+$ are always defined as in the regular case. This will be further discussed in section 7.

Let us now return to the estimation sequence $\{x^{(k)} \mid k \in \mathbb{Z}\}$. The following theorem ensures that no zeros are being lost when we move along the sequence $\{W^{(k)}\}$ from $k = 0$ through negative k.

THEOREM 5.6. *If* Π *is a matrix of zero directions of* $W^{(k)}$*, it is also a matrix of zero directions for* $W^{(k-j)}$ *for* $j = 0, 1, 2, \ldots$*. Moreover, the zeros are preserved.*

Proof. Since Π is a zero direction of $W^{(k)}$, there is a matrix Λ such that

$$\Pi\Gamma^{(k)} = \Lambda\Pi,$$

and therefore, in view of (5.7),

$$\Pi x^{(k)}(t) - \Lambda\Pi x^{(k)}(t-1) - \Pi K^{(k)} y(t-1) = 0, \tag{5.16}$$

because $B^{(k)}(D^{(k)})'[D^{(k)}(D^{(k)})']^{-1} = K^{(k)}$. Consequently, by (4.37),

$$\begin{aligned}
&\Pi x^{(k-1)}(t+1) - \Lambda\Pi x^{(k-1)}(t) - \Pi K^{(k-1)} y(t) \\
&\quad = \Pi\Gamma^{(k)} x^{(k)}(t) - \Lambda\Pi\Gamma^{(k)} x^{(k)}(t-1) - \Lambda\Pi K^{(k)} y(t-1) \\
&\quad = \Lambda \left[\Pi x^{(k)}(t) - \Lambda K^{(k)}(t-1) - \Pi K^{(k)} y(t-1) \right],
\end{aligned}$$

which is zero by (5.16). This together with (5.16) establishes that not only the zero directions but also the zeros are preserved, since the same matrix Λ can be used in each step. ∎

By symmetry we also have the following theorem.

THEOREM 5.6′. *If* $\bar{\Pi}$ *is a matrix of zero directions of* $W^{(k)}$*, it is also a matrix of zero directions for* $\bar{W}^{(k+j)}$ *for* $j = 0, 1, 2, \ldots$*. Moreover, the zeros are preserved.*

We observe that $W^{(k)}$ and $\bar{W}^{(k)}$ have the same zeros in view of Remark 5.3. Theorems 5.6 and 5.6′ show that there is no loss of zeros when we apply a forward or

backward Kalman filter step in (4.37) or (4.40). By the invertibility condition (3.32), all the elements in the sequence $\{W^{(k)} \mid k \in \mathbb{Z}\}$ must then have the *same* zeros. It is also easy to see that the zero directions are being preserved.

These results illustrate the fact that, in the regular case, all internal minimal splitting subspaces are fixed points of the operators σ and $\bar{\sigma}$. In fact, if X is internal, then so are σX and $\bar{\sigma} X$ by construction. Hence they have square spectral factors [16], which, by Theorems 5.6 and 5.6′, have the same zeros. Hence X, σX, and $\bar{\sigma} X$ must be the same. This analysis and the fact that in general there may be $X \in \mathfrak{X}_0$ which are not fixed points show that, in the nonregular case, the zeros may change as you move along the estimation sequence. The precise manner in which this happens is the topic of section 8.

Remark 5.7. Note that Theorem 5.6 and 5.6′ imply that in the regular case the stable and unstable zero directions, i.e., the subspaces $\ker(P^{(-k)} - P_-)$, $\ker(P_+ - P^{(-k)})$ and $\ker(\bar{P}^{(k)} - \bar{P}_+)$, $\ker(\bar{P}_+ - \bar{P}^{(k)})$, remain unchanged as k tends to ∞ in the forward and backward Riccati equations (4.38) and (4.41). In other words, the solutions of the Riccati recursions remain constant in the zero directions, providing a possibility of reducing the size of the Riccati equation. In fact, choosing coordinates so that the last basis vectors span

$$\ker(P^{(-k)} - P_-) \vee \ker(P_+ - P^{(-k)}) \subset \mathbb{R}^n,$$

the matrices $\{P^{(-k)}\}$ in the solution of the Riccati recursion (4.38) take the form

$$P^{(-k)} = \begin{bmatrix} P_{11}^{(-k)} & P_{12} \\ P_{12}' & P_{22} \end{bmatrix}, \tag{5.17}$$

where only the upper left matrix block varies with k. Then, substituting (5.17) into (4.38) we obtain a reduced-order Riccati equation of dimension $\nu \times \nu$ where $\nu = n - \dim(X \cap H_0)$. A completely symmetric argument can be applied to the backward Riccati recursion (4.41).

6. Output-induced subspaces. We have just seen that the matrices Γ and $\bar{\Gamma}$ play an important role in the analysis of the estimation sequence $x^{(k)}$. We have also pointed out that they are easily defined only in the regular case. Therefore, in this section we shall consider only their coordinate-free versions, G and $\bar{G}$, which have natural definitions in the general case.

In the regular case, considered in section 5, the zero dynamics operators G and $\bar{G}$ of a splitting subspace $X \in \mathfrak{X}$ were defined on all of its internal subspace $X \cap H_0$. This is possible due to the direct sum decompositions (5.12) and (5.13). In the nonregular case these decompositions will fail to exist as we demonstrate in section 7. Therefore, we must shrink the domains of the zero dynamics operators.

As demonstrated in [29], G can always be defined on $X \cap X_-$, yielding only the stable zeros (including those on the unit circle), and $\bar{G}$ can always be defined on $X \cap X_+$, producing only the antistable zeros (including those on the unit circle and those at infinity). In fact, this can also be seen from the following representations. (Also see [29, Lemma 5.1].)

LEMMA 6.1. *Let* $X \in \mathfrak{X}$. *Then*

$$U^{-1}(X \cap X_+) \subset (X \cap X_+) + \{y(-1)\} \tag{6.1}$$

and

$$U(X \cap X_-) \subset (X \cap X_-) + \{y(0)\}. \tag{6.2}$$

Proof. We prove (6.2). Then (6.1) follows by symmetry. Obviously,

$$U(X \cap X_-) \subset UX_- \subset H^- \vee \{y(0)\}.$$

Also $X \cap X_- \subset \bar{S} \cap \bar{S}_-$, which is U-invariant. Therefore,

$$U(X \cap X_-) \subset (H^- \vee \{y(0)\}) \vee (\bar{S} \cap \bar{S}_-).$$

But $\{y(0)\} \subset H^+ \subset \bar{S} \cap \bar{S}_-$ and $H^- \cap \bar{S} \cap \bar{S}_- = X \cap X_-$, implying (6.2). ∎

In this paper, however, we would like to define G and $\bar{G}$ on the largest possible spaces. We show that this can be done in such a way that the eigenvalues of G are precisely the finite zeros of X, and the eigenvalues of $\bar{G}$ are the reciprocals of the nonzero zeros of X (using the definition $1/\infty = 0$). Moreover, we want to know on which subspaces G and $\bar{G}$ are invertible so that they can be directly related to each other. This leads to the topic of *output-induced subspaces*, introduced in [13] in the continuous-time setting. We now define it in the discrete-time case. Since, in the nonregular discrete-time case, the covariance matrix of the observation noise of the model (4.1) may be singular, the definition used in the continuous-time case must be somewhat modified.

DEFINITION 6.2. *Let X be a Markovian splitting subspace. A subspace $Y \subset X$ is called* output induced *if*

(i) $Y \subset H_0$,

(ii) $UY \subset Y \vee \{y(0), y(1), \dots, y(k)\}$ *for some* $k \geq 0$,

(iii) $U^{-1}Y \subset Y \vee \{y(-1), y(-2), \dots, y(-k-1)\}$ *for some* $k \geq 0$.

We say that Y is strictly output induced *if it is output induced and k can be chosen to be zero in* (ii) *and* (iii).

The following proposition is an immediate consequence of the definition and the finite dimension of X.

PROPOSITION 6.3. *The sum of two output-induced (strictly output-induced) subspaces is also output induced (strictly output induced). There exist a maximal output-induced (strictly output-induced) subspace in the sense of subspace inclusion.*

Since any output-induced subspace Y satisfies

$$Y \subset X \cap H_0 = (X \cap X_-) \vee (X \cap X_+),$$

let us first consider the subspaces $X \cap X_-$ and $X \cap X_+$. These, of course, trivially satisfy condition (i), and, by Lemma 6.1, they also satisfy one of the conditions (ii) and (iii) with $k = 0$, as required in the definition of being strictly output induced. Next, we show that these subspaces also satisfy the remaining condition so that they are output induced, and we investigate under what conditions they are actually strictly output induced.

THEOREM 6.4. *Let $X \in \mathcal{X}$. Then the subspaces $X \cap X_+$ and $X \cap X_-$ are output induced. Moreover, $X \cap X_+$ is strictly output induced if and only if*

$$(\sigma X) \cap X_+ = X \cap X_+, \tag{6.3}$$

and $X \cap X_-$ is strictly output induced if and only if

$$(\bar{\sigma} X) \cap X_- = X \cap X_-. \tag{6.4}$$

Proof. First we prove that $X \cap X_+$ is output induced. To this end, in view of (6.1), it is enough to check that there exists a $k \leq \dim X$ such that

$$U(X \cap X_+) \subset (X \cap X_+) \vee \{y(0), y(1), \dots, y(k)\}. \tag{6.5}$$

Since

$$X \cap X_+ = X_{0-} \cap X_+,$$

where X_{0-} is the tightest lower internal bound [18, 13], we may without loss of generality assume that X is internal. By Theorem 3.3, there is a $k \leq \dim X$ such that

$$\sigma^k X = \sigma^{k+1} X.$$

Consequently,

$$U^{-k} S \subset S^{(-k)} \subset S^{(-k-1)} = H^- \vee (U^{-(k+1)} S),$$

from which we have

$$US \subset S \vee U^{k+1} H^-.$$

Taking intersection with H^+ in both sides and noting that $U(S \cap H^+) \subset (US) \cap H^+$, we have

$$\begin{aligned} U(S \cap H^+) &\subset [S \vee \{y(0), \dots, y(k)\}] \cap H^+ \\ &= (S \cap H^+) \vee \{y(0), \dots, y(k)\}. \end{aligned}$$

Then (6.5) follows from the fact that $X \cap X_+ = S \cap H^+$. In the same way, we prove that there is an $\ell \leq \dim X$ such that

$$U^{-1}(X \cap X_-) \subset (X \cap X_-) \vee \{y(-1), y(-2), \dots, y(-\ell - 1)\}, \tag{6.6}$$

implying together with (6.2) that $X \cap X_-$ is output induced.

To characterize the strictly output-induced property we prove that

$$(\sigma X) \cap X_+ = (X \cap X_+) \cap \left[\{y(-1)\} \vee U^{-1}(X \cap X_+)\right] \tag{6.7}$$

and that

$$(\bar{\sigma} X) \cap X_- = (X \cap X_-) \cap \left[\{y(0)\} \vee U(X \cap X_-)\right]. \tag{6.8}$$

To this end, let $X \sim (S, \bar{S})$, and note that $S^{(-1)} := H^- \vee U^{-1} S \subset S$. Hence

$$\begin{aligned} (\sigma X) \cap X_+ &= S^{(-1)} \cap X_+ = S^{(-1)} \cap S \cap X_+ \\ &= S^{(-1)} \cap X \cap X_+ = \left[\{y(-1)\} \vee U^{-1} S\right] \cap X \cap X_+ \end{aligned}$$

But since $U^{-1} S = U^{-1} X \oplus U^{-1} \bar{S}^\perp$ and $\{y(-1)\} \subset U^{-1} \bar{S} \perp U^{-1} \bar{S}^\perp$,

$$\begin{aligned} (\sigma X) \cap X_+ &= \left[(\{y(-1)\} \vee U^{-1} X) \oplus U^{-1} \bar{S}^\perp\right] \cap X \cap X_+ \\ &= (\{y(-1)\} \vee U^{-1} X) \cap X \cap X_+, \end{aligned}$$

because $X_+ \subset \bar{S} \subset U^{-1} \bar{S} \perp U^{-1} \bar{S}^\perp$. Moreover, if $\xi \in (\{y(-1)\} \vee U^{-1} X) \cap X_+$, then $\xi = \alpha + \beta$, where $\alpha \in \{y(-1)\} \subset U^{-1} H^+$ and $\beta \in U^{-1} X$. Since $\xi \in H^+ \subset U^{-1} H^+$, we must have $\beta \in U^{-1} H^+$ so that $\beta \in U^{-1}(X \cap H^+) = U^{-1}(X \cap X_+)$. Therefore (6.7) follows. A symmetric argument yields (6.8).

Now, (6.7) and (6.8) immediately imply that

$$(\sigma X) \cap X_+ = X \cap X_+ \iff U(X \cap X_+) \subset (X \cap X_+) \vee \{y(0)\} \tag{6.9}$$

and

$$(\bar{\sigma} X) \cap X_- = X \cap X_- \iff U^{-1}(X \cap X_-) \subset (X \cap X_-) \vee \{y(-1)\}, \tag{6.10}$$

concluding the proof. □

COROLLARY 6.5. *The subspace $X \cap H_0$ is the maximal output-induced subspace of $X \in \mathcal{X}$.*

Proof. This follows immediately from Theorem 6.4 and (5.8). □

COROLLARY 6.6. *The subspace $X_- \cap X_+$ is always strictly output induced.*

Proof. This follows either from Lemma 6.1 or from (6.3) and the fact that $\sigma X_- = X_-$. □

We are now in a position to connect the concept strictly output-induced subspaces to fixed points of σ and $\bar{\sigma}$.

COROLLARY 6.7. *An $X \in \mathcal{X}_0$ is a fixed point of σ if and only if $X \cap X_+$ is strictly output induced. Likewise, $X \in \mathcal{X}_0$ is a fixed point of $\bar{\sigma}$ if and only if $X \cap X_-$ is strictly output induced.*

Proof. In the end of the proof of Theorem 3.3 we pointed out that the internal Markovian splitting subspaces are uniquely determined by $X \cap X_+$. Observe that if $X \in \mathcal{X}_0$, then $\sigma X \in \mathcal{X}_0$. Consequently, Theorem 6.4 implies that $\sigma X = X$ if and only if $X \cap X_+$ is strictly output induced, and the rest follows by a symmetric argument. □

As we shall see in section 8, these conditions can be formulated in terms of the stable and unstable zeros of the spectral factor (1.2) corresponding to the splitting subspace X.

The notion of strictly output-induced subspaces enables us in some cases to characterize the limits $X^{(-\infty)}$ and $X^{(\infty)}$ of the sequence $\{X^{(k)} \mid k \in \mathbb{Z}\}$ defined in section 3. To this end, let us recall [18] that the *tightest internal bounds*, X_{0-} and X_{0+}, are the closest internal X such that

$$X_{0-} \leq X \leq X_{0+}.$$

More precisely,

$$X_{0-} := \sup\{X_0 \in \mathcal{X}_0 \mid X_0 \leq X\}$$

and

$$X_{0+} := \inf\{X_0 \in \mathcal{X}_0 \mid X \leq X_0\}.$$

COROLLARY 6.8. *Let $X \in \mathcal{X}$, and let X_{0-} and X_{0+} be its tightest internal bounds. Then*

$$X^{(-\infty)} = X_{0-}$$

if and only if $X \cap X_+$ is strictly output induced, and

$$X^{(\infty)} = X_{0+}$$

if and only if $X \cap X_-$ is strictly output induced.

Proof. Let us first recall that

$$S_{0-} = S \cap H_0 = H^- \vee (X \cap X_+)$$

(cf. [18, Lemma 6.11], [13]). Therefore Theorem 6.4 implies that X_{0-} is the lower tightest internal bound of σX also if and only if $X \cap X_+$ is strictly output induced. By induction, we then have that $\sigma^{-k}X \geq X_{0-}$ for $k = 0, 1, 2, \dots$ and hence that $X^{(-\infty)} \geq X_{0-}$. But $X^{(-\infty)} \in \mathfrak{X}_0$ (Theorem 3.3), and consequently $X^{(-\infty)} = X_{0-}$ follows from the tightness of the bound. The proof for the upper bound is analogous. □

Another consequence of Theorem 6.4 is that the splitting subspaces in the sequence $\{X^{(k)} \mid k \in \mathbb{Z}\}$ have the same tightest local frame [18] if and only if the internal subspace $X \cap H_0$ is strictly output induced. As we shall see in the next section, this is only true in the regular case.

COROLLARY 6.9. *A necessary and sufficient condition for all splitting subspaces in the family $\{X^{(k)} \mid k \in \mathbb{Z}\}$ to have the same tightest internal bounds is that $X \cap H_0$ is strictly output induced.*

Proof. The proof follows immediately from Corollary 6.8, Proposition 6.3, and (5.8). □

Theorem 6.4 also yields the following alternative characterizations of regularity.

COROLLARY 6.10. *The following conditions are equivalent to the regularity conditions of Propositions* 3.11 *and* 3.11′.

(vii) *X_+ is strictly output induced.*
(viii) *X_- is strictly output induced.*
(vii)′ *$X \cap X_+$ is strictly output induced for all $X \in \mathfrak{X}_0$.*
(viii)′ *$X \cap X_-$ is strictly output induced for all $X \in \mathfrak{X}_0$.*
(ix)′ *All $X \in \mathfrak{X}_0$ are strictly output induced.*
(x)′ *The internal subspace $X \cap H_0$ is strictly output induced for all $X \in \mathfrak{X}$.*

Proof. By Corollary 6.7, (vii)′ and (viii)′ are equivalent to conditions (v)′ and (vi)′ of Proposition 3.11′, and (vii) and (viii) are equivalent to conditions (v) and (vi) of Proposition 3.11. In view of Proposition 6.3, (ix)′ follows from (vii)′, (viii)′ and (3.13), and (x)′ follows from (vii)′, (viii)′, and (5.8). Clearly either (vii) or (viii) implies (ix)′ and (x)′. □

7. Invariant directions and the maximal strictly output-induced subspace. Proposition 6.3 states that, to each $X \in \mathfrak{X}$, there exists a maximal strictly output-induced subspace Y^*. In this section we construct Y^* explicitly. Let us recall that $Y \subset X \cap H_0$ is said to be strictly output induced if

$$UY \subset Y \vee \{y(0)\} \tag{7.1}$$

and

$$U^{-1}Y \subset Y \vee \{y(-1)\}. \tag{7.2}$$

To determine Y^*, we first construct the subspaces $Y, \bar{Y} \subset X \cap H_0$ satisfying (7.1) and (7.2), respectively, which are maximal in the sense of subspace inclusion and show that Y^* is precisely the intersection of these.

To this end, we design a procedure which is akin to the one used in geometric control theory [31] to construct the maximal output-nulling subspace. More precisely, define two sequences of subspaces $\{Y_0, Y_1, Y_2, \dots\}$ and $\{\bar{Y}_0, \bar{Y}_1, \bar{Y}_2, \dots\}$ by

$$Y_k = (\sigma^k X) \cap X \cap H_0 \tag{7.3}$$

and

$$\bar{Y}_k = (\bar{\sigma}^k X) \cap X \cap H_0 \tag{7.4}$$

and show that they converge monotonically to Y and $\bar{Y}$, respectively, in finitely many steps. As will be seen below these are precisely the largest spaces on which the zero dynamics operators may be defined. Obviously, $Y_0 = \bar{Y}_0 = X \cap H_0$. We now give alternative characterizations of these sequences and obtain iterative solutions of (7.1) and (7.2), respectively.

LEMMA 7.1. *For each $k = 1, 2, 3, \ldots$ the subspaces* (7.3) *and* (7.4) *can be written*

$$Y_k = \{\xi \in X \cap H_0 \mid U^k \xi \in (X \cap H_0) \vee \{y(0), \ldots, y(k-1)\}\} \tag{7.5}$$

and

$$\bar{Y}_k = \{\xi \in X \cap H_0 \mid U^{-k} \xi \in (X \cap H_0) \vee \{y(-1), \ldots, y(-k)\}\}, \tag{7.6}$$

respectively. Moreover, the sequences $\{Y_k\}$ and $\{\bar{Y}_k\}$ satisfy the recursions

$$Y_{k+1} = \{\xi \in Y_k \mid U\xi \in Y_k \vee \{y(0)\}\} \tag{7.7}$$

and

$$\bar{Y}_{k+1} = \{\xi \in \bar{Y}_k \mid U^{-1}\xi \in \bar{Y}_k \vee \{y(-1)\}\} \tag{7.8}$$

for $k = 1, 2, 3, \ldots$.

Proof. We prove only (7.5) and (7.7), (7.6) and (7.8) following by a symmetric argument.

To prove (7.5), observe that

$$\sigma^k X = \mathrm{E}^{(U^{-k}S) \vee H^-} X \subset U^{-k} X \vee \{y(-1), \ldots, y(-k)\} \tag{7.9}$$

in view of the decomposition

$$(U^{-k}S) \vee H^- = \left[(U^{-k}X) \vee \{y(-1), \ldots, y(-k)\}\right] \oplus U^{-k}\bar{S}^\perp$$

and the fact that $U^{-k}\bar{S}^\perp \subset \bar{S}^\perp \perp X$.

Consequently, if $\xi \in Y_k$, then $\xi \in X \cap H_0$ and

$$U^k \xi \in [X \vee \{y(0), \ldots, y(k-1)\}] \cap H_0 = (X \cap H_0) \vee \{y(0), \ldots, y(k-1)\}.$$

Conversely, if $\xi \in X \cap H_0$ and $U^k \xi \in (X \cap H_0) \vee \{y(0), \ldots, y(k-1)\}$, then

$$\mathrm{E}^{(U^{-k}S) \vee H^-} \xi = \mathrm{E}^{(U^{-k}X) \vee \{y(-1), \ldots, y(-k)\}} \xi = \xi,$$

proving that $\xi \in \sigma^k X$ so that $\xi \in Y_k$.

Concerning the proof of (7.7), first consider a $\xi \in Y_k$ such that $U\xi \in Y_k \vee \{y(0)\}$. By (7.5) we have

$$\begin{aligned} U^{k+1}\xi = U^k(U\xi) &\in U^k Y_k \vee \{y(k)\} \\ &\subset (X \cap H_0) \vee \{y(0), \ldots, y(k)\}, \end{aligned}$$

proving that $\xi \in Y_{k+1}$, as can be seen from (7.5).

Conversely, if $\xi \in Y_{k+1}$, then (7.5) implies that $U^{k+1}\xi$ has the representation

$$U^{k+1}\xi = \zeta + \lambda_0 + \lambda_1,$$

where $\zeta \in X \cap H_0$, $\lambda_0 \in \{y(0), \dots, y(k-1)\}$ and $\lambda_1 \in \{y(k)\}$. We want to prove that $U\xi - U^{-k}\lambda_1 \in Y_k$, which implies (7.7). To this end, we note that

$$U\xi - U^{-k}\lambda_1 = U^{-k}\zeta + U^{-k}\lambda_0.$$

The left member of this belongs to $\bar{S}$, while the right member belongs to S, implying that they are in X and hence in $X \cap H_0$. Moreover, in view of (7.5), the identity

$$U^k(U\xi - U^{-k}\lambda_1) = \zeta + \lambda_0$$

implies that $U\xi - U^{-k}\lambda_1 \in Y_k$ concluding the proof of (7.7). □

An immediate consequence of Lemma 7.1 is that

$$X \cap H_0 = Y_0 \supset Y_1 \supset Y_2 \supset \cdots \tag{7.10}$$

and that

$$UY_{k+1} \subset Y_k \vee \{y(0)\}. \tag{7.11}$$

Dually, we also have

$$X \cap H_0 = \bar{Y}_0 \supset \bar{Y}_1 \supset \bar{Y}_2 \supset \cdots \tag{7.12}$$

and

$$U^{-1}\bar{Y}_{k+1} \subset \bar{Y}_k \vee \{y(-1)\}. \tag{7.13}$$

Since $X \cap H_0$ is finite dimensional, the chain of inclusions (7.10) implies that there is a $k \leq \dim(X \cap H_0)$ such that $Y_{k+1} = Y_k$. Then (7.7) implies that $Y_\ell = Y_k$ for all $\ell \geq k$. Since $\dim(X \cap H_0) \leq \dim X := n$, we may refer to this subspace as Y_n. Clearly

$$UY_n \subset Y_n \vee \{y(0)\}.$$

Similarly, $\bar{Y}_n$ is the limit of $\{\bar{Y}_k\}$ and satisfies

$$U^{-1}\bar{Y}_n \subset \bar{Y}_n \vee \{y(-1)\}.$$

THEOREM 7.2. *The subspace Y_n is the maximal subspace of $X \cap H_0$ with the property*

$$UY \subset Y \vee \{y(0)\}, \tag{7.14}$$

and $\bar{Y}_n$ is the maximal subspace in $X \cap H_0$ such that

$$U^{-1}\bar{Y} \subset \bar{Y} \vee \{y(-1)\}. \tag{7.15}$$

In the regular case, $Y_n = \bar{Y}_n = X \cap H_0$.

Proof. We have already proved that Y_n and $\bar{Y}_n$ satisfy (7.14) and (7.15), respectively. To prove maximality, consider a $Y \subset X \cap H_0 = Y_0$ satisfying (7.14). We prove

by induction that $Y \subset Y_k$ for $k = 0, 1, 2, \ldots$. To this end, assume that $Y \subset Y_i$ and show that $Y \subset Y_{i+1}$. If $\xi \in Y$, then

$$U\xi \in Y \vee \{y(0)\} \subset Y_i \vee \{y(0)\}.$$

Consequently, in view of (7.7), $\xi \in Y_{i+1}$, as claimed. The maximality of $\bar{Y}_n$ is proven in the same way. The last statement follows from (5.12) and (5.13). □

Remark 7.3. Applying the orthogonal projection operator E^X to (7.14), we obtain

$$FY \subset Y \vee \mathrm{E}^X\{y(0)\}, \tag{7.16}$$

where F is the compressed shift operator $F := \mathrm{E}^X U|_X$. From the systems equations (1.1) one can infer that F has the matrix representations A' in the corresponding basis and that $\mathrm{E}^X\{y(0)\}$ has the matrix representations C'. Therefore, analogously to the continuous-time case [13], (7.16) is a stochastic version of (A', C')-invariance in geometric control theory [31, 4]. This connection to geometric control theory is elaborated upon in [29]. In this context, we note that a similar application of E^X to (7.11) yields

$$FY_{k+1} \subset Y_k \vee \mathrm{E}^X\{y(0)\},$$

which should be compared to the algorithm in geometric control theory to determine the maximal output-nulling subspace $\mathcal{V}^*$.

Now, referring back to the regular case and (5.12) and (5.13), we recall that, in this case, $X \cap H_0$ satisfies (7.14) and (7.15) with direct sum. This enabled us to define the operators G and $\bar{G}$. In the general case $X \cap H_0 \cap \{y(0)\}$ and $X \cap H_0 \cap \{y(-1)\}$ may be nontrivial subspaces. Nevertheless, as we will prove below, Y_n and $\bar{Y}_n$ satisfy (7.14) and (7.15) with direct sum decomposition in the right member. This requires a deeper analysis of so-called *invariant directions* of a system representation (1.1) of X [5, 25, 26, 22].

More precisely, there are two kinds of invariant directions. An $a \in \mathbb{R}^n$ is a *predictable direction* if there is a positive integer k such that

$$a'x(0) \in \{y(-1), y(-2), \ldots, y(-k)\}. \tag{7.17}$$

The smallest k with this property is called the *order* of the invariant direction a. If a satisfies (7.17), the Kalman filter estimate $\hat{x}$ takes the form

$$a'\hat{x}(t) = a'x(t) = \sum_{i=1}^{r} c_i' y(t-i)$$

in that direction so that the estimation error becomes zero. This manifests itself in that the filtering Riccati equation can be reduced in dimension after a finite number of steps. A similar reduction occurs in the fast filtering algorithm [11]; see in particular [12]. It can be shown [22] that a is a predictable direction if and only if, for some $k \geq 0$,

$$a \in \ker(\Gamma_-')^k \cap \ker(P - P_-). \tag{7.18}$$

Dually, $a \in \mathbb{R}^n$ is a *smoothable direction* if there is a positive integer k such that

$$a'\bar{x}(0) \in \{y(0), y(1), \ldots, y(k-1)\}, \tag{7.19}$$

causing a reduction in the backward Kalman filtering algorithms. Again the smallest k with this property is the order of the invariant direction a, and

$$a \in \ker(\bar{\Gamma}'_+)^k \cap \ker(P_+ - P) \tag{7.20}$$

for some k is a necessary and sufficient condition for a to be a smoothable direction.

It can be seen from (7.18) and (7.20) that the order of an invariant direction cannot be larger than the dimension of X. Although the definition of invariant directions depends on the particular choice of coordinates in X, $a'x(0)$ and $a'\bar{x}(0)$ in the definitions (7.17) and (7.19) are independent of the coordinate system. Therefore we shall refer to these elements of X as the invariant directions of X.

Now, let $H^\square$ be the *frame space*

$$H^\square = X_- \vee X_+, \tag{7.21}$$

i.e., the closed linear hull of all internal subspaces $X \cap H_0$ as X ranges over $\mathfrak{X}$, and define the subspace

$$H_{0+} = H^\square \cap \{y(-n), \dots, y(n-1)\}.$$

In analogy with the continuous-time case [7], H_{0+} is called the *germ space* [22], since it contains all differences of y up to order n at $t = 0$.

PROPOSITION 7.4. *The germ space has the direct sum decomposition*

$$H_{0+} = X_- \cap \{y(-1), \dots, y(-n)\} + X_+ \cap \{y(0), \dots, y(n-1)\}. \tag{7.22}$$

Moreover, X_- contains no smoothable and X_+ no predictable directions.

Proof. The inclusion $\supset$ is trivial. To prove the other direction, note that, since y is purely nondeterministic, the two terms in (7.22) have a zero intersection, and every $\xi \in H_{0+}$ has a unique representation $\xi = \xi_- + \xi_+$ such that $\xi_- \in \{y(-1), \dots, y(-n)\} \subset H^-$ and $\xi_+ \in \{y(0), \dots, y(n-1)\} \subset H^+$. But in view of decomposition (4.20), ξ_- has an orthogonal decomposition $\xi_- = \hat{\xi}_- + \tilde{\xi}_-$ such that $\hat{\xi}_- \in X_-$ and $\tilde{\xi}_- \in H^- \cap (H^+)^\perp$, and ξ_+ can be written $\xi_+ = \hat{\xi}_+ + \tilde{\xi}_+$, where $\hat{\xi}_+ \in X_+$ and $\tilde{\xi}_+ \in H^+ \cap (H^-)^\perp$. Therefore, since

$$H_0 = \left[H^- \cap (H^+)^\perp\right] \oplus H^\square \oplus \left[H^+ \cap (H^-)^\perp\right],$$

the fact that $\xi = \tilde{\xi}_- + (\hat{\xi}_- + \hat{\xi}_+) + \tilde{\xi}_+ \in H^\square$ shows that $\tilde{\xi}_- = \tilde{\xi}_+ = 0$. Hence $\xi_- \in X_-$ and $\xi_+ \in X_+$, establishing the inclusion $\subset$. □

Consequently, the germ space is spanned by the predictable invariant directions in X_- and the smoothable invariant directions in X_+. Moreover, y is regular if and only if it has a trivial germ space.

PROPOSITION 7.5. *Let $X \in \mathfrak{X}$. Then*

$$X \cap H_{0+} = X \cap \{y(-1), \dots, y(-n)\} + X \cap \{y(0), \dots, y(n-1)\}; \tag{7.23}$$

i.e., $X \cap H_{0+}$ is spanned by the invariant directions of X. Moreover,

$$X \cap \{y(-1), \dots, y(-n)\} \subset X_- \cap \{y(-1), \dots, y(-n)\} \tag{7.24}$$

and

$$X \cap \{y(0), \dots, y(n-1)\} \subset X_+ \cap \{y(0), \dots, y(n-1)\}. \tag{7.25}$$

Proof. Let $X \sim (S, \bar{S})$. Relations (7.24) and (7.25) follow from the fact that $X \cap H^- = X \cap X_-$ and that $X \cap H^+ = X \cap X_+$, respectively. In view of this and Proposition 7.4, the inclusion $\supset$ in (7.23) is immediate. To prove $\subset$, take $\xi \in X \cap H_{0+}$. By Proposition 7.4, there is a unique decomposition $\xi = \xi_- + \xi_+$ such that $\xi_- \in X_- \cap \{y(-1), \dots, y(-n)\}$ and $\xi_+ \in X_+ \cap \{y(0), \dots, y(n-1)\}$. Hence it just remains to prove that $\xi_- \in X$ and $\xi_+ \in X$. To this end, note that $\xi_- \in H^- \subset S$ and $\xi \in X \subset S$, so we must have $\xi_+ \in S$. But $\xi_+ \in H^+ \subset \bar{S}$, so $\xi_+ \in S \cap \bar{S} = X$. Since $\xi \in X$, we must have $\xi_- \in X$ also. ∎

We have thus proved that all invariant directions of X_- are predictable and all the invariant directions of X_+ are smoothable, while an arbitrary X can have invariant directions of either kind. In view of (7.18), the predictable directions of X are also among the predictable directions of X_-. In the same way, (7.20) implies that the smoothable directions of X form a subspace of the smoothable directions of X_+ [22]. We call $X \cap \{y(-1), \dots, y(-n)\}$ the *predictable subspace* and $X \cap \{y(0), \dots, y(n-1)\}$ the *smoothable subspace* of X.

THEOREM 7.6. *Let $X \in \mathfrak{X}$. Then we have the following.*

(i) *The internal subspace $X \cap H_0$ of X has the direct-sum decomposition*

$$X \cap H_0 = Y_n + X \cap \{y(0), \dots, y(n-1)\}. \tag{7.26}$$

Moreover,

$$X \cap X_- \subset Y_n. \tag{7.27}$$

In particular, Y_n contains the predictable directions $X \cap \{y(-1), \dots, y(-n)\}$ of X.

(ii) *The internal subspace $X \cap H_0$ of X has the direct-sum decomposition*

$$X \cap H_0 = \bar{Y}_n + X \cap \{y(-1), \dots, y(-n)\}. \tag{7.28}$$

Moreover,

$$X \cap X_+ \subset \bar{Y}_n. \tag{7.29}$$

In particular, $\bar{Y}_n$ contains the smoothable directions $X \cap \{y(0), \dots, y(n-1)\}$ of X.

(iii) *The maximal strictly output-induced subspace of X is given by*

$$Y^* = Y_n \cap \bar{Y}_n. \tag{7.30}$$

Moreover,

$$Y_n = Y^* + X \cap \{y(-1), \dots, y(-n)\} \tag{7.31}$$

and

$$\bar{Y}_n = Y^* + X \cap \{y(0), \dots, y(n-1)\}. \tag{7.32}$$

In the regular case, $Y^* = X \cap H_0$ for all $X \in \mathfrak{X}$.

This theorem, the proof of which we defer to the end of the section, shows in particular that the internal subspace $X \cap H_0$ can be decomposed as

$$X \cap H_0 = X \cap \{y(-1), \dots, y(-n)\} + Y^* + X \cap \{y(0), \dots, y(n-1)\}, \tag{7.33}$$

i.e., as the direct sum of the subspace of predictable directions, the maximal strictly output-induced subspace, and the subspace of smoothable directions of X. In view

of Proposition 7.5, $X \cap H_0$ is also the direct sum of Y^* and the germ subspace of X, i.e.,

$$X \cap H_0 = Y^* + X \cap H_{0+}. \tag{7.34}$$

This has the following consequence.

COROLLARY 7.7. *The process y is regular if and only if no $X \in \mathfrak{X}$ has invariant directions.*

Remark 7.8. An immediate consequence of the definitions of Y_n, $\bar{Y}_n$, and Y^* is that

$$Y^* \subset Y_n \subset (\sigma^k X) \cap H_0 \quad \text{and} \quad Y^* \subset \bar{Y}_n \subset (\bar{\sigma}^k X) \cap H_0$$

for all $k = 0, 1, 2, \dots$, showing that the maximal strictly output-induced subspace of any $X \in \mathfrak{X}$ is contained in each of the internal subspaces of the corresponding sequence of splitting subspaces $\{X^{(k)}; k \in \mathbb{Z}\}$.

From Theorems 7.2 and 7.6(i) we also have the following corollary.

COROLLARY 7.9. *The following inclusions hold:*

$$UY_n \subset Y_n + \{y(0)\}, \tag{7.35}$$

$$U^{-1}\bar{Y}_n \subset \bar{Y}_n + \{y(-1)\}. \tag{7.36}$$

Let us recall that, in the regular case, (5.12) and (5.13) enabled us to define the zero dynamics operators G and $\bar{G}$ on all of the internal subspace $X \cap H_0$ of $X \in \mathfrak{X}$. In the general case, Corollary 7.9 provides the appropriate counterparts of (5.12) and (5.13).

DEFINITION 7.10 (general case). *Let the zero dynamics operators $G : Y_n \to Y_n$ and $\bar{G} : \bar{Y}_n \to \bar{Y}_n$ be defined as*

$$G = \pi U|_{Y_n} \tag{7.37}$$

and

$$\bar{G} = \bar{\pi} U^{-1}|_{\bar{Y}_n}, \tag{7.38}$$

where $\pi : Y_n + \{y(0)\} \to Y_n$ and $\bar{\pi} : \bar{Y}_n + \{y(-1)\} \to \bar{Y}_n$ are the oblique projectors projecting parallel to $\{y(0)\}$ and $\{y(-1)\}$, respectively.

Note that in the regular case the definitions of π and $\bar{\pi}$ coincide with that of section 4. In fact, we recall from Theorem 7.2 that $Y_n = \bar{Y}_n = X \cap H_0$ in this case. Consequently, the present definitions of G and $\bar{G}$ are merely straightforward generalizations of those in section 4.

THEOREM 7.11. *The maximal strictly output-induced subspace Y^* is both G- and $\bar{G}$-invariant. Also $G|_{Y^*}$ is invertible, and its inverse is $\bar{G}|_{Y^*}$. Furthermore, the subspaces $X \cap \{y(-1)\}$ and $X \cap \{y(0)\}$ are the null spaces of G and $\bar{G}$, respectively. More generally,*

$$\ker G^k = X \cap \{y(-1), \dots, y(-k)\} \quad \textit{for } k = 1, 2, 3, \dots$$

and

$$\ker \bar{G}^k = X \cap \{y(0), \dots, y(k)\} \quad \textit{for } k = 1, 2, 3, \dots.$$

In particular, $\ker G^n$ *is the space of predictable and* $\ker \bar{G}^n$ *the space of smoothable invariant directions.*

Proof. The fact that Y^* is strictly output induced and thus satisfies

$$UY^* \subset Y^* + \{y(0)\} \quad \text{and} \quad U^{-1}Y^* \subset Y^* + \{y(-1)\}$$

implies that Y^* is both G- and $\bar{G}$-invariant. To prove the last statement, take $\xi \in Y^*$. Then

$$U\xi = G\xi + \lambda,$$

where $G\xi \in Y^*$ and $\lambda \in \{y(0)\}$, and therefore $U^{-1}G\xi = \xi - U^{-1}\lambda$. Consequently, since $U^{-1}\lambda \in \{y(-1)\}$,

$$\bar{G}G\xi = \xi,$$

proving that $G|_{Y^*} = \left[\bar{G}|_{Y^*}\right]^{-1}$. To prove the statement concerning the kernel of G, observe that $\xi \in \ker G$ if and only if $\xi \in Y_n$ and $U\xi \in \{y(0)\}$. Therefore, since $\{y(-1)\} \cap X \subset Y_n$, $\ker G = X \cap \{y(-1)\}$. Similar arguments prove the rest. ∎

This again illustrates the fact that G and $\bar{G}$ are defined and invertible on all of $X \cap H_0$ in the regular case and only in the regular case. The following theorem, the proof of which is deferred to Appendix A, gives an upper bound for the number of invariant directions of any $X \in \mathcal{X}$. This is a generalization of Theorem 3.8 in [22], which deals with the internal case.

THEOREM 7.12. *The space of predictable directions in* X_- *has the same dimension* μ *as the space of smoothable directions in* X_+*, i.e.,*

$$\mu := \dim(X_- \cap \{y(-1), \dots, y(-n)\}) = \dim(X_+ \cap \{y(0), \dots, y(n-1)\}).$$

Moreover, the dimension of the space of invariant directions of any $X \in \mathcal{X}$ *is no larger than* μ*, i.e.,*

$$\dim(X \cap \{y(-1), \dots, y(-n)\} + X \cap \{y(0), \dots, y(n-1)\}) \leq \mu.$$

If X *is internal, there is equality in this relation.*

Proof of Theorem 7.6. We first prove (i). In the same way as in the proof of Theorem 3.3 we observe that

$$S^{(-k)} \cap H_0 = H^- \vee (X^{(-k)} \cap X_+),$$

formed analogously to (3.14), and that $X^{(-k)} \cap X_+$ converges in a finite number of steps which cannot exceed n. Consequently,

$$S^{(-n)} \cap H_0 = S^{(-2n)} \cap H_0.$$

Therefore, since $S^{(-k)} = (U^{-k}S) \vee H^-$ by definition, we obtain

$$(U^n S) \cap H_0 \subset (S \cap H_0) \vee U^{2n}H^- = (S \cap H_0) \vee \{y(0), \dots, y(2n-1)\}.$$

Now, taking the intersection with H^+ and using the shift invariance of H^+, we see that

$$U^n(S \cap H^+) \subset (U^n S) \cap H^+ \subset (S \cap H^+) \vee \{y(0), \dots, y(2n-1)\},$$

i.e.,

$$U^n(X \cap X_+) \subset (X \cap X_+) \vee \{y(0), \dots, y(2n-1)\}. \tag{7.39}$$

At the same time, (6.2) in Lemma 6.1 implies that

$$U^n(X \cap X_-) \subset (X \cap X_-) \vee \{y(0), \dots, y(n-1)\},$$

which together with (7.39) and (5.8) yields

$$U^n(X \cap H_0) \subset (X \cap H_0) \vee \{y(0), \dots, y(2n-1)\}.$$

Consequently,

$$X \cap H_0 \subset \big[U^{-n}(X \cap H_0) \vee \{y(-1), \dots, y(-n)\}\big] \vee \{y(0), \dots, y(n-1)\}, \tag{7.40}$$

where the first summand is a subspace of S, while the second is contained in $H^+ \subset \bar{S}$. Now, observing that

$$\big[U^{-n}(X \cap H_0) \vee \{y(-1), \dots, y(-n)\}\big] \cap \bar{S} \subset S \cap H_0 \cap \bar{S} \subset X \cap H_0,$$

(7.5) implies that the left member equals Y_n. Therefore, taking the intersection with $\bar{S}$ in (7.40), we obtain

$$X \cap H_0 \subset Y_n \vee \{y(0), \dots, y(n-1)\},$$

which after intersection with X yields

$$X \cap H_0 = Y_n \vee (X \cap \{y(0), \dots, y(n-1)\}),$$

the opposite inclusion being trivial. It remains to show that this is a direct sum. To see this, consider a $\xi \in Y_n \cap X \cap \{y(0), \dots, y(n-1)\}$. Then, by (7.5), $U^n\xi$ has the representation

$$U^n\xi = \zeta + \lambda, \tag{7.41}$$

where $\zeta \in X \cap H_0$ and $\lambda \subset \{y(0), \dots, y(n-1)\}$. Therefore,

$$\zeta = U^n\xi - \lambda \in X \cap \{y(0), \dots, y(2n-1)\}$$

is a smoothable direction of X, so we have that $\zeta \in X \cap \{y(0), \dots, y(n-1)\}$. Consequently, by (7.41), $U^n\xi \in \{y(0), \dots, y(n-1)\}$. However, by assumption, $U^n\xi \in \{y(n), \dots, y(2n-1)\}$, which, by virtue of the strictly nondeterministic property of y, implies that $\xi = 0$, concluding the proof (7.26). Since $X \cap X_-$ satisfies (7.14) (Lemma 6.1), the maximality of Y_n implies that $X \cap X_- \subset Y_n$. The last statement now follows from the fact that $X \cap H^- = X \cap X_-$. This concludes the proof of (i). A symmetric argument yields (ii).

Next we prove that Y^* defined by (7.30) satisfies the inclusion

$$UY^* \subset Y^* + \{y(0)\}. \tag{7.42}$$

To this end, consider a $\xi \in Y^*$. In view of Lemma 7.1, using the fact $Y_n = Y_{n+1}$, ξ is seen to have the two representations

$$\xi = U^{-n-1}\zeta + \lambda = U^n\bar{\zeta} + \bar{\lambda},$$

where $\zeta, \bar{\zeta} \in X \cap H_0$, $\lambda \in \{y(-1), \dots, y(-n-1)\}$, and $\bar{\lambda} \in \{y(0), \dots, y(n)\}$. Consequently,

$$U\xi = U^{-n}\zeta + \lambda_1 + \lambda_0,$$

where $\lambda_0 \in \{y(0)\}$ and $\lambda_1 \in \{y(-1), \dots, y(-n)\}$ and $\lambda_1 + \lambda_0 = U\lambda$. We prove that $U^{-n}\zeta + \lambda_1 \in Y^*$, thereby proving (7.42). Since for Y_n the inclusion (7.11) holds and $\xi \in Y_n$, we obtain that

$$U^{-n}\zeta + \lambda_1 \in Y_n.$$

On the other hand,

$$U^{-n}\zeta + \lambda_1 = U^{2n+1}\bar{\zeta} + U\bar{\lambda} - \lambda_0 \in U^{2n+1}(X \cap H_0) \vee \{y(0), \dots, y(n)\},$$

which, in view of Lemma 7.1, gives

$$U^{-n}\zeta + \lambda_1 \in Y_n,$$

concluding the proof of (7.42).

A similar argument shows that

$$U^{-1}Y^* \subset Y^* + \{y(-1)\}. \tag{7.43}$$

Equations (7.42) and (7.43) establish that Y^* is a strictly output-induced subspace. Since the maximality of Y_n and that of $\bar{Y}_n$ imply that any strictly output-induced subspace must be included in both Y_n and $\bar{Y}_n$, the intersection Y^* is the maximal strictly output-induced subspace.

Taking intersections with $\bar{Y}_n$ on both sides in (7.26) and observing that $X \cap \{y(0), \dots, y(n-1)\} \subset \bar{Y}_n$ by (ii), we immediately obtain (7.32).

In the same way, (7.31) follows by intersecting (7.28) with Y_n and observing the fact that $X \cap \{y(-1), \dots, y(-n)\} \subset Y_n$ by (i). □

8. The change of zero dynamics under σ and $\bar{\sigma}$. In Definition 7.10 we assigned to each $X \in \mathcal{X}$ two operators G and $\bar{G}$, defined on the appropriate subspaces of X. Now we will relate the eigenvalues of G and $\bar{G}$ to the zeros of W and $\bar{W}$, the spectral factors corresponding to X, justifying the name zero dynamics operators. Next we analyze the connections between the zero dynamics operators belonging to different splitting subspaces. Finally, using these operators, we describe completely the change in the zero structure when applying the prediction operators σ and $\bar{\sigma}$.

To this end, we recall from [20] that the finite zeros of W and the corresponding zero directions are characterized by the solutions of (5.5), i.e.,

$$\begin{bmatrix} \Pi & -M \end{bmatrix} \begin{bmatrix} A & B \\ C & D \end{bmatrix} = \begin{bmatrix} \Lambda\Pi & 0 \end{bmatrix}, \tag{8.1}$$

in the sense that the eigenvalues of Λ are zeros of W and the rows of Π span the subspace of the corresponding generalized zero directions. In order to describe all finite zeros we need to consider a maximal solution of (8.1) in the sense that Π has maximal rank or in the sense that the subspace generated by the row vectors of Π is maximal.

We now give an alternative proof of a generalization of Lemma 5.2 which also works in the nonregular case.

LEMMA 8.1. *A matrix* Π *satisfies* (8.1) *if and only if there are matrices* Λ *and* M *such that*

$$\Pi x(t+1) = \Lambda\Pi x(t) + My(t). \tag{8.2}$$

Proof. Equation (8.1) is equivalent to

$$\begin{bmatrix}\Pi & -M\end{bmatrix}\begin{bmatrix}A & B\\ C & D\end{bmatrix}\begin{bmatrix}x(t)\\ u(t)\end{bmatrix} = \begin{bmatrix}\Lambda\Pi & 0\end{bmatrix}\begin{bmatrix}x(t)\\ u(t)\end{bmatrix},$$

where x is the state process and u is the driving noise of the stochastic model (1.1). This is seen by observing that the covariance matrix of $\begin{bmatrix}x(t)\\ u(t)\end{bmatrix}$ is full rank. Together with the systems equations (1.1) this yields (8.2). □

A coordinate-free version of (8.2) is provided by

$$UY \subset Y \vee \{y(0)\}, \tag{8.3}$$

where Y consists of the random variables of the form $b'\Pi x(0)$. This observation allows us to characterize the zeros of W in terms of the eigenvalues of G.

PROPOSITION 8.2. *The eigenvalues of* G *are precisely the finite zeros of* W. *Similarly, the eigenvalues of* $\bar{G}$ *are the finite zeros of* $\bar{W}(z^{-1})$.

Proof. Consider the maximal solution of (8.1). Then the eigenvalues of Λ are the finite zeros of W. Moreover, since Y_n is the maximal subspace satisfying (8.3) (Theorem 7.2), $z := \Pi x(0)$ is a basis in Y_n, the space on which G is defined. It then follows from (8.2) that

$$Gz_i = \sum_j \Lambda_{ij} z_j,$$

and consequently Λ' is a matrix representation of G in the basis of z, thus having the same eigenvalues. This concludes the proof of the first statement. The second statement follows by symmetry. □

This together with Theorem 7.11 illustrates that we have zeros at zero and/or infinity precisely in the nonregular case and that such zeros are connected to invariant directions. More precisely, predictable directions correspond to zeros at zero and smoothable directions to zeros at infinity.

It was proven in [20] and [29], independently and with different methods, that W and $\bar{W}$ have the same zeros also in the nonregular case. (A modification of the argument in Remark 5.3 could also be used to see this.) Therefore, any statement about the zeros of W also holds for $\bar{W}$, and vice versa.

Recall that

$$X \cap H_0 = (X \cap X_-) \vee (X \cap X_+), \tag{8.4}$$

where the sum is direct if and only if $X_- \cap X_+ = 0$ or, equivalently, $H^- \cap H^+ = 0$. Only in the regular case can we define G and $\bar{G}$ on all of $X \cap H_0$, but in view of Theorem 7.6(i), $X \cap X_-$ is always contained in the domain of G and $X \cap X_+$ in the domain of $\bar{G}$.

THEOREM 8.3. *Let* G_- *be the* G*-operator of* X_- *and* $\bar{G}_+$ *be the* $\bar{G}$*-operator of* X_+. *Let* $(W, \bar{W})$ *be the spectral factors of* $X \in \mathfrak{X}$. *Then*

$$G|_{X\cap X_-} = G_-|_{X\cap X_-} \tag{8.5}$$

and

$$\bar{G}|_{X\cap X_+} = \bar{G}_+|_{X\cap X_+}. \tag{8.6}$$

Consequently, the eigenvalues of $G|_{X\cap X_-}$ *are the stable zeros of* W *(including those on the unit circle) and that the eigenvalues of* $\bar{G}|_{X\cap X_+}$ *are the reciprocals of the antistable zeros (including those on the unit circle and at infinity). Finally,*

$$G|_{X_-\cap X_+} = \left[\bar{G}|_{X_-\cap X_+}\right]^{-1}, \tag{8.7}$$

and its eigenvalues are precisely the zeros on the unit circle.

Proof. Referring to (6.2) we see that $X \cap X_-$ is invariant under both G and G_- and hence (8.5) follows. In the same way, (6.1) implies that $X \cap X_+$ is invariant under both $\bar{G}$ and $\bar{G}_+$, implying (8.6). This is in harmony with Theorem 5.4 and implies the statements on stable and unstable zeros. Finally, by Corollary 6.6, $X_- \cap X_+$ is strictly output induced and is thus contained in Y^*, on which space G is invertible (Theorem 7.6). Consequently, the last statement follows. ∎

In particular, we have the following observation, which was previously reported in [10] and [9]. (In the latter paper the proof is somewhat incomplete, since the multiplicities are not counted properly.)

COROLLARY 8.4. *All minimal spectral factors have the same number of zeros on the unit circle* (*counting multiplicity*)*, namely,* $\dim X_- \cap X_+ = \dim H^- \cap H^+$.

In section 5 we showed that in the regular case, the zeros, as well as the zero directions, are preserved as the operators σ and $\bar{\sigma}$ are applied. In general this is not true in the nonregular case. In view of Theorem 8.3, the following two theorems, relating σ and $\bar{\sigma}$ to the operators G and $\bar{G}$, show what happens.

THEOREM 8.5. *Let* $X \in \mathfrak{X}$. *Then*

$$(\sigma X) \cap X_+ = \bar{G}_+(X \cap X_+) = \bar{G}(X \cap X_+) \tag{8.8}$$

and

$$(\bar{\sigma} X) \cap X_- = G_-(X \cap X_-) = G(X \cap X_-). \tag{8.9}$$

Proof. We prove only (8.8). Then a symmetric argument yields (8.9). First we show that

$$\sigma X \subset \{y(-1)\} \vee U^{-1}X. \tag{8.10}$$

To this end, observe that $H^- \vee U^{-1}S = \{y(-1)\} \vee U^{-1}S$, which, in view of the fact that $\{y(-1)\} \subset U^{-1}H^+ \subset U^{-1}\bar{S}$, equals $(\{y(-1)\} \vee U^{-1}X) \oplus U^{-1}\bar{S}^\perp$. However, $X \perp \bar{S}^\perp \supset U^{-1}\bar{S}^\perp$, and consequently (8.10) follows from the definition (3.1). Now, consider $\zeta \in (\sigma X) \cap X_+$. In view of (8.10), we have the representation

$$\zeta = \eta + U^{-1}\xi,$$

where $\eta \in \{y(-1)\}$ and $\xi \in X$. On the other hand, since $U^{-1}\xi = \zeta - \eta \in U^{-1}H^+$, we see that $\xi \in H^+ \cap X = X \cap X_+$. From the definition of the operator $\bar{G}_+$, we have

$$\zeta = \bar{G}_+\xi, \quad \text{where } \xi \in X \cap X_+.$$

Conversely, if $\xi \in X \cap X_+$, $\bar{G}_+\xi \in X_+$ and $\bar{G}_+\xi - U^{-1}\xi \in \{y(-1)\}$, implying that $\bar{G}_+\xi \in H^- \vee U^{-1}S = S^{(-1)}$. Hence,

$$\bar{G}_+\xi \in (\sigma X) \cap X_+,$$

which together with (8.6) concludes the proof of the theorem. ∎

THEOREM 8.6. *Let $X \in \mathcal{X}$. Then*

$$(\sigma X) \cap X_- = \{\xi \in X_- \mid G_-\xi \in X \cap X_-\} \tag{8.11}$$

and

$$(\bar{\sigma} X) \cap X_+ = \{\xi \in X_+ \mid \bar{G}_+\xi \in X \cap X_+\}. \tag{8.12}$$

Proof. We prove (8.11); then (8.12) follows by symmetry. Let $\xi \in X_-$. Since $\mathrm{E}^{H^-}X = X_-$ [18, Lemma 4.6 and Theorem 4.10] and $\ker \mathrm{E}^{H^-}|_X = X \cap (H^-)^\perp = 0$ (see section 2), there is a unique $\zeta \in X$ such that $\xi = \mathrm{E}^{H^-}\zeta$. By the definition of G_-,

$$U\xi = G_-\xi + \eta,$$

where $\eta \in \{y(0)\}$, and therefore

$$U(\zeta - \xi) = U\zeta - \eta - G_-\xi.$$

Since $G_-\xi \in X_- \subset S$, $U\zeta \in U\bar{S} \subset \bar{S}$, and $\eta \in H^+ \subset \bar{S}$, the splitting property (2.8) yields

$$\mathrm{E}^S U(\zeta - \xi) = \mathrm{E}^X(U\zeta - \eta) - G_-\xi. \tag{8.13}$$

Now, suppose $\xi \in (\sigma X) \cap X_-$. Then by definition (3.1), $\xi = \mathrm{E}^{H^- \vee U^{-1}S}\lambda$ for some $\lambda \in X$. But then, since $\xi \in H^-$, $\xi = \mathrm{E}^{H^-}\lambda$, so by uniqueness we must have $\lambda = \zeta$. Consequently, $\zeta - \xi \perp H^- \vee U^{-1}S$, which in particular implies that $U(\zeta - \xi) \perp S$. Hence, it follows from (8.13) that $G_-\xi \in X$, proving that $G_-\xi \in X \cap X_-$.

Conversely, suppose that $G_-\xi \in X \cap X_-$. Then, by (8.13),

$$\mathrm{E}^S U(\zeta - \xi) \in X. \tag{8.14}$$

But, since $\xi = \mathrm{E}^{H^-}\zeta$,

$$U(\zeta - \xi) \perp UH^- \supset H^-. \tag{8.15}$$

Therefore, since $S = H^- \oplus S \cap (H^-)^\perp$ by (4.20), we have

$$\mathrm{E}^S U(\zeta - \xi) = \mathrm{E}^{S \cap (H^-)^\perp} U(\zeta - \xi) \in S \cap (H^-)^\perp. \tag{8.16}$$

Since $X \cap S \cap (H^-)^\perp = X \cap (H^-)^\perp = 0$ (see section 2), it follows from (8.14) and (8.16) that $\mathrm{E}^S U(\zeta - \xi) = 0$, and hence

$$U(\zeta - \xi) \perp S,$$

which together with (8.15) yields

$$\zeta - \xi \perp H^- \vee U^{-1}S.$$

Consequently, $\xi = \mathrm{E}^{H^- \vee U^{-1}S}\zeta \in \sigma X$, and so $\xi \in (\sigma X) \cap X_-$ as claimed. ∎

In particular, Theorems 8.5 and 8.6 show that

$$(\bar{\sigma} X) \cap X_- \subset X \cap X_- \subset (\sigma X) \cap X_-; \tag{8.17}$$

i.e., stable zeros may be lost as we apply $\bar{\sigma}$ and gained as we apply σ. In the same way,

$$(\sigma X) \cap X_+ \subset X \cap X_+ \subset (\bar{\sigma} X) \cap X_+, \tag{8.18}$$

showing that antistable zeros may be lost when applying σ and gained when applying $\bar{\sigma}$. This is in agreement with Proposition 3.4 and formulas (6.7) and (6.8).

To determine what zeros are being lost and gained under these operations, we observe from Theorems 8.5 and 8.6 that the subspaces being added or subtracted from $X \cap X_-$ and $X \cap X_+$ must be contained in the kernel of some G- or $\bar{G}$-operator. Consequently, by Theorem 7.11, the corresponding zero directions are invariant directions.

We may therefore formulate an amplification of statement (8.17), namely, that zeros at zero together with the corresponding predictable directions may be gained when applying σ and lost when applying $\bar{\sigma}$. In the same way, (8.18) and Theorem 7.11 show that zeros at infinity together with the corresponding smoothable directions may be lost when applying σ and gained when applying $\bar{\sigma}$.

The following corollary is an immediate consequence of Theorems 8.5 and 8.6.

COROLLARY 8.7. *Let $X \in \mathcal{X}$, and let Y_n and $\bar{Y}_n$ be defined as in section 7. Then*

$$\begin{aligned}(\sigma^k X) \cap H_0 &= Y_n \vee \{\textit{predictable directions in } X_-\} \\ &= Y^* + \{\textit{predictable directions in } X_-\}\end{aligned}$$

and

$$\begin{aligned}(\bar{\sigma}^k X) \cap H_0 &= \bar{Y}_n \vee \{\textit{smoothable directions in } X_+\} \\ &= Y^* + \{\textit{smoothable directions in } X_+\}\end{aligned}$$

for $k = n, n+1, \ldots$.

Remark 8.8. Theorem 7.11, Remark 7.8, and Corollary 8.7 enable us to generalize the statement in Remark 5.7 to the nonregular case. The same construction that was used in this remark to reduce the Riccati equations can be applied here with modifications which take into account the fact that the internal subspace $X^{(k)} \cap H_0$ is no longer constant along the sequence of splitting subspaces $\{X^{(k)}\}$ in the nonregular case. In view of Remark 7.8, the solutions of the Riccati recursions are constant from the start in the zero directions of Y_n, while they become constant only after a finite number of steps in the remaining predictable directions by Corollary 8.7. Consequently, after a finite number of steps the size of the reduced Riccati equations is $\nu \times \nu$ where $\nu = n - \dim(\sigma^n X) \cap H_0$. In view of Corollary 8.7 and Theorem 7.12, the backward Riccati equation can be reduced to the same size.

9. Conclusions. In this paper we discuss the very rich and intricate structure of discrete-time linear stochastic systems in the context of an interpolation-type problem, namely, to reconstruct lost state information on a finite interval using the whole history of the output process and the remaining state information. We show that, at each time, the (least squares) state estimate can be written as a linear combination of two filter estimates, which are generated by (forward respectively backward) Kalman filtering-type recursions with the initial condition being itself a state. Remarkably, these Kalman filtering recursions generate sequences of state processes from different stochastic realizations which are totally ordered. When $k \to \infty$ and when $k \to -\infty$, the sequence of splitting subspaces $X^{(k)}$ converge to limits which are internal splitting

subspaces. These limits are determined by the zero structure of the spectral factor (1.2).

In the regular case, when there are no zeros at the origin and at infinity, the set of zeros and zero directions of the spectral factors $W^{(k)}$ corresponding to the splitting subspaces $X^{(k)}$ remain invariant during these recursions giving a set of invariants. We show that in the nonregular case the whole set of zeros is no longer invariant but the finite zeros with finite reciprocals still are.

We have shown that the computational burden of determining the interpolation estimate depends on the dimension of the internal subspace $X \cap H_0$, i.e., on the number of zeros. This leads to the study of output-induced and strictly output-induced subspaces and zero dynamics operators. In particular, if $a'x(0) \in X \cap H_0$, then, in the regular case, the solutions of the Riccati equations (4.38) and (4.41) for the interpolation problem becomes constant in the direction a, allowing for a reduction in size of the Riccati equations. In fact, if $\dim X \cap H_0 = n - \nu$, we need only to solve Riccati equations of dimension $\nu \times \nu$ rather than $n \times n$ in the regular case (Remark 5.7). In the nonregular case the reduction may be even larger after a finite number of steps (Remark 8.8).

What makes the discrete-time case more complicated than the continuous-time case is the possibility that the *predictable subspace* $X \cap \{y(-1), \dots, y(-n)\}$ and the *smoothable subspace* $X \cap \{y(0), \dots, y(n-1)\}$ are nontrivial. In fact, if these spaces are zero spaces (the regular case), the structure of the problem is very much like the continuous-time coercive case, studied in [13], and $X \cap H_0$ is itself strictly output induced. If they are not, the matrix D will lose rank, and the matrix Riccati equations of forward and backward Kalman filtering will become constant in the directions a for which $a'x(0)$ is an element of these spaces, thus influencing the implementation of the filtering algorithms, as explained above. These a are called *invariant directions*. Nonregularity, and hence invariant directions, are connected with zeros at zero and at infinity.

In particular, we have demonstrated that $X \cap H_0$ can be decomposed as a direct sum of the predictable subspace, the smoothable subspace and the maximal strictly output-induced subspace, corresponding to the zeros at zero, the zeros at infinity, and the remaining zeros, respectively. The maximal strictly output-induced subspace Y^* equals $X \cap H_0$ in the regular case and plays the role of $X \cap H_0$ in the nonregular case. We have given several geometric characterizations of regularity (Propositions 3.11 and 3.11′ and Corollaries 6.10 and 7.7). We have also shown that Y^* can be determined by algorithms akin to that used in geometric control theory for determining the maximal output-nulling subspace.

On the maximal strictly output-induced subspace Y^* the forward and backward zero dynamics operators G and $\bar{G}$, respectively, are inverses of each other. The eigenvalues are the finite zeros with finite reciprocals. The operators G and $\bar{G}$ can be separately extended to a larger subspace. On these subspaces (in the nonregular case) these operators are in general singular and the invariant directions determine the kernel of these operators.

Appendix A. Proof of Theorem 7.12.

Let us denote by $I_p(X)$ the predictable directions in $X \in \mathfrak{X}$ under the natural isomorphism $a \mapsto a'x(0)$ and by $I_s(X)$ the smoothable directions under the same isomorphism. Then Theorem 7.11 implies that

$$I_p(X) = \ker(P - P_-) \cap \ker(\Gamma'_-)^n \quad \text{and} \quad I_s(X) = \ker(P_+ - P) \cap \ker P_+^{-1}(\bar{\Gamma}'_+)^n P_+.$$

In particular,

$$I_p(X_-) = \ker(\Gamma'_-)^n \quad \text{and} \quad I_s(X_+) = \ker P_+^{-1}(\bar{\Gamma}'_+)^n P_+.$$

As in [28, p. 53], a straightforward but somewhat tedious calculation yields the identity

$$\bar{\Gamma}_+ P_+^{-1}(P_+ - P_-) = P_+^{-1}(P_+ - P_-)\Gamma'_-. \tag{A.1}$$

First, we prove that

$$\dim I_p(X_-) = \dim I_s(X_+).$$

To this end, observe that (A.1) implies that

$$(P_+ \bar{\Gamma}_+^n P_+^{-1})(P_+ - P_-) = (P_+ - P_-)(\Gamma'_-)^n. \tag{A.2}$$

Since it follows from Theorem 8.3 that $\ker(P_+ - P_-) \cap \ker \Gamma'_- = 0$, we obtain that if $\xi \in \ker(\Gamma'_-)^n$ is nonzero, then $(P_+ - P_-)\xi \neq 0$. Thus

$$(P_+ - P_-) I_p(X_-) \subset \ker P_+^{-1}(\bar{\Gamma}_+)^n P_+,$$

implying that

$$\dim I_p(X_-) \leq \dim I_s(X_+).$$

A symmetric argument yields the reverse inequality, proving the first statement in the theorem and also that

$$\ker P_+^{-1}(\bar{\Gamma}_+)^n P_+ = (P_+ - P_-) I_p(X_-).$$

Consequently, $Y := \operatorname{Im} P_+^{-1}(\bar{\Gamma}'_+)^n P_+$, the counterpart of the maximal strictly output-induced subspace Y^* of X_+ under the natural isomorphism, is the orthogonal complement of $(P_+ - P_-) I_p(X_-)$, i.e.,

$$Y = \{a \in \mathbb{R}^n \mid a'(P_+ - P_-)b = 0 \quad \text{for} \quad b \in I_p(X_-)\}.$$

Thus, again invoking that $(P_+ - P_-)$ is nonsingular on $I_p(X_-)$, we obtain the direct-sum decomposition

$$I_p(X_-) + Y = \mathbb{R}^n,$$

where the two summands are "orthogonal" in the inner product defined by $(P_+ - P_-)$. In view of the direct-sum decomposition

$$I_s(X_+) + Y = \mathbb{R}^n \tag{A.3}$$

we see that both the predictable directions and the smoothable directions under the natural isomorphisms are mapped to subspaces which are complementary to Y.

Now observe that if $a \in \ker(P - P_-)$ and $b \in \ker(P_+ - P)$, then $a'(P_+ - P_-)b = 0$, i.e., $\ker(P - P_-)$ and $\ker(P_+ - P)$, are also orthogonal in the inner product determined by $(P_+ - P_-)$. This inner product is nonsingular on $I_p(X_-)$, so we can consider a $(P_+ - P_-)$-orthogonal complement Z of $I_p(X)$ in $I_p(X_-)$, i.e.,

$$I_p(X) + Z = I_p(X_-). \tag{A.4}$$

Then the $(P_+ - P_-)$-orthogonal complement of $I_p(X)$ in $\mathbb{R}^n$ is $Z + Y$, the latter obviously containing $\ker(P_+ - P)$. Consequently

$$I_s(X) \subset (Z + Y) \cap I_s(X_+). \tag{A.5}$$

In the internal case we have equality in this inclusion. Now, the identity (A.3) implies that

$$\dim(Z + Y) \cap I_s(X_+) = \dim Z,$$

which together with (A.4) and (A.5) yields

$$\dim I_p(X) + \dim I_s(X) \leq \dim I_p(X_-) = \mu$$

with equality in the internal case, concluding the proof of the theorem.

Appendix B. Zero direction of σX and $\bar{\sigma} X$.

Theorems 8.5 and 8.6 can be reformulated in terms of (generalized) zero directions.

THEOREM B.1. *The antistable zero directions of σX are described by*

$$(\sigma X) \cap X_+ = \{a' \bar{x}_+(0) \mid a = \bar{\Gamma}'_+ b, \text{ where } b \in \ker(\bar{P} - \bar{P}_+)\}. \tag{B.1}$$

Similarly, the stable zero directions of $\bar{\sigma} X$ are given by

$$(\bar{\sigma} X) \cap X_- = \{a' \bar{x}_-(0) \mid a = \bar{\Gamma}'_- b, \text{ where } b \in \ker(P - P_-)\}. \tag{B.2}$$

THEOREM B.2. *The stable zero directions of σX are described by*

$$(\sigma X) \cap X_- = \{a' x_-(0) \mid a \in \mathbb{R}^n, \Gamma'_- a \in \ker(P - P_-)\}. \tag{B.3}$$

Similarly, the antistable zero directions of $\bar{\sigma} X$ are given by

$$(\bar{\sigma} X) \cap X_+ = \{a' \bar{x}_+(0) \mid a \in \mathbb{R}^n, \bar{\Gamma}'_+ a \in \ker(\bar{P} - \bar{P}_+)\}. \tag{B.4}$$

These theorems follow directly from Theorems 8.5 and 8.6, identifying G_- and $\bar{G}_+$ with Γ'_- and $\bar{\Gamma}'_+$ and $X \cap X_-$ and $X \cap X_+$ with $\ker(P - P_-)$ and $\ker(\bar{P} - \bar{P}_+)$, respectively. (Also see [13].) However, we also have the following independent coordinate-dependent proofs.

Proof of Theorem B.1. We prove (B.1). Then (B.2) follows by symmetry. The proof of this theorem runs parallel to that of Theorem 8.6. In view of the definition (3.1) of σX, we need to characterize all $\xi \in X$ such that

$$\mathrm{E}^{H^- \vee U^{-1} S} \xi = \mathrm{E}^{\{y(-1)\} \vee U^{-1} X} \xi \in X_+,$$

or, in other words, $\xi = d' \bar{x}_+(0)$ such that

$$\mathrm{E}^{\{y(-1)\} \vee U^{-1} X} d' \bar{x}(0) = c' y(-1) + b' \bar{x}(-1) = a' \bar{x}_+(0)$$

for appropriate vectors a, b, c, and d. The equations connecting a, b, c, and d are

$$d' \begin{bmatrix} \bar{P}\bar{C}' & \bar{P}A \end{bmatrix} = \begin{bmatrix} c' & b' \end{bmatrix} \begin{bmatrix} \Lambda_0 & C \\ C' & P \end{bmatrix}, \tag{B.5}$$

$$\begin{bmatrix} c' & b' \end{bmatrix} \begin{bmatrix} \bar{C} \\ A' \end{bmatrix} = a', \tag{B.6}$$

and

$$\begin{bmatrix} c' & b' \end{bmatrix} \begin{bmatrix} \Lambda_0 & C \\ C' & P \end{bmatrix} = a' \bar{P}_+ a. \tag{B.7}$$

Here the first two equations are projection formulas projecting ξ onto $\{y(-1)\} \vee U^{-1}X$ and $c'y(-1) + b'\bar{x}(-1)$ onto X_+, respectively, and the third equation expresses that in the latter projection the error is zero. Now, insert (B.6) into (B.7) and rearrange terms to obtain

$$\begin{bmatrix} c' & b' \end{bmatrix} \begin{bmatrix} \Lambda_0 - \bar{C}\bar{P}_+\bar{C}' & C - \bar{C}\bar{P}_+A \\ C' - A'\bar{P}_+\bar{C}' & \bar{P} - A'\bar{P}_+A \end{bmatrix} \begin{bmatrix} c \\ b \end{bmatrix} = 0.$$

Using the facts that $\Lambda_0 - \bar{C}\bar{P}_+\bar{C}' = \bar{D}_+\bar{D}'_+$, $C - \bar{C}\bar{P}_+A = \bar{D}_+\bar{B}'_+$, and also $\bar{P} - A'\bar{P}_+A = \bar{B}_+\bar{B}'_+ + (\bar{P} - \bar{P}_+)$, we obtain

$$\begin{bmatrix} c' & b' \end{bmatrix} \begin{bmatrix} \bar{D}_+\bar{D}'_+ & \bar{D}_+\bar{B}'_+ \\ \bar{B}_+\bar{D}'_+ & \bar{B}_+\bar{B}'_+ + (\bar{P} - \bar{P}_+) \end{bmatrix} \begin{bmatrix} c \\ b \end{bmatrix} = 0.$$

Since $\bar{P} \geq \bar{P}_+$, this is clearly a positive semidefinite quadratic form, and therefore

$$\begin{bmatrix} \bar{D}_+\bar{D}'_+ & \bar{D}_+\bar{B}'_+ \\ \bar{B}_+\bar{D}'_+ & \bar{B}_+\bar{B}'_+ + (\bar{P} - \bar{P}_+) \end{bmatrix} \begin{bmatrix} c \\ b \end{bmatrix} = 0.$$

The first block equation together with the fact that $\bar{D}_+\bar{D}'_+$ is invertible yields

$$c = -(\bar{D}_+\bar{D}'_+)^{-1}\bar{D}_+\bar{B}'_+ b. \tag{B.8}$$

Inserting this into the second block equation we get

$$(\bar{P} - \bar{P}_+)b = 0.$$

This shows that, if $a'\bar{x}_+(0) \in (\sigma X) \cap X_+$, then there is a $b \in \mathbb{R}^n$ such that $b \in \ker(\bar{P} - \bar{P}_+)$ and $a = \bar{\Gamma}'_+ b$. Conversely, assume that this is satisfied, define c by (B.8), and set $d := a$. Now straightforward calculations show that (B.5), (B.6), and (B.7) are satisfied. □

Proof of Theorem B.2. We prove (B.3); then (B.4) follows by symmetry. Recall that $\sigma X = \mathrm{E}^{H^- \vee U^{-1}S}X$. Therefore, if $a'x_-(0) \in (\sigma X) \cap X_-$, there exists a $\xi \in X$ such that

$$\mathrm{E}^{H^- \vee U^{-1}S}\xi = a'x_-(0).$$

Apply E^{H^-} to this to see that $\mathrm{E}^{H^-}\xi = a'x_-(0)$. Hence, by uniqueness of the uniform choice of bases, $\xi = a'x(0)$. On the other hand,

$$H^- \vee U^{-1}S = \{y(-1)\} \vee U^{-1}S = (\{y(-1)\} \vee U^{-1}X) \oplus U^{-1}\bar{S}^\perp,$$

since $\{y(-1)\} \in U^{-1}H^+ \subset U^{-1}\bar{S} \perp U^{-1}\bar{S}^\perp$ and $S = X \oplus \bar{S}^\perp$. Hence, since $\xi \in X \perp \bar{S}^\perp \supset U^{-1}\bar{S}^\perp$,

$$\mathrm{E}^{H^- \vee U^{-1}S}\xi = \mathrm{E}^{\{y(-1)\} \vee U^{-1}X}\xi.$$

But $a'x(0) - a'x_-(0) \perp H^- \supset \{y(-1)\}$, so the space $(\sigma X) \cap X_-$ is completely characterized by the condition

$$a'x(0) - a'x_-(0) \perp \{x(-1)\},$$

or, in other words,

$$\mathrm{E}\{[a'x(0) - a'x_-(0)]x(-1)'\} = 0. \tag{B.9}$$

To compute this covariance, note that the error process $x(t) - x_-(t)$ satisfies the forward state equation

$$x(t+1) - x_-(t+1) = \Gamma_-[x(t) - x_-(t)] + (B - B_-D_-^{-1}D)u(t)$$

so that

$$\mathrm{E}\{[x(0) - x_-(0)][x(-1) - x_-(-1)]'\} = \Gamma_-(P - P_-).$$

However, since $x_-(0) = \mathrm{E}^{H^-}x(0)$ and $a'x_-(-1) \in H^-$, (B.9) yields $a'\Gamma_-(P-P_-) = 0$ as claimed. □

REFERENCES

[1] B. D. O. Anderson, *The inverse problem of stationary covariance generation*, J. Statist. Phys., 1 (1969), pp. 133–147.

[2] F. Badawi, *Structures and Algorithms in Stochastic Realization Theory and the Smoothing Problem*, Ph.D. dissertation, University of Kentucky, 1980.

[3] F. Badawi, A. Lindquist, and M. Pavon, *A stochastic realization approach to the smoothing problem*, IEEE Trans. Automat. Control, AC-24 (1979), pp. 878–888.

[4] G. Basile and G. Marro, *Controlled and conditioned invariant subspaces in linear system theory*, J. Optim. Theory Appl., 3 (1973), pp. 306–316.

[5] R. S. Bucy, D. Rappaport, and L. M. Silverman, *Correlated noise filtering and invariant directions for the Riccati equation*, IEEE Trans. Automat. Control, AC-15 (1970), pp. 535–540.

[6] P. E. Caines, *Linear Stochastic Systems*, Wiley, Chichester, UK, 1988.

[7] H. Dym and H. P. McKean, *Gaussian Processes, Function Theory and the Inverse Spectral Problem*, Academic Press, New York, 1976.

[8] P. Faurre, M. Clerget, and F. Germain, *Opérateurs Rationnels Positifs*, Dunod, Paris, 1979.

[9] M. Green, *Balanced stochastic realizations*, Linear Algebra Appl., 98 (1988), pp. 211–247.

[10] E. J. Hannan and D. S. Poskitt, *Unit canonical correlations between future and past*, Ann. Statist., 16 (1988), pp. 784–790.

[11] A. Lindquist, *A new algorithm for optimal filtering of discrete-time stationary processes*, SIAM J. Control Optim., 12 (1974), pp. 736–746.

[12] A. Lindquist, *Some reduced-order non-Riccati equations for linear least-squares estimation: The stationary, single-output case*, Int. J. Control, 24 (1976), pp. 821–842.

[13] A. Lindquist, Gy. Michaletzky, and G. Picci, *Zeros of spectral factors, the geometry of splitting subspaces, and the algebraic Riccati inequality*, SIAM J. Control Optim., 33 (1995), pp. 365–401.

[14] A. Lindquist and M. Pavon, *On the structure of state-space models for discrete-time stochastic vector processes,* IEEE Trans. Automat. Control, AC-29 (1984), pp. 418–432.

[15] A. Lindquist, M. Pavon, and G. Picci, *Recent trends in stochastic realization theory,* in Prediction Theory and Harmonic Analysis: The Pesi Masani Volume, V. Mandrekar and H. Salehi, eds., North-Holland, Amsterdam, 1983.

[16] A. Lindquist and G. Picci, *On the stochastic realization problem,* SIAM J. Control Optim., 17 (1979), pp. 365–389.

[17] A. Lindquist and G. Picci, *Realization theory for multivariate stationary Gaussian processes,* SIAM J. Control Optim., 23 (1985), pp. 809–857.

[18] A. Lindquist and G. Picci, *A geometric approach to modelling and estimation of linear stochastic systems,* J. Math. Systems Estim. Control, 1 (1990), pp. 241–333.

[19] Gy. Michaletzky, *Zeros of (non-square) spectral factors and canonical correlations*, in Proc. 11th IFAC World Congress, Tallinn, Estonia, 1992, pp. 167–172.

[20] Gy. Michaletzky and A. Ferrante, *Splitting subspaces and acausal spectral factors,* J. Math. Systems Estim. Control, 5 (1995), pp. 363–366 (summary; full paper retrieval code: 89459).

[21] Gy. Michaletzky and G. Tusnády, *Representation of inner products and stochastic realization,* in Topics in Stochastic Systems: Modelling, Estimation and Adaptive Control, L. Gerencsér and P. E. Caines, eds., Springer-Verlag, New York, 1991.

[22] M. Pavon, *Stochastic realization and invariant directions of the matrix Riccati equation,* SIAM J. Control Optim., 18 (1980), pp. 155–180.

[23] M. Pavon, *New results on the interpolation problem for continuous-time stationary increments processes,* SIAM J. Control Optim., 22 (1984), pp. 133–142.

[24] M. Pavon, *A new algorithm for optimal interpolation of discrete-time stationary processes,* in Analysis and Optimization of Systems, A. Bensoussan and J. L. Lions, eds., Springer-Verlag, New York, 1982, pp. 701–718.

[25] D. Rappaport, *Constant directions of the Riccati equation*, Automatica J. IFAC, 8 (1972), pp. 175–186.

[26] D. Rappaport and L. M. Silverman, *Structure and stability of discrete-time optimal systems*, IEEE Trans. Automat. Control, AC-16 (1971), pp. 227–232.

[27] N. I. Rozanov, *Stationary Random Processes*, Holden–Day, San Francisco, 1963.

[28] G. Ruckebuch, *Représentations Markoviennes de processus Gaussiens Stationaires*, Thesis for Doctorat de 3ème cycle, Université de Paris VI, 1975.

[29] J.-Å. Sand, *Zeros of discrete-time spectral factors and the internal part of a Markovian splitting subspace*, J. Math. Systems Estim. Control, 6 (1996), pp. 351–354 (summary; full paper retrieval code: 01844).

[30] B. Sz.-Nagy and C. Foias, *Harmonic Analysis of Operators on Hilbert Space,* North-Holland, Amsterdam, 1970.

[31] W. M. Wonham, *Linear Multivariable Control: a Geometric Approach*, Springer-Verlag, New York, 1977.

SIAM J. CONTROL OPTIM.
Vol. 35, No. 3, pp. 860–875, May 1997

006

ON THE PUISEUX SERIES EXPANSION OF THE LIMIT DISCOUNT EQUATION OF STOCHASTIC GAMES*

W. W. SZCZECHLA†, S. A. CONNELL‡, J. A. FILAR‡, AND O. J. VRIEZE§

Abstract. In this paper we give a new proof of the existence of Puiseux series expansion of the limit discount equation of finite state stochastic games. Unlike the original proof, due to Bewley and Kohlberg [*Math. Oper. Res.*, 3 (1976), pp. 197–208], our proof is not algebraic and does not invoke Tarski's principle. Instead we use only the theory of functions of complex variables and complex analytic varieties.

Key words. stochastic games, Puiseux series, limit discount equation

AMS subject classification. 90D15

PII. S0363012995284138

Introduction. Perhaps the first minimax theorem of modern game theory is due to von Neumann in 1928 (see [5]). In that classical paper it is shown that a matrix game possesses a "value": an optimal gain/loss for players I and II when they are playing in an antagonistic (zero-sum) fashion. In 1950 Weyl (see [8]) simplified von Neumann's proof and demonstrated that if the entries of a matrix game belong to an ordered field, then the value belongs to the same ordered field.

An important generalization to the infinite horizon dynamic case was proposed in 1953 by Shapley (see [6]). Shapley established the existence of the value vector in a class of games analogous to what are now called discounted stochastic games. In these games the rewards at future stages are discounted by a factor $\beta \in [0, 1)$. Shapley also observed that even when all of the data of these games lie in the field of rational numbers, the entries of the value need not be in the same field.

The asymptotic behavior of the value vector as $\beta \nearrow 1$ has been studied by a number of authors. In an important contribution Bewley and Kohlberg [1] viewed Shapley's "optimality condition" as an elementary sentence in formal logic over the closed ordered field of real Puiseux series. These authors invoke a powerful theorem from mathematical logic, known as Tarski's principle, to conclude that in some neighborhood of $\beta = 1$, the value vector belongs to the field of real Puiseux series. The essentially algebraic nature of Bewley and Kohlberg's approach and the use of Tarski's principle, while ingenious, do not give insight into the manner in which fractional power series solutions arise naturally in this problem. Arguably, this has proved to be a difficulty for researchers in stochastic games, because Bewley and Kohlberg's result has become an important building block in subsequent developments. For instance, Mertens and Neyman [4] used it to prove the existence of the value vector in the limiting average stochastic games.

*Received by the editors March 31, 1995; accepted for publication (in revised form) March 28, 1996.

http://www.siam.org/journals/sicon/35-3/28413.html

†Instytut Matematyki, Uniwersytet Warszawski, ul. Banacha 2, 02–097 Warszawa, Poland, and Université de Bourgogne, Département de Mathématiques—Laboratoire de Topologie, B.P. 138, 21004 Dijon Cedex, France (witold@mimuw.edu.pl). The research of this author was supported by Polish KBN grant 2 P03A 022 08 and a fellowship granted by the Conseil Régional de Bourgogne.

‡School of Mathematics, University of South Australia, The Levels, Pooraka 5095, South Australia, Australia (scott.connell@unisa.edu, majaf@levels.unisa.edu.au).

§Rijksuniversiteit Limburg, Maastricht, the Netherlands.

In this paper we offer an alternative proof of the Puiseux series expansion of the value vector of discounted stochastic games. Our approach is based on the fact that Shapley's optimality equation can be viewed as a system of polynomial equations in several variables. Solution sets of such systems of equations, over the complex numbers, are called complex analytic varieties. We exploit the geometric structure of the variety that arises in our case, thereby enabling us to reduce the polynomial system to one where each polynomial is in the form to which a classical result of Puiseux (see [2]) can be applied. In this way we obtain the desired fractional power series expansion. In our approach the Puiseux series play only an auxiliary role, being used mainly to translate our results on varieties into the language of Bewley and Kohlberg's paper [1]. In particular, note that the existence of the limit of the value vector as $\beta \nearrow 1$ is now proven as a consequence of the geometric structure of the varieties and without invoking the Puiseux series expansion. This may be the first use of techniques from the theory of several complex variables to solve a problem in game theory. As such, considerable preparation is still required to follow the proof. However, we believe that our approach is more natural and may shed light on as yet unsolved problems. In fact Theorem 5.2 already extends Bewley and Kohlberg's result to every point in the entire interval $[0, 1]$, and Theorem 4.3 contains geometric insight that is not apparent in an application of Tarski's principle.

1. Definitions and preliminaries of matrix games. Any $m \times n$ real matrix $A = (a_{ij})_{i,j=1}^{m,n}$ can be regarded as a two-person, zero-sum *matrix game* with a_{ij} denoting the amount player II will pay player I if II chooses an action $j \in \{1, 2, \ldots, n\}$ and I chooses an action $i \in \{1, 2, \ldots, m\}$. A *mixed* (or *randomized*) *strategy* for player I (II) in such a game is an $m(n)$-component probability vector $x(y)$ whose ith (jth) entry $x_i(y_j)$ denotes the probability that player I (II) will choose an action $i(j)$. It was von Neumann [5] who proved the celebrated "minimax theorem" for matrix games. It is a consequence of this theorem that there always exists a strategy pair (x^o, y^o) satisfying

$$x^T A y^o \leq (x^o)^T A y^o \leq (x^o)^T A y \tag{1.1}$$

for all mixed strategies $x(y)$ of player I (II). The strategies x^o, y^o are then called *optimal strategies*, and the real number $\mathrm{val}(A) := (x^o)^T A y^o$ is called the *value* of the matrix game A. It is well known that if $b_{ij} = ka_{ij} + c$, for all $i, j,\ k > 0$, and $B = (b_{ij})_{i,j=1}^{m,n}$, then $\mathrm{val}B = \mathrm{val}A + c$.

A matrix game A is called *completely mixed* if all of its optimal strategies are strictly positive in every component. Extending earlier results of Kaplansky [3], Shapley and Snow [7] have demonstrated the following result.

PROPOSITION 1.1. *If A is a matrix game and* $\mathrm{val}\, A \neq 0$, *then A has a square invertible submatrix $\overline{A}$, called a Shapley–Snow kernel, such that*

(K1) $\mathrm{val}\, \overline{A} = \mathrm{val}\, A = \det \overline{A} / \sum_{ij} \overline{A}^{ij} = (\mathbf{1}^T \overline{A}^{-1} \mathbf{1})^{-1}$, *where $\overline{A}^{ij}$ denotes the (i, j)th cofactor of $\overline{A}$.*

(K2) *There is a pair (x^0, y^0) of strategies for $\overline{A}$ which is optimal for A (after inserting zeroes) and satisfies*

$$(x^0)^T = (\mathrm{val}\, A)\mathbf{1}^T \overline{A}^{-1} \quad \textit{and} \quad y^0 = (\mathrm{val}\, A)\overline{A}^{-1}\mathbf{1},$$

where $\mathbf{1} = (1, \ldots, 1)^T$.

LEMMA 1.2. *If A is a matrix game and* $\mathrm{val}\, A \neq 0$, *then A has a square invertible submatrix $\overline{A}$, which will be called a* cmv-kernel, *such that*

(K0) *The matrix game* $\overline{A}$ *is completely mixed, and* $\overline{A}$ *satisfies condition* (K1) *of Proposition* 1.1.

Proof. Among the submatrices X of A satisfying $\mathrm{val}X = \mathrm{val}A$, let $\overline{A}$ be minimal with respect to matrix inclusion. First note that $\overline{A}$ is completely mixed. Indeed, if (p, q) is a nonmixed pair of optimal strategies for $\overline{A}$, then the matrix A', corresponding to the strictly positive entries of p and q, is a proper submatrix of $\overline{A}$. By the minimax condition (1.1), (p, q) is optimal for A'. Hence $\mathrm{val}A' = p^T A' q = p^T \overline{A} q = \mathrm{val}\overline{A}$, which contradicts the minimality of $\overline{A}$.

Since the matrix game $\overline{A}$ is completely mixed, Proposition 1.1 (K2) implies that it has a unique Shapley–Snow kernel equal to $\overline{A}$. Hence (K1) follows. □

By the above proof the cmv-kernels are simply the minimal "value-wise Shapley–Snow kernels." Their utility comes from the fact, proved in Lemma 1.3 below, that the completely mixed matrix games form an open set in the space of the entries. In particular the single algebraic formula $\mathrm{val}(M) = (\mathbf{1}M^{-1}\mathbf{1})^{-1}$ is insensitive to small perturbations of a completely mixed matrix game M. This fact will be used in the proof of Lemma 4.1.

LEMMA 1.3. *Let* M *be a completely mixed matrix game. Then every matrix game of the same size whose entries are sufficiently close to the corresponding entries of* M*, is also completely mixed.*

Proof. Suppose that (M_i) is a sequence of not completely mixed matrix games of the same size, whose entries converge to the corresponding entries of M. For every i choose a pair of nonmixed optimal strategies for M_i and extract a subsequence converging to a pair of vectors (p, q). Since the set of all probability vectors (of a suitable dimension) having a zero in some entry is closed, and since the minimax condition is closed, then (p, q) is nonmixed and optimal for M, which is a contradiction. □

2. Discounted stochastic games.

2.1. Definitions and preliminaries of stochastic games. A *stochastic game* as formulated by Shapley [6] is played in stages. At each stage, the game is in one of finitely many *states*, $s = 1, 2, \ldots, N$, in which players I and II are obliged to play a matrix game $R(s) = (r(s, a, b))_{a,b=1}^{m_s,n_s}$, once. The "law of motion" is defined by the probabilities $p(s'|s, a, b)$, where the event $\{s'|s, a, b\}$ is the event that the game will enter state s' at the next stage given that at the current stage the state of the game is s, and that players I and II choose the ath row and the bth column of $R(s)$, respectively.

In general, players' strategies will depend on complete past histories. In this paper, however, we shall only be concerned with *stationary strategies*. We may represent a typical stationary strategy μ for player I by a "composite" vector, $\mu = (\mu(1), \mu(2), \ldots, \mu(N))$, where each $\mu(s)$ is a probability distribution on $\{1, 2, \ldots, m_s\}$. Player II's stationary strategies ν are similarly defined.

It should be clear that once we specify the initial state s and a strategy pair (μ, ν) for players I and II, we implicitly define a probability distribution over all sequences of states and actions which can occur during the game and consequently over all sequences of payoffs to player I. In particular, if the random variable $\mathcal{R}_t$ denotes the payoff to player I for stage t, then the expected value of $\mathcal{R}_t$ given s and (μ, ν)

$$E_{\mu\nu s}(\mathcal{R}_t) := E\{\mathcal{R}_t|\mu, \nu, s\}$$

is well defined. The *β-discounted stochastic game* Γ_β is then the game in which the overall payoff, normalized by a factor of $1 - \beta$, resulting from the strategy pair (μ, ν)

and a starting state s is evaluated according to

$$v_\beta(\mu,\nu,s) := \sum_{t=1}^{\infty} \beta^{t-1}(1-\beta)E_{\mu\nu s}(\mathcal{R}_t),$$

where $\beta \in (0,1)$ is called the *discount factor*. A number $v_s(\beta)$ is called the *value* of the game Γ_β starting in state s if $v_s(\beta) = \sup_\mu \inf_\nu v_\beta(\mu,\nu,s) = \inf_\nu \sup_\mu f_\beta(\mu,\nu,s)$. The vector $\mathbf{v}(\beta) = (v_1(\beta), v_2(\beta), \ldots, v_N(\beta))$ is called the *value vector*. Furthermore, the pair (μ^o, ν^o) is called an *optimal strategy pair* for players I and II if

$$v_s(\beta) = v_\beta(\mu^o, \nu^o, s).$$

The existence of the value vector and of a pair of *optimal stationary strategies* was proved in 1953 in Shapley's seminal paper on the subject [6]. A key element in Shapley's proof was the construction of N auxiliary matrix games $R_\beta(s,\mathbf{u})$ that depend on an arbitrary vector $\mathbf{u} = (u_1, u_2, \ldots, u_N) \in \mathbf{R^N}$ according to

$$R_\beta(s,\mathbf{u}) = \left[(1-\beta)r(s,a,b) + \beta \sum_{s'=1}^{N} p(s'|s,a,b)u_{s'}\right]_{a,b=1}^{m_s,n_s}. \tag{2.1}$$

In view of the fact that the value of a matrix game always exists, it is possible to define, for each $\beta \in (0,1)$, an operator $T_\beta : \mathbf{R^N} \to \mathbf{R^N}$, the sth component of which is given by

$$[T_\beta(\mathbf{u})]_s := \mathrm{val}[R_\beta(s,\mathbf{u})]. \tag{2.2}$$

This operator is a contraction operator in the sup-norm with contraction constant $\leq \beta$; see [6]. It therefore follows from Banach's fixed-point theorem that there exists a unique fixed point $\mathbf{v}(\beta)$ of T_β; that is,

$$\mathbf{v}(\beta) = T_\beta(\mathbf{v}(\beta)). \tag{2.3}$$

This can be shown to be an equivalent definition of the value vector $\mathbf{v}(\beta)$ introduced above. Also, any optimal strategy pairs for $R_\beta(s,\mathbf{v}(\beta))$ can be shown to form an optimal strategy pair for Γ_β.

Since $\{T_\beta\}$ is a continuous family of contractions we have the following result.

LEMMA 2.1. *The function $\beta \mapsto \mathbf{v}(\beta)$ is bounded and continuous on $[0,1)$.*

Proof. To prove boundedness, note that if $M_1 \leq r(s,i,j) \leq M_2$ for all s, i, and j, then T_β maps the hypercube $K = [M_1, M_2]^N$ into itself. Hence T_β has a fixed point in K, which must be $\mathbf{v}(\beta)$.

To prove continuity of $\mathbf{v}(\beta)$ at β_0, let $\beta' < \beta_0 < \beta'' < 1$ (for $\beta_0 \neq 0$; if $\beta_0 = 0$, then $\beta' = \beta_0$) and note that for all $\beta \in [\beta', \beta'']$ the map T_β is a contraction with a constant β''. Now for any $\epsilon > 0$ take the closed ball $\overline{B}(\epsilon) = \overline{B}(\mathbf{v}(\beta_0), \epsilon)$ in $\mathbf{R}^N$, and observe that T_{β_0} maps $\overline{B}(\epsilon)$ into itself. Moreover,

$$T_\beta(\overline{B}(\epsilon)) \subseteq \overline{B}\left(T_\beta(\mathbf{v}(\beta_0)), \beta''\epsilon\right) \tag{2.4}$$

for all $\beta \in [\beta', \beta'']$. By the continuous dependence of val$M$ on the entries of matrix M the map

$$\beta \mapsto T_\beta(\mathbf{v}(\beta_0))$$

(which moves the center of $\overline{B}(\epsilon)$) is continuous. This and (2.4) imply that, for all β close enough to β_0, T_β maps $\overline{B}(\epsilon)$ into itself (just like T_{β_0} does). Hence, for all β close enough to β_0, both T_β and T_{β_0} have their fixed points in $\overline{B}(\epsilon)$. Since those fixed points are $\mathbf{v}(\beta)$ and $\mathbf{v}(\beta_0)$ and since $\epsilon > 0$ is arbitrary, this implies that $\mathbf{v}$ is continuous at β_0. □

Assuming $v_s(\beta) \neq 0$ for all β and s, in view of Lemma 1.3 we know that for each fixed $\beta \in (0,1)$ and each $\mathbf{u}$ close enough to $\mathbf{v}(\beta)$ there exist cmv-kernels $\overline{R_\beta(s,\mathbf{u})}$ such that the fixed-point equation above locally reduces to

$$v_s(\beta) = \frac{|\overline{R_\beta(s,\mathbf{v}(\beta))}|}{\sum_i \sum_j \left[\overline{R_\beta(s,\mathbf{v}(\beta))}\right]_{ij}} \quad \text{for each } s = 1,2,\ldots,N,$$

where $[V]_{ij}$ denotes the (i,j)th cofactor of a matrix V.

If we now transform the above equations to

$$v_s(\beta)\left\{\sum_i \sum_j \left[\overline{R_\beta(s,\mathbf{v}(\beta))}\right]_{ij}\right\} - |\overline{R_\beta(s,\mathbf{v}(\beta))}| = 0 \quad \forall s = 1,2,\ldots,N, \tag{2.5}$$

then for each fixed combination of the locations of these kernels we can regard this system as being a system of polynomials in the variables $x_0 := \beta, x_1 := v_1(\beta), x_2 := v_2(\beta), \ldots, x_N := v_N(\beta)$ of the form

$$\begin{aligned} f_1(x_0,\ldots,x_N) &= 0, \\ &\vdots \\ f_N(x_0,\ldots,x_N) &= 0. \end{aligned} \tag{2.6}$$

2.2. Asymptotic result of Bewley and Kohlberg. The behavior of the value vector $\mathbf{v}(\beta)$ as $\beta \nearrow 1$ is extremely important in the analysis of the so-called *limiting-average stochastic game* Γ_α, where the overall payoff resulting from the strategy pair (μ,ν) and starting state s is evaluated according to

$$v_\alpha(\mu,\nu,s) := \liminf_{\tau\to\infty}\left(\frac{1}{\tau+1}\right)\sum_{t=o}^{\tau} E_{\mu\nu s}(R_t).$$

In particular, the limit of $\mathbf{v}(\beta)$ as $\beta \nearrow 1$ is the natural candidate for the value vector of Γ_α. Indeed, this was proved by Mertens and Neyman [4], who exploited an interesting characteristic of $\mathbf{v}(\beta)$ due to Bewley and Kohlberg [1]. We give a direct proof of the existence of this limit in section 5.

The result of Bewley and Kohlberg [1] can be interpreted as saying that *the solutions to the system of equations* (2.5) *are given by Puiseux series over the field of real numbers* (see section 3.1 below). As mentioned earlier, for β sufficiently near 1, their proof relies on a result in formal logic known as Tarski's principle (see [1]). However, systems of polynomial equations such as (2.6) are precisely the objects of study in the field of algebraic geometry. This is an old branch of mathematics, and it is in this setting that Puiseux series originally arose. In fact the use of fractional power series goes back to Newton. One would therefore expect that the tools of algebraic geometry would allow a "natural" proof of Bewley and Kohlberg's theorem. This indeed turns out to be the case. However, in order to do this some results from algebraic geometry will be needed.

3. Puiseux series and systems of polynomial equations.

3.1. Puiseux series. *Puiseux series* over a field $\mathbf{K}$, equal to either $\mathbf{R}$ or $\mathbf{C}$, are "fractional power series" of the form $\sum_{\nu=k}^{\infty} c_\nu z^{\nu/m}$, where $c_\nu \in \mathbf{K}$, $k \in \mathbf{Z}$, $m \in \mathbf{N}\setminus\{0\}$, and the series converges in some annulus, $\{z \in \mathbf{K} | 0 < |z| < R,\ R > 0\}$. The Puiseux series over K form a field (in fact the Puiseux series over $\mathbf{R}$ form an ordered field). Puiseux series over $\mathbf{C}$ are simply called Puiseux series. They arise as the solutions of particular polynomial equations. Specifically, there is the following classical result due to Puiseux (see [2, Chapter 1, Theorem 8.14]).

THEOREM 3.1 (Puiseux). *Let $\mathbf{C}\{\{z\}\}$ be the field of all Laurent series with finite principal part*

$$\phi(z) = \sum_{\nu=k}^{\infty} c_\nu z^\nu, \quad k \in \mathbf{Z}, \quad c_\nu \in \mathbf{C},$$

converging on some punctured disc $\{z \in \mathbf{C} : 0 < |z| < r\}$, where r may depend on the element $\phi \in \mathbf{C}\{\{z\}\}$. Let

$$F(z, w) = w^n + a_1(z)w^{n-1} + \cdots + a_n(z)$$

be a polynomial in w of degree n which is irreducible over the field $\mathbf{C}\{\{z\}\}$. Then there exists a Laurent series

$$\phi(\zeta) = \sum_{\nu=k}^{\infty} c_\nu \zeta^\nu \in \mathbf{C}\{\{\zeta\}\}$$

such that

$$F(\zeta^n, \phi(\zeta)) = 0$$

as an element of $\mathbf{C}\{\{\zeta\}\}$. In other words the equation

$$F(z, w) = 0$$

can be solved by a Puiseux series

$$w = \phi(z^{1/n}) = \sum_{\nu=k}^{\infty} c_\nu z^{\nu/n},$$

where $z^{1/n}$ is a branch of the nth root function.

The case where $F(z, w)$ is not necessarily irreducible over $\mathbf{C}\{\{z\}\}$ is covered by the following result. This result also states that locally *all* solutions to $F(z, w) = 0$ are given by Puiseux series.

COROLLARY 3.2. *Let*

$$F(z, t) = z^m + \sum_{k=0}^{m-1} z^k g_k(t),$$

where the g_k are holomorphic (resp., meromorphic) in a disc $D(0, r')$. Then there is $r \in (0, r')$, a positive integer M, and functions $\varphi_1, \ldots, \varphi_m$ holomorphic (resp., meromorphic) in $D(0, r^{1/M})$ such that

$$F(z, t^M) = \prod_{k=1}^{m} z - \varphi_k(t).$$

Note. Only the holomorphic variant will be used.

Proof. It suffices to prove the meromorphic version. Indeed, for each fixed t, the $\varphi_k(t)$ are roots of the monic polynomial $F(z, t^M)$. If the coefficients g_k are holomorphic near $t = 0$, they are bounded, and hence the roots stay in a bounded set; it follows that each φ_k has a removable singularity at 0.

We will write $\mathcal{M}$ for $\mathbf{C}\{\{t\}\}$. Each g_k will be regarded as an element of $\mathcal{M}$. Thus F is a polynomial over the field $\mathcal{M}$, that is, $F \in \mathcal{M}[z]$. Note that $\mathcal{M}$ (and $\mathbf{C}\{\{t\}\}$) can also be identified with the field $\mathcal{M}_0$ of germs of meromorphic functions at 0.

Observe that if $F = F_1F_2$, where $F_1, F_2 \in \mathcal{M}[z]$ are nonconstant monic polynomials satisfying the assertion with

$$F_j\left(z, t^{M_j}\right) = \prod_{k=1}^{m_j}(z - \varphi_{j,k}(t)) \quad (j = 1, 2),$$

then F also satisfies the assertion, since

$$F\left(z, t^{M_1M_2}\right) = \prod_{k=1}^{m_1}\left(z - \varphi_{1,k}\left(t^{M_2}\right)\right)\prod_{k=1}^{m_2}\left(z - \varphi_{2,k}\left(t^{M_1}\right)\right).$$

Also observe that, if the assertion is valid for the polynomial $F(z, t^n)$, where n is a positive integer, then it is also valid for $F(z, t)$, since

$$F\left(z, t^{M_n}\right) = \prod_{k=1}^{m}\left(z - \varphi_k\left(t^M\right)\right).$$

We will use induction with respect to the degree of F, the assertion being trivial if $\deg F = 1$. Let $G \in \mathcal{M}[z]$ be an irreducible factor of F. Theorem 3.1 says that the polynomial $G(z, t^n) \in \mathcal{M}[z]$, where $n = \deg G$, has a root $\varphi \in \mathcal{M}$. Hence $F(z, t^n) = (z - \varphi(t))F_1(z, t)$, where $F_1 \in \mathcal{M}[z]$. Since F_1 is a monic polynomial of degree $n - 1 < \deg F$, it satisfies the assertion. By the previous observations, the assertion is satisfied by $F(z, t^n)$ and hence by F. □

Remark. Another way of stating Puiseux's theorem is to say that the field of *real* Puiseux series is *real closed* (that is, no proper algebraic extension of the field of real Puiseux series is ordered). This is precisely the condition Bewley and Kohlberg needed in order to apply Tarski's principle and show that the value of a discounted, two-person, zero-sum stochastic game is given by a Puiseux series over $\mathbf{R}$ (see [1, section 10]).

As it stands, Puiseux's theorem does not apply to the system (2.6) for several reasons. First, (2.6) is a system of equations over $\mathbf{R}$ and not $\mathbf{C}$. Second, Puiseux's theorem applies to polynomial equations of a very special form, and there is no guarantee that the polynomials in (2.6) will have this form. In particular Puiseux's theorem involves *two* variables z and w, whereas the system (2.6) involves $N + 1$ variables $\beta, v_1(\beta), \ldots, v_N(\beta)$. Finally, the form of the polynomial system (2.6) depends on the location of the kernels and hence may be different for different values of β. Of these problems, the last two can be helped to be overcome by invoking a more general point of view, namely, that of algebraic geometry.

3.2. Algebraic and analytic varieties. Generally speaking, algebraic geometry over a field $\mathbf{F}$ (usually taken to be $\mathbf{R}$ or $\mathbf{C}$) is the study of the solution sets of

systems of polynomial equations of the form

$$\begin{aligned} f_1(z_0,\dots,z_N) &= 0,\\ &\vdots\\ f_k(z_0,\dots,z_N) &= 0, \end{aligned} \tag{3.1}$$

where, for $i = 1,\dots,k$, f_i is a polynomial of degree d_i over $\mathbf{F}$ and $(z_0,\dots,z_N) \in \mathbf{F}^{N+1}$. The solution set of (3.1) is called an *algebraic variety* over $\mathbf{F}$. Only the case $\mathbf{F} = \mathbf{C}$ will concern us at this stage. In this case the variety V defined by (3.1) will be a subset of $\mathbf{C}^{\mathbf{N+1}}$. The study of algebraic varieties over $\mathbf{C}$ uses powerful tools from both abstract algebra and complex analysis. The problem at hand will require the use of complex analytic methods. In particular the *local* properties of holomorphic functions are exploited.

A few preliminary definitions and results will be needed. Most of the definitions and theorems in this section are taken from [9]. A subset V of $\mathbf{C}^{N+1}$ is *analytic near* a point $p \in \mathbf{C}^{N+1}$ or is a *variety near* p if there is a neighborhood U of p and functions $f_1,\dots,f_k$ holomorphic on U such that $V \cap U$ is the zero set of these functions. If V is analytic near each of its points, then V is called *locally analytic* or a *local variety*. If H is open in $\mathbf{C}^{N+1}$ and $V \subseteq H$ is a local variety, then V is called a *variety in* H if V is closed in H (that is, if V is a variety near every point of H). If V and V' are local varieties such that $V' \subseteq V$ and V' is closed in V, then V' is called a *subvariety* of V (that is, V' is a variety near every point of V). Note that V is a variety in an open set H iff V is a subvariety of H. (Compare [9, Chapter 2, section 1].) We will be mainly interested in the case $H = \mathbf{C}^{N+1}$.

It is easy to see that any intersection or union of a family of varieties in H (or subvarieties of V) is also a variety in H (resp., subvariety of V), provided that the family is locally finite in H (resp., in V).

A point p of a local variety V is called a *regular* or *smooth point* of V of *dimension* d if V is a submanifold of dimension d near p. That is, if V is given in some neighborhood of p by the zero set of a collection of holomorphic functions $f_1,\dots,f_k$ whose Jacobian matrix has rank k; in this case $d = n - k$. The set of regular points of V of dimension d is denoted $\operatorname{Reg}_d V$, where $d = 0, 1, 2, \dots$. The set of regular points of any dimension is denoted $\operatorname{Reg} V$. It follows from Proposition 3.3 below that $\operatorname{Reg} V$ is dense in V. The *dimension* of V is defined by

$$\dim V = \max\{d : \operatorname{Reg}_d V \neq \emptyset\}.$$

If $\operatorname{Reg}_d V = \operatorname{Reg} V$ then we say that V is *purely d-dimensional* (or of *constant dimension* d) and write $\dim V \equiv d$ (see [9, Chapter 2, section 2]).

A local variety V is *irreducible* if it cannot be written as the union of two local varieties neither of which is equal to V. Proposition 3.3 below shows how every local variety can be written as a union of its irreducible subvarieties.

PROPOSITION 3.3 (see [9, Chapter 3, Theorem 1G]). *Let V be a local variety, let $M_1, M_2, \dots$ be the connected pieces of $\operatorname{Reg} V$, and let V_i denote the V-closure of M_i. Then V is the union of the V_i. Also*

(a) *each V_i is an irreducible subvariety of V of constant dimension equal to $\dim M_i$.*
(b) *the V_i form a locally finite collection of sets in V (i.e., any point $p \in V$ has a neighborhood in $\mathbf{C}^{N+1}$ which intersects only finitely many V_i's).*

The varieties V_i are called the *irreducible components* or *ircomps* of V. We will be mostly interested in the 1-dimensional ircomps. Also note that every 0-dimensional

local variety consists of isolated points (that is, $V = \mathrm{Reg}_0 V$). Indeed, the connected pieces of $\mathrm{Reg}_0 V$ are single isolated points, hence their closures are single isolated points and, by Proposition 3.3, their union is V.

PROPOSITION 3.4. *If $X = \bigcup_\alpha V_\alpha$, where the V_α are purely 1-dimensional varieties in $\mathbf{C}^{N+1}$ and the union is locally finite in $\mathbf{C}^{N+1}$, then X is a purely 1-dimensional variety in $\mathbf{C}^{N+1}$.*

Proof. This follows from [9, Chapter 2, Lemma 9J] since, by the local finiteness, the union is countable and X is locally analytic and closed in $\mathbf{C}^{N+1}$. □

PROPOSITION 3.5. *If V and W are varieties in $\mathbf{C}^{N+1}$, V is irreducible 1-dimensional, and V is not contained in W, then the set $V \cap W$ is discrete (or empty) in $\mathbf{C}^{N+1}$.*

Proof. The set $V \cap W$ is analytic in $\mathbf{C}^{N+1}$. Since $V \cap W$ is a proper subvariety of V, [9, Chapter 3, Theorem 1J] states that $\dim V \cap W < \dim V$. Hence $V \cap W$ is either (-1)-dimensional (that is, empty) or 0-dimensional, in which case it consists of isolated points. It is also closed in $\mathbf{C}^{N+1}$. □

It is convenient to introduce some notation to be used throughout the rest of the paper. For any $c \in \mathbf{C}$ define

$$H_c := \left\{(z_0, \ldots, z_N) \in \mathbf{C}^{N+1} : z_0 = c\right\}.$$

This is a hyperplane (dimension N variety) in $\mathbf{C}^{N+1}$. It is in the case $W = H_c$ that Proposition 3.6 will be applied.

The next result is essential in our approach to the connection with Puiseux series. It is known in the literature as the Remmert–Stein representation theorem (see [9, Chapter 3, section 3]). The 1-dimensional case of this theorem states that if a local variety V has constant dimension 1 and $p \in V$, then V can be locally represented by a set of equations to which Puiseux's theorem is applicable. However, V is not arbitrary: near p it must have a "good" location in the system of coordinates. Our formulation will follow [9, Theorems 3A(c) and 3D(a)].

PROPOSITION 3.6. *Let V be a local variety of constant dimension 1, and let $p = (c_0, \ldots, c_N) \in V$. Also, suppose that p is isolated in $V \cap H_{c_0}$. Then we may find*

(1) *a neighborhood $U = D_0 \times \cdots \times D_N$ of p, where the D_j are open discs;*
(2) *positive integers m_j;*
(3) *holomorphic functions $f_{j,k} : D_0 \to \mathbf{C}$,*

with the following property: for every point $(z_0, \ldots, z_N) \in V \cap U$ one has

$$z_j^{m_j} + \sum_{k=0}^{m_j - 1} z_j^k f_{j,k}(z_0) = 0 \quad \text{for} \quad j = 1, \ldots, N. \tag{3.2}$$

Proof. We may assume that p is the origin 0 since we can replace z_j with $z_j + c_j$ without affecting the leading term $z_i^{m_j}$ in (3.2). Thus, $0 \in V$ is isolated in $V \cap \mathbf{C}^N$ but is not isolated in $V \cap \mathbf{C}^{N+1}$ because $0 \notin \mathrm{Reg}_0 V = \emptyset$ (by 1-dimensionality). Now the proposition follows from [9, Chapter 2, Theorem 7A(b)] (with $n = N + 1$ and $\kappa = N$). □

The following result reinterprets the preceding proposition in terms of Puiseux series.

PROPOSITION 3.7. *Let V be a local variety of constant dimension 1, and let $p = (c_0, \ldots, c_N) \in V$. Also, suppose that p is isolated in $V \cap H_{c_0}$. Then there are a disc $D_0' = D(c_0, r)$, a positive integer M, and a finite set $\mathcal{S}$ of functions holomorphic in the disc $D(0, r^{1/M})$ with the following property. If*

$$(c_0 + t^M, z_1, \ldots, z_N) \in V \cap (D_0' \times D_1 \times \cdots \times D_N),$$

then for each $j = 1, \dots, N$ there is $\varphi \in \mathcal{S}$ such that

$$z_j = \varphi(t).$$

Proof. Use Proposition 3.6 and Corollary 3.2 for each j with $z = z_j$, $t = z_0 - c_0$ and $g_k(t) = f_{j,k}(z_0)$ to find corresponding integers M_j and sets $\mathcal{S}_j$. Then set $M = \prod M_j$ and

$$\mathcal{S} = \bigcup_j \left\{ \varphi(t^{M/M_j}) : \ \varphi \in \mathcal{S}_j \right\}. \qquad \square$$

4. Proof of main results.

4.1. Outline. Briefly, our approach is as follows. (Detailed proofs of statements below will be given in subsequent sections.) We want to use the Shapley–Snow equations to define a finite collection of varieties V_j in $\mathbf{C}^{N+1}$, each corresponding to one location of N square submatrices in the N $m_s \times n_s$ matrices $R(s, \mathbf{v}(\beta))$. Since equations (2.6) are polynomial, each such variety V_j is a variety in $\mathbf{C}^{N+1}$. The existence of the cmv-kernels implies that for each β there is j such that $(\beta, \mathbf{v}(\beta)) \in V_j$. Next, using the contraction property of the Shapley operator for the stochastic game, restricted to the cmv-kernels of the games $R_\beta(s, \mathbf{v}(\beta))$, we show that in fact $(\beta, \mathbf{v}(\beta)) \in \mathrm{Reg}_1 V_j$ ($j = j(\beta)$). We let V be the union of the closures of those connected components of the $\mathrm{Reg}_1 V_j$, which contain a corresponding point of the form $(\beta, \mathbf{v}(\beta))$; then, by Proposition 3.4, V is a variety in $\mathbf{C}^{N+1}$ of constant dimension 1. By Proposition 3.5, the set where V meets the hyperplane $H_c = \{(z_0, \dots, z_N) \in \mathbf{C}^{N+1} | z_0 = c\}$ is discrete (for all $c \in \mathbf{C}$). For $c = 1$ this, together with continuity of $\mathbf{v}$, directly implies the existence of $\lim_{\beta \to 1} \mathbf{v}(\beta)$. This also makes it possible to apply Proposition 3.7 which, along with the Puiseux theorem, readily implies that $\mathbf{v}$ has one-sided Puiseux series developments at every point of the closed interval $[0, 1]$. (In particular, it is analytic on $[0, 1]$ except for a finite set.)

Next we use the fact that whether a given submatrix is a Shapley–Snow kernel is determined by a finite number of inequalities involving meromorphic functions of the entries. This fact, along with Puiseux developments of $\mathbf{v}$, implies that Shapley–Snow kernels of $R_\beta(s, \mathbf{v}(\beta))$ can be selected piecewise constant on the interval $[0, 1]$. Therefore, by Proposition 1.1 (K2), optimal strategies can be selected to have one-sided Puiseux developments near all points of the interval $[0, 1]$.

4.2. Algebraic varieties and stochastic games. Since we want to compute the components of $\mathbf{v}(\beta)$ using the cmv-kernels, we need the condition

$$\operatorname{val} R_\beta(s, \mathbf{v}(\beta)) \neq 0 \quad (\beta \in (0, 1), \quad s = 1, \dots, N). \tag{4.1}$$

This can be achieved, without loss of generality, by replacing every reward $r(s, i, j)$ with $r'(s, i, j) = r(s, i, j) + r$, where r is large enough. Since $\mathbf{v}(\beta)$ is a bounded function of β (by Lemma 2.1), if r is large enough, we have $v_s'(\beta) > 0$ for all s and β. Moreover, adding the constant r does not affect the existence of $\lim_{\beta \to\ 1} \mathbf{v}(\beta)$ or of the Puiseux series expansion. In what follows we assume that Γ itself satisfies (4.1).

Consider any fixed collection of nonempty subsets of actions $K_s \subseteq \{1, \dots, m_s\}$ and $L_s \subseteq \{1, \dots, n_s\}$ such that $\operatorname{card} K_s = \operatorname{card} L_s$ for all $s = 1, \dots, N$. Thus we have a finite sequence of sets $\kappa = (K_1, L_1, \dots, K_N, L_N)$. By restricting the players' actions in each state s to K_s and L_s, we obtain a discounted stochastic game, $\Gamma_{\beta,\kappa}$ with auxiliary Shapley games

$$R_{\beta,\kappa}(s,\mathbf{x}) = [R_\beta(s,\mathbf{x})_{ij}]_{i\in K_s, j\in L_s}\,.$$

Let $T_{\beta,\kappa}:\mathbf{R}^N\to\mathbf{R}^N$ denote the Shapley operator (2.2) for $\Gamma_{\beta,\kappa}$.

In the original game Γ_β, for each β and $\mathbf{x}$ such that for each s $\operatorname{val}R_\beta(s,\mathbf{x})\neq 0$ select some cmv-kernel of each matrix game $R_\beta(s,\mathbf{x})$, given by

$$\overline{R_\beta(s,\mathbf{x})} = [R_\beta(s,\mathbf{x})_{ij}]_{i\in K_s(\beta,\mathbf{x}),\ j\in L_s(\beta,\mathbf{x})}\,,$$

where $K_s(\beta,\mathbf{x})\subseteq\{1,\dots,m_s\}$ and $L_s(\beta,\mathbf{x})\subseteq\{1,\dots,n_s\}$. Recall that $\overline{R_\beta(s,\mathbf{x})}$ are completely mixed square matrix games. By (4.1) and by Lemma 1.2, these kernels are defined in particular, for all those $(\beta,\mathbf{x})$ lying on the graph of $\mathbf{v}$. Also, let

$$\kappa(\beta) = (K_1(\beta,\mathbf{v}(\beta)), L_1(\beta,\mathbf{v}(\beta)),\dots,K_N(\beta,\mathbf{v}(\beta)), L_N(\beta,\mathbf{v}(\beta)))\,.$$

In other words $\kappa(\beta)$ correspond to the cmv-kernels of $R_\beta(s,\mathbf{v}(\beta))$.

Now, for a fixed κ, consider the following four systems of N equations in the $N+1$ variables $\beta,x_1,\dots,x_N$:

(V0) $x_s = \operatorname{val}R_\beta(s,x_1,\dots,x_N)\quad(s=1,\dots,N)$,

(V1) $x_s = \operatorname{val}R_{\beta,\kappa}(s,x_1,\dots,x_N)\quad(s=1,\dots,N)$,

(V2) $x_s = \frac{\det R_{\beta,\kappa}(s,x_1,\dots,x_N)}{\sum_{ij}R^{ij}_{\beta,\kappa}(s,x_1,\dots,x_N)}\quad(s=1,\dots,N)$,

(V3) $x_s\sum_{ij}R^{ij}_{\beta,\kappa}(s,x_1,\dots,x_N) - \det R_{\beta,\kappa}(s,x_1,\dots,x_N) = 0\ (s=1,\dots,N)$,

where the R^{ij} denote cofactors. To begin with, we make some simple observations concerning these systems.

(i) The systems $(V0)$ and $(V1)$ are defined in $(0,1)\times\mathbf{R}^N$. (V0) is a reformulation of $\mathbf{x}=T_\beta(\mathbf{x})$, and (V1) is a reformulation of $\mathbf{x}=T_{\beta,\kappa}(\mathbf{x})$. Let $CMV_\kappa\subseteq(0,1)\times\mathbf{R}^N$ be the set of parameters $(\beta,\mathbf{x})$, where $R_{\beta,\kappa}(s,\mathbf{x}) = \overline{R_\beta(s,\mathbf{x})}$. Then (V0) and (V1) are equivalent on CMV_κ.

(ii) The system (V2) is defined and analytic in an open subset $U_\kappa\subseteq\mathbf{C}^{N+1}$. Let $CM^+_\kappa\subseteq(0,1)\times\mathbf{R}^N$ denote the set of parameters $(\beta,\mathbf{x})$, where all the matrix games $R_{\beta,\kappa}(s,x_1,\dots,x_N)$ $(s=1,\dots,N)$ are completely mixed and have a nonzero value. Then

$$CMV_\kappa\subseteq CM^+_\kappa\subseteq U_\kappa.$$

Also, on CM^+_κ, (V2) is a simple restatement of (V1), for, by Proposition 1.1 (K2), on CM^+_κ each matrix $R_{\beta,\kappa}(s,\mathbf{x})$ is a Shapley–Snow kernel of itself and hence (K1) can be applied with $A=\overline{A}=R_{\beta,\kappa}(s,\mathbf{x})$.

(iii) The system (V3) is defined and analytic (in fact, algebraic) in $\mathbf{C}^{N+1}$. It thus defines an analytic variety in $\mathbf{C}^{N+1}$. Call this variety V_κ. Clearly, (V3) is equivalent to (V2) whenever the denominators in (V2) are nonzero.

From Lemma 1.3 it follows that CM^+_κ is an open subset of $(0,1)\times\mathbf{R}^N$.

Note that the CMV_κ are not necessarily open (or closed). However, for all $\beta\in(0,1)$ we have, by definition,

$$(\beta,\mathbf{v}(\beta))\in CMV_{\kappa(\beta)}$$

and hence

$$\bigcup_\beta(\beta,\mathbf{v}(\beta))\subseteq\bigcup_\kappa V_\kappa.$$

LEMMA 4.1. *If $p \in V_\kappa \cap CM_\kappa^+$, then $p \in \mathrm{Reg}_1 V_\kappa$. Moreover, in some neighborhood of p in $\mathbf{C}^{N+1}$, the variety V_κ is the graph of a holomorphic function from an open subset of $\mathbf{C}$ into $\mathbf{C}^N$.*

Proof. We use the fact that $T_{\beta,\kappa}$ is a contraction in $\mathbf{R}^N$. Let $p = (\beta^0, x_1^0, \dots, x_N^0)$. Near p, namely, in the set U_κ—which is a neighborhood of p as $p \in CM_\kappa^+ \subseteq U_\kappa$—the variety V_κ is, in view of remark (iii) above, defined by the system (V2). First rewrite (V2) directly in the form

$$F_\kappa(\beta, x_1, \dots, x_N) - (x_1, \dots, x_N) = 0.$$

Thus, $F_\kappa : U_\kappa \to \mathbf{C}^N$ is an analytic (in fact, rational) function, and by remark (ii) above,

$$F_\kappa(\beta, \cdot) \equiv T_{\beta,\kappa}(\cdot) \qquad \text{for} \qquad (\beta, \cdot) \in CM_\kappa^+. \tag{4.2}$$

By the complex implicit function theorem it is enough to show that the $N \times N$ complex matrix

$$D^{\mathbf{C}}_{x_1,\dots,x_N} F_\kappa(\beta^0, x_1^0, \dots, x_N^0) - I \tag{4.3}$$

is nonsingular, where $D^{\mathbf{C}}_{x_1,\dots,x_N}$ denotes complex differentiation along the complex hyperplane $H_0 = \{(x_0, \dots, x_N) \in \mathbf{C}^{N+1} | x_0 = 0\}$. By (4.2), the map $F_\kappa(\beta^0, \cdot)$ is a contraction (in the real sup-norm) from the set $\{\mathbf{x} : (\beta^0, \mathbf{x}) \in CM_\kappa^+\} \subseteq \mathbf{R}^N$ into $\mathbf{R}^N$. Let $D^{\mathbf{R}}$ denote differentiation along the real directions (and here, as the target-space is $\mathbf{R}^N$, this is equivalent to real differentiation). Since CM_κ^+ is open, and since $p \in CM_\kappa^+$, it follows that $D^{\mathbf{R}}_{x_1,\dots,x_N} F_\kappa(p)$ is also a contraction, and hence the matrix

$$D^{\mathbf{R}}_{x_1,\dots,x_N} F_\kappa(\beta^0, x_1^0, \dots, x_N^0) - I$$

is nonsingular over $\mathbf{R}$. But since F_κ is a holomorphic extension of a real function and $p \in \mathbf{R}^{N+1}$, we have $D^{\mathbf{C}} F_\kappa(p) = D^{\mathbf{R}} F_\kappa(p)$. (Alternatively, it is enough to note that, since F_κ is a rational function, both Jacobians can be computed algebraically in the same way.) Since a real nonsingular matrix is also nonsingular as a complex matrix, we find that (4.3) is indeed nonsingular and the proof is complete. □

We now prove that $(\beta, \mathbf{v}(\beta))$ lies in the 1-dimensional regular part of V_κ (for a suitable κ).

LEMMA 4.2. *For all $\beta \in (0,1)$ we have $(\beta, \mathbf{v}(\beta)) \in \mathrm{Reg}_1 V_{\kappa(\beta)}$, with $\kappa(\beta)$ defined above, corresponding to some arbitrarily selected cmv-kernels of $R_\beta(s, \mathbf{v}(\beta))$. More precisely, $\mathrm{Reg}_1 V_{\kappa(\beta)}$ meets the hyperplane H_β transversally at the point $(\beta, \mathbf{v}(\beta))$.*

Proof. By Lemma 4.1 it is enough to show that $(\beta, \mathbf{v}(\beta)) \in V_{\kappa(\beta)} \cap CM^+_{\kappa(\beta)}$. Since $\kappa(\beta)$ corresponds to the cmv-kernels, we have $(\beta, \mathbf{v}(\beta)) \in CMV_{\kappa(\beta)} \subseteq CM^+_{\kappa(\beta)}$. To see that $(\beta, \mathbf{v}(\beta)) \in V_{\kappa(\beta)}$, recall that by (i), (ii), (iii) on $CMV_{\kappa(\beta)}$, (V3) is equivalent to (V0) which, evaluated at $(\beta, \mathbf{v}(\beta))$, has the form

$$\mathbf{v}(\beta) = T_\beta(\mathbf{v}(\beta)), \tag{4.4}$$

which is true by definition. The fact that $\mathrm{Reg}_1 V_{\kappa(\beta)}$ meets H_β transversally at $(\beta, \mathbf{v}(\beta))$ follows from the second statement of Lemma 4.1. □

We are now in a position to define a variety V associated to a discounted stochastic game.

THEOREM 4.3. *There exists a variety $V = V(\Gamma)$ in $\mathbf{C}^{N+1}$ with the following properties*

(a) *V has constant dimension* 1.
(b) *For every $c \in \mathbf{C}$ the intersection of V with the hyperplane $H_c = \{(z_0, \ldots, z_N) \in \mathbf{C}^{N+1} | z_0 = c\}$ is a discrete subset of $\mathbf{C}^{N+1}$.*
(c) *For every $\beta \in (0,1)$ we have $(\beta, \mathbf{v}(\beta)) = (\beta, v_1(\beta), \ldots, v_N(\beta)) \in V$.*

Proof. By Lemma 4.2, for every $\beta \in (0,1)$ there is $\kappa(\beta)$ such that $(\beta, \mathbf{v}(\beta)) \in \mathrm{Reg}_1 V_{\kappa(\beta)}$. Let V_β be the closure of the connected component of $\mathrm{Reg}_1 V_{\kappa(\beta)}$, containing the point $(\beta, \mathbf{v}(\beta))$. (Note that it may frequently happen that $V_{\beta'} = V_{\beta''}$ with $\beta \neq \beta''$.) Then, by Proposition 3.3(a), V_β is an irreducible purely 1-dimensional variety in $\mathbf{C}^{N+1}$. Also, V_β is not contained in H_c; otherwise $c = \beta$, but in this case the transversality property in Lemma 4.2 leads to a contradiction. Since H_c is a variety in $\mathbf{C}^{N+1}$ (defined by the equation $x_0 = c$), Proposition 3.5 states that the set $H_c \cap V_\beta$ is discrete. Let $V = V(\Gamma)$ be defined by

$$V = \bigcup_{\beta \in (0,1)} V_\beta.$$

By construction, V is a union of certain 1-dimensional ircomps of the V_κ. By Proposition 3.3(b), and since there are only a finite number of κ's, the union is locally finite. Hence, Proposition 3.4 implies that V is a purely 1-dimensional variety in $\mathbf{C}^{N+1}$. Also, by local finiteness the set $H_c \cap V$ is discrete. ∎

Remark. The variety V constructed in Theorem 4.3 may not necessarily be unique since it may depend on the choice of $\kappa(\beta)$. However, this fact is not relevant to the application of Theorem 4.3. The notation $V = V(\Gamma)$ is therefore a little misleading. This can be easily remedied if we define $V(\Gamma)$ as the smallest V satisfying Theorem 4.3. In fact it can be shown that the intersection of all the V's satisfying Theorem 4.3 also satisfies that theorem.

THEOREM 4.4. *The value vector $\mathbf{v}(\beta)$ converges in $\mathbf{C}^N$ as $\beta \nearrow 1$.*

Proof. Let $V = V(\Gamma)$ and H_1 be as in Theorem 4.3 (i.e., $H_1 = H_c$, where $c = 1$). Since $\mathbf{v}$ is a bounded function of β (by Lemma 2.1), take a compact set $K \subseteq \mathbf{C}^{N+1}$ such that $(\beta, \mathbf{v}(\beta)) \in K$ for every β. Then, by Theorem 4.3(b), the set $P = H_1 \cap V \cap K$ is discrete in $\mathbf{C}^{N+1}$. Also, since K is compact, P is finite. Let $P = \{p_1, p_2, \ldots p_r\}$.

Next, we claim that

$$\mathrm{dist}((t,x), P) \to 0 \quad \text{as} \quad (t,x) \in V \cap K, \quad t \to 1 \quad \text{and} \quad x \in \mathbf{C}^N. \tag{4.5}$$

Suppose that it is not so. Then there exist $\delta > 0$ and a sequence $(a_k) \equiv (t_k, x_k)$ with $t_k \to 1$ and $a_k \in V \cap K$ such that $\mathrm{dist}(a_k, P) > \delta$ for each k. There is a subsequence of (t_k, x_k) converging to a point $p = (1, x_\infty)$, and $p \in V \cap K$ since V and K are closed. It follows that $p \in N$. Thus, δ cannot exist.

From (4.5) and Theorem 4.3(c) it follows that

$$\mathrm{dist}((\beta, \mathbf{v}(\beta)), P) \to 0 \quad \text{as} \quad \beta \to 1. \tag{4.6}$$

In particular, this implies that P is not empty.

Now, since P is finite, take a fixed $\epsilon > 0$ such that the balls $B_j = B(p_j, \epsilon)$ are disjoint. By (4.6), there exists $c < 1$ such that for all $\beta > c$, $(\beta, \mathbf{v}(\beta)) \in B_k$ for some $k = k(\beta)$. But since $\mathbf{v}(\beta)$ is continuous on the interval $(c, 1)$ (Lemma 2.1), $k(\beta)$ must be the same for all $\beta > c$; otherwise the preimages under the continuous function $\beta \mapsto (\beta, \mathbf{v}(\beta))$ of the B_k would partition the interval (c,1) into at least two disjoint nonempty open sets. Thus it follows from (4.6) that $\lim_{\beta \to 1} \mathbf{v}(\beta) = p_k$. ∎

COROLLARY 4.5. *The function $\mathbf{v}$ can be extended continuously over the interval $[0,1]$. Also, $(\beta, \mathbf{v}(\beta)) \in V(\Gamma)$ for all $\beta \in [0,1]$.*

Proof. The first assertion follows directly from Theorem 4.4 (having Lemma 2.1 in mind). The second follows from Theorem 4.4 and the fact that $V(\Gamma)$ is closed. □

4.3. The connection with Puiseux series. We are now in a position to apply Proposition 3.7 to the variety V at the point $p = (1, v_1(1), \dots, v_N(1))$. In fact we prove a slightly more general result.

THEOREM 4.6. *There is a finite set $\mathcal{N} \subseteq [0,1]$ such that $\mathbf{v}$ is analytic on $[0,1]\backslash\mathcal{N}$ and develops (component-wise) into a real Puiseux series in one-sided neighborhoods of each point of $\mathcal{N}$.*

More precisely, for any $c_0 \in (0,1]$ there exist $\epsilon > 0$ and $M \in \mathbf{N}$ such that each v_j is an analytic function of $u = \sqrt[M]{c_0 - \beta}$ for $\beta \in (c_0 - \epsilon, c_0]$, and a similar statement holds for each $c_0 \in [0,1)$, $u = \sqrt[M]{\beta - c_0}$, and $\beta \in [c_0, c_0 + \epsilon)$.

Proof. It is sufficient to prove the second part of the theorem. The fact that V is analytic outside a finite set $\mathcal{N}$ can be deduced from this second part by selecting a suitable finite subcover of $[0,1]$ and noting that a Puiseux function is analytic except at the base point. Let $V = V(\Gamma)$ be as in Theorem 4.3. Let $c_0 \in [0,1]$ be fixed, and denote

$$p := (c_0, \mathbf{v}(\beta)) = (c_0, v_1(c_0), \dots, v_N(c_0)) = (c_0, c_1, \dots, c_N).$$

By Theorem 4.3(b) the set $V \cap H_{c_0}$ is discrete. (In fact, $V \cap H_c$ is discrete for every $c \in \mathbf{C}$.) Hence p is isolated in $V \cap H_{c_0}$. Hence Proposition 3.7 is applicable, and we will follow its notation.

Since $\mathbf{v}$ is continuous at c_0, there is $\epsilon_1 > 0$ such that if $\beta \in [0,1]$ and $|\beta - c_0| < \epsilon_1$, then $(\beta, \mathbf{v}(\beta)) \in V \cap (D_0' \times \cdots \times D_N)$.

Let $t = \alpha u$, where α is chosen so that $t^M = \beta - c_0$; say, $\alpha = 1$ if we consider $\beta \geq c_0$ and $\alpha = e^{2\pi i/M}$ if we consider $\beta \leq c_0$. Thus, by Proposition 3.7, for any $j \in \{1, \dots, N\}$ we have

$$\text{if } |t^M| < \epsilon_1, \text{ then } (\exists \varphi \in \mathcal{S}) \quad v_j(c_0 + t^M) = \varphi(t) = \varphi(\alpha u). \tag{4.7}$$

Since the elements of $\mathcal{S}$ are holomorphic, the zero set of $\varphi - \phi$, where $\varphi, \phi \in \mathcal{S}$ and $\varphi \not\equiv \phi$, cannot have 0 as a limit point. Hence, there is $\epsilon_2 > 0$ with $\epsilon_2 \leq \epsilon_1$ such that

$$\text{if } \varphi, \phi \in \mathcal{S} \text{ and } \varphi \not\equiv \phi \text{ and } 0 < |t^M| < \epsilon_2, \text{ then } \varphi(t) \neq \phi(t) \tag{4.8}$$

By the continuity of v_j and of the $\varphi \in \mathcal{S}$, it follows that for each j the choice of the function $\varphi \in \mathcal{S}$ in (4.7) must be locally constant for $|t^M| \in (0, \epsilon_2)$. It follows that the choice of φ must be constant for $\beta \in I^+ = (c_0, c_0 + \epsilon_2) \cap [0,1]$ and for $\beta \in I^- = (c_0 - \epsilon_2, c_0) \cap [0,1]$, because the intervals I^+ and I^- are connected. Thus, by (4.7),

$$v_j(\beta) = f(u) \quad \text{on } I^+, \tag{4.9}$$

where $f(x) = \varphi(\alpha x)$ and φ is a fixed element of $\mathcal{S}$. An analogous result holds on I^-. However, the choice of φ (a fixed element of $\mathcal{S}$) may be different on I^-. It suffices to consider the case where $I^+ \neq \emptyset$ (resp., $I^- \neq \emptyset$). Since v_j is real valued on I^+ (resp., I^-), the power series of f has real coefficients. By continuity, (4.9) holds for $u = 0$ as well. By formula (4.9) the theorem is proven. □

5. Connection with analytic matrix games, Shapley–Snow kernels, and optimal strategies. In this subsection we derive from Theorem 4.6 some results on Shapley–Snow kernels and optimal strategies. First we present a lemma on analytic matrix games.

LEMMA 5.1. *Let $A(t)$ $(-r \le t \le r)$ be a family of nonzero-value matrix games of constant dimension $m \times n$, whose entries are holomorphic in the disc $D(0, r')$ with $r' > r$. Then it is possible to choose a Shapley–Snow kernel of $A(t)$ for each $t \in [-r, r]$ so that the location of the kernels is piecewise constant.*

More precisely, there exist a finite partition $[-r, r] = \cup_k I_k$, where the I_k are open or degenerate disjoint intervals, and sets $K_k \subseteq \{1, \dots, m\}$ and $L_k \subseteq \{1, \dots, n\}$ such that $[A(t)_{ij}]_{i \in K_k, j \in L_k}$ is a Shapley–Snow kernel of $A(t)$ for all $t \in I_k$.

Proof. Let $A(t) = [a_{ij}(t)]_{i,j=1}^{m,n}$ and consider any submatrix $B(t) = [a_{ij}(t)]_{i \in K, j \in L}$, where the sets K and L are fixed. The necessary and sufficient conditions for $B(t)$ to be a Shapley–Snow kernel of $A(t)$ are the following:

$$\begin{aligned} &\det B(t) \neq 0, \mathbf{1}^T B(t)^{-1} \mathbf{1} \neq 0, \text{ and} \\ &(p(t)^o, q(t)^o) \text{ is a pair of optimal strategies for } A(t), \end{aligned} \tag{5.1}$$

where

$$p(t) = \frac{\mathbf{1}^T B(t)^{-1}}{\mathbf{1}^T B(t)^{-1} \mathbf{1}} \text{ and } q(t) = \frac{B(t)^{-1} \mathbf{1}}{\mathbf{1}^T B(t)^{-1} \mathbf{1}} \tag{5.2}$$

and x^o denotes a vector of dimension m or n obtained by inserting zeroes in entries that do not appear in x. (Then one also automatically has val $A(t) =$ val $B(t) = (\mathbf{1}^T B(t)^{-1} \mathbf{1})^{-1}$.)

The conditions of (5.1) are equivalent to

$$\begin{aligned} &\det B(t) \neq 0, \mathbf{1}^T B(t)^{-1} \mathbf{1} \neq 0, \\ &p_i(t) \ge 0 \text{ and } q_j(t) \ge 0 \quad (i = 1, \dots, m, j = 1, \dots, n), \text{ and} \\ &e_i A(t) q(t)^o \le p(t)^o A(t) q(t)^o \le p(t)^o A(t) e_j \quad (i = 1, \dots, m, j = 1, \dots, n), \end{aligned} \tag{5.3}$$

where the e_k are unit coordinate vectors of appropriate dimension. (One also automatically has $\sum_i p_i(t) = \sum_j q_j(t) = 1$ by (5.2).)

We may leave out the case where $\det B(t) \equiv 0$ on $D(0, r')$. Then the entries of $B(t)^{-1}$ are meromorphic on $D(0, r')$. Hence the formulae in (5.4) involve functions meromorphic in $t \in D(0, r')$. Consequently, each of these formulas becomes an equality either for all $t \in [-r, r]$ or at most for a finite number of values of $t = t_0, \dots, t_M \in [-r, r]$, and we may assume $-r = t_0 < \cdots < t_M = r$. On each interval (t_k, t_{k+1}) these functions are continuous, so the relations in (5.3) become $>$ or $<$ or $=$ and are fixed.

Consequently, if $B(t)$ is a Shapley–Snow kernel for some $t \in (t_k, t_{k+1})$, then it is so for all $t \in (t_k, t_{k+1})$. Since a Shapley–Snow kernel exists for each $t \in [-r, r]$ (by Proposition 1.1), if we set $I_0 = \{t_0\}$, $I_1 = (t_0, t_1)$, $I_2 = \{t_1\}$, $I_3 = (t_1, t_2)$, etc., then the submatrices $K_k \times L_k$ can be found. ∎

THEOREM 5.2. *There is a finite partition of the interval $[0, 1]$ into subintervals I_k, open or degenerate, such that the Shapley–Snow kernels of the matrix games $R_\beta(s, \mathbf{v}(\beta))$ can be chosen for each $\beta \in [0, 1]$ in such a way that the location of these kernels is constant on each I_k.*

Proof. By Theorem 4.6 and by compactness, the interval $[0, 1]$ can be divided into subintervals $I_k = [t_k, t_{k+1}]$ $(0 = t_0 < \cdots < t_s = 1)$ so that the entries of the

$R_\beta(s, \mathbf{v}(\beta))$ are given by analytic functions of $u = \sqrt[M]{\beta - t_k}$ or $u = \sqrt[M]{t_{k+1} - \beta}$, having holomorphic extensions over a disc $D(0, r')$ with $r' > \sqrt[M]{t_{k+1} - t_k}$. By Lemma 5.1, the interval $[0, \sqrt[M]{t_{k+1} - t_k}]$ has the desired partition, which can be pulled back to I_k as the real Mth root function $\sqrt[M]{}$ is increasing. Now it is enough to use these partitions, letting each t_k form a degenerated interval. □

COROLLARY 5.3. *One can choose some optimal stationary strategy pair* $(u^0(\beta), v^0(\beta))$ *of* Γ_β *for each* $\beta \in [0, 1)$ *in such a way that there exists a set* $\{c_k\}_{k=0}^n$, $0 = c_0 < \cdots < c_n = 1$, *satisfying the following:*

(i) *For* $\beta \in [0, 1] \backslash \{c_k\}_{k=0}^n$, u^0 *and* v^0 *are analytic functions of* β.
(ii) *For all* $k = 1, \ldots, n$, *there exists* $\epsilon > 0$ *such that for* $\beta \in (c_k - \epsilon, c_k)$ *each component of* u^0 *and* v^0 *is a real Puiseux function of* $c_k - \beta$.
(iii) *For all* $k = 0, \ldots, n-1$, *there exists* $\epsilon > 0$ *such that for* $\beta \in (c_k, c_k + \epsilon)$ *each component of* u^0 *and* v^0 *is a real Puiseux function of* $\beta - c_k$.

Proof. Let a set $\{c_0, \ldots, c_M\}$ with $0 = c_0 < \cdots < c_n = 1$ contain all the partition points implied by Theorems 5.2 and 4.6. Let $A(s, \beta)$ denote some corresponding Shapley–Snow kernels of $R_\beta(s, v(\beta))$, whose location, by Theorem 5.2, may be assumed to be constant on each (c_k, c_{k+1}).

By subsection 2.1, a pair $(u^0(\beta), v^0(\beta))$ will be optimal for Γ_β if

$$u^0(\beta) = (p(1, \beta), \ldots, p(N, \beta)) \text{ and } v^0(\beta) = (q(1, \beta), \ldots, q(N, \beta)),$$

where each $(p(s, \beta), q(s, \beta))$ is an optimal strategy pair for $R_\beta(s, v(\beta))$.

By Proposition 1.1 we can put

$$p(s, \beta) = \mathbf{1}^T A(s, \beta)^{-1} \mathrm{val} A(s, \beta),$$

$$q(s, \beta) = A(s, \beta)^{-1} \mathbf{1} \mathrm{val} A(s, \beta), \tag{5.4}$$

$$\text{where } \mathrm{val} A(s, \beta) = (\mathbf{1}^T A(s, \beta)^{-1} \mathbf{1})^{-1}.$$

For $\beta \in (c_k, c_{k+1})$ the entries of $A(s, \beta)$ depend analytically on β (since $\mathbf{v}(\beta)$ does, by Theorem 4.6). For $\beta \in (c_k - \epsilon, c_k)$ $(k = 1, \ldots, n)$ or $\beta \in (c_k, c_k + \epsilon)$ $(k = 0, \ldots, n-1)$ the entries of $A(s, \beta)$ are Puiseux functions of $|\beta - c_k|$ (by the same Theorem 4.6). By the formulas in (5.4), the same can be said about $p(s, \beta)$ and $q(s, \beta)$. In addition, here all functions are real. □

REFERENCES

[1] T. BEWLEY AND E. KOHLBERG, *The asymptotic theory of stochastic games*, Math. Oper. Res., 3 (1976), pp. 197–208.
[2] O. FORSTER, *Lectures on Riemann Surfaces*, Graduate Texts in Math. 81, Springer-Verlag, New York, 1981.
[3] I. KAPLANSKY, *A contribution to von Neumann's theory of games*, Ann. of Math., 46 (1945), pp. 474–479.
[4] J. F. MERTENS AND A. NEYMAN, *Stochastic games*, Internat. J. Game Theory, 10 (1981), pp. 53–56.
[5] J. VON NEUMANN, *Zur Theorie der Gesellschaftsspiele*, Math. Ann., 100 (1928), pp. 295–320.
[6] L. S. SHAPLEY, *Stochastic games*, Proc. Nat. Acad. Sci. USA, 39 (1953), pp. 1095–1100.
[7] L. S. SHAPLEY AND R. N. SNOW, *Basic Solutions of Discrete Games*, Ann. of Math. Stud. 24, Princeton University Press, Princeton, NJ, 1952.
[8] H. WEYL, *Elementary Proof of a Minimax Theorem due to von Neumann*, Ann. of Math. Stud. 24, Princeton University Press, Princeton, NJ, 1950.
[9] H. WHITNEY, *Complex Analytic Varieties*, Addison-Wesley, Reading, MA, 1972.

SIAM J. CONTROL OPTIM.
Vol. 35, No. 3, pp. 876–900, May 1997

007

CONSTRAINED REGULAR LQ-CONTROL PROBLEMS*

G. STEFANI† AND P. ZEZZA‡

Abstract. In this paper we give a complete description of how the Jacobi theory of conjugate points can be extended to a regular linear-quadratic control problem where the state end-points are jointly constrained to belong to a subspace of $\mathbf{R}^n \times \mathbf{R}^n$ and there is a linear pointwise state-control constraint. We introduce also the definition of semiconjugate point which describes a distinctive feature of these problems and state the corresponding necessary and sufficient conditions for the quadratic form to be nonnegative or coercive. In the case in which the constraints and the costs act separately on the initial and final points we give equivalent characterizations of the coercivity of the quadratic form by means of the solutions of an associated Riccati equation in both the controllable and uncontrollable case.

Key words. linear-quadratic problems, conjugate points, end-point constraints, pointwise state-control constraints, necessary and sufficient conditions, Riccati equation

AMS subject classifications. 49N10, 49N15, 49N35

PII. S0363012995286848

1. Introduction.

Although linear-quadratic (LQ) problems have been studied since the birth of control theory, there are still interesting questions which wait for a complete answer. This is the case for the problem we analyze in this paper. It is an LQ-control problem whose main features are that the state end-points are jointly constrained to belong to a subspace of $\mathbf{R}^n \times \mathbf{R}^n$ and that there is a linear pointwise state-control constraint. These kinds of problems arise while studying second-order optimality conditions in nonlinear problems with mixed state-control constraints of both equality and inequality type. Second-order conditions are needed in nonlinear problems both when the candidate solution is not unique (see [13] for a numerical example) and when an existence theorem cannot be applied (see Example 4.4 and [13] for numerical examples).

The aim of this paper is to give a complete description of how the Jacobi theory of conjugate points can be extended to this kind of regular LQ-control problems with a possibly indefinite cost; here, regular means that a suitable version of the Legendre condition is fulfilled. We are then able to state necessary and sufficient conditions for the quadratic form to be nonnegative or coercive. This issue is interesting because these problems have properties different from the usual ones. For this reason we introduce the definition of semiconjugate point because the points which characterize the coercivity and the nonnegativity of the quadratic form are different when the problem is not controllable or when it does not have at least one fixed end-point.

In the case when the constraints and the costs act separately on the initial and final points we give an equivalent characterization of the coercivity of the quadratic form by means of the solutions of the associated Riccati equation in both the controllable and the uncontrollable case. In the case of controllable systems the result is sharper, and it allows us to solve the associated affine problem; see [22]. These

*Received by the editors May 31, 1995; accepted for publication (in revised form) March 28, 1996. This research was supported in part by Ministero Università e Ricerca Scientifica research grants "Teoria dei sistemi e del controllo" and "Equazioni differenziali ordinarie e applicazioni."
http://www.siam.org/journals/sicon/35-3/28684.html

†Dipartimento di Matematica e Applicazioni, Via Mezzocannone 8, 80134 Napoli, Italy (stefani@facec.cce.unifi.it).

‡DiMaDEFAS, Via C. Lombroso 6/17, 50134 Firenze, Italy (pzezza@facec.cce.unifi.it).

Riccati-type conditions can be applied also to problems with boundary conditions given jointly on both end-points, for example, periodic, by transforming the problem into another one in double dimension. Here we give a sufficient condition which can be simply stated and proved.

The previous results, together with the results in [23] (announced in [18]), give necessary and/or sufficient conditions for a constrained nonlinear optimal control problem to have a weak local minimum. For a survey, see [21].

Our results extend to the control, setting some classical results for problems in the calculus of variations, where an important role is played by the Legendre and Jacobi conditions. They concern the quadratic form obtained as the second variation of a nonlinear problem. The original results are about the simplest problem in the calculus of variations, where both end-points are fixed (Jacobi, 1837), and they have been generalized by Kneser, 1896, to the case of one fixed end-point. The statements are based on the definition of conjugate (both fixed end-points) or focal point (one fixed end-point). Until the crucial work by Hestenes [7] and Morse [17], the improvements concerned only the regularity of the data and of the optimal solution while they developed an index theory for quadratic forms in Hilbert spaces, and they explain the connection between the index of the quadratic form and the conjugate (or focal) points.

Conjugate points and the Riccati equation have been widely studied in connection with regular LQ problems, especially for problems with at least one end-point fixed. It is impossible to quote all the results published since the seminal papers by Kalman [9] and Hestenes [7]. Here we mention the contributions in [5, 6, 8] and [15], where the Hestenes approach has been used to consider optimal control problems; those in [3, 10], where the classical approach has been extended to the optimal control setting but in a rather informal and nonrigorous approach; and those in [4, 16], where an approach to the Riccati equation closer to ours is used.

As far as conjugate points are concerned, the problem when both end-points are variable has been analyzed in [26, 27], where, under suitable controllability assumptions, the authors introduce the concept of coupled point, and they state the corresponding necessary conditions. The calculus of variations case with separate end-point conditions is studied in [24], where necessary and sufficient conditions for the coercivity of the quadratic form are given by means of coupled points and, equivalently, by the properties of a solution of the Riccati equation. In [28] the author gives necessary and sufficient conditions for the case of joint end-points constraints, still with suitable controllability assumptions.

LQ problems with constraints can be seen as those defining the accessory minimization problem associated with nonlinear optimization problems with state-control and end-point equality and inequality constraints. There is a large amount of literature on these problems; for some recent results see, for example, [14, 19, 23, 25] and the references therein.

In [25], while studying a nonlinear control problem with state-control pure inequality constraints and with fixed initial point, Zeidan analyzes the accessory LQ problem, assuming that the linearized system is controllable. Zeidan gives necessary conditions for the nonnegativity of the quadratic form by means of conjugate points and sufficient conditions for its coercivity by the existence of a solution of the Riccati equation satisfying certain boundary conditions.

In the appendix we summarize an example of an application to an economic model (Example 4.4). Techniques similar to those that we study here are used in [13], and

in other papers by the same authors, to analyze some applied examples. The results we state in this paper widen the class of problems to which these techniques can be applied and allow us to use either conjugate points or Riccati-type results to check both necessary and sufficient conditions.

In [29] an abstract Jacobi theory has been presented. We are going to use the results therein to develop a conjugate point approach to LQ problems in the presence of constraints. The abstract theory points out that the different definitions of conjugate, focal, and coupled point can be seen as corresponding to the same object in different situations. For this reason we go back to the original name of conjugate point, including in this definition all the previously mentioned cases. To take into account the situation related to the existence of a focal interval or table we introduce the definition of semiconjugate point.

The abstract theory analyzes the sign of a form J on an Hilbert space H by means of its restriction to a family of subspaces depending on a parameter c. The changes of the index or the nullity of the form are investigated through a value function V which is the minimum of the form J on a level set of a weakly continuous positive quadratic form in a subspace belonging to the family. A kind of "continuation principle" is given in the sense that, roughly speaking, the Jacobi condition states that if V is positive at level $c = 0$ and it does not become zero along the family, then at the final level, which corresponds to the whole space, it is still positive and hence the form is coercive. Analogously, if the function V is nonnegative at level $c = 0$ and it does not change sign, then the form is nonnegative. A detailed analysis is needed to characterize these points by the properties of the solutions of the Jacobi system.

The Hilbert space we consider is given by the couples made by the initial state and the control which satisfy the constraints; this is because the initial condition plays the role of a further control. From this point of view it is quite natural to consider the family of subspaces obtained by taking as parameter c the time and by setting the control equal to a reference one after c. This reference control is chosen in a feed-back form such that the state-control constraint is satisfied for every initial point. With this choice the subspace at level $c = 0$ is finite dimensional so that it is easy to check if the form is nonnegative there. This approach leads to the definition of semiconjugate point which can usually be given through the solutions of the Jacobi system and through the transversality conditions which can be obtained by setting the first derivative of the form constrained to the subspace equal to zero. The transversality conditions include an extra term which does not appear when the moving end-point is constrained to be zero. Let us remark that here we consider as fixed the left end-point of the time interval $t = 0$, and we take the other one as a parameter. We could very well do the opposite, doubling all the statements and obtaining symmetric results.

Differently from previous works, we do not assume any kind of controllability with respect to the boundary conditions here, but we need a surjectivity assumption on the state-control constraint (Assumption 2.2) and a suitable version of the Legendre–Clebsch condition (Assumption 2.3) which allows us to give the Jacobi system only by means of the trajectory and the adjoint covector. Theorems 2.5 and 2.6 give a complete characterization of the coercivity and of the nonnegativity of the studied quadratic form by means of semiconjugate and conjugate points, including and extending all the previous results on the subject. Preliminary results has been announced in [20].

These results concerning conjugate points are crucial in proving the necessary and sufficient conditions in the framework of Riccati theory. Here we fully analyze only the case when costs and boundary conditions are given separately on the two end-points.

Theorem 2.9 states that the coercivity of the form plus a controllability assumption is equivalent to the existence of a solution of the Riccati equation on the half open interval $(0,T]$ with suitable boundary conditions. While, without any controllability assumption, Theorem 2.10 states the equivalence between the coercivity of the form and the existence of a solution of the Riccati equation on the closed interval $[0,T]$ with suitable inequality-type boundary conditions. Under stronger controllability assumptions it is possible to characterize also the nonnegativity of the quadratic form by a solution of the Riccati equation, Theorem 2.11. These theorems include and extend the previous results on the subject.

For the case of joint costs and boundary conditions we give here only some simple, partial results in Theorem 2.12 and Corollary 2.13; a complete description will be the subject of a forthcoming paper. The plan of the paper is the following. In section 2 we describe the main assumptions and results. Subsection 2.1 is dedicated to the conjugate points, and subsection 2.2 concerns the Riccati equation. All the proofs and all the needed lemmas are in section 3. In the appendix there are some related results and examples.

2. Statement of the problem and main results. On a given compact interval $[0,T]$ we consider the quadratic form

$$\begin{aligned} I(\eta,u) &= \frac{1}{2}(\eta^\top(0),\eta^\top(T))\Gamma\begin{pmatrix}\eta(0)\\ \eta(T)\end{pmatrix} \\ &+\frac{1}{2}\int_0^T \left\{\eta^\top(s)P(s)\eta(s)+2u^\top(s)Q(s)\eta(s)+u^\top(s)R(s)u(s)\right\}ds, \end{aligned} \tag{2.1}$$

defined on the subspace of $AC([0,T],\mathbf{R}^n)\times L^2([0,T],\mathbf{R}^m)$ of the couples (η,u) satisfying

$$\dot\eta(t)=A(t)\eta(t)+B(t)u(t) \text{ almost everywhere (a.e.) } t\in[0,T]. \tag{2.2}$$

This subspace is an Hilbert space since it can be identified with

$$U=\mathbf{R}^n\times L^2([0,T],\mathbf{R}^m)$$

by $(\eta,u)\mapsto(\eta(0),u)$. We study the restriction of the quadratic form I to the closed subspace K of U given by two different types of constraints. The first one is a boundary condition on the state end-points

$$N\begin{pmatrix}\eta(0)\\ \eta(T)\end{pmatrix}=0, \tag{2.3}$$

whereas the second one is a pointwise state-control constraint which is an infinite dimensional constraint

$$C(t)\eta(t)+D(t)u(t)=0 \text{ a.e. } t\in[0,T]. \tag{2.4}$$

Γ is a symmetric matrix in $M_{2n\times 2n}$, $N\in M_{p\times 2n}$. We assume that the data satisfy the following standard minimal assumptions

$$\begin{gathered} A,P\in L^1([0,T],M_{n\times n}),\ B\in L^2([0,T],M_{n\times m}),\ Q\in L^2([0,T],M_{m\times n}),\\ R\in L^\infty([0,T],M_{m\times m}),\ C\in L^2([0,T],M_{k\times n}),\ D\in L^\infty([0,T],M_{k\times m}) \end{gathered}$$

and that $P(t)$ and $R(t)$ are symmetric matrices. When any of this time-dependent matrix acts on a function, it becomes an operator between L^p spaces, which will be denoted by the same capital letter. In the following all the equalities between L^p functions are assumed to hold a.e.

If the quadratic form describes an accessory minimization problem coming from inequality constraint, then the set of active constraint depends on t. To include this possibility we have to allow the ranges of the maps C, D to change. This can be done by assuming the following.

Assumption 2.1. There exists $\delta_i \in L^\infty([0,T],\mathbf{R})$, $\delta_i(t) \in \{0,1\}, i = 1,\dots,k$, such that the orthogonal projection defined by $\Delta(t) = \operatorname{diag}\{\delta_1(t),\delta_2(t),\dots,\delta_k(t)\}$ satisfies

$$\Delta C = C \quad \text{and} \quad \Delta D = D.$$

Our aim is to give necessary and sufficient conditions for the quadratic form to be coercive or nonnegative on K. Combining these results with those in [23], we obtain a complete set of necessary and sufficient conditions for a weak local minimum in presence of equality-type constraints.

Besides the above regularity assumptions on the data, we make two main assumptions which have been considered by several authors also in the nonlinear case [11, 12]. The first one concerns the regularity of the infinite-dimensional constraint (2.4). By Assumption 2.1, the map $(\eta(0),u) \mapsto C\eta + Du$ is a map from U into Range $\Delta \subset L^2([0,T],\mathbf{R}^k)$, and it is onto if and only if the following assumption is satisfied (see Lemma 4.1 in the Appendix). An analogous result has been proved in [19] for the L^∞ case.

Assumption 2.2. There is $h > 0$ such that

$$D\,D^\top + (Id - \Delta) \geq h\,Id.$$

Under this assumption we can define the operator $D^\sharp \in L^\infty([0,T], M_{m\times k})$:

$$D^\sharp \equiv D^\top (DD^\top + (Id - \Delta))^{-1}.$$

Since Δ is an orthogonal projection, Assumption 2.1 yields

$$(Id - \Delta)(DD^\top + (Id - \Delta))^{-1} = (Id - \Delta);$$

hence

$$D\,D^\sharp = \Delta. \tag{2.5}$$

The maps

$$\Pi_1 = Id - D^\sharp D, \quad \Pi_2 = D^\sharp D$$

are orthogonal projections in $L^2([0,T],\mathbf{R}^m)$, and they give an orthogonal decomposition of U:

$$U_1 = \mathbf{R}^n \times \operatorname{Range}\Pi_1,\ U_2 = \{0\} \times \operatorname{Ker}\Pi_1.$$

The other main assumption we are going to make is a suitable version of the strengthened Legendre–Clebsch condition on K; see also [15].

Assumption 2.3. The quadratic form I satisfies the strengthened Legendre–Clebsch condition on K; that is, one of the following equivalent statements holds:

$$\text{there are } h_0, h_1 > 0 \text{ such that } R + h_1\Pi_2 \geq h_0\, Id, \tag{2.6}$$

$$\text{there is } h > 0 \text{ such that } \Pi_1 R \Pi_1 + \Pi_2 \geq h\, Id, \tag{2.7}$$

$$\text{there is } h > 0 \text{ such that } \Pi_1 R \Pi_1 \geq h\Pi_1. \tag{2.8}$$

The proof of the equivalence among the statements is in Lemma 4.2 in the appendix. Notice that when the number of active constraints is equal to the number of controls, then $\Pi_1 \equiv 0$ and Assumption 2.3 is empty.

In what follows we use a natural generalization of the definition of controllability for the linear system (2.2) constrained by (2.4); namely, we will say the following.

DEFINITION 2.1. *The constrained system* (2.2) *and* (2.4) *is* N*-controllable at time* t_1 *if the map*

$$(x, u) \mapsto N \begin{pmatrix} \eta(0) \\ \eta(t_1) \end{pmatrix}$$

is onto Range N*, where* η *is the solution of* (2.2) *corresponding to* u*, with* $\eta(0) = x$ *and* (η, u) *satisfying* (2.4)*. For*

$$N = \begin{pmatrix} L & 0 \\ 0 & Id \end{pmatrix},$$

the above definition coincides with the usual definition of controllability from a subspace, and we say that the constrained system (2.2) *and* (2.4) *is controllable from* Ker L *at time* t_1.

By Assumption 2.1, if we take controls of the form

$$u = \Pi_1 u + \Pi_2 u = \Pi_1 u - D^\sharp C\eta,$$

then the constraint (2.4) is verified independently from the starting point of the trajectory. Thus the solutions of the constrained control system (2.2) and (2.4) correspond to the solutions of

$$\dot{\xi}(t) = \bar{A}(t)\xi + B(t)\Pi_1(t)u(t), \tag{2.9}$$

where

$$\bar{A} \equiv A - BD^\sharp C. \tag{2.10}$$

This is because if (η, u) is a solution of (2.2) and (2.4), then $\xi = \eta$ is a solution of (2.9) with control $\Pi_1 u$; see the beginning of section 3 for a detailed analysis.

Remark 2.1. By the above properties we can give an equivalent definition of N-controllability without a direct reference to the constraint (2.4). In fact, the system (2.2) and (2.4) is N-controllable at time t_1 if and only if the system (2.9) is N-controllable at time t_1.

We can now describe how the Jacobi theory of conjugate points can be extended to these problems. Following [29] we define the conjugate points by means of a family of problems which are parametrized by the time $c \in [0, T]$. This family is obtained by considering the restriction of the functional I to a family of subspaces of K increasing with c. They are defined by establishing the control values on the interval $[c, T]$ to be

equal to the feedback control (2.11) which is taken as reference one for system (2.2). Hence a special role is played by the solutions of (2.2) corresponding to the feedback control

$$u = -D^\sharp C\eta, \tag{2.11}$$

which are the solutions of (2.9) corresponding to the null control; they can be described through the solution $\Phi(t,c)$ of the matrix equation

$$\dot\Phi(t) = \bar A(t)\Phi(t)\ ,\ \Phi(c) = Id.$$

Since the control values are fixed on the interval $[c,T]$, then the problems of the family can be considered as problems on $[0,c]$ with new boundary conditions and end-point cost at time c given by

$$N_c = N\begin{pmatrix} Id & 0 \\ 0 & \Phi(T,c)\end{pmatrix},$$

$$\Gamma_c = \begin{pmatrix} Id & 0 \\ 0 & \Phi^\top(T,c)\end{pmatrix}\Gamma\begin{pmatrix} Id & 0 \\ 0 & \Phi(T,c)\end{pmatrix} + \begin{pmatrix} 0 & 0 \\ 0 & \displaystyle\int_c^T \Phi^\top(s,c)\bar P(s)\Phi(s,c)ds\end{pmatrix}, \tag{2.12}$$

where

$$\bar P \equiv P - C^\top(D^\sharp)^\top Q - Q^\top D^\sharp C + C^\top(D^\sharp)^\top R D^\sharp C. \tag{2.13}$$

We consider also the case $c=0$ that is the restriction of the quadratic functional I to the subspace of K corresponding to the control which is feedback on the whole $[0,T]$. This subspace can be identified with the subspace of $\mathbf{R}^n$ given by

$$K_0 \equiv \left\{ x\in\mathbf{R}^n : N_0\begin{pmatrix} x \\ x\end{pmatrix} = 0\right\}.$$

Notice that K_0 may be nontrivial if no end-point is fixed. The restriction of I to K_0 can be written as a finite-dimensional quadratic form on K_0:

$$I_0 : x \mapsto \frac{1}{2}(x^\top, x^\top)\Gamma_0\begin{pmatrix} x \\ x\end{pmatrix}.$$

We are going to define an Hamiltonian which differs from the usual one because we have to take into account the functional constraint (2.4). Let us emphasize that the Hamiltonian is the same for all the problems in the family because the new boundary conditions, N_c, and the new cost, Γ_c, on the state end-points do not play any role in its definition. Thanks to Assumption 2.3, define

$$S = \Pi_1(\Pi_1 R\Pi_1 + \Pi_2)^{-1}\Pi_1, \tag{2.14}$$

which plays the role usually played by R^{-1}, and set

$$\bar Q \equiv Q - RD^\sharp C, \tag{2.15}$$

$$E = \bar Q^\top S\bar Q - \bar P\ ,\ F = \bar A - BS\bar Q\ ,\ G = -BSB^\top.$$

The *true* (minimized) Hamiltonian $\mathcal{H} : \mathbf{R}\times(\mathbf{R}^n)^*\times\mathbf{R}^n \to \mathbf{R}$ is given by

$$\mathcal{H}(t,\gamma,x) \equiv -\frac{1}{2}x^\top E(t)x + \gamma F(t)x + \frac{1}{2}\gamma G(t)\gamma^\top. \tag{2.16}$$

Notice that the presence of extra terms in the Hamiltonian depends on the pointwise constraint on the state; in fact, they disappear if $C = 0$. Two systems are associated with the Hamiltonian $\mathcal{H}$: the Jacobi system on $\mathbf{R}^n \times (\mathbf{R}^n)^*$

$$\begin{aligned} \dot{\zeta}(t) &= F(t)\zeta(t) + G(t)\lambda^\top(t), \\ \dot{\lambda}(t) &= \zeta^\top(t)E(t) - \lambda(t)F(t), \end{aligned} \tag{2.17}$$

and the corresponding matrix system

$$\begin{aligned} \dot{Z}(t) &= F(t)Z(t) + G(t)\Lambda^\top(t), \\ \dot{\Lambda}(t) &= Z^\top(t)E(t) - \Lambda(t)F(t). \end{aligned} \tag{2.18}$$

The solutions of the Jacobi system which satisfy appropriate transversality conditions describe the extremals, and they correspond to the critical points of the problems in the family; see Lemma 3.8. Let us introduce the following.

DEFINITION 2.2. *For $c \in [0, T]$, an arc ζ is called a c-transversal extremal if there exists an arc λ such that*

$$(\zeta, \lambda) : [0, T] \to \mathbf{R}^n \times (\mathbf{R}^n)^*$$

is a solution of the Jacobi system (2.17) *satisfying the following transversality conditions:*

$$N_c \begin{pmatrix} \zeta(0) \\ \zeta(c) \end{pmatrix} = 0,$$

$$(-\lambda(0), \lambda(c)) = (\zeta^\top(0), \zeta^\top(c))\Gamma_c + \sigma N_c \text{ for some } \sigma \in (\mathbf{R}^p)^*.$$

Notice that this definition is given also for $c = 0$, and it corresponds to the existence of a critical point $x = \zeta(0)$ of I_0 restricted to K_0. $\lambda(0)$ is the associated Lagrange multiplier; see Lemma 3.8.

Remark 2.2. Differently from other authors in the literature, we do not define the extremals as the solutions of the Jacobi system because there can exist a nontrivial solution of the Jacobi system corresponding to the zero extremal without any controllability assumption, as is shown by Example 4.1. There is a one-to-one correspondence between c-transversal extremals and c-transversal solutions of the Jacobi system if and only if the system (2.9) is N_c-controllable at time c; see Lemmas 3.8 and 3.9.

The existence of nontrivial c-transversal extremals which are also solutions of (2.2) corresponding to the feedback control is typical of control problems with nonzero boundary conditions; hence, we need the following definition which characterizes them (see also Remark 2.3).

DEFINITION 2.3. *A c-transversal extremal ζ is said to be degenerate on the interval $[\alpha, \beta]$ containing c if*

$$\Pi_1(t)(\bar{Q}(t)\zeta(t) + B^\top(t)\lambda^\top(t)) = 0, \quad t \in [\alpha, \beta],$$

or, equivalently, for $t \in [\alpha, \beta]$

$$\begin{aligned} \dot{\zeta}(t) &= \bar{A}(t)\zeta(t), \\ \dot{\lambda}(t) &= -\lambda(t)\bar{A}(t) - \zeta^\top(t)\bar{P}(t). \end{aligned}$$

2.1. Conjugate points. This subsection contains our main results concerning the extension of the Jacobi theory of conjugate points to our problem. Let us first introduce the definitions of conjugate and semiconjugate points. This distinction is needed because of the existence of c-transversal degenerate extremals.

DEFINITION 2.4. *A point $c \in [0,T]$ is called semiconjugate with zero if there exists a nontrivial c-transversal extremal. A point $c \in [0,T)$ will be called conjugate with zero if there exists a nontrivial c-transversal extremal which is not degenerate on $[c,\beta]$, $c < \beta \leq T$.*

Remark 2.3. When $I_0 \geq 0$ on K_0, the interval between the first semiconjugate point and the first conjugate point has been called a focal interval in [15] or a table in [5]. Degenerate extremals do not exist when the right end-point is fixed and the linear system (2.9) is controllable from $\{0\}$ on any subinterval (see Lemma 3.10). In this last case the semiconjugate points are also conjugate and the focal interval (if it exists) reduces to a point. Recall that the controllability assumption is always satisfied in the calculus of variations problems.

We are now able to state our main results concerning conjugate points which correspond to the classical Jacobi necessary and sufficient conditions. Under Assumptions 2.2 and 2.3 the following theorems hold.

THEOREM 2.5. *The quadratic form I is nonnegative if and only if I_0 is nonnegative on K_0 and there is no point $c \in [0,T)$ conjugate with zero.*

THEOREM 2.6. *The quadratic form I is coercive if and only if I_0 is positive on K_0 and there is no point $c \in (0,T]$ semiconjugate with zero.*

These definitions naturally generalize the previous ones of conjugate and focal point, while the definition of a point coupled with zero, which has been introduced in [26] for the calculus of variations and in [27] for an optimal control problem under controllability assumptions, in our setting can be generalized by the following.

DEFINITION 2.7. *A point $c \in (0,T]$ is coupled with zero if there exists a nontrivial c-transversal extremal which is not degenerate on $[c,T]$. This last condition is empty when $c = T$.*

Because we want to emphasize the point of view of index theory, we prefer Definition 2.4, since a conjugate point corresponds to the point where the index increases while a semiconjugate point corresponds to an arc in the nullity.

Lemma 3.13 shows that the existence of a coupled point on the half-closed interval $(0,T]$ is equivalent to the existence of a semiconjugate point on the same interval. This is analogous for the existence of coupled points and conjugate points on the open interval $(0,T)$. Thus Theorems 2.5 and 2.6 could have been stated by means of coupled points.

COROLLARY 2.8. *The quadratic form I is nonnegative (coercive) if and only if I_0 is nonnegative (positive) on K_0 and there is no point $c \in [0,T)$ ($c \in (0,T]$) coupled with zero.*

This result includes the previous results stated by means of coupled points, in particular those in [25] where necessary conditions for the nonnegativity of the quadratic form I are given under controllability assumptions and with fixed initial point.

2.2. The Riccati equation. This section contains our main results concerning Riccati-type results for our problem. The following classical Riccati differential equation is associated with the Hamiltonian $\mathcal{H}$:

$$\dot{W}(t) = -W(t)F(t) - F^{\top}(t)W(t) - W(t)G(t)W(t) + E(t). \tag{2.19}$$

Notice that when the number of active constraints is equal to the number of controls then $G(t) \equiv 0$ and the Riccati equation becomes linear.

To obtain Riccati-type results for our problem, we consider first the problem with separate costs and end-points constraints; i.e., we set

$$N = \begin{pmatrix} N_0 & 0 \\ 0 & N_T \end{pmatrix}, \qquad \Gamma = \begin{pmatrix} \Gamma_0 & 0 \\ 0 & \Gamma_T \end{pmatrix}.$$

Later, we will state some results for the general case. Let us assume, without loss of generality, that N_0, N_T are orthogonal projections in $\mathbf{R^n}$.

Let (Z_0, Λ_0) be the solution of (2.18) defined by the initial conditions

$$Z_0(0) = Id - N_0, \qquad \Lambda_0(0) = -(Id - N_0)\Gamma_0(Id - N_0) + N_0. \tag{2.20}$$

Our main results relate the existence of a solution of the Riccati equation to the coercivity of I, and they correspond to Theorem 2.6, while there is no analogue to Theorem 2.5 in the general case. Under Assumptions 2.2 and 2.3 the following hold.

THEOREM 2.9. *The following statements are equivalent:*

(1) *I is coercive on K and the system* (2.9) *is controllable for every $c > 0$ from* Ker N_0.

(2) *$I_0 > 0$ on K_0 and there is a symmetric solution W of the Riccati equation* (2.19) *on $(0, T]$ such that*

$$\lim_{t \to 0^+} W(t)Z_0(t) = -(Id - N_0)\Gamma_0(Id - N_0) + N_0 \tag{2.21}$$

and

$$\Gamma_T - W(T) > 0 \text{ on } \operatorname{Ker} N_T. \tag{2.22}$$

In the case of unconstrained initial point (i.e., $N_0 = 0$), the controllability assumption is always satisfied and W is defined on $[0, T]$, so that the limit condition (2.21) *becomes*

$$W(0) = -\Gamma_0.$$

It is also possible to state necessary and sufficient conditions without any controllability hypothesis.

THEOREM 2.10. *The following statements are equivalent:*

(1) *The quadratic form I is coercive on K.*

(2) *There is a symmetric solution W of* (2.19) *on $[0, T]$ and $\beta \in \mathbf{R}$ such that*

$$W(0) = -\Gamma_0 - \beta N_0 \quad \text{and} \quad \Gamma_T - W(T) > 0 \text{ on } \operatorname{Ker} N_T.$$

(3) *There is a symmetric solution W of* (2.19) *on $[0, T]$ such that*

$$W(0) + \Gamma_0 > 0 \quad \text{on} \quad \operatorname{Ker} N_0 \quad \text{and} \quad \Gamma_T - W(T) > 0 \text{ on } \operatorname{Ker} N_T.$$

Let us remark that when an end-point is fixed, the corresponding inequality condition for the solution of the Riccati equation is empty.

The existence of a solution of the Riccati equation (or Riccati inequality) has been used to prove sufficient conditions for the nonlinear problem [25, 14].

Although general results for the nonnegativity of I on K cannot be stated, we can prove the following theorem in the case when the are no degenerate c-transversal extremals.

THEOREM 2.11. *Assume that the system* (2.9) *is controllable from* $\{0\}$ *on each subinterval and that the right end-point is fixed; then the following statements are equivalent:*

(1) *I is nonnegative on K.*

(2) $I_0 > 0$ *on* K_0, *and there is a symmetric solution* W *of the Riccati equation* (2.19) *on* $(0,T)$ *such that*

$$\lim_{t\to 0^+} W(t)Z_0(t) = -(Id - N_0)\Gamma_0(Id - N_0) + N_0.$$

(3) *There is a symmetric solution* W *of* (2.19) *on* $[0,T)$ *and* $\beta \in \mathbf{R}$ *such that*

$$W(0) = -\Gamma_0 - \beta N_0.$$

(4) *There is a symmetric solution* W *of* (2.19) *on* $[0,T)$ *such that*

$$W(0) + \Gamma_0 \geq 0 \text{ on } \operatorname{Ker} N_0.$$

Theorem 2.11 is of some interest because it includes the well-studied case of problems in the calculus of variations with one fixed end-point. The analogue of statement (3) in Theorem 2.10 does not hold with strict inequality, as is shown by Example 4.3.

The previous theorems are stated for the case when the boundary conditions and the costs are given separately on the initial and final points. They can be applied also to problems with boundary conditions given jointly on both end-points, periodic, for example, by transforming the problem into another one in double dimension. This can be done by considering an extended system (ξ, ν), where ν satisfies $\dot\nu(t) = 0$ subject to the constraint $\xi(0) = \nu(0)$. Then any condition involving $\xi(0)$ may be imposed on $\nu(T)$, so that the original cost and constraint are imposed on the final values of the extended system, while the new boundary conditions involve only the initial point. These necessary and sufficient conditions are stated by means of a $2n \times 2n$ solution W of a Riccati equation in double dimension,

$$W = \begin{pmatrix} W_{11} & W_{12} \\ W_{12}^\top & W_{22} \end{pmatrix}.$$

The three matrices are a solution of a cascade of differential equations given by a Riccati differential equation, a linear equation, and an integrator. The properties of this system will be described in a forthcoming paper. Here we give a sufficient condition which has a simple formulation and proof. It is not a necessary condition, as is shown by Example 4.2 in the appendix.

THEOREM 2.12. *If there is a symmetric solution* W *of* (2.19) *on* $[0,T]$ *such that*

$$\Gamma + \begin{pmatrix} W(0) & 0 \\ 0 & -W(T) \end{pmatrix} > 0 \text{ on } \operatorname{Ker} N, \tag{2.23}$$

then the quadratic form I *is coercive on the subspace* K.

Of some interest is the case of periodic boundary conditions, i.e., $N = (Id, -Id)$, when the end-point cost can be imposed on one of the end-points, i.e., $\Gamma = \begin{pmatrix} \Gamma_0 & 0 \\ 0 & 0 \end{pmatrix}$. From Theorem 2.12 we obtain the following.

COROLLARY 2.13. *If there is a symmetric solution* W *of* (2.19) *on* $[0,T]$ *such that*

$$\Gamma_0 + W(0) - W(T) > 0,$$

then the quadratic form I *is coercive on the subspace* K *defined by* $\eta(0) = \eta(T)$ *and by* (2.4).

3. Proofs of the results. As we said in section 2, we consider the quadratic form I acting on the Hilbert space U of the couples $(x,u) = (\textit{initial state}, \textit{control})$.

Thanks to Assumptions 2.1 and 2.2, our problem can be equivalently written in a simpler form. Let us consider a couple (x,u) such that (η,u) satisfies (2.2) and (2.4). From Assumption 2.2 and by (2.5) we deduce that

$$u = \Pi_1 u + \Pi_2 u = \Pi_1 u - D^\sharp C\eta.$$

Thus η solves also (2.9) with $\xi(0) = \eta(0)$, as it can be checked by direct substitution, and we can write $u = \Pi_1 u - D^\sharp C\xi$. For the quadratic form J defined by

$$J(x,u) = \frac{1}{2}(x^\top, \xi^\top(T))\Gamma\begin{pmatrix} x \\ \xi(T)\end{pmatrix}$$
$$+\frac{1}{2}\int_0^T \left\{\xi^\top(s)\bar{P}(s)\xi(s) + 2u^\top(s)\Pi_1(s)\bar{Q}(s)\xi(s) + u^\top(s)\Pi_1(s)R(s)\Pi_1(s)u(s)\right\} ds, \tag{3.1}$$

where $\bar{A}$, $\bar{P}$, $\bar{Q}$ are defined, respectively, in (2.10), (2.13), (2.15), we obtain that

$$I(x,u) = J(x,u) \text{ on } K. \tag{3.2}$$

Instead of studying the quadratic form I on K we study the quadratic form J on the closed subspace H of U defined by

$$N\begin{pmatrix} x \\ \xi(T)\end{pmatrix} = 0, \tag{3.3}$$

$$Du \equiv 0. \tag{3.4}$$

The equivalence between these two problems is stated in the following.

LEMMA 3.1. *The quadratic form I is coercive (nonnegative) on K if and only if the quadratic form J is coercive (nonnegative) on H.*

Proof. Since $J(x,u) = J(x,\Pi_1 u)$, the statement on nonnegativity is straightforward from (3.2). To prove the statement relative to the coercivity, by (3.2) it is sufficient to prove that J coercive on H implies that there exists $k > 0$ such that

$$J(x,u) \geq k(||x||^2 + ||\Pi_1 u||^2 + ||D^\sharp C\xi||^2).$$

It is not difficult to see that the linear map $\Xi : U_1 \to L^2([0,T],\mathbf{R}^m)$ defined as $(x,\Pi_1 u) \mapsto D^\sharp C\xi$, where ξ is the solution of (2.9) such that $\xi(0) = x$, is continuous. The result follows from standard computations. □

In the following we use some general properties of quadratic forms on Hilbert spaces stated in [7]. We recall first the basic definition of the *Legendre* form, sometimes called *elliptic* [15].

DEFINITION 3.2. *The quadratic form J is said to be Legendre on H if it is weakly lower semicontinuous on H and if, for every sequence $\{h_n\}$ which converges weakly to $h \in H$ and such that $\{J(h_n)\}$ converges to $J(h)$, $\{h_n\}$ converges strongly to $h \in H$.*

In order to prove Theorems 2.5 and 2.6 we use the results in [29], which we briefly recall. The results concern a Legendre quadratic form J defined on an Hilbert space H and a family of closed subspaces with some continuity properties by which it is possible to define a Jacobi-type condition.

DEFINITION 3.3. *A one-parameter family $\{H_c\}$, $c \in [0,T]$, of closed subspaces of H is said to be continuous if it has the following properties:*

(1) *for $0 \leq c_1 \leq c_2 \leq T$, $\quad H_{c_1} \subset H_{c_2}$ and $H_T = H$.*

(2) *for* $c_0 \in (0,T]$, cl $(\bigcup_{0\le c<c_0} H_c) = H_{c_0}$.
(3) *for* $c_0 \in [0,T)$, $\quad \bigcap_{c_0<c\le T} H_c = H_{c_0}$.

For a given continuous family H_c and for a positive weakly continuous quadratic form $\mathcal{K}$ define

$$\Omega_c = \{e \in H_c : \mathcal{K}(e) = 1\}.$$

The study of a Legendre form J can be pursued through the following "value function" $V : [0,T] \to (-\infty, +\infty]$ defined as

$$V(c) = \min_{e\in\Omega_c} J(e),$$

as usually $V(c) = +\infty$ if $\Omega_c = \emptyset$. The use of the min operator instead of the infimum in the definition of the function V is possible because a Legendre form attains its minimum on Ω_c and hence we have

- J is coercive on $H_c \iff V(c) > 0$,
- J is nonnegative on $H_c \iff V(c) \ge 0$.

The main result which summarizes the properties of the function V is the following [29].

THEOREM 3.4. *Assume that J is Legendre on H. The function V is nonincreasing and continuous as an extended-real-valued function*

$$V : [0,T] \to \mathbf{R} \cup \{+\infty\}.$$

Remark 3.1. For an optimal control problem we can choose $\mathcal{K}(x,u) = \|x\|^2 + \|\omega\|_2^2$, where $w : t \mapsto \int_0^t u(s)ds$. This choice is the usual one in the calculus of variations and leads us to characterize $V(c)$ as the first eigenvalue of the associated Euler–Lagrange differential operator.

From the properties of the function V we can deduce that J is coercive on H if and only if J is positive on H_0 and there is no point $c \in (0,T]$ at which V becomes zero. Moreover, J is nonnegative on H if and only if J is nonnegative on H_0 and there is no point $c \in [0,T)$ at which V changes sign. In our case, using a suitable family of subspaces, we can link the existence of such points with semiconjugate and conjugate points.

To apply the above theory it is convenient to see the Hilbert space H as the Kernel of the linear operator $\Sigma : U_1 \to \mathbf{R}^p$, defined by

$$\Sigma(x,u) = N\begin{pmatrix} x \\ \xi(T) \end{pmatrix},$$

where ξ is the solution of equation (2.9) with initial condition $\xi(0) = x$.

As a first step we show that our quadratic form is Legendre.

THEOREM 3.5. *Under Assumption* 2.2*, the quadratic form J is Legendre on H if and only if Assumption* 2.3 *is satisfied.*

Proof. Let

$$E(x,u) = \|x\|^2 + \int_0^T u^\top(s)\Pi_1(s)R(s)\Pi_1(s)u(s)ds.$$

The quadratic form J can be written as the sum of a weakly continuous form $J - E$ plus the form E. $J - E$ is weakly continuous because of the compactness of the solution operator associated to a linear differential equation. By Theorem 11.5 in [7] the quadratic form J is Legendre on H if and only if the form E is Legendre on H. On

the other hand the orthogonal complement of $H = \operatorname{Ker}\Sigma$ in U_1 is finite dimensional as the range of Σ; therefore it follows from Theorem 11.4 in [7] that E is Legendre on H if and only if it is on U_1. It is clear that the strengthened Legendre–Clebsch condition implies that the form E is coercive and hence Legendre on U_1.

Let us prove the converse. If E is Legendre on U_1, by Theorem 5.2 in [7] its nullity and index are finite. Since it is nonnegative, the index must be zero. But, for this particular form, the nullity also has to be zero because if $E(x,u) = 0$, then for all $\phi \in C([0,T],\mathbf{R})$ also $E(\|\phi\|x, \phi u) = 0$; hence the form E is positive on U_1. By Theorem 11.1 in [7] this yields that it is coercive, so that (2.8) holds. □

In order to define the family of subspaces $\{H_c\}$ for our problem let us introduce the following notation. For $c \in [0,T]$ set

$$U_c = \{(x,u) \in U_1 : u(t) \equiv 0 \ \text{ for } \ t \in [c,T]\},$$

and consider the following one-parameter family of subspaces:

$$H_c = U_c \bigcap \operatorname{Ker}\Sigma.$$

A subspace H_c is isomorphic to a subspace of $\mathbf{R}^n \times L^2([0,c],\mathbf{R}^m)$; in particular, H_0 can be identified with the subspace K_0 of $\mathbf{R}^n$ defined in section 1 and, moreover, $H_T = H$.

Since we do not have any controllability assumption on the system (2.9),

$$\mathcal{R}_c \equiv \Sigma\,(U_c)$$

need not be the whole space $\mathbf{R}^p$. Notice that for $(x,u) \in U_c$ we have that

$$\Sigma(x,u) = N\begin{pmatrix} x \\ \xi(T) \end{pmatrix} = N_c\begin{pmatrix} x \\ \xi(c) \end{pmatrix},$$

and hence, with these notations, the system (2.9) is N_c-controllable at time c if and only if $\mathcal{R}_c = \operatorname{Range} N$.

The following lemma describes a property of $\mathcal{R}_c$ which is crucial to prove the continuity of the family $\{H_c\}$.

LEMMA 3.6. *Under Assumption* 2.2*, there is a finite partition* $c_0 = 0 < c_1 < \cdots < c_s = T$ *such that* $\mathcal{R}_c$ *is constant for all* $c \in (c_i, c_{i+1}]$.

Proof. It follows immediately from the definition that the spaces U_c's have the following properties:

$$U_c \subseteq U_d,\ 0 \le c \le d \le T,\ \operatorname{cl}\left(\bigcup_{0\le c<c_0} U_c\right) = U_{c_0}.$$

Therefore, by the continuity of Σ the corresponding sets $\mathcal{R}_c$'s have the same properties as the sets U_c's, and since they also are finite dimensional, the statement follows immediately. □

We can now prove that the family $\{H_c\}$ satisfies the properties required by Definition 3.3.

LEMMA 3.7. *Under Assumption* 2.2 *the above-defined family* $\{H_c\}$ *is continuous.*

Proof. Properties 1 and 3 are immediate.

Property 2. Since H_{c_0} is a closed subspace, it is clear that $\operatorname{cl}\left(\bigcup_{0\le c<c_0} H_c\right) \subseteq H_{c_0}$; we have to prove the converse inclusion. Let $c_0 \in (0,T]$; by Lemma 3.6, there is $c_1 < c_0$

such that $\mathcal{R}_c = \mathcal{R}_{c_0}$ for all $c \in [c_1, c_0]$. Let $\Psi : \mathcal{R}_{c_0} \to U_{c_1}$ be the right inverse of $\Sigma : U_{c_1} \to \mathcal{R}_{c_1} = \mathcal{R}_{c_0}$. Choose a fixed $(x, u) \in H_{c_0}$. For $c_1 < c < c_0$ consider the truncation of the control u given by

$$u_c(t) = u(t) \text{ for } t \in [0, c] \text{ and } u_c(t) = 0 \text{ for } t \in [c, T].$$

(x, u_c) need not belong to H_c, but we can define

$$(x_c, v_c) = \Psi\, \Sigma(0, u_c - u) \in U_{c_1}$$

so that $\Sigma(x_c, v_c) = \Sigma(0, u_c - u)$ and hence

$$(y_c, w_c) = (x - x_c, u_c - v_c)$$

belongs to H_c. It is easy to see that as $c \to c_0$ then $(x_c, v_c) \to 0$ in H_{c_1}. Let us compute

$$\begin{aligned}\|(y_c, w_c) - (x, u)\|^2 &= \|x_c\|^2 + \|u_c - v_c - u\|^2 \\ &= \|x_c\|^2 + \int_0^c \|v_c(s)\|^2 ds + \int_c^{c_0} \|u(s)\|^2 ds\end{aligned}$$

so that $(y_c, w_c) \to (x, u)$ as $c \to c_0$, and the statement is proved. □

Since we have shown that the abstract theory in [29] applies to our case, we have now to determine the relation between the zeros of V and the semiconjugate and conjugate points. Let us consider the expression of the restriction of J to U_c written as a quadratic form on $[0, c]$. It will be denoted by J_c. For $(x, u) \in U_c$ we have

$$\begin{aligned}J(x, u) = J_c(x, u) = \frac{1}{2}(x^\top, \xi^\top(c))\Gamma_c \begin{pmatrix} x \\ \xi(c) \end{pmatrix} \\ + \frac{1}{2}\int_0^c \left\{\xi^\top(s)\bar{P}(s)\xi(s) + 2u^\top(s)\Pi_1(s)\bar{Q}(s)\xi(s) + u^\top(s)\Pi_1(s)R(s)\Pi_1(s)u(s)\right\} ds,\end{aligned}$$

where ξ is a solution of (2.9) and Γ_c is given by (2.12). Let us remark that $J_0 = I_0$.

LEMMA 3.8. *(x, u) is a critical point for J_c restricted to H_c if and only if there exists a c-transversal extremal ζ such that*

$$(x, u) = (\zeta(0), -S(\bar{Q}\zeta + B^\top\lambda^\top)) \quad \text{on} \quad [0, c]. \tag{3.5}$$

(ζ, λ) is uniquely determined if and only if the system (2.9) is N_c-controllable at time c. Moreover ζ is degenerate on $[\alpha, \beta]$ if and only if the corresponding control is identically zero on $[\alpha, \beta]$.

Proof. (x, u) is critical for J_c restricted to H_c if and only if $DJ_c(x, u)(y, v) = 0$ for all $(y, v) \in U_c$ such that $\Sigma(y, v) = 0$. For $c = 0$ the above extremality condition can be easily characterized by the usual Lagrange multiplier rule for a finite-dimensional problem. That is, there exists a multiplier $\sigma \in (\mathbf{R}^p)^*$ such that

$$(x^\top, x^\top)\Gamma_0 \begin{pmatrix} Id \\ Id \end{pmatrix} + \sigma N_0 \begin{pmatrix} Id \\ Id \end{pmatrix} = 0,$$

which can be equivalently written as

$$(x^\top, x^\top)\Gamma_0 + \sigma N_0 = (-y, y)$$

for some $y \in (\mathbf{R}^n)^*$, which is equivalent to the existence of a zero-transversal extremal with $\lambda(0) = y$.

Let us now examine the case $c > 0$. We can consider $\mathcal{R}_c$ as the range of Σ so that the constraint can be thought as regular and the above extremality condition is equivalent to the existence of a unique multiplier $\sigma \in (\mathcal{R}_c)^*$ which can be extended to a multiplier $\sigma \in (\mathbf{R}^p)^*$ such that $DJ_c(x,u) + \sigma\Sigma = 0$ on U_c. Such an extension is unique up to a component in the orthogonal of Range N if and only if $\mathcal{R}_c = \text{Range } N$. We can use an equivalent characterization of the critical points by means of the Hamiltonian $H : \mathbf{R} \times (\mathbf{R}^n)^* \times \mathbf{R}^n \times \mathbf{R}^m \to \mathbf{R}$:

$$\begin{aligned} H(t,\gamma,x,w) &\equiv \gamma(\bar{A}(t)x + B(t)\Pi_1(t)w) \\ &+ \frac{1}{2}(x^\top \bar{P}(t)x + 2w^\top \Pi_1(t)\bar{Q}(t)x + w^\top \Pi_1(t)R(t)\Pi_1(t)w); \end{aligned} \tag{3.6}$$

see, e.g., Lemma 5.5 in [23]. The couple (x,u) is a critical point of J_c restricted to H_c if and only if there exists a multiplier $\sigma \in (\mathbf{R}^p)^*$ and a solution of the adjoint equation

$$-\dot{\lambda}(t) = \lambda(t)\bar{A}(t) + \xi^\top(t)\bar{P}(t) + u^\top(t)\Pi_1(t)\bar{Q}(t)$$

satisfying the corresponding transversality conditions

$$(-\lambda(0), \lambda(c)) = (\xi^\top(0), \xi^\top(c))\Gamma_c + \sigma N_c$$

such that

$$\begin{aligned} &D_4 H(t, \lambda(t), \xi(t), \Pi_1(t)u(t)) \\ &= (u^\top(t)\Pi_1(t)R(t) + \lambda(t)B(t) + \xi^\top(t)\bar{Q}^\top(t))\Pi_1(t) \equiv 0. \end{aligned}$$

Since the operator S defined in (2.14) is such that

$$S_{|\text{Range}\,\Pi_1} = (\Pi_1 R \Pi_{1|\text{Range}\,\Pi_1})^{-1} \quad \text{and} \quad S\,\Pi_1 = \Pi_1 S = S,$$

we obtain

$$u = \Pi_1 u = -S(\bar{Q}\xi + B^\top \lambda^\top).$$

Substituting the above expression for u in the Hamiltonian H and in the equations defining ξ and λ, we obtain a solution of the Jacobi system satisfying the transversality conditions and hence a c-transversal extremal. Notice that two multipliers σ_1, σ_2 such that $\sigma_1 - \sigma_2 \in (\text{Range}\, N)^\perp$ identify the same λ, and hence if the system (2.9) is N_c-controllable at time c, then the extremal is uniquely determined.

Let us now prove that ζ is degenerate on $[\alpha,\beta]$ if and only if the corresponding control is identically zero on $[\alpha,\beta]$. Since S is invertible on Range Π_1, u is zero on $[\alpha,\beta]$ if and only if $\Pi_1(\bar{Q}\zeta + B^\top \lambda^\top) = 0$. □

Remark 3.2. By the strengthened Legendre–Clebsch condition, the Hamiltonian $\mathcal{H}$ in (2.16) can be obtained by minimization from the Hamiltonian H in (3.6). Namely,

$$\hat{w}(\gamma,x)(t) = -S(t)(\bar{Q}(t)x + B^\top(t)\gamma^\top) \tag{3.7}$$

is the unique minimizer of $H(t,\gamma,x,\cdot)$ and

$$\mathcal{H}(t,\gamma,x) \equiv H(t,\gamma,x,\hat{w}(\gamma,x)(t)) = \min_w H(t,\gamma,x,w). \tag{3.8}$$

LEMMA 3.9. *A c-transversal extremal ζ is trivial if and only if the corresponding couple (x,u) is zero. Moreover any nonzero c-transversal couple (ζ,λ) corresponds to a nontrivial c-transversal extremal if the system* (2.9) *is N_c-controllable at time c.*

Proof. If $(x,u)=(0,0)$, then ζ is identically zero. Conversely, let $\zeta=0$; then $\dot{\zeta}(t)=-B(t)S(t)B^\top(t)\lambda^\top(t)\equiv 0$ yields

$$\lambda(t)B(t)S(t)B^\top(t)\lambda^\top(t)\equiv 0.$$

From the properties of S we obtain that $\Pi_1 B^\top\lambda^\top\equiv 0$ and hence $SB^\top\lambda^\top=u=0$. By Lemma 3.8, the c-transversal extremal corresponding to $(0,0)$ is uniquely determined if and only if Σ is onto Range N; hence the only trivial c-transversal extremal is the one that is zero, if and only if Σ is onto Range N. ∎

LEMMA 3.10. *Assume that the right end-point is fixed, that is,* $N=\begin{pmatrix}\Gamma_0 & 0\\ 0 & Id\end{pmatrix}$, *and that the system* (2.9) *is controllable from* $\{0\}$ *on any subinterval. Then there are no degenerate c-transversal extremals.*

Proof. Assume that ζ is a c-transversal degenerate extremal. Since the right end-point is fixed, the transversality condition implies that $\zeta(c)=0$. From the characterization of degenerate extremal it follows that there exists an interval $[\alpha,\beta]$ containing c such that $\zeta(t)\equiv 0$ on $[\alpha,\beta]$ with a corresponding λ satisfying

$$\dot{\lambda}(t)=-\lambda(t)\bar{A},\qquad \Pi_1 B^\top\lambda^\top=0$$

on the same interval. By the controllability assumption λ has to be zero on that interval and hence the couple (ζ,λ) is trivial. ∎

LEMMA 3.11. *If* $V(c)=0$, *then there is a nontrivial c-transversal extremal.*

Proof. If $V(c)=0$, then 0 is the global minimum for J_c restricted to H_c and the minimum is attained at a nonzero (x,u). Lemmas 3.8 and 3.9 imply that there exists a nontrivial c-transversal extremal. ∎

LEMMA 3.12. *A* c_0*-transversal extremal* ζ *which is degenerate on* $[\alpha,\beta]$ *is a c-transversal extremal for every* $c\in[\alpha,\beta]$.

Proof. The proof can be carried out by direct computation, taking into account that for $c\in[\alpha,\beta]$ we have

$$\zeta(c)=\Phi(c,c_0)\zeta(c_0),$$
$$\lambda(c)=\lambda(c_0)\Phi(c_0,c)-\zeta^\top(c)\int_{c_0}^{c}\Phi^\top(s,c)\bar{P}(s)\Phi(s,c)\,ds.$$ ∎

We are now able to prove the main results.

Proof of Theorem 2.5. Assume that J_0 is nonnegative on H_0 (i.e., $V(0)\geq 0$) and that there is no point $c_0\in(0,T)$ conjugate with zero. We show the result by proving that there is no point $c_1\in[0,T]$ such that $V(c_1)<0$. Assume by contradiction that there is a point $c_1\in[0,T]$ such that $V(c_1)<0$; we can find a point c_0 at which V changes its sign. Since $V(c_0)=0$, then by Lemma 3.11 there exists a nontrivial c_0-transversal extremal. Since $V(c)<0$ for $c>c_0$, the index of the quadratic form J_c is positive. By Lemma 16.3 in [7] there exists a nontrivial critical point $(x,u)\in H_{c_0}$ which is not a critical point for $c>c_0$. By Lemma 3.12 the corresponding c_0-transversal extremal cannot be degenerate on $[c_0,\beta]$, with $\beta>c_0$; that is, c_0 is conjugate with zero yielding a contradiction.

Conversely, assume that the form J is nonnegative on H; then the function V is nonnegative on $[0,T]$. It is clear that J_0 is nonnegative on H_0. Let us show that there is no point $c_0\in[0,T)$ conjugate with zero. Assume by contradiction that $c_0\in[0,T)$ is conjugate with zero. By Lemma 3.8 it follows that J_{c_0} has a critical point in H_{c_0}, so that $V(c_0)\leq 0$. If $V(c_0)<0$ we have a contradiction. Otherwise $V(c_0)=0$ and hence any nontrivial c_0-transversal extremal ζ corresponds to a nonzero $(\bar{x},\bar{u})\in H_{c_0}$ which is a minimizer for J_{c_0}. It is clear that $J_T(\bar{x},\bar{u})=J_{c_0}(\bar{x},\bar{u})=0$, so that, since V

is identically zero on the interval $[c_0, T]$, $(\bar{x}, \bar{u})$ is a minimizer also for J_T. Therefore by Lemmas 3.8 and 3.9 ζ is degenerate on $[c_0, T]$, a contradiction. ∎

Proof of Theorem 2.6. Assume that J_0 is positive on H_0 (i.e., $V(0) > 0$) and that there is no point $c_0 \in (0, T]$ semiconjugate with zero. Let us show that there is no point $c \in (0, T]$ such that $V(c) = 0$. Assume by contradiction that $V(c) = 0$; then by Lemma 3.11, there exists a nontrivial c-transversal extremal, a contradiction.

Conversely, assume that the quadratic form J is coercive on H. Then the function V is strictly positive on $[0, T]$. This ensures that quadratic form J_c restricted to H_c does not have any nonzero critical point; by Lemmas 3.8 and 3.9 the statement follows. ∎

LEMMA 3.13. *There exists a point $c \in (0, T]$ coupled with zero if and only if there exists a point semiconjugate with zero on the same interval. There exists a point $c \in (0, T)$ coupled with zero if and only if there is a point conjugate with zero on the same interval.*

Proof. Assume that there exists a point $c \in (0, T]$ coupled with zero; then by definition there exists a point semiconjugate with zero. Conversely, if there exists a point $c \in (0, T]$ semiconjugate with zero, then there are two possibilities: either the extremal is not degenerate on $[c, T]$ and hence the point c is coupled with zero or it is degenerate until T and in this case T is coupled with zero.

Let us now prove the second statement. Assume that there exists a point $c \in (0, T)$ coupled with zero. Since the corresponding extremal is not degenerate until T, there exists a maximal interval $[c, c_1]$ with $c \le c_1 < T$ on which it is degenerate. By Lemma 3.12 it is a c_1-transversal extremal corresponding to a conjugate point. The converse follows immediately from the definitions. ∎

The proof of Corollary 2.8 is now straightforward.

Let us now prove the statements concerning the Riccati theory. For sake of completeness we prove some standard properties of the solutions of Jacobi system and Riccati equation. In the following we use the fact that if (ζ, λ) is a solution of (2.17), then ζ is a solution of (2.9) corresponding to the control $\hat{w}(\zeta, \lambda)$, where $\hat{w}$ is defined in (3.7). Set

$$g(t, x, w) = x^\top \bar{P}(t)x + 2w^\top \Pi_1(t)\bar{Q}(t)x + w^\top \Pi_1(t)R(t)\Pi_1(t)w.$$

LEMMA 3.14. *If (ζ, λ) is a solution of* (2.17)*, then the following equality holds:*

$$g(t, \zeta(t), \hat{w}(\zeta, \lambda)(t)) = -\frac{d(\lambda\zeta)}{dt}(t).$$

Proof. From the definition of the Hamiltonian H and its minimum property (3.8) it follows that

$$\begin{aligned} g(t, \zeta(t), \hat{w}(\zeta, \lambda)(t)) &= 2H(t, \lambda(t), \zeta(t), \hat{w}(\zeta, \lambda)(t)) - 2\lambda(t)\dot{\zeta}(t) \\ &= 2\mathcal{H}(t, \lambda(t), \zeta(t)) - 2\lambda(t)\dot{\zeta}(t). \end{aligned}$$

Moreover, since $\mathcal{H}$ is quadratic,

$$\begin{aligned} &g(t, \zeta(t), \hat{w}(\zeta, \lambda)(t)) \\ &= \lambda(t)\frac{\partial \mathcal{H}}{\partial \gamma}(t, \lambda(t), \zeta(t)) + \frac{\partial \mathcal{H}}{\partial x}(t, \lambda(t), \zeta(t))\,\zeta(t) - 2\lambda(t)\dot{\zeta}(t) \\ &= -\dot{\lambda}(t)\zeta(t) - \lambda(t)\dot{\zeta}(t). \end{aligned}$$

The statement is proved. ∎

LEMMA 3.15. *Let ξ be a solution of* (2.9) *corresponding to the control u and let W be a solution of the Riccati equation* (2.19) *on an interval $I \subseteq [0,T]$. Then*

$$g(t,\xi(t),u(t)) \geq -\frac{d\,(\xi^\top W\xi)}{dt}(t),\ t \in I.$$

Moreover, the equality holds if and only if

$$u = -S(\bar{Q} + B^\top W)\xi, \tag{3.9}$$

and this happens if and only if the couple $(\xi, \xi^\top W)$ is a solution of the Jacobi system (2.17).

Proof. Set $\lambda(t) = \xi^\top(t)W(t)$. Easy computations show that

$$-\frac{1}{2}\xi^\top(t)\dot{W}(t)\xi(t) = \mathcal{H}(t,\lambda(t),\xi(t)).$$

By the minimum properties of the Hamiltonian $\mathcal{H}$, (3.8), we have that

$$\frac{1}{2}g(t,\xi(t),u(t)) \geq \mathcal{H}(t,\lambda(t),\xi(t)) - \lambda(t)\dot{\xi}(t) = -\frac{1}{2}\frac{d\,(\xi^\top W\xi)}{dt}(t).$$

Since the Hamiltonian H attains its minimum in a unique point, the equality holds if and only if

$$u = -S(\bar{Q}\zeta + B^\top\lambda^\top) = -S(\bar{Q} + B^\top W)\xi.$$

A direct computation proves the last statement. □

LEMMA 3.16. *Let (Z_0,Λ_0) be the solution of* (2.18) *defined by the initial conditions* (2.20). *If $Z_0(t)$ is invertible for $t \in [a,b]$, then $W = \Lambda_0{}^\top Z_0^{-1}$ is a symmetric solution of the Riccati equation* (2.19) *on $[a,b]$.*

Proof. A direct computation proves that W is a solution of (2.19). Since the derivative of $\Lambda_0 Z_0 - Z_0{}^\top\Lambda_0{}^\top$ is zero, then from the symmetry of $\Lambda_0(0)Z_0(0)$, it follows that $\Lambda_0 Z_0 = Z_0{}^\top\Lambda_0{}^\top$. Multiplying the equality by $(Z_0{}^\top)^{-1}$ from the left and by Z_0^{-1} from the right, we obtain the statement. □

Proof of Theorem 2.9. Let (Z_0,Λ_0) be the solution of (2.18) defined by the initial conditions (2.20). It is easy to show that each solution (ζ,λ) of the Jacobi system (2.17) which satisfies the transversality conditions at $t=0$ can be uniquely expressed as $(\zeta,\lambda) = (Z_0 y, y^\top\Lambda_0),\ y \in \mathbf{R}^n$.

Let us now prove that $1 \Rightarrow 2$. If J is coercive, then clearly J_0 is positive on H_0. Let us prove that $Z_0(t)$ is invertible for $t \neq 0$. Let $(\zeta,\lambda) = (Z_0 y, y^\top\Lambda_0)\ ,\ y \in \mathbf{R}^n$. By Lemma 3.14 we obtain

$$J_c(\zeta(0),\hat{w}(\zeta,\lambda)) = y^\top(Z_0{}^\top(c)\Gamma_T^c - \Lambda_0(c))Z_0(c)y, \tag{3.10}$$

where

$$\Gamma_T^c = \Phi^\top(T,c)\Gamma_T\Phi(T,c) + \int_c^T \Phi^\top(s,c)\bar{P}(s)\Phi(s,c)ds.$$

Assume that $Z_0(c)y = 0$; then (3.10) yields $J_c(\zeta(0),\hat{u}(\zeta,\lambda)) = 0$ and hence from the coercivity assumption it follows that $(\zeta(0),\hat{u}(\zeta,\lambda)) = 0$ on $[0,c]$, and it corresponds to a global minimum for J_c. ζ is a c-transversal extremal for the problem with fixed right end-point, and by applying Lemma 3.9, whose controllability assumptions are

satisfied, we deduce that (ζ, λ) is identically zero on $[0, c]$. If we evaluate the initial conditions we obtain $\zeta(0) = (Id - N_0)y = 0$ and $0 = \lambda(0) = -y^\top(Id - N_0)\Gamma_0(Id - N_0) + y^\top N_0 = y^\top N_0$ and hence $y = 0$, so that $Z_0(c)$ is invertible. Let us define $W = \Lambda_0{}^\top Z_0^{-1}$, which is a symmetric solution of (2.19) by Lemma 3.16. The limit condition (2.21) is straightforward. Since $Z_0(T)$ is invertible, any $w \in \operatorname{Ker} N_T$ can be written as $w = Z_0(T)y$. From (3.10) evaluated at $c = T$, by the coercivity assumption on J, one obtains immediately (2.22).

Let us now prove that $2 \Rightarrow 1$. By contradiction assume that J is not coercive. Then from Theorem 2.6 there is a semiconjugate point $c \in (0, T]$ and by Lemma 3.8 there is a nonzero $(x, u) \in H_c$ such that

$$0 = J_c(x, u) = J(x, u) = \frac{1}{2}x^\top \Gamma_0 x + \frac{1}{2}\xi^\top(T)\Gamma_T \xi(T) + \frac{1}{2}\int_0^T g(s, \xi(s), u(s))\, ds.$$

ξ coincides with a c-transversal extremal on $[0, c]$ so that there exists $y \in \mathbf{R}^n$ such that $\xi = Z_0 y$ on $[0, c]$. From Lemma 3.15 it follows that

$$\begin{aligned} 0 &= J_c(x, u) \\ &\geq \frac{1}{2}x^\top \Gamma_0 x + \frac{1}{2}\xi^\top(T)\Gamma_T \xi(T) + \frac{1}{2}\lim_{t \to 0^+} y^\top Z_0{}^\top(t) W(t) Z_0(t) y - \frac{1}{2}\xi^\top(T) W(T) \xi(T) \\ &= \frac{1}{2}\xi^\top(T)(\Gamma_T - W(T))\xi(T) \geq 0. \end{aligned}$$

The two opposite inequalities yield $\xi(T) = 0$. Moreover, the above inequality is indeed an equality and Lemma 3.15 implies that the couple $(\xi, \xi^\top W)$ is a solution of (2.17). Since it has zero final value, it is the zero solution, a contradiction.

Let us now prove that statement 2 implies that $Z_0(c)$ is invertible for $c \in (0, T]$ and hence the constrained control system is controllable at any time c from $\operatorname{Ker} N_0$. For $c \in (0, T]$ and $y \in \mathbf{R}^n$, assume that $Z_0(c)y = 0$; hence the trajectory

$$\xi(t) = \begin{cases} Z_0(t)y & \text{if } t \in [0, c], \\ 0 & \text{if } t \in [c, T], \end{cases} \qquad u(t) = \begin{cases} \hat{u}(Z_0 y, y^\top \Lambda_0) & \text{if } t \in [0, c], \\ 0 & \text{if } t \in [c, T] \end{cases}$$

is admissible. Lemma 3.14 yields $J(\xi(0), u) = 0$, and since we have shown that the quadratic form is coercive, $Z_0(0)y = (Id - N_0)y = 0$. On the other hand, by Lemma 3.15 one obtains

$$0 = J(\xi(0), u) \geq \frac{1}{2}y^\top(Id - N_0)\Gamma_0(Id - N_0)y + \frac{1}{2}\lim_{t \to 0^+} y^\top Z_0{}^\top(t) W(t) Z_0(t) y = 0.$$

Hence, again from Lemma 3.15, since the equality holds, $(\xi, \xi^\top W)$ is a solution of (2.17) which has value zero at $t = c$. Therefore it is identically zero, so that $\lambda(0) = N_0 y = 0$ also and hence $y = 0$, which proves that $Z_0(c)$ is invertible.

In the case of unconstrained initial point, i.e., $N_0 = 0$, $Z_0(0) = Id$ is invertible; therefore $W = \Lambda_0{}^\top Z_0^{-1}$ exists on the whole interval and $W(0) = -\Gamma_0$. □

In the next proofs we will reduce our problem to one without boundary constraints. This can be done by applying Theorems 13.2 and 13.3 of [7] to obtain the following.

LEMMA 3.17. *Let J be a Legendre form and Q be a nonnegative w-continuous form such that J is coercive (nonnegative) where Q is zero; then there exists a constant $\beta > 0$ such that $J + \frac{1}{2}\beta Q$ is coercive (nonnegative) on the whole space.*

Proof of Theorem 2.10. We can apply Lemma 3.17 with

$$Q(x,u) = x^\top N_0 x + \xi^\top(T) N_T \xi(T),$$

which is a w-continuous form since it involves only the state end-points. If we apply Theorem 2.9 to this special case, we obtain the first formulation of the statement taking into account that $\Gamma_T + \beta N_T - W(T)$ is positive on $\mathbf{R}^n$ if and only if $\Gamma_T - W(T)$ is positive on Ker N_T. To prove the other equivalent form of the statement let us first notice that elementary results in perturbation theory yield that, for $\epsilon > 0$ sufficiently small, there exists a solution of the Riccati equation such that

$$W(0) + \Gamma_0 + \beta N_0 = \epsilon Id \text{ and } \Gamma_T - W(T) > 0 \text{ on } \operatorname{Ker} N_T.$$

To prove the converse statement it is sufficient to use Lemma 3.15 in a way analogous to the proof of Theorem 2.9. □

Proof of Theorem 2.11. It is a consequence of Lemma 3.10 that every semiconjugate point is also a conjugate point. Hence the function V changes sign at the point where it becomes zero. The form J is then nonnegative on H if and only if it is coercive on H_c for $c \in [0,T)$. The prove of the equivalence among the first three statements can be carried out as in the proofs of Theorems 2.9 and 2.10. Since statement 4 is an obvious consequence of statement 3, we need only to prove that if statement 4 holds then the form J is nonnegative on H. This can be done by direct computation on J_c using Lemma 3.15. Let us underline that the perturbation argument in Theorem 2.10 does not hold. □

Let us now go back to the case of mixed boundary conditions and cost.

Proof of Theorem 2.12. Let (x,u) belong to H; then the couple $(x,\xi(T)) \in \operatorname{Ker} N$ and by Lemma 3.15 we obtain

$$\begin{aligned} 2J(x,u) &\geq (x^\top, \xi^\top(T))\Gamma \begin{pmatrix} x \\ \xi(T) \end{pmatrix} + x^\top W(0) x - \xi^\top(T) W(T) \xi(T) \\ &= (x^\top, \xi^\top(T)) \left(\Gamma + \begin{pmatrix} W(0) & 0 \\ 0 & -W(T) \end{pmatrix} \right) \begin{pmatrix} x \\ \xi(T) \end{pmatrix} \geq 0. \end{aligned}$$

Let us now show that J is not semidefinite but indeed coercive. If $J(x,u) = 0$, then $(x,\xi(T)) = 0$. Moreover, again by Lemma 3.15, the equality holds if and only if $(\xi, \xi^\top W)$ is a solution of the Jacobi equation. Since it is zero at the initial point, it is identically zero. By (3.9) the couple (x,u) is also trivial; thus J is coercive on H. □

4. Appendix. This section contains two technical results we have been using and the examples.

Lemma 4.1 states that Assumption 2.2 is equivalent to the regularity of the infinite-dimensional constraint (2.4) under Assumption 2.1.

LEMMA 4.1. *Assume that Assumption* 2.1 *holds, and let the map* $\Psi : U \to W = \operatorname{Range} \Delta$ *be given by* $(x,u) \mapsto C\eta + Du$, *where* η *is the solution of* (2.2). Ψ *is onto if and only if Assumption* 2.2 *holds.*

Proof. Assume that Assumption 2.2 holds; the right inverse can be written explicitly through $D^\sharp$ as in [19], taking into account the presence of the projection. Assume that Ψ is onto. Since it is a map between Hilbert spaces this is equivalent to the existence of a continuous right inverse of Ψ denoted by $\Psi^\sharp$. Let us first prove that this

implies that there exists a positive constant h such that $\Psi\Psi^\top \geq h\, Id$. By contradiction assume that there exists a sequence $\{u_n\} \subset U$ such that

$$\|u_n\| = 1, \qquad \left\|\Psi^\top u_n\right\|^2 \to 0.$$

Hence

$$(\Psi^\sharp)^\top \Psi^\top u_n = (\Psi\Psi^\sharp)^\top u_n = u_n \xrightarrow{s} 0,$$

a contradiction. Easy computations show that the map $\Psi^\top : W \subset L^2([0,T], \mathbf{R}^k) \to U$ is given by $v \mapsto (\nu(0), B^\top\nu + D^\top v)$, where $\dot\nu(t) = -A^\top(t)\nu(t) - C^\top(t)v(t),\ \nu(T) = 0$. Hence

$$\left\|\Psi^\top v\right\|^2 = \nu(0)^\top\nu(0) + \int_0^T (\nu^\top BB^\top\nu + 2\nu^\top BD^\top v + v^\top DD^\top v)(s)\, ds.$$

All the addends where the state ν appears give weakly continuous quadratic forms; moreover $\Psi\Psi^\top$ is coercive and hence Legendre. By Theorem 11.5 in [7] the nonnegative quadratic form $Q : v \mapsto \left\|D^\top v\right\|^2$ is Legendre. The nullity of Q coincides with Dim Ker $D^\top$ and therefore is either 0 or $+\infty$. Since Q is Legendre its nullity is finite dimensional and hence it is 0; that is, Q is positive and hence coercive by Theorem 11.1 in [7]. Since Δ is an orthogonal projection, the quadratic form on $L^2([0,T], \mathbf{R}^k)$ given by

$$w \mapsto \left\|D^\top \Delta w\right\|^2 + \|(Id - \Delta)w\|^2$$

is coercive. Finally, by Theorem 4.4 in [7] we obtain Assumption 2.2. □

Lemma 4.2 shows that the three statements in Assumption 2.3 are equivalent.

LEMMA 4.2. *The three statements in Assumption* 2.3 *are equivalent.*

Proof. From (2.6) it follows that

$$u^\top \Pi_1 R \Pi_1 u \geq h_0 \|\Pi_1 u\|^2;$$

therefore, since the Π_i's are orthogonal projections,

$$u^\top(\Pi_1 R \Pi_1 u + \Pi_2 u) \geq h_0\|\Pi_1 u\|^2 + \|\Pi_2 u\|^2 \geq h\|u\|^2.$$

Hence (2.6) yields (2.7). (2.7) implies (2.8), obviously.

Let us now show that (2.8) implies (2.6). Let the L^∞ norm of the matrix $R(t)$ be bounded by M; from (2.8), by easy computations, we obtain that for all $h_0, h_1 > 0$

$$\begin{aligned} &u^\top(R + h_1\Pi_2 - h_0 Id)u \\ &= u^\top(\Pi_1 R\Pi_1 + \Pi_2 R\Pi_2 + 2\Pi_1 R\Pi_2 + h_1\Pi_2 - h_0 Id)u \\ &\geq u^\top(h\Pi_1 - M\Pi_2 + h_1\Pi_2 - h_0 Id)u - 2M\|\Pi_1 u\|\|\Pi_2 u\| \\ &= (h - h_0)\|\Pi_1\|^2 + (h_1 - h_0 - M)\|\Pi_2\|^2 - 2M\|\Pi_1 u\|\|\Pi_2 u\|. \end{aligned}$$

By a suitable choice of h_0, h_1 the sum in the brackets can be made nonnegative, proving (2.6). □

In the following example we have a coercive quadratic form with a nonzero c-transversal solution of the Jacobi system which corresponds to the zero extremal.

Example 4.1. Let us consider the quadratic form

$$I(u) = \frac{1}{2}\int_0^2 u^2(s)\, ds$$

on the subspace H of the controls u such that

$$\dot{\xi}(t) = b(t)u(t), \quad t \in [0,2],$$

$$\xi(0) = 0, \quad \xi(2) = 0,$$

where the function b is identically zero on $[0,1]$ and positive elsewhere.

I is clearly coercive on H and the solutions of the Jacobi system are given by

$$(\zeta(t), \lambda(t)) = \left(-\lambda_0 \int_0^t b^2(s)\, ds, \lambda_0 \right).$$

For any $c \in [0,1]$ we have nonzero c-transversal solutions (ζ, λ), but they correspond to the trivial c-transversal extremal so that c is not a point semiconjugate with zero.

The next example shows that the sufficient conditions in Theorem 2.11 are not necessary.

Example 4.2. Let us consider the quadratic form

$$I(x,u) = \xi(0)\xi(T) + \frac{1}{2}\int_0^T u^2(s)\, ds$$

on the subspace H of the couples (ξ, u) such that $\dot{\xi}(t) = 0$ and $\xi(0) = x$. The state is constant, and hence the form J is clearly coercive because it can be written as $I(x,u) = \|\xi(0)\|^2 + \frac{1}{2}\|u\|^2$, but the conditions of Theorem 2.12 are not satisfied. In fact the solutions of the Riccati equations are also constant, but there is no solution W which satisfies

$$\begin{pmatrix} 0 & 1 \\ 1 & 0 \end{pmatrix} + \begin{pmatrix} W & 0 \\ 0 & -W \end{pmatrix} > 0 \text{ on } \mathbf{R}^2.$$

Let us remark that if we write

$$I(x,u) = \xi^2(0) + \frac{1}{2}\int_0^T u^2(s)\, ds,$$

which is clearly an equivalent expression for I on H, we can then apply Theorems 2.9 and 2.10, whose conditions are satisfied for $W = -\frac{1}{2}$.

The next example shows why in statement 4 of Theorem 2.11 we do not have strict inequality as in Theorem 2.10. The perturbation argument cannot be applied to a solution defined on a half-open interval $[0,T)$, and hence a solution of the Riccati equation satisfying $W(0) + \Gamma_0 > 0$ may not exist on the whole interval $[0,T)$.

Example 4.3. Let us consider the following form in the calculus of variations

$$I(x,\dot{\xi}) = \frac{1}{2}\xi^2(0) + \frac{1}{2}\int_0^{\frac{3\pi}{4}} \left\{ \dot{\xi}^2(s) - \xi(s) \right\} ds$$

with $\xi(T) = 0$. This form is positive semidefinite, and it has value zero along the arc $\xi(t) = \sin(t) + \cos(t)$. All the assumptions of Theorem 2.11 are satisfied, and there exists a solution of the corresponding Riccati equation which satisfies the suitable initial condition, namely $W(t) = \tan(t - \frac{\pi}{4})$, defined on the interval $[0, \frac{3\pi}{4})$. If we want $W(0) + \Gamma_0 > 0$ then we have to perturb the initial conditions by a positive constant ϵ. The solution becomes $W_\epsilon(t) = \tan(t - \frac{\pi}{4} + \epsilon)$, which does not exist on the whole interval $[0, \frac{3\pi}{4})$. Clearly if we want $W(0) + \Gamma_0 \geq 0$, it is enough to take $\epsilon = 0$.

We end this section with an example where these results have been applied to study an economic theory model (see [1, 2]).

Example 4.4. The model concerns the classical optimal saving problem for a consumer with increasing marginal utility in an economy characterized by irreversible investments. If the consumer has a quadratic utility function then the problem over a finite horizon can be seen as the following optimal control problem:

$$\text{Maximize } J_T(c) = \frac{1}{2}\int_0^T e^{-\rho s}c^2(s)\,ds,$$
$$\dot{k}(t) = \alpha k(t) - c(t), \quad k(0) = k_0 > 0,$$
$$0 \le c(t) \le \alpha k(t),$$

where

- c is consumption,
- k is capital,
- ρ is discount factor, $\rho > 0$,
- α is instantaneous rate of return, $0 \le \alpha \le 1$.

This example is of some interest because the usual existence theorems do not apply; hence the candidate optimal solution obtained by the Pontryagin maximum principle needs to be tested by second-order conditions in order to prove its optimality.

In [1] the following candidate optimal solution is singled out:

$$\hat{c}(t) = \begin{cases} 0, & t \in [0, t^*), \\ \alpha\hat{k}(t), & t \in [t^*,\ T], \end{cases} \qquad \hat{k}(t) = \begin{cases} k_0 e^{\alpha t}, & t \in [0, t^*), \\ k_0 e^{\alpha t^*}, & t \in [t^*,\ T], \end{cases}$$

where

$$t^* = \begin{cases} 0 & \text{if } \rho \ge 2\alpha \text{ or if } \rho < 2\alpha \text{ and } T \le \bar{t}, \\ T - \bar{t} & \text{if } \rho < 2\alpha \text{ and } T > \bar{t} \end{cases}$$

and

$$\bar{t} = \frac{1}{\rho}\ln\left(\frac{2\alpha}{2\alpha - \rho}\right).$$

By adding some extra controls this problem can be reduced to one with equality constraints. In [2] it is shown that there are no semiconjugate points in the accessory problem; therefore the candidate solution provides a local maximum which is also a global optimum.

The solution has an interesting economic meaning. If this *hedonistic* consumer has a discount rate that is small compared with the instantaneous rate of return, $\rho < 2\alpha$, and he/she has a long enough life, $T > \bar{t}$, then he or she accumulates the returns from investments during a certain period of time and afterwards he or she consumes all the surplus. Otherwise he or she will consume all the investment returns immediately. Hence a hedonistic consumer, i.e., with increasing marginal utility, might also be willing to save.

REFERENCES

[1] E. Barucci and P. Zezza, *Optimality conditions for control systems and economic applications*, Rivista Internazionale di Scienze Economiche e Sociali, XLII (1995), pp. 257–283.

[2] E. Barucci and P. Zezza, *Does a life cycle exist for a hedonistic consumer?*, Math. Social Sci., 32 (1996), pp. 57–69.

[3] J. Breakwell and Y. Ho, *On the conjugate point condition for the control problem*, Internat. J. Engrg. Sci., 2 (1965), pp. 565–579.

[4] W. A. Coppel, *Linear-quadratic optimal control*, Proc. Roy. Soc. Edinburgh Sect. A, 73 (1974/75), pp. 271–289.
[5] A. Dmitruk, *The Euler-Jacobi equation in the calculus of variations*, Mat. Zametki, 20 (1976), pp. 847–858 (in Russian). (English translation Math. Notes, 20 (1976), pp. 1032–1038.)
[6] A. Dmitruk, *Conditions of Jacobi-type for Bolza's problem with inequalities*, Mat. Zametki, 35 (1984), pp. 813–827 (in Russian). (English translation Math. Notes, 35 (1984), pp. 427–435.)
[7] M. R. Hestenes, *Applications of the theory of quadratic forms in Hilbert space to calculus of variations*, Pacific J. Math., 1 (1951), pp. 525–581.
[8] M. R. Hestenes, *On quadratic control problems*, in Calculus of Variations and Control Theory, D. L. Russell, ed., Academic Press, New York, 1976, pp. 289–304.
[9] R. E. Kalman, *Contributions to the theory of optimal control*, Bol. Soc. Mat. Mexicana, 5 (1960), pp. 102–119.
[10] H. Kelley, *Guidance theory and extremal fields*, IEEE Trans. Automat. Control, 7 (1962), pp. 75–82.
[11] K. Malanowski, *Two-norm approach in stability and sensitivity analysis of optimization and optimal control problems*, Adv. Math. Sci. Appl., 2 (1993), pp. 397–443.
[12] K. Malanowski, *Regularity of solutions in stability analysis of optimization and optimal control problems*, Control Cybernet., 23 (1994), pp. 61–86.
[13] H. Maurer and H. Pesch, *Solution differentiability for parametric nonlinear control problems with control-state constraints*, Control Cybernet., 23 (1994), pp. 201–227.
[14] H. Maurer and S. Pickenhain, *Second order sufficient conditions for optimal control problems with mixed control-state constraints*, J. Optim. Theory Appl., 86 (1995), pp. 649–667.
[15] E. Mikami, *Focal points in a control problem*, Pacific J. Math., 35 (1970), pp. 473–485.
[16] B. P. Molinari, *Nonnegativity of a quadratic functional*, SIAM J. Control, 13 (1975), pp. 792–806.
[17] M. Morse, *The Calculus of Variations in the Large*, AMS Colloquium Publications 18, New York, 1934.
[18] G. Stefani and P. Zezza, *A new type of sufficient optimality conditions for a nonlinear constrained optimal control problem*, in Proc. Nonlinear Control System Design Symposium, M. Fliess, ed., Bordeaux, 1992, pp. 713–719.
[19] G. Stefani and P. Zezza, *Optimal control problems with mixed state-control constraints: Necessary conditions*, J. Math. Systems Estim. Control, 2 (1992), pp. 155–189.
[20] G. Stefani and P. Zezza, *The Jacobi condition for LQ-control problems with constraints*, in Proc. Second European Control Conference, 1993, J.W. Nieuwenhuis, C. Praagman, H.L. Trentelman, eds., pp. 1003–1007.
[21] G. Stefani and P. Zezza, *Minima in control problems with constraints*, in Geometry in Nonlinear Control and Differential Inclusions, Banach Center Publications 32, Warszawa, 1995, pp. 361–378.
[22] G. Stefani and P. Zezza, *The Riccati equation for regular LQ-control problems with constraints*, in Proc. Nonlinear Control System Design Symposium, Lake Tahoe, International Federation on Automatic Control, 1995, pp. 145–149.
[23] G. Stefani and P. Zezza, *Optimality conditions for a constrained optimal control problem*, SIAM J. Control Optim., 34 (1996), pp. 635–659.
[24] V. Zeidan, *Sufficiency conditions for variational problems with variable endpoints: Coupled points*, Appl. Math. Optim., 27 (1993), pp. 191–209.
[25] V. Zeidan, *The Riccati equation for optimal control problems with mixed state-control constraints : necessity and sufficiency*, SIAM J. Control Optim., 32 (1994), pp. 1297–1321.
[26] V. Zeidan and P. Zezza, *Coupled points in the calculus of variations and applications to periodic problems*, Trans. Amer. Math Soc., 1 (1989), pp. 323–335.
[27] V. Zeidan and P. Zezza, *Coupled points in optimal control*, IEEE Trans. Automat. Control, 36 (1991), pp. 1276–1281.
[28] P. Zezza, *The Jacobi condition in optimal control*, in Control Theory, Stochastic Analysis and Applications, S. Chen and J. Yong, eds., World Scientific, Singapore, 1991, pp. 137–149.
[29] P. Zezza, *The Jacobi condition for elliptic forms in Hilbert spaces*, J. Optim. Theory Appl., 76 (1993), pp. 357–380.

SIAM J. CONTROL OPTIM.
Vol. 35, No. 3, pp. 901–929, May 1997

008

OPTIMAL CONTROL FOR HOLONOMIC AND NONHOLONOMIC MECHANICAL SYSTEMS WITH SYMMETRY AND LAGRANGIAN REDUCTION*

WANG-SANG KOON† AND JERROLD E. MARSDEN‡

Abstract. In this paper we establish necessary conditions for optimal control using the ideas of Lagrangian reduction in the sense of reduction under a symmetry group. The techniques developed here are designed for Lagrangian mechanical control systems with symmetry. The benefit of such an approach is that it makes use of the special structure of the system, especially its symmetry structure, and thus it leads rather directly to the desired conclusions for such systems.

Lagrangian reduction can do in one step what one can alternatively do by applying the Pontryagin maximum principle followed by an application of Poisson reduction. The idea of using Lagrangian reduction in the sense of symmetry reduction was also obtained by Bloch and Crouch [*Proc. 33rd CDC*, IEEE, 1994, pp. 2584–2590] in a somewhat different context, and the general idea is closely related to those in Montgomery [*Comm. Math. Phys.*, 128 (1990), pp. 565–592] and Vershik and Gershkovich [*Dynamical Systems* VII, V. Arnold and S. P. Novikov, eds., Springer-Verlag, 1994]. Here we develop this idea further and apply it to some known examples, such as optimal control on Lie groups and principal bundles (such as the ball and plate problem) and reorientation examples with zero angular momentum (such as the satellite with moveable masses). However, one of our main goals is to extend the method to the case of nonholonomic systems with a nontrivial momentum equation in the context of the work of Bloch, Krishnaprasad, Marsden, and Murray [*Arch. Rational Mech. Anal.*, (1996), to appear]. The snakeboard is used to illustrate the method.

Key words. constraints, Lagrangian reduction, mechanical systems with symmetry, nonholonomic, optimal control

AMS subject classifications. 49K99, 49R99, 58F05, 70E99, 70H99, 93B29

PII. S0363012995290367

1. Introduction.

Recently several papers have appeared exploring the symmetry reduction of optimal control problems on configuration spaces such as Lie groups and principal bundles. The mechanical systems which they have modeled vary widely, ranging from the falling cat and the rigid body with two oscillators to the plate-ball system as well as the (airport) landing tower problem. Since the Pontryagin maximum principle is such an important and powerful tool in optimal control theory, it is frequently employed as a first step in finding necessary conditions for the optimal controls. Finally, different variants of Poisson reduction on the cotangent bundle T^*Q of the configuration space Q are used to obtain the reduced equations of motion for the optimal trajectories.

This paper develops a Lagrangian alternative to the method of Pontryagin maximum principle and Poisson reduction used in many of the above studies. More importantly, our method can handle the optimal control of nonholonomic mechanical system such as the snakeboard, which has a nontrivial evolution equation for its nonholonomic momentum. Our key idea is to link the method of Lagrange multipliers with Lagrangian reduction. This procedure, which will be referred to as "reduced

*Received by the editors August 14, 1995; accepted for publication (in revised form) March 30, 1996. This research was supported in part by the NSF and DOE.
http://www.siam.org/journals/sicon/35-3/29036.html
†Department of Mathematics, University of California, Berkeley, CA 94720-3840 (koon@math.berkeley.edu). Current address: Control and Dynamical Systems, California Institute of Technology 116-81, Pasadena, CA 91125.
‡Control and Dynamical Systems, California Institute of Technology 116-81, Pasadena, CA 91125 (marsden@cds.caltech.edu).

Lagrangian optimization," is able to handle all the above cases, including the snakeboard. We hope that it will complement other existing methods and may also have the advantage that it is easier to use in many situations and can solve many new problems. In the optimal control problems we deal with in this paper, one encounters degenerate Lagrangians; fortunately this does not cause problems with the technique of Lagrangian reduction. For more information on these degeneracies, see Bloch and Crouch [1994].

Our objectives in this paper are limited to presenting reduced Lagrangian optimization in the context of both holonomic and nonholonomic systems that may have conservation laws or nontrival momentum equations. We use this approach as an alternative to the Pontryagin maximum principle and Poisson reduction. Although an assumption of controllability underlies most optimal control problems, we are concerned here with finding necessary conditions for optimality and so do not discuss controllability explicitly. We do not extensively develop the geometry of the situation in much detail, and we restrict our attention to regular extremals throughout the paper without explicit mention. Of course all of these points are of interest in themselves.

In the course of working on this paper, we have found some related ideas in Montgomery [1990], Vershik and Gershkovich [1994], Bloch and Crouch [1994, 1995], Kelly and Murray [1995], and Bloch, Crouch, and Ratiu [1994]. The paper by Bloch, Krishnaprasad, Marsden, and Murray [1996] provides a useful framework for the present work.

Outline of the paper. In section 2, we recall some basic facts about both holonomic and nonholonomic mechanical systems with symmetry. We set up a class of optimal control problems for holonomic mechanical systems on a (trivial) principal bundle as was done in Montgomery [1990, 1993] and in Krishnaprasad, Yang, and Dayawansa [1991]. We also set up the corresponding problems for nonholonomic systems. We will call these "Lagrangian optimal control problems."

In section 3 we review some aspects of the theory of Lagrangian reduction and use it to solve the Lagrangian optimal control problem in the holonomic case, showing that an optimal trajectory is a solution of Wong's equations (at least for regular extremals). This provides an alternative derivation to the approach (based on methods of sub-Riemannian geometry) in Montgomery [1990] and the approach (based on the Pontryagin maximum principle and Poisson reduction) in Krishnaprasad, Yang, and Dayawansa [1991].

In section 4 we generalize these results to the case of nonholonomic systems. Notice in particular that our techniques allow for nonzero values of the momentum map, which is interesting even for the holonomic case. In section 5 we consider a number of examples, such as the ball on a plate (as in Bloch, Krishnaprasad, Marsden, and Murray [1995]) and the snakeboard. We also consider optimal control problems for systems on Lie groups such as the landing tower problem (see Krishnaprasad [1993] and Walsh, Montgomery, and Sastry [1994]) and the plate-ball problem considered in Jurdjevic [1993, 1996].

In the conclusion, we give a few remarks on future research directions.

2. Lagrangian mechanical systems with symmetry. In this section we shall review, for the convenience of the reader, some notation and results for mechanical systems with symmetry. We will begin with the case of holonomic systems and then study the nonholonomic case.

2.1. Holonomic systems with symmetry.

Notation. A simple Lagrangian system with symmetry consists of a configuration manifold Q, a metric tensor (the mass matrix) $\langle\langle\, , \rangle\rangle$, a symmetry group G (a Lie group), and a Lagrangian L. Assume that G acts on Q by isometries and that the Lagrangian L is of the form kinetic minus potential energy; i.e.,

$$L(q,v) = \frac{1}{2}\|v\|_q^2 - V(q),$$

where $\|\cdot\|_q$ denotes the norm on T_qQ and V is a G-invariant potential. For more information, see, for example, Marsden [1992] and Marsden and Ratiu [1994]. Examples of such systems are the falling cat (Montgomery [1990, 1991]) and the rigid body with two oscillators (Krishnaprasad, Yang, and Dayawansa [1991]).

The associated equivariant momentum map $J : TQ \to \mathfrak{g}^*$ for a simple Lagrangian system with symmetry is given by

$$\langle J(q,v), \xi\rangle = \langle\langle v, \xi_Q(q)\rangle\rangle = \frac{\partial L}{\partial \dot{q}^i}(\xi_Q)^i, \tag{2.1}$$

where $\mathfrak{g}^*$ is the dual of the Lie algebra $\mathfrak{g}$ of G, ξ_Q is the infinitesimal generator of $\xi \in \mathfrak{g}$ on Q, and $\langle\, , \rangle$ is the pairing between $\mathfrak{g}^*$ and $\mathfrak{g}$ (other natural pairings between spaces and their duals are also denoted $\langle\, , \rangle$ in this paper).

Assume that G acts freely and properly on Q, so we can regard $Q \to Q/G$ as a principal G-bundle (Q, B, π, G), where $B = Q/G$ is called the base (or shape) space and $\pi : Q \to B$ is the bundle projection. On this bundle, we construct the mechanical connection $\mathcal{A}$ as follows: for each $q \in Q$, let the locked inertia tensor be the map $\mathbb{I}(q) : \mathfrak{g} \to \mathfrak{g}^*$ defined by

$$\langle \mathbb{I}(q)\eta, \xi\rangle = \langle\langle \eta_Q(q), \xi_Q(q)\rangle\rangle.$$

The terminology comes from the fact that for a coupled rigid body, particle, or elastic system, $\mathbb{I}(q)$ is the classical moment of inertia tensor of the instantaneous rigid system. The mechanical connection is the map $\mathcal{A} : TQ \to \mathfrak{g}$ that assigns to each (q,v) the "angular velocity of the locked system"

$$\mathcal{A}(q,v) = \mathbb{I}(q)^{-1} J(q,v). \tag{2.2}$$

When there is danger of confusion, we will write the mechanical connection as $\mathcal{A}^{\text{mec}}$ (additional connections will be introduced later in the paper). The map $\mathcal{A}$ is a connection on the principal G-bundle $Q \to Q/G$; that is, $\mathcal{A}$ is G-equivariant and satisfies $\mathcal{A}(\xi_Q(q)) = \xi$, both of which are readily verified. The horizontal space of the connection $\mathcal{A}$ is given by

$$\text{hor}_q = \{(q,v) \mid J(q,v) = 0\},$$

i.e., the space orthogonal to the G-orbits. The vertical space consists of vectors that are tangent to the group orbits; i.e.,

$$\text{ver}_q = T_q(\text{Orb}(q)) = \{\xi_Q(q) \mid \xi \in \mathfrak{g}\}.$$

For later use, we would like to say a few words about a general principal connection and its expression in a local trivialization. As stated above, a principal connection is

a $\mathfrak{g}$-valued one-form $\mathcal{A} : TQ \to \mathfrak{g}$ such that $\mathcal{A}(g \cdot v) = \mathrm{Ad}_g \mathcal{A}(v)$ and $\mathcal{A}(\xi_Q(q)) = \xi$ for each $\xi \in \mathfrak{g}$. For example, if $Q = G$, there is a canonical connection given by the right invariant one-form which equals the identity at $g = e$. That is, for $v \in T_gG$, we let $\mathcal{A}_G : TG \to \mathfrak{g}$, $\mathcal{A}_G(v) = TR_{g^{-1}} \cdot v$. In a local trivialization, where we can locally write $Q = B \times G$ and the action of G is given by left translation on the second factor, a connection $\mathcal{A}$ as a one-form has the form

$$\mathcal{A}(r, g) = \mathcal{A}_{\mathrm{loc}}(r, g)dr + \mathcal{A}_G$$

and

$$\mathcal{A}(r, g)(\dot{r}, \dot{g}) = \mathcal{A}_{\mathrm{loc}}(r, g)\dot{r} + \dot{g}g^{-1} = \mathrm{Ad}_g(\mathcal{A}_{\mathrm{loc}}(r, e)\dot{r} + g^{-1}\dot{g}),$$

where $(\dot{r}, \dot{g})$ is the tangent vector at each point $q = (r, g)$. With abuse of notation, we denote $\mathcal{A}_{\mathrm{loc}}(r, e) = \mathcal{A}_{\mathrm{loc}}(r)$. Hence, for a principal connection, we can write

$$\mathcal{A}(r, g)(\dot{r}, \dot{g}) = \mathrm{Ad}_g(g^{-1}\dot{g} + \mathcal{A}_{\mathrm{loc}}(r)\dot{r}). \tag{2.3}$$

Holonomic optimal control problems. Now we are ready to formulate an optimal control problem for a holonomic system on a trivial bundle $(B \times G, B, \pi, G)$. As in Montgomery [1990, 1991] and Krishnaprasad, Yang, and Dayawansa [1991], let us assume that the control is internal to the system, which leaves invariant the conserved momentum map J and that there is no drift; i.e., $\mu = J(q, v) = 0$. Assume further that the velocity $\dot{r}$ of the path in the base space B can be directly controlled; then an associated control problem can be set up as

$$\begin{cases} \dot{r} = u, \\ g^{-1}\dot{g} = -\mathcal{A}_{\mathrm{loc}}(r)u \end{cases} \tag{2.4}$$

because, from the results above, the constraint that $\mu = 0$ is nothing but $(\dot{r}, \dot{g}) \in \mathrm{hor}_{(r,g)}$, which is equivalent to $g^{-1}\dot{g} + \mathcal{A}_{\mathrm{loc}}(r)\dot{r} = 0$. Here $u(\cdot)$ is a vector-valued function.

Let C be a cost function which usually is a positive definite quadratic function in u and hence C can be written as the square of a metric on B. Then we can formulate an optimal control problem on $Q = B \times G$ as follows.

Optimal control problem for holonomic systems. *Given two points q_0, q_1 in Q, find the optimal controls $u(\cdot)$ which steer from q_0 to q_1 and minimize $\int_0^1 C(u)dt$ subject to the constraints $\dot{r} = u$, $g^{-1}\dot{g} = -\mathcal{A}_{\mathrm{loc}}(r)u$.*

Clearly the above optimal control problem is equivalent to the following constrained variational problem.

Constrained variational problem for holonomic systems. *Among all curves $q(t)$ such that $\dot{q}(t) \in \mathrm{hor}_{q(t)}$, $q(0) = q_0$, $q(1) = q_1$, find the optimal curves $q(t)$ such that $\int_0^1 C(\dot{r})dt$ is minimized, where $r = \pi(q)$.*

For example, Krishnaprasad, Yang, and Dayawansa [1991] considered a rigid body with two (driven) oscillators, which was used to model the drift observed in the Hubble Space Telescope due to thermoelastically driven shape changes of the solar panels arising from the day–night thermal cycling during orbit. The bundle used was $(\mathbb{R}^2 \times \mathrm{SO}(3), \mathbb{R}^2, \pi, \mathrm{SO}(3))$, and the corresponding optimal control problem was as follows.

Optimal control for a rigid body with two oscillators. *Find the control $u(\cdot) = (u^1(\cdot), u^2(\cdot))$ that minimizes $\int_0^1((u^1)^2 + (u^2)^2)dt$ subject to $\dot{r} = u$, $\dot{g} = -g\mathcal{A}_{\mathrm{loc}}(r)u$ for $r^1(0) = r^1(1) = r^2(0) = r^2(1) = 0$, $g(0) = g_0$, and $g(1) = g_1 \in \mathrm{SO}(3)$.*

For more details on the derivation of this model, see Krishnaprasad, Yang, and Dayawansa [1991]. Below we will take this optimal control problem as given and focus on finding the necessary conditions for its optimal trajectories; see Montgomery [1990, 1991] for additional examples.

2.2. Simple nonholonomic mechanical systems with symmetry. Next we recall some basic ideas and results from Bloch, Krishnaprasad, Marsden, and Murray [1995] which will help to set the overall context for the optimal control of a simple nonholonomic system. Assume that we have data as before, namely a configuration manifold Q, a Lagrangian of the form kinetic minus potential, and a symmetry group G that leaves the Lagrangian invariant. However, now we also assume we have a distribution $\mathcal{D}$ that describes the kinematic nonholonomic constraints. Thus, $\mathcal{D}$ is a collection of linear subspaces denoted $\mathcal{D}_q \subset T_qQ$, one for each $q \in Q$. We assume that G acts on Q by isometries and leaves the distribution invariant; i.e., the tangent of the group action maps $\mathcal{D}_q$ to $\mathcal{D}_{gq}$. Moreover, we assume that we are in the principal case where the constraints and the orbit directions span the entire tangent space to the configuration space: $\mathcal{D}_q + T_q(\mathrm{Orb}(q)) = T_qQ$ for each $q \in Q$.

As discussed in Bloch, Krishnaprasad, Marsden, and Murray [1996], the dynamics of a nonholonomically constrained mechanical system is governed by the Lagrange–d'Alembert principle. This principle states that (at least in the case of homogeneous linear constraints) the equations of motion of a curve $q(t)$ in configuration space are obtained by setting to zero the variations in the integral of the Lagrangian subject to variations lying in the constraint distribution and that the velocity of the curve $q(t)$ itself satisfies the constraints.

The momentum equation. In the case of a simple holonomic mechanical system, setting up an optimal control problem uses the momentum map J, the mechanical connection $\mathcal{A}$ as well as the reconstruction of path on Q given a path in Q/G. For the case of a simple nonholonomic mechanical system, we shall need similar notions, and they are recalled in the following discussion.

Let the intersection of the tangent to the group orbit and the distribution at a point $q \in Q$ be denoted

$$\mathcal{S}_q = \mathcal{D}_q \cap T_q(\mathrm{Orb}(q)).$$

Define, for each $q \in Q$, the vector subspace $\mathfrak{g}^q$ to be the set of Lie algebra elements in $\mathfrak{g}$ whose infinitesimal generators evaluated at q lie in $\mathcal{S}_q$:

$$\mathfrak{g}^q = \{\xi \in \mathfrak{g} : \xi_Q(q) \in S_q\}.$$

Then $\mathfrak{g}^{\mathcal{D}}$ is the corresponding bundle over Q whose fiber at the point q is given by $\mathfrak{g}^q$. The nonholonomic momentum map J^{nh} is the bundle map taking TQ to the bundle $(\mathfrak{g}^{\mathcal{D}})^*$ (whose fiber over the point q is the dual of the vector space $\mathfrak{g}^q$) that is defined by

$$\langle J^{\mathrm{nh}}(v_q), \xi\rangle = \frac{\partial L}{\partial \dot{q}^i}(\xi_Q)^i, \tag{2.5}$$

where $\xi \in \mathfrak{g}^q$.

As examples such as the snakeboard show, in general the tangent space to the group orbit through q intersects the constraint distribution at q nontrivially.

Notice that the nonholonomic momentum map may be viewed as giving just some of the components of the ordinary momentum map, namely along those symmetry directions that are consistent with the constraints.

It is proven in Bloch, Krishnaprasad, Marsden, and Murray [1996] that if the Lagrangian L is invariant under the group action and that if ξ^q is a section of the bundle $\mathfrak{g}^{\mathcal{D}}$, then any solution $q(t)$ of the Lagrange–d'Alembert equations for a nonholonomic system must satisfy, in addition to the given kinematic constraints, the momentum equation

$$\frac{d}{dt}\left(J^{\mathrm{nh}}(\xi^{q(t)})\right) = \frac{\partial L}{\partial \dot{q}^i}\left[\frac{d}{dt}(\xi^{q(t)})\right]_Q^i. \tag{2.6}$$

When the momentum map is paired with a section in this way, we will just refer to it as the momentum. Examples show that the nonholonomic momentum map may or may not be conserved.

The momentum equation in an orthogonal body frame. Let a local trivialization (r, g) be chosen on the principal bundle $\pi : Q \to Q/G$. Let $\eta \in \mathfrak{g}^q$ and $\xi = g^{-1}\dot{g}$. Since L is G invariant, we can define a new function l by writing $L(r, g, \dot{r}, \dot{g}) = l(r, \dot{r}, \xi)$. Define $J^{\mathrm{nh}}_{\mathrm{loc}} : TQ/G \to (\mathfrak{g}^{\mathcal{D}})^*$ by

$$\left\langle J^{\mathrm{nh}}_{\mathrm{loc}}(r, \dot{r}, \xi), \eta \right\rangle = \left\langle \frac{\partial l}{\partial \xi}, \eta \right\rangle.$$

As with connections, J^{nh} and its version in a local trivialization are related by the Ad map; i.e., $J^{\mathrm{nh}}(r, g, \dot{r}, \dot{g}) = \mathrm{Ad}^*_{g^{-1}} J^{\mathrm{nh}}_{\mathrm{loc}}(r, \dot{r}, \xi)$.

Choose a q-dependent basis $e_a(q)$ for the Lie algebra such that the first m elements span the subspace $\mathfrak{g}^q$. We require the basis to be such that the infinitesimal generators of the first m basis elements are orthogonal in the kinetic energy metric to the generators of the last $k - m$ basis elements. In a local trivialization, one chooses, for each r, such a basis at the identity element, and we denote it by

$$e_1(r), e_2(r), \ldots, e_m(r), e_{m+1}(r), \ldots, e_k(r).$$

Define the orthogonal body frame by

$$e_a(r, g) = \mathrm{Ad}_g \cdot e_a(r);$$

thus, by G invariance, the first m elements span the subspace $\mathfrak{g}^q$. In this basis, we have

$$\left\langle J^{\mathrm{nh}}(r, g, \dot{r}, \dot{g}), e_b(r, g) \right\rangle = \left\langle \frac{\partial l}{\partial \xi}, e_b(r) \right\rangle := p_b, \tag{2.7}$$

which defines p_b, a function of r, $\dot{r}$, and ξ. It is proven in Bloch, Krishnaprasad, Marsden, and Murray [1995] that in such an orthogonal body frame, the momentum equation can be written in the following form:

$$\dot{p} = \dot{r}^T H(r)\dot{r} + \dot{r}^T K(r)p + p^T D(r)p. \tag{2.8}$$

Note that in this body representation, the functions p_b are *invariant* rather than equivariant, as is usually the case with the momentum map, and the momentum equation is independent of, that is, decouples from, the group variables g.

The nonholonomic connection. Recall that in the case of holonomic mechanical systems, the mechanical connection $\mathcal{A}$ is defined by $\mathcal{A}(v_q) = \mathbb{I}(q)^{-1}J(v_q)$ or equivalently by the fact that its horizontal space at q is orthogonal to the group orbit at q. For the case of a simple nonholonomic mechanical system where the Lagrangian is of the form kinetic minus potential energy and G acts on Q by isometries and leaves $\mathcal{D}$ invariant, the result turns out to be quite similar.

As Bloch, Krishnaprasad, Marsden, and Murray [1996] point out, in the principal case where the constraints and the orbit directions span the entire tangent space to the configuration space (that is, $\mathcal{D}_q + T_q(\mathrm{Orb}(q)) = T_qQ$), the nonholonomic connection $\mathcal{A}^{\mathrm{nh}}$ is a principal connection on the bundle $Q \to Q/G$ whose horizontal space at the point $q \in Q$ is given by the orthogonal complement to the space $\mathcal{S}_q$ within the space $\mathcal{D}_q$. Moreover, Bloch, Krishnaprasad, Marsden, and Murray [1996] develop formulas for $\mathcal{A}^{\mathrm{nh}}$ similar to those for the mechanical connection; namely,

$$\mathcal{A}^{\mathrm{nh}}(v_q) = \mathbb{I}^{\mathrm{nh}}(q)^{-1}J^{\mathrm{nh}}(v_q), \tag{2.9}$$

where $\mathbb{I}^{\mathrm{nh}} : \mathfrak{g}^{\mathcal{D}} \to (\mathfrak{g}^{\mathcal{D}})^*$ is the locked inertia tensor defined in a way similar to that given above for holonomic systems. In an orthogonal body frame, (2.9) can be written as

$$\mathrm{Ad}_g(g^{-1}\dot{g} + \mathcal{A}^{\mathrm{nh}}_{\mathrm{loc}}(r)\dot{r}) = \mathrm{Ad}_g(\mathbb{I}^{\mathrm{nh}}_{\mathrm{loc}}(r)^{-1}p), \tag{2.10}$$

where $\mathcal{A}^{\mathrm{nh}}_{\mathrm{loc}}$ and $\mathbb{I}^{\mathrm{nh}}_{\mathrm{loc}}$ are the representations of $\mathcal{A}^{\mathrm{nh}}$ and $\mathbb{I}^{\mathrm{nh}}$ in a local trivialization. For simplicity in what follows, we shall omit the subscript "loc."

Control systems in momentum equation form. With the help of the momentum equations and the nonholonomic mechanical connection, Bloch, Krishnaprasad, Marsden, and Murray [1996] provide a framework for studying the general form of nonholonomic mechanical control systems with symmetry that may have a nontrivial evolution of their nonholonomic momentum. The dynamics of such a system can be described by a system of equations of the form of a reconstruction equation for a group element g, an equation for the nonholonomic momentum p (no longer conserved in the general case), and the equations of motion for the reduced variables r which describe the "shape" of the system. In terms of these variables, the equations of motion have the functional form

$$\begin{cases} g^{-1}\dot{g} = -\mathcal{A}^{\mathrm{nh}}(r)\dot{r} + \Gamma(r)p, \\ \dot{p} = \dot{r}^T H(r)\dot{r} + \dot{r}^T K(r)p + p^T D(r)p, \\ M(r)\ddot{r} = \delta(r,\dot{r},p) + \tau, \end{cases} \tag{2.11}$$

where $(\Gamma(r) = \mathbb{I}^{\mathrm{nh}}(r)^{-1})$.

The first equation describes the motion in the group variables as the flow of a left invariant vector field determined by the internal shape r, its velocity $\dot{r}$, as well as the generalized momentum p. The term $g^{-1}\dot{g} + \mathcal{A}^{\mathrm{nh}}(r)\dot{r} = \Gamma(r)^{-1}p$ is interpreted as the local representation of the body angular velocity. This is nothing more than the vertical part of the bundle velocity. The momentum equation describes the evolution of p and is bilinear in $(\dot{r}, p)$. Finally, the bottom (second-order) equation for $\ddot{r}$ describes the motion of the variables which describe the configuration up to a symmetry (i.e., the shape). The variable τ represents the external forces applied to the system, which we assume here only affect the shape variables; i.e., the external forces are G invariant. Note that the evolution of the momentum p and the shape r decouple from the group variables.

The optimal control problem for nonholonomic systems on a trivial bundle. Assume that we have a simple nonholonomic mechanical system with symmetry; thus, assume we have data $(Q, \mathcal{D}, \langle\langle\, , \rangle\rangle, G, L)$, where the Lagrangian L is G invariant and of the form kinetic minus potential energy, the distribution $\mathcal{D}$ is G invariant, and we are in the principal case where the constraints and the orbit directions span the tangent space to the configuration space. Let us also assume in this section that the principal bundle $\pi : Q \to Q/G$ is trivial; all the examples we consider (including the snakeboard) have a trivial principal bundle structure. We consider this simplification as a first step to the general case because in a local trivialization any principal bundle is a trivial bundle $(B \times G, B, \pi, G)$. Furthermore, we will assume the following.

(1) Any control forces applied to the system affect only the shape variables, which leaves the generalized momenta and the momentum equation unchanged. Indeed, such forces would be invariant under the action of the Lie group G and so would be annihilated by the variations taken to derive the momentum equation.
(2) We have full control of the shape variables; that is, the curve $r(t)$ in the shape space B can be specified arbitrarily using a suitable control force τ.

Given a cost function C which is a positive definite quadratic function of $\dot{r}(t)$ (so can be written as the square of a metric on the shape space B), we can formulate an optimal control problem on $Q = B \times G$ as follows.

OPTIMAL CONTROL PROBLEM FOR NONHOLONOMIC SYSTEMS. *Given two points $q_0, q_1 \in Q$, find the curves $r(t) \in B$ which steer the system from q_0 to q_1 and which minimize the total cost $\int_0^1 C(\dot{r})dt$, where $r = \pi(q)$, subject to the constraints $g^{-1}\dot{g} = -\mathcal{A}^{\rm nh}(r)\dot{r} + \Gamma(r)p$ and to the momentum equation $\dot{p} = \dot{r}^T H(r)\dot{r} + \dot{r}^T K(r)p + p^T D(r)p$.*

This optimal control problem is clearly equivalent to the following constrained variational problem.

CONSTRAINED VARIATIONAL PROBLEM FOR NONHOLONOMIC SYSTEMS. *Among all curves $q(t)$ with $q(0) = q_0, q(1) = q_1$ and satisfying $g^{-1}\dot{g} = -\mathcal{A}^{\rm nh}(r)\dot{r} + \Gamma(r)p$, where $\dot{p} = \dot{r}^T H(r)\dot{r} + \dot{r}^T K(r)p + p^T D(r)p$, find the curves $q(t)$ such that $\int_0^1 C(\dot{r})dt$ is minimized, where $r = \pi(q)$.*

Now we are ready to use the method of Lagrange multipliers and Lagrangian reduction to find necessary conditions for optimal trajectories.

3. Optimal control and Lagrangian reduction for holonomic systems.

In this section we consider reduced Lagrangian optimization in the context of holonomic systems.

3.1. A review of Lagrangian reduction.

We first recall some facts about Lagrangian reduction theory for systems with holonomic constraints (see Marsden and Scheurle [1993a, 1993b]).

Rigid body reduction. Let $R \in \mathrm{SO}(3)$ denote the time-dependent rotation that gives the current configuration of a rigid body. The body angular velocity Ω is defined in terms of R by

$$R^{-1}\dot{R} = \hat{\Omega},$$

where $\hat{\Omega}$ is the three-by-three skew matrix defined by $\hat{\Omega}v := \Omega \times v$. Denoting by I the (time-independent) moment of inertia tensor, the Lagrangian thought of as a function of R and $\dot{R}$ is given by $L(R, \dot{R}) = \frac{1}{2}\langle I\Omega, \Omega\rangle$, and when we think of it as a function of Ω alone, we write $l(\Omega) = \frac{1}{2}\langle I\Omega, \Omega\rangle$.

The following statements are equivalent.

(1) $(R, \dot{R})$ satisfies the Euler–Lagrange equations on SO(3) for L.
(2) Hamilton's principle on SO(3) holds:

$$\delta \int L\, dt = 0.$$

(3) Ω satisfies the Euler equations

$$I\dot{\Omega} = I\Omega \times \Omega.$$

(4) The reduced variational principle holds on $\mathbb{R}^3$:

$$\delta \int l\, dt = 0,$$

where variations in Ω are restricted to be of the form $\delta\Omega = \dot{\eta} + \eta \times \Omega$, with η an arbitrary curve in $\mathbb{R}^3$ satisfying $\eta = 0$ at the temporal endpoints.

An important point is that when one reduces the standard variational principle from SO(3) to its Lie algebra $\mathfrak{so}(3)$, one ends up with a variational principle in which the *variations are constrained*; that is, one has a principle of Lagrange–d'Alembert type. In this case, the term η represents the infinitesimal displacement of particles in the rigid body. Note that the same phenomenon of constrained variations occurs in the case of nonholonomic systems.

The Euler–Poincaré equations. Let $\mathfrak{g}$ be a Lie algebra and let $l : \mathfrak{g} \to \mathbb{R}$ be a given Lagrangian. Then the Euler–Poincaré equations are

$$\frac{d}{dt}\frac{\partial l}{\partial \xi} = \mathrm{ad}^*_\xi \frac{\partial l}{\partial \xi}$$

or, in coordinates,

$$\frac{d}{dt}\frac{\partial l}{\partial \xi^a} = C^b_{da}\xi^d \frac{\partial l}{\partial \xi^b},$$

where the structure constants are defined by $[\xi, \eta]^a = C^a_{de}\xi^d\eta^e$. If G is a Lie group with Lie algebra $\mathfrak{g}$, we let $L : TG \to \mathbb{R}$ be the left invariant extension of l and let $\xi = g^{-1}\dot{g}$. In the case of the rigid body, ξ is $\hat{\Omega}$, where Ω is the body angular velocity.

The basic fact regarding the Lagrangian reduction leading to these equations is as follows.

THEOREM 3.1 (Euler–Poincaré reduction). *A curve $(g(t), \dot{g}(t)) \in TG$ satisfies the Euler–Lagrange equations for L if and only if ξ satisfies the Euler–Poincaré equations for l.*

In this situation, the reduction is implemented by the map $(g, \dot{g}) \in TG \mapsto g^{-1}\dot{g} =: \xi \in \mathfrak{g}$.

One proof of this theorem is of special interest, as it shows how to drop variational principles to the quotient (see Marsden and Scheurle [1993b] and Bloch, Krishnaprasad, Marsden, and Ratiu [1996] for more details). Namely, we transform

$$\delta \int L\, dt = 0$$

under the map $(g, \dot{g}) \mapsto g^{-1}\dot{g}$ to give the reduced variational principle for the Euler–Poincaré equations: ξ satisfies the Euler–Poincaré equations if and only if

$$\delta \int l\, dt = 0,$$

where the variations are all those of the form

$$\delta\xi = \dot{\eta} + [\xi, \eta]$$

and where η is an arbitrary curve in the Lie algebra satisfying $\eta = 0$ at the endpoints. Variations of this form are obtained by calculating what variations are induced by variations on the Lie group itself.

One obtains the Lie–Poisson equations on $\mathfrak{g}^*$ by the Legendre transformation:

$$\mu = \frac{\partial l}{\partial \xi}, \quad h(\mu) = \mu \cdot \xi - l(\xi).$$

Dropping the variational principle this way is the analogue of Lie–Poisson reduction in which one drops the Poisson bracket from T^*G to the Lie–Poisson bracket on $\mathfrak{g}^*$.

The reduced Euler–Lagrange equations. The Euler–Poincaré equations can be generalized to the situation in which G acts freely on a configuration space Q to obtain the *reduced Euler–Lagrange equations*. This process starts with a G-invariant Lagrangian $L : TQ \to \mathbb{R}$, which induces a reduced Lagrangian $l : TQ/G \to \mathbb{R}$. The Euler–Lagrange equations for L induce the reduced Euler–Lagrange equations on TQ/G. To compute them in coordinates, it is useful to introduce a principal connection on the bundle $Q \to Q/G$. Although any can be picked, a common choice is the mechanical connection.

Thus, assume that the bundle $Q \to Q/G$ has a given (principal) connection $\mathcal{A}$. Divide variations into horizontal and vertical parts—this breaks up the Euler–Lagrange equations on Q into two sets of equations that we now describe. Let r^α be coordinates on shape space Q/G and Ω^a be coordinates for vertical vectors in a local bundle chart. Drop L to TQ/G to obtain a reduced Lagrangian $l : TQ/G \to \mathbb{R}$ in which the group coordinates are eliminated. We can represent this reduced Lagrangian in a couple of ways. First, if we choose a local trivialization as we have described earlier, we obtain l as a function of the variables $(r^\alpha, \dot{r}^\alpha, \xi^a)$. However, it will also be convenient to change variables from ξ^a to the local version of the locked angular velocity, i.e., the body angular velocity, namely $\Omega = \xi + \mathcal{A}_{\text{loc}}\dot{r}$, or in coordinates,

$$\Omega^a = \xi^a + \mathcal{A}^a_\alpha(r)\dot{r}^\alpha.$$

We will write $l(r^\alpha, \dot{r}^\alpha, \Omega^a)$ for the local representation of l in these variables.

THEOREM 3.2 (Lagrangian reduction theorem). *A curve $(q^i, \dot{q}^i) \in TQ$ satisfies the Euler–Lagrange equations if and only if the induced curve in TQ/G with coordinates given in a local trivialization by $(r^\alpha, \dot{r}^\alpha, \Omega^a)$ satisfies the reduced* Euler–Lagrange equations:

$$\frac{d}{dt}\frac{\partial l}{\partial \dot{r}^\alpha} - \frac{\partial l}{\partial r^\alpha} = \frac{\partial l}{\partial \Omega^a}\left(-\mathcal{B}^a_{\alpha\beta}\dot{r}^\beta + \mathcal{E}^a_{\alpha d}\Omega^d\right), \tag{3.1}$$

$$\frac{d}{dt}\frac{\partial l}{\partial \Omega^b} = \frac{\partial l}{\partial \Omega^a}(-\mathcal{E}^a_{\alpha b}\dot{r}^\alpha + C^a_{db}\Omega^d), \tag{3.2}$$

where

$$\mathcal{B}^b_{\alpha\beta} = \frac{\partial \mathcal{A}^b_\alpha}{\partial r^\beta} - \frac{\partial \mathcal{A}^b_\beta}{\partial r^\alpha} - C^b_{ac}\mathcal{A}^a_\beta\mathcal{A}^c_\alpha$$

are the coordinates of the curvature $\mathcal{B}$ of $\mathcal{A}$ and $\mathcal{E}^a_{\alpha d} = C^a_{bd}\mathcal{A}^b_\alpha$.

The first of these equations is similar to the Lagrange–d'Alembert equations for a nonholonomic system written in terms of the constrained Lagrangian, and the second is similar to the momentum equation. It is useful to note that the first set of equations results from Hamilton's principle by restricting the variations to be horizontal relative to the given connection.

If one uses the variables $(r^\alpha, \dot{r}^\alpha, p_a)$, where p is the body angular momentum, so that $p = \mathbb{I}_{\text{loc}}(r)\Omega = \partial l/\partial\Omega$, then the equations become (using the same letter l for the reduced Lagrangian, an admitted abuse of notation)

$$\frac{d}{dt}\frac{\partial l}{\partial \dot{r}^\alpha} - \frac{\partial l}{\partial r^\alpha} = p_a\left(-\mathcal{B}^a_{\alpha\beta}\dot{r}^\beta + \mathcal{E}^a_{\alpha d}I^{de}p_e\right) - p_d\frac{\partial I^{de}}{\partial r^\alpha}p_e, \tag{3.3}$$

$$\frac{d}{dt}p_b = p_a(-\mathcal{E}^a_{\alpha b}\dot{r}^\alpha + C^a_{db}I^{de}p_e), \tag{3.4}$$

where I^{de} denotes the inverse of the matrix I_{ab}.

Connections are also useful in control problems with feedback. For example, Bloch, Krishnaprasad, Marsden, and Sánchez de Alvarez [1992] found a feedback control that stabilizes rigid body dynamics about its middle axis using an internal rotor. This feedback controlled system can be described in terms of connections (Marsden and Sánchez de Alvarez [1996]); a shift in velocity (change of connection) turns the free Euler–Poincaré equations into the feedback-controlled Euler–Poincaré equations.

3.2. Reduced Lagrangian optimization for holonomic systems. Let us assume for the moment that we are dealing with a holonomic system on a trivial bundle and that the momentum map vanishes. Since we would like to use the method of Lagrange multipliers to relax the constraints, we define a new Lagrangian by $\mathcal{L}$:

$$\mathcal{L} = C(\dot{r}) + \langle\lambda(t), \xi + \mathcal{A}_{\text{loc}}(r)\dot{r}\rangle \tag{3.5}$$

for some $\lambda(t) \in \mathfrak{g}^*$, where $\xi = g^{-1}\dot{g} \in \mathfrak{g}$. Clearly $\mathcal{L}$ is G invariant and induces a function l on $(TQ/G) \times \mathfrak{g}^*$, where

$$l = C(\dot{r}) + \langle\lambda(t), \xi + \mathcal{A}_{\text{loc}}(r)\dot{r}\rangle \tag{3.6}$$

THEOREM 3.3 (reduced Lagrangian optimization for holonomic systems). *Assume that $q(t) = (r(t), g(t))$ is a (regular) optimal trajectory for the above optimal control problem; then there exists a $\lambda(t) \in \mathfrak{g}^*$ such that the reduced curve $(r(t), \dot{r}(t), \xi(t)) \in TQ/G$ with coordinates given by $(r^\alpha, \dot{r}^\alpha, \xi^a)$ satisfies the constraints $\xi = -\mathcal{A}_{\text{loc}}(r)\dot{r}$, as well as the reduced Euler–Lagrange equations*

$$\frac{d}{dt}\frac{\partial l}{\partial \dot{r}^\alpha} - \frac{\partial l}{\partial r^\alpha} = 0, \tag{3.7}$$

$$\frac{d}{dt}\frac{\partial l}{\partial \xi^b} = \frac{\partial l}{\partial \xi^a}C^a_{db}\xi^d, \tag{3.8}$$

where $l = C(\dot{r}) + \langle\lambda(t), \xi + \mathcal{A}_{\text{loc}}(r)\dot{r}\rangle$.

Proof. If $(r(t), g(t))$ is a (regular) optimal trajectory, then by the method of Lagrange multipliers, it solves the following variational problem:

$$\delta\int_0^1 \mathcal{L}dt = \delta\int_0^1 (C(\dot{r}) + \langle\lambda(t), \xi + \mathcal{A}_{\text{loc}}(r)\dot{r}\rangle)dt = 0$$

for some $\lambda(t) \in \mathfrak{g}^*$.

Since $B\times G \to B$ is trivial, we can put a trivial connection on this bundle and use it to split the variations into the horizontal and vertical parts. Then by the Lagrangian reduction method recalled above, the reduced curve $(r(t), \dot{r}(t), \xi(t)) \in TQ/G$ with coordinates given by $(r^\alpha, \dot{r}^\alpha, \xi^a)$ satisfies the reduced Euler–Lagrange equations stated above. (When using a trivial connection, the coefficients of $\mathcal{A}$ and $\mathcal{B}$ vanish and the reduced Euler–Lagrange equations are called Hamel's equations). □

Now we are ready to generalize one of the results in Krishnaprasad, Yang, and Dayawansa [1991]. Define the components $\mathcal{A}^a_\alpha$ of the mechanical connection by $\mathcal{A}_{\rm loc}(r)\dot{r} = \mathcal{A}^a_\alpha \dot{r}^\alpha e_a$, where $\{e_a\}$ is the basis of $\mathfrak{g}$ and $\{e^a\}$ is its dual basis. Here α runs from 1 to $n-k$ and a runs from 1 to k, where $n-k$ is the dimension of the base space B and k is the dimension of the Lie algebra $\mathfrak{g}$. The result deals with the following problem.

ISOHOLONOMIC PROBLEM FOR TRIVIAL BUNDLES. *Minimize* $\int_0^1 C(\dot{r})\,dt$, *subject to* $\dot{r} = u, \dot{g} = -g\mathcal{A}_{\rm loc}u = -g\mathcal{A}^a_\alpha(r)u^\alpha e_a$, *for given boundary conditions*

$$(r(0), g(0)) = (\mathbf{0}, g_0), \qquad (r(1), g(1)) = (\mathbf{0}, g_1).$$

COROLLARY 3.4. *Let the cost function C be quadratic in u, say,*

$$C = \sum_1^{n-k} c_\alpha (u^\alpha)^2.$$

If $(r(t), g(t))$ is a (regular) optimal trajectory with the control $\bar{u}(t)$ for the isoholonomic problem, then there exist $\rho(t) \in T^*B$ and $\lambda(t) \in \mathfrak{g}^*$ satisfying $\dot{r}^\alpha = \bar{u}^\alpha$, $\xi^a = -\mathcal{A}^a_\alpha(x)\bar{u}^\alpha$, and the following ordinary differential equations:

$$\dot{\rho}_\beta = \lambda_a \frac{\partial \mathcal{A}^a_\alpha}{\partial r^\beta}\bar{u}^\alpha,$$
$$\dot{\lambda}_b = -C^a_{db}\lambda_a \mathcal{A}^d_\alpha \bar{u}^\alpha,$$

where

$$\bar{u}_\beta = \frac{1}{2c_\beta}(\rho_\beta - \lambda_a \mathcal{A}^a_\beta)$$

with boundary conditions $r(0) = \mathbf{0}, g(0) = g_0, r(1) = \mathbf{0}, g(1) = g_1$.

Proof. According to Theorem 3.3, there exists some $\lambda(t) \in \mathfrak{g}^*$ such that the reduced curve $(r(t), \dot{r}(t), \xi(t))$ satisfies the reduced Euler–Lagrange equations for

$$l = c_\alpha(\dot{r}^\alpha)^2 + \langle \lambda_a e^a, (\xi^a + \mathcal{A}^a_\alpha \dot{r}^\alpha)e_a\rangle = c_\alpha(\dot{r}^\alpha)^2 + \lambda_a(\xi^a + \mathcal{A}^a_\alpha \dot{r}^\alpha).$$

After some computations, we find

$$\frac{\partial l}{\partial \dot{r}^\beta} = 2c_\beta \dot{r}^\beta + \lambda_a \mathcal{A}^a_\beta,$$
$$\frac{\partial l}{\partial r^\beta} = \lambda_a \frac{\partial \mathcal{A}^a_\alpha}{\partial r^\beta}\dot{r}^\alpha,$$
$$\frac{\partial l}{\partial \xi^b} = \lambda_b.$$

Now let

$$\rho_\beta = \frac{\partial l}{\partial \dot{r}^\beta} = 2c_\beta \dot{r}^\beta + \lambda_a \mathcal{A}^a_\beta$$

and solve for $\dot{r}$ to give

$$\dot{r}_\beta = \frac{1}{2c_\beta}(\rho_\beta - \lambda_a \mathcal{A}^a_\beta).$$

Moreover, the reduced Euler–Lagrange equations (3.7) and (3.8) give

$$\dot{\rho}_\beta = \frac{d}{dt}\frac{\partial l}{\partial \dot{r}^\beta} = \frac{\partial l}{\partial r^\beta} = \lambda_a \frac{\partial \mathcal{A}^a_\alpha}{\partial r^\beta}\dot{r}^\alpha,$$
$$\dot{\lambda}_b = \frac{d}{dt}\frac{\partial l}{\partial \xi^b} = \frac{\partial l}{\partial \xi^a} C^a_{db}\xi^d = C^a_{db}\lambda_a \xi^d.$$

After substituting

$$\dot{r}^\alpha = \bar{u}^\alpha \qquad \xi^d = -\mathcal{A}^d_\alpha \bar{u}^\alpha,$$

we get the desired equations. □

Remarks.

(1) This corollary generalizes the result of Krishnaprasad, Yang, and Dayawansa [1991] for the trivial principal bundle $(\mathbb{R}\times\mathbb{R}\times \mathrm{SO}(3), \mathbb{R}\times\mathbb{R}, \pi, \mathrm{SO}(3))$ (see Theorem 3.3 and Remark 3.2 in Krishnaprasad, Yang, and Dayawansa [1991]).

(2) The reduced equations of motion for ρ_β and λ_b can be written in intrinsic form as a special case of Wong's equations in r_β and λ_b (see the following subsection).

3.3. Optimal control of a holonomic system on a principal bundle. While the above method seems to work only for the case where the principle bundle is trivial, it can be easily generalized to an arbitrary principle bundle. In fact, the proof of the Lagrangian reduction theorem stated above provides all the necessary techniques. Recall that Marsden and Scheurle [1993b] arrived at the general reduced Euler–Lagrange equations in two steps:

(1) One first gets the Hamel equations in a local bundle trivialization:

$$\frac{d}{dt}\frac{\partial l}{\partial \dot{r}^\alpha} - \frac{\partial l}{\partial r^\alpha} = 0,$$

$$\frac{d}{dt}\frac{\partial l}{\partial \xi^b} = \frac{\partial l}{\partial \xi^a} C^a_{db}\xi^d.$$

(2) One introduces an arbitrary principal connection $\mathcal{A}$ (which is not necessarily the mechanical connection) to split the original variational principle intrinsically and globally relative to horizontal and vertical parts of the variation δq and derived the general form from the above form by means of a velocity shift replacing ξ by the vertical part relative to this connection:

$$\Omega^a = \mathcal{A}^a_\alpha \dot{r}^\alpha + \xi^a.$$

Here $\mathcal{A}^a_\alpha$ are the local coordinates of the connection $\mathcal{A}$. The resulting reduced Euler–Lagrange equations are then as given earlier.

Now we are ready to state a general theorem for the constrained variational problem on a principal bundle. This problem is as follows.

ISOHOLONOMIC PROBLEM FOR GENERAL BUNDLES (THE FALLING CAT PROBLEM). *Among all curves $q(t)$ such that $q(0) = q_0, q(1) = q_1$, and $\dot{q}(t) \in \mathrm{hor}_{q(t)}$*

(*horizontal with respect to the mechanical connection* $\mathcal{A}^{\text{mec}}$), *find the optimal curves* $q(t)$ *such that* $\int_0^1 C(\dot{r})dt$ *is minimized, where* $r = \pi(q)$.

Observe that although this problem is set up using the mechanical connection $\mathcal{A}^{\text{mec}}$, when applying the Lagrangian reduction theorem, one may use an arbitrary connection $\mathcal{A}$ to split the variational principle. This observation is used in the proof of the following result.

THEOREM 3.5. *If* $q(t)$ *is a* (*regular*) *optimal trajectory for the isoholonomic problem for general bundles, then there exists a* $\lambda(t) \in \mathfrak{g}^*$ *such that the reduced curve in* TQ/G *with coordinates given in a local trivialization by* $(r^\alpha, \dot{r}^\alpha, \Omega^\alpha)$ *satisfies the constraints* $\xi^a = -(\mathcal{A}^{\text{mec}})^a_\alpha \dot{r}^\alpha$ *as well as the reduced Euler–Lagrange equations* (3.1) *and* (3.2), *where*

$$l = C(\dot{r}) + \langle \lambda(t), \xi + \mathcal{A}_{\text{loc}}(r)\dot{r} \rangle$$

and

$$\Omega^a = \mathcal{A}^a_\alpha \dot{r}^\alpha + \xi^a.$$

Proof. The proof proceeds as in the proof in Marsden and Scheurle [1993b] in the present context. The needed modifications of what we have done before are minor and so are omitted. □

COROLLARY 3.6. *In Theorem* 3.5, *if we use the mechanical connection* $\mathcal{A}^{\text{mec}}$ *to split the variational principle, then the reduced Euler–Lagrange equations coincide with Wong's equations* (*see Montgomery* [1984, 1993] *and references therein*):

$$\begin{aligned} \dot{p}_\alpha &= -\lambda_a \mathcal{B}^a_{\alpha\beta} \dot{r}^\beta - \frac{1}{2} \frac{\partial g^{\beta\gamma}}{\partial r^\alpha} p_\beta p_\gamma, \\ \dot{\lambda}_b &= -\lambda_a C^a_{db} \mathcal{A}^d_\alpha \dot{r}^\alpha, \end{aligned}$$

where $g_{\alpha\beta}$ *is the local representation of the metric on the base space* Q/G, *that is,*

$$C(\dot{r}) = \frac{1}{2} g_{\alpha\beta} \dot{r}^\alpha \dot{r}^\beta,$$

$g^{\beta\gamma}$ *is the inverse of the matrix* $g_{\alpha\beta}$, *and* p_α *is defined by*

$$p_\alpha = \frac{\partial C}{\partial \dot{r}^\alpha} = g_{\alpha\beta} \dot{r}^\beta,$$

and where we write the components of $\mathcal{A}^{\text{mec}}$ *simply as* $\mathcal{A}^b_\alpha$ *and similarly for its curvature.*

Proof. Apply Theorem 3.5 to the function l, where

$$\begin{aligned} l &= C(\dot{r}) + \langle \lambda(t), \xi + A_{\text{loc}}(r)\dot{r} \rangle \\ &= C(\dot{r}) + \langle \lambda(t), \Omega \rangle \\ &= C(\dot{r}^\alpha) + \lambda_a \Omega^a. \end{aligned}$$

Clearly,

$$\begin{aligned} \frac{\partial l}{\partial \dot{r}^\alpha} &= \frac{\partial C}{\partial \dot{r}^\alpha} = g_{\alpha\beta} \dot{r}^\beta, \\ \frac{\partial l}{\partial r^\alpha} &= \frac{\partial C}{\partial r^\alpha} = \frac{1}{2} \frac{\partial g^{\beta\gamma}}{\partial r^\alpha} \dot{r}^\beta \dot{r}^\gamma, \\ \frac{\partial l}{\partial \Omega^a} &= \lambda_a. \end{aligned}$$

Since $\xi^a = -\mathcal{A}^a_\alpha \dot{r}^\alpha$ (the constraints) and $\Omega^a = \mathcal{A}^a_\alpha \dot{r}^\alpha + \xi^a$, we have $\Omega^a = 0$, and the reduced Euler–Lagrange equations become

$$\begin{aligned}\frac{d}{dt}\frac{\partial C}{\partial \dot{r}^\alpha} - \frac{\partial C}{\partial r^\alpha} &= -\lambda_a(\mathcal{B}^a_{\alpha\beta}\dot{r}^\beta),\\ \frac{d}{dt}\lambda_b &= -\lambda_a(\mathcal{E}^a_{\alpha b}\dot{r}^\alpha) = -\lambda_a C^a_{db}\mathcal{A}^d_\alpha \dot{r}^\alpha.\end{aligned}$$

But

$$\begin{aligned}\frac{d}{dt}\frac{\partial C}{\partial \dot{r}^\alpha} - \frac{\partial C}{\partial r^\alpha} &= \dot{p}_\alpha - \frac{1}{2}\frac{\partial g_{\beta\gamma}}{\partial r^\alpha}\dot{r}^\beta \dot{r}^\gamma\\ &= \dot{p}_\alpha + \frac{1}{2}\frac{\partial g^{\kappa\sigma}}{\partial r^\alpha} g_{\kappa\beta} g_{\sigma\gamma}\dot{r}^\beta \dot{r}^\gamma\\ &= \dot{p}_\alpha + \frac{1}{2}\frac{\partial g^{\kappa\sigma}}{\partial r^\alpha} p_\kappa p_\sigma\\ &= \dot{p}_\alpha + \frac{1}{2}\frac{\partial g^{\beta\gamma}}{\partial r^\alpha} p_\beta p_\gamma,\end{aligned}$$

and so we have the desired equations. □

Remark. Recall that in Corollary 3.4, we have the reduced equations

$$\begin{aligned}\dot{\rho}_\beta &= \lambda_a \frac{\partial \mathcal{A}^a_\alpha}{\partial r^\beta}\dot{r}^\alpha,\\ \dot{\lambda}_b &= -C^a_{db}\lambda_a \mathcal{A}^d_\alpha \dot{r}^\alpha.\end{aligned}$$

But $\rho_\beta = \lambda_a \mathcal{A}^a_\alpha + 2c_\beta \dot{r}^\beta$ and hence

$$\dot{\rho}_\beta = 2c_\beta \ddot{r}^\beta + \dot{\lambda}_a \mathcal{A}^a_\beta + \lambda_a \frac{\partial \mathcal{A}^a_\beta}{\partial r^\alpha}\dot{r}^\alpha = \lambda_a \frac{\partial \mathcal{A}^a_\alpha}{\partial r^\beta}\dot{r}^\alpha.$$

Therefore,

$$\begin{aligned}2c_\beta \ddot{r}^\beta &= \lambda_a \frac{\partial \mathcal{A}^a_\alpha}{\partial r^\beta}\dot{r}^\alpha - \lambda_a \frac{\partial \mathcal{A}^a_\beta}{\partial r^\alpha}\dot{r}^\alpha - (-C^a_{db}\lambda_a \mathcal{A}^d_\alpha \dot{r}^\alpha)\mathcal{A}^b_\beta\\ &= \lambda_a\left(\frac{\partial \mathcal{A}^a_\alpha}{\partial r^\beta} - \frac{\partial \mathcal{A}^a_\beta}{\partial r^\alpha} - C^a_{bd}\mathcal{A}^d_\alpha \mathcal{A}^b_\beta\right)\dot{r}^\alpha\\ &= -\lambda_a \mathcal{B}^a_{\beta\alpha}\dot{r}^\alpha.\end{aligned}$$

That is, the reduced equations in Corollary 3.4 (and those in Krishnaprasad, Yang, and Dayawansa [1991]) can be written intrinsically as Wong's equations after a change of variables. This should not surprise us, because Marsden and Scheurle derived the general reduced Euler–Lagrange equations from the Hamel equations using a suitable change of variables from local trivialization variables to those in which the Lie algebra variable is replaced by the vertical part of the bundle velocity.

4. Optimal control and Lagrangian reduction for nonholonomic systems. Now we are ready to use the method of Lagrange multipliers and Lagrangian reduction to find the necessary conditions for optimal trajectories of nonholonomic systems in the case of a trivial bundle.

4.1. The general theorem for optimization. In Bloch, Krishnaprasad, Marsden, and Murray [1996], the reconstruction process may be seen in a two-step fashion: given an initial condition and a path $r(t)$ in the base space, we first integrate the momentum equation to determine $p(t)$ for all time and then use $r(t)$ and $p(t)$ jointly to determine the motion $g(t)$ in the fiber. But in studying the optimal control problem, it is better to treat p as a set of independent variables and the momentum equation as an additional set of constraints. With this viewpoint, it is possible to write down the reduced equations of motion for the optimal trajectories.

Since we would like to use the method of Lagrange multipliers to relax the constraints, we define a new Lagrangian $\mathcal{L}$:

$$\begin{aligned}\mathcal{L} = C(\dot{r}) &+ \langle \lambda(t), \xi + \mathcal{A}(r)\dot{r} - \Gamma(r)p \rangle \\ &+ \left\langle \kappa(t), \dot{p} - \dot{r}^T H(r)\dot{r} - \dot{r}^T K(r)p - p^T D(r)p \right\rangle\end{aligned} \tag{4.1}$$

for some $\lambda(t) \in \mathfrak{g}^*$ and for some $\kappa(t) \in \mathbb{R}^m$, where m is the number of momentum functions p_b. For simplicity of notation we have written $\mathcal{A}$ for $\mathcal{A}^{\text{nh}}$. Clearly $\mathcal{L}$ is G invariant and induces a function on $(T(Q \times \mathbb{R}^m)/G) \times \mathfrak{g}^* \times \mathbb{R}^m$ which is also denoted $\mathcal{L}$.

We formulate the main problem to be studied as follows.

ISOHOLONOMIC PROBLEM FOR NONHOLONOMIC SYSTEMS. *Among all curves $q(t)$ such that $q(0) = q_0, q(1) = q_1$, $\dot{q}(t) \in \mathcal{D}_{q(t)}$ and in which $g^{-1}\dot{g} + \mathcal{A}(r)\dot{r} = \Gamma(r)p$ and the momentum equation, find the optimal curves $q(t)$ such that $\int_0^1 C(\dot{r})dt$ is minimized, where $r = \pi(q)$.*

Before we state the theorem and do some computations, we want to make sure that the readers understand the index convention used in this section.

(1) The first batch of indices is denoted $a, b, c, ...$ and range from 1 to k corresponding to the symmetry direction ($k = \dim \mathfrak{g}$).
(2) The second batch of indices will be denoted $i, j, k, ...$ and range from 1 to m corresponding to the symmetry direction along constraint space (m is the number of momentum functions).
(3) The indices $\alpha, \beta, ...$ on the shape variables r range from 1 to $n - k$ ($n - k = \dim (Q/G)$, i.e., the dimension of the shape space).

THEOREM 4.1 (reduced Lagrangian optimization for nonholonomic systems). *If $q(t) = (r(t), g(t))$ is a (regular) optimal trajectory for the above optimal control problem, then there exist a $\lambda(t) \in \mathfrak{g}^*$ and a $\kappa(t) \in \mathbb{R}^m$ such that the reduced curve $(r(t), \dot{r}(t), \xi(t)) \in TQ/G$ with coordinates $(r^\alpha, \dot{r}^\alpha, \xi^\alpha)$ satisfies the reduced Euler–Lagrange equations*

$$\begin{aligned}\frac{d}{dt}\frac{\partial \mathcal{L}}{\partial \dot{r}^\beta} - \frac{\partial \mathcal{L}}{\partial r^\beta} &= 0, \\ \frac{d}{dt}\frac{\partial \mathcal{L}}{\partial \xi^b} &= \frac{\partial \mathcal{L}}{\partial \xi^a} C^a_{db}\xi^d, \\ \frac{d}{dt}\frac{\partial \mathcal{L}}{\partial \dot{p}_j} - \frac{\partial \mathcal{L}}{\partial p_j} &= 0,\end{aligned}$$

as well as

$$\begin{aligned}\xi &= -\mathcal{A}(r)\dot{r} + \Gamma(r)p, \\ \dot{p} &= \dot{r}^T H(r)\dot{r} + \dot{r}^T K(r)p + p^T D(r)p.\end{aligned}$$

Here C^a_{db} are the structure coefficients of the Lie algebra $\mathfrak{g}$ and

$$\mathcal{L} = C(\dot r) + \langle \lambda(t), \xi + \mathcal{A}(r)\dot r - \Gamma(r)p\rangle + \left\langle \kappa(t), \dot p - \dot r^T H(r)\dot r - \dot r^T K(r)p - p^T D(r)p\right\rangle. \tag{4.2}$$

Proof. If $(r(t), g(t))$ is a (regular) optimal trajectory, then by the method of Lagrange multipliers it solves the following variational problem:

$$\delta \int_0^1 \mathcal{L} dt = 0$$

for some $\lambda(t) \in \mathfrak{g}^*$ and some $\kappa(t) \in \mathbb{R}^m$. Since the bundle is trivial, we can put a flat connection on this bundle and use it to split the variations into horizontal and vertical parts. Then by the Lagrange reduction theorem, the reduced curve $(r(t), \dot r(t), \xi(t)) \in TQ/G$ satisfies the reduced Euler–Lagrange equations stated above. ∎

4.2. The optimality conditions in coordinates. Now let us work out everything in detail in bundle coordinates. Since

$$\mathcal{L} = \frac{1}{2} C_\alpha (\dot r^\alpha)^2 + \lambda_a(\xi^a + \mathcal{A}^a_\alpha \dot r^\alpha - \Gamma^{ai} p_i) + \kappa^i(\dot p_i - H_{\alpha\gamma i}\dot r^\alpha \dot r^\gamma - K^l_{i\alpha}\dot r^\alpha p_l - D^{lk}_i p_l p_k), \tag{4.3}$$

we find after some computations that

$$\frac{\partial \mathcal{L}}{\partial \dot r^\beta} = C_\beta \dot r^\beta + \lambda_a \mathcal{A}^a_\beta - \kappa^i(2H_{\alpha\beta i}\dot r^\alpha + K^l_{i\beta} p_l),$$
$$\frac{\partial \mathcal{L}}{\partial r^\beta} = \lambda_a \left(\frac{\partial \mathcal{A}^a_\alpha}{\partial r^\beta}\dot r^\alpha - \frac{\partial \Gamma^{ai}}{\partial r^\beta} p_i\right) - \kappa^i \left(\frac{\partial H_{\alpha\gamma i}}{\partial r^\beta}\dot r^\alpha \dot r^\gamma + \frac{\partial K^l_{i\alpha}}{\partial r^\beta}\dot r^\alpha p_l + \frac{\partial D^{lk}_i}{\partial r^\beta} p_l p_k\right).$$

Also, we have

$$\frac{\partial \mathcal{L}}{\partial \xi^b} = \lambda_b,$$
$$\frac{\partial \mathcal{L}}{\partial \dot p_j} = \kappa^j,$$
$$\frac{\partial \mathcal{L}}{\partial p_j} = -\lambda_a \Gamma^{aj} - \kappa^i(K^j_{i\alpha}\dot r^\alpha + 2D^{lj}_i p_l).$$

By Theorem 4.1, we know that the reduced curve $(r(t), \dot r(t), \xi(t))$ must satisfy the following system of differential equations for the given boundary conditions $q(0) = (r_0, g_0), q(1) = (r_1, g_1)$:

$$\frac{d}{dt}[C_\beta \dot r^\beta + \lambda_a \mathcal{A}^a_\beta - \kappa^i(2H_{\alpha\beta i}\dot r^\alpha + K^l_{i\beta} p_l)] = \lambda_a \left(\frac{\partial \mathcal{A}^a_\alpha}{\partial r^\beta}\dot r^\alpha - \frac{\partial \Gamma^{ai}}{\partial r^\beta} p_i\right) - \kappa^i \left(\frac{\partial H_{\alpha\gamma i}}{\partial r^\beta}\dot r^\alpha \dot r^\gamma + \frac{\partial K^l_{i\alpha}}{\partial r^\beta}\dot r^\alpha p_l + \frac{\partial D^{lk}_i}{\partial r^\beta} p_l p_k\right)$$

and

$$\dot\kappa^j = -\lambda_a \Gamma^{aj} - \kappa^i(K^j_{i\alpha}\dot r^\alpha + 2D^{lj}_i p_l),$$
$$\dot\lambda_b = C^a_{db}\lambda_a \xi^d = C^a_{db}\lambda_a(-\mathcal{A}^d_\alpha \dot r^\alpha + \Gamma^{di} p_i),$$
$$\dot p_i = H_{\alpha\gamma i}\dot r^\alpha \dot r^\gamma + K^l_{i\alpha}\dot r^\alpha p_l + D^{lk}_i p_l p_k.$$

Remarks.

(1) The first set of equations can be simplified somewhat as follows:

$$\begin{aligned}\frac{d}{dt}&\left[C_\beta \dot{r}^\beta - \kappa^i(2H_{\alpha\beta i}\dot{r}^\alpha + K^l_{i\beta}p_l)\right]\\ &= \lambda_a \mathcal{B}^a_{\beta\alpha}\dot{r}^\alpha - \lambda_a\left(\frac{\partial \Gamma^{ai}}{\partial r^\beta} + C^a_{db}\mathcal{A}^b_\beta \Gamma^{di}\right)p_i\\ &\qquad -\kappa^i\left(\frac{\partial H_{\alpha\gamma i}}{\partial r^\beta}\dot{r}^\alpha\dot{r}^\gamma + \frac{\partial K^l_{i\alpha}}{\partial r^\beta}\dot{r}^\alpha p_l + \frac{\partial D^{lk}_i}{\partial r^\beta}p_l p_k\right),\end{aligned}$$

where $\mathcal{B}^a_{\beta\alpha}$ are the coordinates of the curvature $\mathcal{B}$ of the nonholonomic connection $\mathcal{A}$, which is used to set up the constrained variational problem. Clearly more work is needed to establish a better form of the first set of equations as well as the geometry behind them. However, for the snakeboard, the reduced equations of motion for the optimal trajectories turn out to be rather simple.

(2) In proving the above theorem, although variations with fixed endpoints for $r(t)$ can be used, we generally can only hold the initial endpoint fixed for the variations of $p(t)$ and leave their final endpoints free (which is called "free endpoint problem" in the language of calculus of variations). However, we will obtain the same system of differential equations (namely the reduced Euler–Lagrange equations) except the need to impose some kind of transversality condition at $t = 1$; e.g., in this case we need to have $\kappa(1) = 0$.

In the following section, we will apply the method of reduced Lagrangian optimization developed in this section to some examples, especially the snakeboard.

5. Examples.

5.1. Optimal control of a homogeneous ball on a rotating plate. Bloch, Krishnaprasad, Marsden, and Murray [1995] also study a well-known example, namely the model of a homogeneous ball on a rotating plate, and write down its equations of motion in a form that is suitable for the application of control theory. (For more information, also see Naimark and Fufaev [1972] and Yang [1992] for the affine case and Bloch and Crouch [1992], Brockett and Dai [1992], and Jurdjevic [1993] for the linear case.)

Fix coordinates in inertial space and let the plane rotate with constant angular velocity Ω about the z-axis. The configuration space of the sphere is $Q = \mathbb{R}^2 \times \mathrm{SO}(3)$, parameterized by $(x, y, g), g \in \mathrm{SO}(3)$, all measured with respect to the inertial frame. Let $\omega = (\omega_x, \omega_y, \omega_z)$ be the angular velocity vector of the sphere measured also with respect to the inertial frame, let m be the mass of the sphere and mk^2 its inertia about any axis, and let a be its radius.

The Lagrangian of the system is

$$L = \frac{1}{2}m(\dot{x}^2 + \dot{y}^2) + \frac{1}{2}mk^2({\omega_x}^2 + {\omega_y}^2 + {\omega_z}^2),$$

with the affine nonholonomic constraints

$$\begin{aligned}\dot{x} - a\omega_y &= -\Omega y,\\ \dot{y} + a\omega_x &= +\Omega x.\end{aligned}$$

Note that the Lagrangian here is a metric on Q which is bi-invariant on $\mathrm{SO}(3)$ as the ball is homogeneous. Note also that $\mathbb{R}^2 \times \mathrm{SO}(3)$ is a principal bundle over $\mathbb{R}^2$ with

respect to the right SO(3) action on Q given by

$$(x, y, g) \mapsto (x, y, gh)$$

for $h \in \mathrm{SO}(3)$. The action is on the *right* since the symmetry is a material symmetry.

After some computations, it can be shown (for details, see Bloch, Krishnaprasad, Marsden, and Murray [1996]) that the equations of motion are

$$\begin{aligned} \omega_x + \frac{1}{a}\dot{y} &= \frac{\Omega x}{a}, \\ \omega_y - \frac{1}{a}\dot{x} &= \frac{\Omega y}{a}, \\ \omega_z &= c, \end{aligned}$$

(where c is a constant), together with

$$\begin{aligned} \ddot{x} + \frac{k^2\Omega}{a^2 + k^2}\dot{y} &= 0, \\ \ddot{y} - \frac{k^2\Omega}{a^2 + k^2}\dot{x} &= 0. \end{aligned}$$

Notice that the first set of three equations has the form

$$\dot{g}g^{-1} = -\mathcal{A}_{\mathrm{loc}}(r)\dot{r} + \Gamma_{\mathrm{loc}}(r),$$

where

$$\mathcal{A}_{\mathrm{loc}} = \frac{1}{a}e_1 dy - \frac{1}{a}e_2 dx$$

and

$$\Gamma_{\mathrm{loc}} = \frac{\Omega}{a}xe_1 + \frac{\Omega}{a}ye_2 + ce_3.$$

Here, $r^1 = x, r^2 = y$, and e_1, e_2, e_3 is the standard basis of $so(3)_-$. Also, $\mathcal{A}_{\mathrm{loc}}$ is the expression of nonholonomic connection relative to the (global) trivialization, and Γ_{loc} is the expression of the affine piece of the constraints with respect to the same trivialization (see Bloch, Krishnaprasad, Marsden, and Murray [1995]).

Now we are ready to apply reduced Lagrangian optimization to find the optimal trajectories for a homogeneous ball. Clearly the homogeneous ball on a rotating plate is a simple nonholonomic mechanical system with symmetry as defined earlier which also has a trivial principal bundle structure (except that the constraint is affine which can be dealt with in the same way). Also, we can assume that we have full control over the motion of the center of the ball, i.e., over the shape variables. Now let the cost function be $C(\dot{r}) = \frac{1}{2}[(\dot{x})^2 + (\dot{y}^2)]$ and set $a = 1$ for simplicity; then we can use the method of Lagrange multipliers and Lagrangian reduction to find the necessary conditions for the optimal trajectories of the following optimal control problem.

PLATE-BALL PROBLEM. *Given two points $q_0, q_1 \in \mathbb{R}^2 \times \mathrm{SO}(3)$, find the optimal control curves $(x(t), y(t)) \in \mathbb{R}^2$ that steer the system from q_0 to q_1 and minimize*

$$\int_0^1 \frac{1}{2}[(\dot{x})^2 + (\dot{y})^2]dt\,,$$

subject to the constraints

$$\dot{g}g^{-1} = -\dot{y}e_1 + \dot{x}e_2 + ce_3 + \Omega x e_1 + \Omega y e_2,$$

where, again, e_a is the standard basis of $so(3)_-$.

Following the reduced Lagrangian optimization method developed in the preceding section, we define a new Lagrangian $\mathcal{L}$ by

$$\mathcal{L} = \frac{1}{2}[(\dot{x})^2 + (\dot{y})^2] + \lambda_a \xi^a + \lambda_1 \dot{y} - \lambda_2 \dot{x} - \lambda_3 c - \Omega\lambda_1 x - \Omega\lambda_2 y,$$

where $\lambda(t) \in so(3)^*_-$. (Note that we use the negative Lie–Poisson structure because the right action is used.)

By Theorem 4.1, we know that any reduced optimal curve

$$(x(t), y(t), \dot{x}(t), \dot{y}(t), \xi^a(t))$$

must satisfy the reduced Euler–Lagrangian equations. Simple computations show that

$$\begin{aligned}
\frac{\partial \mathcal{L}}{\partial \dot{x}} &= \dot{x} - \lambda_2 = \rho_1, \\
\frac{\partial \mathcal{L}}{\partial x} &= -\Omega\lambda_1, \\
\frac{\partial \mathcal{L}}{\partial \dot{y}} &= \dot{y} + \lambda_1 = \rho_2, \\
\frac{\partial \mathcal{L}}{\partial y} &= -\Omega\lambda_2, \\
\frac{\partial \mathcal{L}}{\partial \xi^b} &= \lambda_b.
\end{aligned}$$

Therefore,

$$\begin{aligned}
\dot{\rho}_1 &= -\Omega\lambda_1, \\
\dot{\rho}_2 &= -\Omega\lambda_2,
\end{aligned}$$

and

$$\dot{\lambda}_b = C^a_{db}\lambda_a \xi^d;$$

that is,

$$\begin{aligned}
\dot{\lambda}_1 &= \lambda_3\xi^2 - \lambda_2\xi^3 = \lambda_3(\rho_1 + \lambda_2 + \Omega y) - c\lambda_2, \\
\dot{\lambda}_2 &= -\lambda_3\xi^1 + \lambda_1\xi^3 = \lambda_3(\rho_2 - \lambda_1 - \Omega x) + c\lambda_1, \\
\dot{\lambda}_3 &= \lambda_2\xi^1 - \lambda_1\xi^2 = -(\lambda_1\rho_1 + \lambda_2\rho_2) + \Omega(\lambda_2 x - \lambda_1 y).
\end{aligned}$$

In the special case where $c = 0$ (no drift) and $\Omega = 0$ (no rotation) studied in Jurdjevic [1993], we have

$$\begin{aligned}
\dot{\rho}_1 &= 0, \\
\dot{\rho}_2 &= 0, \\
\dot{\lambda}_1 &= \lambda_3(\rho_1 + \lambda_2), \\
\dot{\lambda}_2 &= \lambda_3(\rho_2 - \lambda_1), \\
\dot{\lambda}_3 &= -(\lambda_1\rho_1 + \lambda_2\rho_2),
\end{aligned}$$

which gives the same result as in Jurdjevic [1993] obtained through the application of the Pontryagin maximum principle.

5.2. Optimal control of the snakeboard. The snakeboard is a modified version of a skateboard in which the front and back pairs of wheels are independently actuated. The extra degree of freedom enables riders to generate forward motion by twisting their bodies back and forth, while simultaneously moving the wheels with the proper phase relationship. For details, see Bloch, Krishnaprasad, Marsden, and Murray [1996] and the references listed there. Here we will include the computations shown in that paper for completeness and to make concrete the nonholonomic theory.

The snakeboard is modeled as a rigid body (the board) with two sets of independently actuated wheels, one on each end of the board. The human rider is modeled as a momentum wheel which sits in the middle of the board and is allowed to spin about the vertical axis. Spinning the momentum wheel causes a counter torque to be exerted on the board. The configuration of the board is given by the position and orientation of the board in the plane, the angle of the momentum wheel, and the angles of the back and front wheels. Thus the configuration space is $Q = \mathrm{SE}(2) \times S^1 \times S^1 \times S^1$. Let (x, y, θ) represent the position and orientation of the center of the board, ψ the angle of the momentum wheel relative to the board, and ϕ_1 and ϕ_2 the angles of the back and front wheels, also relative to the board. Take the distance between the center of the board and the wheels to be r.

The Lagrangian for the snakeboard consists only of kinetic energy terms and can be written as

$$L(q, \dot{q}) = \frac{1}{2} m(\dot{x}^2 + \dot{y}^2) + \frac{1}{2} J\dot{\theta}^2 + \frac{1}{2} J_0(\dot{\theta} + \dot{\psi})^2 + \frac{1}{2} J_1(\dot{\theta} + \dot{\phi}_1)^2 + \frac{1}{2} J_2(\dot{\theta} + \dot{\phi}_2)^2,$$

where m is the total mass of the board, J is the inertia of the board, J_0 is the inertia of the rotor, and J_i, $i = 1, 2$, is the inertia corresponding to ϕ_i. The Lagrangian is independent of the configuration of the board and hence it is invariant to all possible group actions.

The rolling of the front and rear wheels of the snakeboard is modeled using nonholonomic constraints which allow the wheels to spin about the vertical axis and roll in the direction that they are pointing. The wheels are not allowed to slide in the sideways direction. This gives constraint one forms

$$\begin{aligned} \omega_1(q) &= -\sin(\theta + \phi_1)dx + \cos(\theta + \phi_1)dy - r\cos\phi_1 d\theta, \\ \omega_2(q) &= -\sin(\theta + \phi_2)dx + \cos(\theta + \phi_2)dy + r\cos\phi_2 d\theta. \end{aligned}$$

These constraints are invariant under the SE(2) action given by

$$(x, y, \theta, \psi, \phi_1, \phi_2) \mapsto (x\cos\alpha - y\sin\alpha + a, x\sin\alpha + y\cos\alpha + b, \theta + \alpha, \psi, \phi_1, \phi_2),$$

where $(a, b, \alpha) \in \mathrm{SE}(2)$. The constraints determine the kinematic distribution $\mathcal{D}_q$:

$$\mathcal{D}_q = \mathrm{span}\left\{ \frac{\partial}{\partial\psi}, \frac{\partial}{\partial\phi_1}, \frac{\partial}{\partial\phi_2}, a\frac{\partial}{\partial x} + b\frac{\partial}{\partial y} + c\frac{\partial}{\partial\theta} \right\},$$

where a, b, and c are given by

$$\begin{aligned} a &= -r(\cos\phi_1\cos(\theta + \phi_2) + \cos\phi_2\cos(\theta + \phi_1)), \\ b &= -r(\cos\phi_1\sin(\theta + \phi_2) + \cos\phi_2\sin(\theta + \phi_1)), \\ c &= \sin(\phi_1 - \phi_2). \end{aligned}$$

The tangent space to the orbits of the SE(2) action is given by

$$T_q(\mathrm{Orb}(q)) = \mathrm{span}\left\{\frac{\partial}{\partial x}, \frac{\partial}{\partial y}, \frac{\partial}{\partial \theta}\right\}.$$

The intersection between the tangent space to the group orbits and the constraint distribution is thus given by

$$\mathcal{D}_q \cap T_q(\mathrm{Orb}(q)) = a\frac{\partial}{\partial x} + b\frac{\partial}{\partial y} + c\frac{\partial}{\partial \theta}.$$

The momentum can be constructed by choosing a section of $\mathcal{D} \cap T\mathrm{Orb}$ regarded as a bundle over Q. Since $\mathcal{D}_q \cap T_q\mathrm{Orb}(q)$ is one dimensional, the section can be chosen to be

$$\xi_Q^q = a\frac{\partial}{\partial x} + b\frac{\partial}{\partial y} + c\frac{\partial}{\partial \theta},$$

which is invariant under the action of SE(2) on Q. The corresponding Lie algebra element in $\mathfrak{se}(2)$, ξ^q, is

$$\xi^q = (a + yc)e_x + (b - xc)e_y + ce_\theta,$$

where e_x is the basis element of the Lie algebra corresponding to translations in the x direction (and whose corresponding infinitesimal generator is $\partial/\partial x$), etc. The nonholonomic momentum map is thus given by

$$\begin{aligned} p = J^{\mathrm{nh}}(\xi^q) &= \frac{\partial L}{\partial \dot{q}^i}(\xi_Q^q)^i \\ &= ma\dot{x} + mb\dot{y} + Jc\dot{\theta} + J_0c(\dot{\theta} + \dot{\psi}) + J_1c(\dot{\theta} + \dot{\phi}_1) + J_2c(\dot{\theta} + \dot{\phi}_2). \end{aligned}$$

In Bloch, Krishnaprasad, Marsden, and Murray [1996] a simplification is made which we shall also assume in this paper, namely, $\phi_1 = -\phi_2$, $J_1 = J_2$. The parameters are also chosen such that $J + J_0 + J_1 + J_2 = mr^2$ (which eliminates some terms in the derivation but does not affect the essential geometry of the problem). Setting $\phi = \phi_1 = -\phi_2$, the constraints plus the momentum are given by

$$\begin{aligned} 0 &= -\sin(\theta + \phi)\dot{x} + \cos(\theta + \phi)\dot{y} - r\cos\phi\dot{\theta}, \\ 0 &= -\sin(\theta - \phi)\dot{x} + \cos(\theta - \phi)\dot{y} + r\cos\phi\dot{\theta}, \\ p &= -2mr\cos^2(\phi)\cos(\theta)\dot{x} - 2mr\cos^2(\phi)\sin(\theta)\dot{y} \\ &\quad + mr^2\sin(2\phi)\dot{\theta} + J_0\sin(2\phi)\dot{\psi}. \end{aligned}$$

Adding, subtracting, and scaling these equations, we can write (away from $\phi = \pi/2$)

$$\begin{bmatrix} \cos(\theta)\dot{x} + \sin(\theta)\dot{y} \\ -\sin(\theta)\dot{x} + \cos(\theta)\dot{y} \\ \dot{\theta} \end{bmatrix} + \begin{bmatrix} -\dfrac{J_0}{2mr}\sin(2\phi)\dot{\psi} \\ 0 \\ \dfrac{J_0}{mr^2}\sin^2(\phi)\dot{\psi} \end{bmatrix} = \begin{bmatrix} \dfrac{-1}{2mr}p \\ 0 \\ \dfrac{\tan\phi}{2mr^2}p \end{bmatrix}. \tag{5.1}$$

These equations have the form

$$g^{-1}\dot{g} + \mathcal{A}_{\mathrm{loc}}(r)\dot{r} = \Gamma(r)p,$$

where

$$\mathcal{A}_{\text{loc}} = -\frac{J_0}{2mr}\sin(2\phi)e_x\,d\psi + \frac{J_0}{mr^2}\sin^2(\phi)e_\theta\,d\psi,$$
$$\Gamma(r) = \frac{-1}{2mr}e_x + \frac{1}{2mr^2}\tan(\phi)\,e_\theta.$$

These are precisely the terms which appear in the nonholonomic connection relative to the (global) trivialization (r, g). The momentum equation, which governs the evolution of p, is given by

$$\begin{aligned}\dot{p} &= \frac{\partial L}{\partial \dot{q}^i}\left[\frac{d}{dt}\xi^q\right]_Q^i \\ &= 4mr\cos(\theta)\cos(\phi)\sin(\phi)\dot{x}\dot{\phi} + 4mr\sin(\theta)\cos(\phi)\sin(\phi)\dot{y}\dot{\phi} \\ &\quad +2J_0\cos(2\phi)\dot{\phi}\dot{\psi} + 2mr^2\cos(2\phi)\dot{\theta}\dot{\phi} \\ &\quad -2mr\cos(\theta)\cos^2(\phi)\dot{y}\dot{\theta} + 2mr\sin(\theta)\cos^2(\phi)\dot{x}\dot{\theta}.\end{aligned}$$

Solving for the group velocities $\dot{x}, \dot{y}, \dot{\theta}$ from the equations which define the nonholonomic connection, the momentum equation can be rewritten as

$$\dot{p} = 2J_0\cos^2(\phi)\,\dot{\phi}\dot{\psi} - \tan(\phi)\,p\dot{\phi}.$$

This version of the momentum equation corresponds to the coordinate form in body representation, but it contains no terms which are quadratic in p due to the fact that $\mathfrak{g}^q$ is one dimensional.

These equations describe how paths in the base space, parameterized by $r \in S^1 \times S^1 \times S^1$ (in fact, the base space is $S^1 \times S^1$ if we assume $\phi_1 = -\phi_2$), are lifted to the fiber SE(2). The utility of these equations is that they greatly simplify the process of solving for the motion of the system given the base space trajectory.

Now we are ready to apply the method of reduced Lagrangian optimization to find the optimal trajectories for the snakeboard. Clearly the snakeboard is a simple nonholonomic mechanical system with symmetry as defined earlier and which also has a trivial principal bundle structure. Moreover, the control forces are only applied to the shape variables for which we have full control. Let the cost function be $C(\dot{r}) = \frac{1}{2}[(\dot{\psi})^2 + (\dot{\phi})^2]$ for simplicity. We can use the method of Lagrange multipliers and Lagrangian reduction to find the necessary conditions for the optimal trajectories of the following optimal control problem.

Optimal control problem for the snakeboard. *Given two points $q_0, q_1 \in \text{SE}(2) \times S^1 \times S^1$, find the optimal control curves $(\psi(t), \phi(t)) \in S^1 \times S^1$ that steer from q_0 to q_1 and minimize $\int_0^1 \frac{1}{2}((\dot{\psi})^2 + (\dot{\phi})^2)dt$, subject to the constraints*

$$g^{-1}\dot{g} + \mathcal{A}_{\text{loc}}(r)\dot{r} = \Gamma(r)p,$$
$$\dot{p} = 2J_0\cos^2(\phi)\dot{\phi}\dot{\psi} - \tan(\phi)p\dot{\phi},$$

where

$$\mathcal{A}_{\text{loc}} = -\frac{J_0}{2mr}\sin(2\phi)e_x\,d\psi + \frac{J_0}{mr^2}\sin^2(\phi)e_\theta\,d\psi,$$
$$\Gamma(r) = \frac{-1}{2mr}e_x + \frac{1}{2mr^2}\tan(\phi)\,e_\theta.$$

Following the general procedures in the previous section, we define a new $\mathcal{L}$ by

$$\begin{aligned}\mathcal{L} = &\frac{1}{2}((\dot\psi)^2 + (\dot\phi)^2) + \lambda_a\xi^a - \frac{J_0}{2mr}\lambda_1 \sin(2\phi)\dot\psi + \frac{J_0}{mr^2}\lambda_3 \sin^2(\phi)\dot\psi \\ &+ \frac{1}{2mr}\lambda_1 p - \frac{1}{2mr^2}\lambda_3 \tan(\phi)p + \kappa\dot p - 2J_0\kappa\cos^2(\phi)\dot\phi\dot\psi + \kappa\tan(\phi)p\dot\phi,\end{aligned}$$

where $\xi = g^{-1}\dot g \in \mathfrak{g}$, $\lambda(t) \in \mathfrak{g}^*$, and $\kappa(t) \in \mathbb{R}^1$ are Lagrange multipliers. Here ξ^a and λ_a are the components of ξ and λ in the standard basis of $\mathfrak{se}(2)$ and $\mathfrak{se}(2)^*$, respectively.

By Theorem 4.1, we know that the reduced optimal curves

$$(\psi(t), \phi(t), \dot\psi(t), \dot\phi(t), \xi^a(t))$$

must satisfy the reduced Euler–Lagrangian equations for $\mathcal{L}$. After some computations, we find

$$\begin{aligned}\frac{\partial\mathcal{L}}{\partial\dot\psi} &= \dot\psi - \frac{J_0}{2mr}\lambda_1\sin(2\phi) + \frac{J_0}{mr^2}\lambda_3\sin^2(\phi) - 2J_0\kappa\cos^2(\phi)\dot\phi,\\
\frac{\partial\mathcal{L}}{\partial\psi} &= 0,\\
\frac{\partial\mathcal{L}}{\partial\dot\phi} &= \dot\phi - 2J_0\kappa\cos^2(\phi)\dot\psi + \kappa\tan(\phi)p,\\
\frac{\partial\mathcal{L}}{\partial\phi} &= -\frac{J_0}{mr}\lambda_1\cos(2\phi)\dot\psi + \frac{J_0}{mr^2}\lambda_3\sin(2\phi)\dot\psi - \frac{1}{2mr^2}\lambda_3\sec^2(\phi)p\\
&\quad +2J_0\kappa\sin(2\phi)\dot\phi\dot\psi + \kappa\sec^2(\phi)p\dot\phi,\\
\frac{\partial\mathcal{L}}{\partial\dot p} &= \kappa,\\
\frac{\partial\mathcal{L}}{\partial p} &= \frac{1}{2mr}\lambda_1 - \frac{1}{2mr^2}\lambda_3\tan(\phi) + \kappa\tan(\phi)\dot\phi,\\
\frac{\partial\mathcal{L}}{\partial\xi^b} &= \lambda_b.\end{aligned}$$

Substitute the above calculations into the reduced Euler–Lagrangian equations and simplify, giving

$$\begin{aligned}&\ddot\psi - \frac{J_0}{2mr}\dot\lambda_1\sin(2\phi) - \frac{J_0}{mr}\lambda_1\cos(2\phi)\dot\phi + \frac{J_0}{mr^2}\lambda_3\sin(2\phi)\dot\phi\\
&\quad + \frac{J_0}{mr^2}\dot\lambda_3\sin^2(\phi) - 2J_0\dot\kappa\cos^2\phi\,\dot\phi + 2J_0\kappa\sin(2\phi)(\dot\phi)^2 - 2J_0\kappa\cos^2(\phi)\ddot\phi = 0,\\
&\ddot\phi - 2J_0\dot\kappa\cos^2(\phi)\dot\psi - 2J_0\kappa\cos^2(\phi)\ddot\psi + \dot\kappa\tan(\phi)p + \kappa\tan(\phi)\dot p\\
&\qquad\qquad = -\frac{J_0}{mr}\lambda_1\cos(2\phi)\dot\psi + \frac{J_0}{mr^2}\lambda_3\sin(2\phi)\dot\psi - \frac{1}{2mr^2}\lambda_3\sec^2(\phi)p.\end{aligned}$$

Also, we have

$$\begin{aligned}\dot\kappa &= \frac{1}{2mr}\lambda_1 - \frac{1}{2mr^2}\lambda_3\tan(\phi) + \kappa\tan(\phi)\dot\phi,\\
\dot\lambda_1 &= \lambda_2\xi^3 = \lambda_2\left(-\frac{J_0}{mr^2}\sin^2(\phi)\dot\psi + \frac{1}{2mr^2}\tan(\phi)p\right),\end{aligned}$$

$$\dot{\lambda}_2 = -\lambda_1 \xi^3 = -\lambda_1 \left(-\frac{J_0}{mr^2} \sin^2(\phi)\dot{\psi} + \frac{1}{2mr^2} \tan(\phi) p \right),$$
$$\dot{\lambda}_3 = -\lambda_2 \xi^1 = -\lambda_2 \left(\frac{J_0}{2mr} \sin(2\phi)\dot{\psi} - \frac{1}{2mr} p \right),$$
$$\dot{p} = 2J_0 \cos^2(\phi)\ \dot{\phi}\dot{\psi} - \tan(\phi)\ p\dot{\phi}.$$

After eliminating $\dot{\lambda}_1, \dot{\lambda}_3, \dot{\kappa}$, and $\dot{p}$ from the first set of two equations, we finally obtain

$$\ddot{\psi} - \frac{J_0}{2mr}\lambda_1(1 + 3\cos(2\phi))\dot{\phi} + \frac{3J_0}{2mr^2}\lambda_3 \sin(2\phi)\dot{\phi} + J_0\kappa \sin(2\phi)(\dot{\phi})^2 - 2J_0\kappa \cos^2(\phi)\ddot{\phi} = 0,$$
$$\ddot{\phi} - \frac{J_0}{mr}\lambda_1 \sin^2\phi\ \dot{\psi} + \frac{1}{2mr}\lambda_1 \tan(\phi) p + \frac{1}{2mr^2}\lambda_3 p - \frac{J_0}{2mr^2}\lambda_3 \sin(2\phi)\dot{\psi} - 2J_0\kappa \cos^2(\phi)\ddot{\psi} = 0.$$

5.3. Optimal control on a Lie group. Krishnaprasad [1993] considered the following optimal control problem on a finite-dimensional Lie group G which has been used to model various problems in several other papers (e.g., the plate-ball problem in Jurdjevic [1993] and the landing tower problem in Walsh, Montgomery, and Sastry [1994]). While it is possible to model this class of problems as a special case of the optimal control of nonholonomic system on a trivial principal bundle and apply reduced Lagrangian optimization, it may be useful to provide in this section a more direct proof that uses simpler machinery.

OPTIMAL CONTROL PROBLEM FOR A LIE GROUP. *Given a left-invariant control system on G, $\dot{g} = g \cdot \xi_u$, where $\xi_u = e_0 + \sum_{i=1}^m u^i(t) e_i$, find the optimal controls $u(\cdot)$ that steer from g_0 to g_1 and minimize $\int_0^1 L(u)dt$.*

Here $\{e_0, e_1, \ldots, e_m\}$ spans an $(m+1)$-dimensional subspace of the whole Lie algebra $\mathfrak{g}$ of G; $m+1 \le n = \dim(\mathfrak{g})$; $u(\cdot)$ is a vector-valued control function with $u^i(t) \in \mathbb{R}$; L is a cost function on $\mathbb{R}^m$, which is the space of values of controls; and $L(u) = \frac{1}{2}\sum_{i=1}^m I_i (u^i)^2$ with $I_i > 0$.

To apply the method of Lagrangian reduction, we recast the above optimal control problem as a constrained variational problem. For simplicity of exposition, we will deal with the vector space case first, where there is no e_0 term, and will take up the affine case later.

Let $\mathcal{C}$ be the m-dimensional subspace of $\mathfrak{g}$ spanned by $\{e_1, \ldots, e_m\}$. We make the following points:

(i) $\xi_u = \sum_{i=1}^m u^i(t) e_i$ lies in $\mathcal{C}$.

(ii) If we define $L_1 = L \circ \phi$, where $L = \frac{1}{2}\sum_{i=1}^m I_i (u^i)^2$ with $I_i > 0$ and $\phi = (e^1, \ldots, e^m)$ with $\{e^1, \ldots, e^m\}$ as the dual basis of $\{e_1, \ldots, e_m\}$, then $L_1 : \mathcal{C} \to \mathbb{R}$ is nothing but one half of the square of a metric on $\mathcal{C}$ which is intrinsically defined and does not depend on the basis chosen.

(iii) We can extend L_1 to be half of the square of a metric $\bar{L}$ on $\mathfrak{g}$ such that $\bar{L} = L_1$ on $\mathcal{C}$. As we will see, the necessary conditions for an optimal control do not depend on how the extension is done.

(iv) For the affine case, we will simply set $\xi_u - e_0 = \sum_{i=1}^m u^i(t) e_i$.

Now it should be clear that the original problem is equivalent to the following constrained variational problem.

CONSTRAINED VARIATIONAL PROBLEM FOR OPTIMAL CONTROL ON LIE GROUPS. *Given an m-dimensional subspace $\mathcal{C}$ of $\mathfrak{g}$, find the optimal control curves $\xi - e_0 \in \mathcal{C}$ such that $g(0) = g_0$, $g(1) = g_1$ and minimize $\int_0^1 \bar{L}(\xi - e_0)\, dt$.*

Since we want to use the method of Lagrange multipliers to relax the constraint on the variations, we define a new Lagrangian

$$\mathcal{L} = \bar{L}(\xi - e_0) + \lambda(t)(\xi - e_0) = \tilde{L}(\xi) + \tilde{\lambda}(t)(\xi), \tag{5.2}$$

where $\lambda(t)$ lies in the annihilator $\mathcal{C}^0$ of $\mathcal{C}$; furthermore, $\tau(\xi) = \xi - e_0$, $\tilde{L} = \bar{L} \circ \tau$, and $\tilde{\lambda} = \lambda \circ \tau$.

THEOREM 5.1 (optimization theorem for nonholonomic systems on Lie groups). *If $\bar{\xi}$ is a (regular) optimal control curve in $\mathcal{C} + e_0 = \{\xi \in \mathfrak{g} : \xi = \xi_c + e_0, \xi_c \in \mathcal{C}\}$, then there exists a $\lambda(t) \in \mathfrak{g}^*$ such that $\bar{\xi}$ satisfies the Euler–Poincaré equation*

$$\frac{d}{dt}\left(\frac{\delta \tilde{L}}{\delta \xi} + \lambda\right) = \mathrm{ad}^*_\xi \left(\frac{\delta \tilde{L}}{\delta \xi} + \lambda\right). \tag{5.3}$$

Proof. If $\bar{\xi}(t)$ is an optimal control curve in $\mathcal{C} + e_0$, then by the Lagrangian reduction method, $\bar{\xi}(t)$ is a solution of the following variational problem:

$$\delta \int_0^1 \mathcal{L}(\xi) dt = \delta \int_0^1 (\tilde{L}(\xi) + \tilde{\lambda}(\xi)) dt = 0$$

for some $\lambda \in \mathfrak{g}^*$, where the variations take the form $\delta\xi = \dot{\Omega} + [\xi, \Omega]$ with $\Omega = g^{-1} \cdot \delta g$ arbitrary except vanishing at the endpoints. Since

$$\begin{aligned}
0 &= \delta \int_0^1 \left(\tilde{L}(\xi) + \tilde{\lambda}(\xi)\right) dt \\
&= \int_0^1 \left(\frac{\delta \tilde{L}}{\delta \xi} \delta\xi + \lambda\,(\delta\xi)\right) dt \\
&= \int_0^1 \left(\frac{\delta \tilde{L}}{\delta \xi} + \lambda\right) \left(\dot{\Omega} + [\xi, \Omega]\right) dt \\
&= \int_0^1 \left(-\frac{d}{dt}\left(\frac{\delta \tilde{L}}{\delta \xi} + \lambda\right) + \mathrm{ad}^*_\xi \left(\frac{\delta \tilde{L}}{\delta \xi} + \lambda\right)\right) \Omega dt,
\end{aligned}$$

we conclude that $\bar{\xi}(t)$ satisfies

$$\frac{d}{dt}\left(\frac{\delta \tilde{L}}{\delta \xi} + \lambda\right) = \mathrm{ad}^*_\xi \left(\frac{\delta \tilde{L}}{\delta \xi} + \lambda\right). \qquad \square$$

COROLLARY 5.2. *Given a left-invariant control system on G, $\dot{g} = g \cdot \xi_u$, where*

$$\xi_u = e_0 + \sum_{i=1}^m u^i(t) e_i.$$

If $\bar{u}(\cdot)$ is an optimal control, then

$$\bar{u}^i(t) = \frac{\mu_i(t)}{I_i},$$

where $i = 1, \ldots, m$ and μ_i, $i = 1, \ldots, m$ is the solution of the following system of differential equations:

$$\dot{\mu}_i = C^k_{ji} \mu_k \xi^j_u,$$

where $i, j, k = 0, \ldots, n-1$ and where C^k_{ij} are the structure constants of $\mathfrak{g}$.

Proof. Extend $\{e_0, e_1, \ldots, e_m\}$ to a basis $\{e_0, \ldots, e_{n-1}\}$, and let $\{e^0, \ldots, e^{n-1}\}$ be its dual basis.

(i) For $i = 1, \ldots, m$ and $\xi_u \in e_0 + \mathcal{C}$, we have

$$\frac{\delta \tilde{L}}{\delta \xi_u^i} = \frac{\partial L}{\partial u^i} = I_i u^i$$

because $\tilde{L}(\xi_u) = L \circ \phi \circ \tau(\xi_u) = L(u)$ and $\xi_u^i = u^i$; furthermore,

$$\lambda_i = 0, \qquad i = 1, \ldots, m$$

because λ lies in the annihilator $\mathcal{C}^0$.

(ii) If we set

$$\mu_i = \frac{\delta \tilde{L}}{\delta \xi_u^i}, \ i = 1, \ldots, m,$$

and

$$\mu_i = \frac{\delta \tilde{L}}{\delta \xi_u^i} + \lambda_i, \ i = m+1, \ldots, n-1, 0,$$

and write out the Euler–Poincaré equation using the above coordinates, we will get the desired system of differential equations. □

Remarks.

(1) From the above computations we can see that the necessary conditions for an optimal control $\bar{u}(\cdot)$ depend only on L and have nothing to do with how the extension is done, because not only $u^i(t) = \mu_i(t)/I_i$ but also $\dot{\mu}_i = C_{ji}^k \mu_k \xi_u^j$ do not depend on $\bar{L}$.

(2) The necessary conditions given in Corollary 5.2 are the same as those in Krishnaprasad [1993]:

$$u^i = \frac{\mu_i}{I_i}, \qquad i = 1, \ldots, m,$$
$$\dot{\mu}_i = -\mu_k C_{ij}^k \frac{\delta h}{\delta \mu_j}, \qquad i, j, k = 0, \ldots, n-1,$$

where

$$h = \mu_0 + \frac{1}{2} \sum_{i=1}^{m} \frac{\mu_i^2}{I_i}.$$

This is because $C_{ji}^k = -C_{ij}^k$ and

$$\frac{\delta h}{\delta \mu_j} = \left\{ \begin{array}{ll} 1, & j = 0, \\ \dfrac{\mu_j}{I_j} = u^j, & j = 1, \ldots, m, \\ 0, & j = m+1, \ldots, n-1 \end{array} \right\} = \xi_u^j.$$

Conclusions. We have found a procedure based on reduced Lagrangian optimization that can be used to directly establish results on

(1) optimal control for left-invariant systems on Lie group with velocity constraints,

(2) optimal control for holonomic systems on principal bundles with the constraint of the vanishing of the momentum map,

(3) optimal control for nonholonomic systems on (trivial) principal bundles that may have a nontrivial evolution of its nonholonomic momentum.

In fact, the first two results can be seen as special cases of the last result even though we have derived each of them in a parallel way. Recall that in the nonholonomic case, we have

$$\begin{aligned}\mathcal{L} = C(\dot{r}) &+ \langle \lambda(t), \xi + \mathcal{A}(r)\dot{r} - \Gamma(r)p \rangle \\ &+ \left\langle \kappa(t), \dot{p} - \dot{r}^T H(r)\dot{r} - \dot{r}^T K(r)p - p^T D(r)p \right\rangle .\end{aligned} \tag{5.4}$$

In the driftless holonomic case, $\mathcal{D}_q = T_q Q$ for each $q \in Q$, the momentum is conserved and assumed to be zero, so the above Lagrangian $\mathcal{L}$ reduces to

$$\mathcal{L} = C(\dot{r}) + \langle \lambda(t), \xi + \mathcal{A}(r)\dot{r} \rangle ,$$

which is exactly the same Lagrangian used in the second case. As for a system on Lie group G with velocity constraint (say, $g^{-1}\dot{g} = \sum_{i=1}^{m} u^i e_i$ for simplicity), it can be seen as a system on (trivial) principal bundle $G \times \mathbb{R}^m$ whose (nonholonomic) connection is independent of the shape variable r; i.e.,

$$\xi^a = \mathcal{A}^a_\alpha \dot{r}^\alpha,$$

where $\mathcal{A}^a_\alpha = 1$ and $\dot{r}^\alpha = u^\alpha$.

Topics for future work.

(1) In the nonholonomic case, we have stated only the result for the case of a trivial principal bundle. While it is true that all examples known to us have a trivial bundle structure, it is of interest to generalize the reduced Lagrangian optimization theorem to the case of an arbitrary principal bundle. Also, we need to understand better the geometry underlying this procedure. We hope to address all of these issues in a follow-up paper.

(2) We need to construct algorithms that can effectively find approximate solutions to the system of differential equations that are obtained through reduced Lagrangian optimization. For example, finite element techniques appear to be appropriate and will be explored.

Acknowledgments. We thank Anthony Bloch, P. S. Krishnaprasad, Naomi Leonard, James Ostrowski, and Richard Montgomery for helpful comments on this paper.

REFERENCES

A. M. Bloch and P. Crouch (1992), *On the dynamics and control of nonholonomic systems on Riemannian manifolds*, in Proc. NOLCOS '92, Bordeaux, France, pp. 368–372.

A. M. Bloch and P. Crouch (1994), *Reduction of Euler Lagrange problems for constrained variational problems and relation with optimal control problems*, in Proc. 33rd CDC, IEEE, pp. 2584–2590.

A. M. Bloch and P. Crouch (1995), *Nonholonomic control systems on Riemannian manifolds*, SIAM J. Control Optim., 33, pp. 126–148.

A. M. Bloch, P. Crouch, and T. S. Ratiu (1994), *Sub-Riemannian optimal control problems*, Fields Inst. Commun., 3, pp. 35–48.

A. M. Bloch, P. S. Krishnaprasad, J. E. Marsden, and R. Murray (1996), *Nonholonomic mechanical systems with symmetry*, Arch. Rational Mech. Anal., to appear.

A. M. Bloch, P. S. Krishnaprasad, J. E. Marsden, and T. S. Ratiu (1996), *The Euler–Poincaré equations and double bracket dissipation*, Comm. Math. Phys., 175, pp. 1–42.

A. M. Bloch, P. S. Krishnaprasad, J. E. Marsden, and G. Sánchez de Alvarez (1992), *Stabilization of rigid body dynamics by internal and external torques*, Automatica J. IFAC, 28, pp. 745–756.

R. W. Brockett and L. Dai (1992), *Nonholonomic kinematics and the role of elliptic functions in constructive controllability*, in Nonholonomic Motion Planning, Z. Li and J. F. Canny, eds., Kluwer International Series in Engineering and Science, Kluwer, Amsterdam, pp. 1–22.

V. Jurdjevic (1993), *The geometry of the plate-ball problem*, Arch. Rational Mech. Anal., 124, pp. 305–328.

V. Jurdjevic (1996), *Optimal control problems on Lie groups: Crossroads between geometry and mechanics*, in Geometry and Feedback Control, Marcel Dekker, New York.

S. D. Kelly and R. M. Murray (1995), *Geometric phases and robotic locomotion*, J. Robotic Systems, to appear.

P. S. Krishnaprasad (1993), *Optimal Control and Poisson Reduction*, Institute for Systems Research Technical report 93-87, University of Maryland, College Park, MD, pp. 1–16.

P. S. Krishnaprasad, R. Yang, and W. Dayawansa (1991), *Control problems on principal bundles and nonholonomic mechanics*, in Proc. 30th CDC, IEEE, pp. 1133–1138.

J. E. Marsden (1992), *Lectures on Mechanics*, London Mathematical Society Lecture Note Series 174, Cambridge University Press.

J. E. Marsden and T. S. Ratiu (1994), *An Introduction to Mechanics and Symmetry*, Texts in Appl. Math. 17, Springer-Verlag, Berlin.

J. E. Marsden and J. Scheurle (1993a), *Lagrangian reduction and the double spherical pendulum*, Z. Angew. Math. Phys., 44, pp. 17–43.

J. E. Marsden and J. Scheurle (1993b), *The reduced Euler–Lagrange equations*, Fields Institute Comm., 1, pp. 139–164.

R. Montgomery (1984), *Canonical formulations of a particle in a Yang-Mills field*, Lett. Math. Phys., 8, pp. 59–67.

R. Montgomery (1990), *Isoholonomic problems and some applications*, Comm. Math. Phys., 128, pp. 565–592.

R. Montgomery (1991), *Optimal control of deformable bodies and its relation to gauge theory*, in The Geometry of Hamiltonian Systems, T. Ratiu, ed., Springer-Verlag, Berlin.

R. Montgomery (1993), *Gauge theory of the falling cat*, Fields Inst. Comm., 1, pp. 193–218.

Ju. I. Naimark and N. A. Fufaev (1972), *Dynamics of Nonholonimic Systems*, Translations of Mathematical Monographs 33, AMS, Providence, RI.

A. M. Vershik and V. Ya. Gershkovich (1994), *Nonholonomic dynamical systems, geometry of distributions and variational problems*, in Dynamical Systems VII, V. Arnold and S. P. Novikov, eds., Springer-Verlag, New York, pp. 1–81.

G. Walsh, R. Montgomery, and S. Sastry (1994), *Optimal Path Planning on Matrix Lie Groups*, preprint.

R. Yang (1992), *Nonholonomic Geometry, Mechanics and Control*, Institute for Systems Research Technical Report, 92-14, University of Maryland, College Park, MD.

SIAM J. CONTROL OPTIM.
Vol. 35, No. 3, pp. 930–952, May 1997

009

INVESTIGATION OF THE DEGENERACY PHENOMENON OF THE MAXIMUM PRINCIPLE FOR OPTIMAL CONTROL PROBLEMS WITH STATE CONSTRAINTS*

ARAM V. ARUTYUNOV† AND SERGEI M. ASEEV‡

Abstract. In this paper we study the degeneracy phenomenon in optimal control problems with state constraints. It is shown that this phenomenon occurs because of the incompleteness of the standard variants of Pontryagin's maximum principle for problems with state constraints. A new maximum principle containing additional information about the behavior of the Hamiltonian at the endtimes is developed. We also obtain some sufficient and necessary conditions for nondegeneracy and pointwise nontriviality of the maximum principle. The results obtained pertain to optimal control problems with systems described by differential inclusions and ordinary differential equations.

Key words. optimal control, state constraints, maximum principle, degeneracy phenomenon, controllability conditions

AMS subject classifications. 49K15, 49J24

PII. S036301299426996X

1. The degeneracy phenomenon. To illustrate the degeneracy phenomenon, let us consider the "simplest" optimal control problem with state constraints:

$$\dot{x} = f(x,t,u), \qquad u \in U; \tag{1.1}$$

$$x(t_1) = x_1; \tag{1.2}$$

$$g(x(t)) \le 0 \quad \forall\, t \in [t_1, t_2]; \tag{1.3}$$

$$J(u(\cdot)) = k_0(x(t_2)) \to \min. \tag{1.4}$$

Here $x \in R^n$, $U \subset R^k$, the time interval $I = [t_1, t_2]$ is fixed, x_1 is a given initial point, and the right endpoint $x(t_2)$ is free. We suppose that the vector function f and the scalar functions g, k_0 are smooth and $\frac{\partial g}{\partial x}(x) \neq 0$ for all x such that $g(x) = 0$. The set of admissible controls consists of all bounded measurable functions u such that $u(t) \in U$ almost everywhere (a.e.) on I.

The problem (1.1)–(1.4) satisfies the hypotheses of several variants of Pontryagin's maximum principle (see [1], [2], [3], [4], [5], [6], [7]). Let us apply the standard one.

We define the Hamilton–Pontryagin function $\mathcal{H}$ and the Hamiltonian H as follows:

$$\mathcal{H}(x,t,u,\psi) = \langle \psi, f(x,t,u) \rangle,$$

$$H(x,t,\psi) = \sup_{u \in U} \mathcal{H}(x,t,u,\psi).$$

Let the pair x_0, u_0 solve the problem (1.1)–(1.4). Then Pontryagin's well-known maximum principle for problems with state constraints, proven by Dubovitskii and

*Received by the editors June 20, 1994; accepted for publication (in revised form) April 1, 1996. This research was supported by the Russian Foundation for Basic Research grant 94-01-00476 and by International Science Foundation grant NBV000.

http://www.siam.org/journals/sicon/35-3/26996.html

†Department of Differential Equation and Functional Analysis, Russian Peoples Friendship University, Ordzonikidze str. 3, Moscow, Russia (arutunov@sa640.cs.msu.su).

‡Steklov Institute of Mathematics, Gubkina str. 8, 117966, Moscow, Russia (aseev@mian.su).

Milutin [1], [3], [7], asserts that there exists a number $\lambda_0 \geq 0$, a left-continuous function ψ of bounded variation and a nonnegative bounded regular Borel measure η on I such that following conditions hold[1]:

$$\psi(t) = -\lambda_0 \frac{\partial k_0}{\partial x}(x_0(t_2)) + \int_t^{t_2} \frac{\partial \mathcal{H}}{\partial x}(x_0(s), s, u_0(s), \psi(s))\, ds \tag{1.5}$$

$$- \int_t^{t_2} \frac{\partial g}{\partial x}(x_0(s))\, d\eta \quad \forall\, t \in [t_1, t_2],$$

$$H(x_0(t), t, \psi(t)) = - \int_t^{t_2} \frac{\partial \mathcal{H}}{\partial t}(x_0(s), s, u_0(s), \psi(s))\, ds \tag{1.6}$$

$$+ H(x_0(t_2), t_2, \psi(t_2)) \quad \forall\, t > t_1,$$

$$\mathcal{H}(x_0(t), t, u_0(t), \psi(t)) \overset{\text{a.e.}}{=} H(x_0(t), t, \psi(t)), \tag{1.7}$$

$$\operatorname{supp} \eta \subset \{t :\ g(x_0(t)) = 0\}, \tag{1.8}$$

$$\lambda_0 + \|\eta\| + \sup_{t_1 \leq t \leq t_2} \|\psi(t)\| \neq 0. \tag{1.9}$$

Here,

$$\|\eta\| = \sup_{\|x\|_{C(I)}=1} \int_{t_1}^{t_2} x(s)\, d\eta.$$

Suppose now that the initial endpoint x_1 belongs to the boundary of the state constraints (1.3), i.e., $g(x_1) = 0$. Then the maximum principle (1.5)–(1.9) degenerates. This means that it is satisfied by a collection of Lagrange multipliers λ_0, ψ, η such that $\lambda_0 = 0$ and $\psi(t) = 0\ \forall\, t \in (t_1, t_2)$. (We shall call such a collection of Lagrange multipliers trivial.) Indeed, one can take $\lambda_0 = 0$; $\eta = \delta_{t_1}$, the Dirac measure concentrated at the point t_1; $\psi(t_1) = -\frac{\partial g}{\partial x}(x_1)$, $\psi(t) = 0\ \forall\, t \in (t_1, t_2]$. Obviously, in this case any admissible trajectory satisfies the maximum principle and hence we get no useful information. Applying to the problem (1.1)–(1.4) any of the above-cited variants of the maximum principle, we get the same result.

Note that in the general case when an optimal control problem incorporates a Bolza-type functional, free time, and full endpoint constraints, the maximum principle may degenerate if one end of an optimal trajectory belongs to the boundary of the state constraints. Degenerate multipliers exist in any problem with one end fixed on the state boundary. Moreover, there exist problems in which the trivial collection of Lagrange multipliers is the only one satisfying the maximum principle.

The following example is due to Dubovitskii and Dubovitskii [8].

[1] Putting

$$\hat{\psi}(t) = \psi(t) - \int_{[t_1, t)} \frac{\partial g}{\partial x}(x_0(s))\, d\eta,$$

one can rewrite the conditions (1.5)–(1.9) in terms of the absolutely continuous function $\hat{\psi}$.

Example 1. Consider the following problem:

$$\begin{aligned}
&\dot{x}_1 = tu, \qquad |u| \le 1;\\
&\dot{x}_2 = u;\\
&x_1(0) = 0, \qquad x_2(0) = 0;\\
&x_1(t) \ge 0 \quad \forall\, t \in I = [0,1];\\
&J(u(\cdot)) = x_2(1) \to \min.
\end{aligned}$$

It is not difficult to show that the unique solution of this problem is $x_0(t) \equiv 0$ and $u_0(t) = 0$ a.e. on I. Let λ_0, ψ, η be a collection of Lagrange multipliers corresponding to x_0. Then due to the maximum condition (1.7) we have

$$|t\psi_1(t) + \psi_2(t)| \overset{\text{a.e.}}{=} tu_0(t)\psi_1(t) + u_0(t)\psi_2(t) \overset{\text{a.e.}}{=} 0.$$

From the adjoint equation (1.5) we have $\psi_2(t) \equiv \psi_2(0)\ \forall\, t \in I$. Let $\psi_2(0) \neq 0$. Then $\psi_1(t) \overset{\text{a.e.}}{=} -\frac{\psi_2(0)}{t}$ fails to be a function of bounded variation. Thus, $\psi_2(t) \equiv 0$ and $\psi_1(t) = 0\ \forall\, t \in (0,1]$. Further, due to (1.5) $\lambda_0 = 0$. So, there is a unique (up to a positive multiplier) collection of Lagrange multipliers corresponding to the optimal pair in this example and it is trivial.

In the present paper we investigate the degeneracy phenomenon. Some results in this direction have been obtained earlier in [8], [9], [10], [11], [12], [13], [14], [15], [16], [17], [18] (see section 6). Our contribution to the subject is the following. We show that the degeneracy phenomenon arises due to the incompleteness of the standard variants of the maximum principle for problems with state constraints and develop a new version containing additional information about the Hamiltonian (a new jump condition at the endtimes). This condition enables one to investigate the degeneracy phenomenon.

The paper is organized as follows. In section 2 we state our main result—a new version of the maximum principle (as announced in [19], [20]). It is important to underline that the Hamiltonian plays the main role in our investigation. That is why we state the result in Hamiltonian form. In addition, the consideration of the problem in intrinsic form makes the presentation more descriptive. For simplicity we consider the problem on a fixed time interval (the free time problem can be investigated analogously [17], [19]). Section 3 contains the technical proof of the main result. In section 4 under some natural controllability assumptions we present sufficient conditions for nondegeneracy and pointwise nontriviality. In the case in which the state constraint set is regular at the endpoints, these conditions of controllability turn out to be precisely the ones necessary for the maximum principle to be informative. In section 5 we consider the free time classical optimal control problem and formulate the results applicable to it. Furthermore, sufficient conditions for the existence of at least one nontrivial collection of Lagrange multipliers corresponding to an optimal trajectory are obtained. Section 6 contains bibliographical comments.

2. The main result. Consider the optimal control problem with state constraints in intrinsic form:

$$\dot{x} \in F(x,t); \tag{2.1}$$

$$x(t_1) \in C_1, \qquad x(t_2) \in C_2; \tag{2.2}$$

$$x(t) \in G \quad \forall\, t \in [t_1, t_2]; \tag{2.3}$$

$$J(x(\cdot)) = k_0(x(t_1), x(t_2)) \to \min. \tag{2.4}$$

Here $x \in R^n$; C_1, C_2, G are nonempty closed subsets of R^n; the time interval $I = [t_1, t_2]$ is fixed; $k_0 : R^{2n} \to R^1$ is a locally Lipschitz function; and the multivalued mapping F is supposed to be locally Lipschitz with nonempty convex compact values. We assume also that Clarke's tangent cone $T_G(x)$ has a nonempty interior at any point $x \in G$ [2]. The class of admissible trajectories consists of all Lipschitz continuous functions x satisfying conditions (2.1)–(2.3).

Denote by $\partial f(x)$ Clarke's generalized gradient [2] of the locally Lipschitz function f at the point x, by $H(x,t,\psi) = \max_{f \in F(x,t)} \langle f, \psi \rangle$ the Hamiltonian of the system (2.1), and by $N_M(x)$ Clarke's normal cone to the set M at the point x [2]. Denote by $\hat{N}_M(x) = \limsup_{x_i \xrightarrow{M} x} \xi_i(x_i) = \{\xi \in R^n : \xi = \lim \xi_i(x_i)$ for some sequences $\xi_i(x_i)$ and $x_i \to x$ in $M\}$ the cone of limiting normals to M at x, where $\xi_i(x_i)$ is a proximal normal to M at x_i [2], [5], [21], [22], and by

$$\hat{\partial} f(x) = \{\xi \in R^n : (\xi, -1) \in \hat{N}_{epif}(x, f(x))\}$$

the set of limiting subgradients of the locally Lipschitz function f at x [5], [21], [22].

Our main result is the following.

THEOREM 1 (maximum principle). *Let x_0 be a solution of the problem (2.1)–(2.4). Then there exists a number $\lambda_0 \geq 0$, an absolutely continuous function ψ, and a bounded regular Borel vector measure η on I such that following conditions hold:*

(a) *The absolute continuity condition: the function $h(t) = H(x_0(t), t, \psi(t) + \int_{t_1}^{t} d\eta)$ is absolutely continuous on I;*

(b) *the jump condition:*

$$H\left(x_0(t), t, \psi(t) + \int_{t_1}^{t} d\eta\right) = H\left(x_0(t), t, \psi(t) + \int_{t_1}^{t} d\eta - \eta(t)\right) \quad \forall\, t \in I;$$

(c) *the measure sign condition: the measure η is nonpositive on the set of all continuous functions y with values in $T_G(x_0(t))$; i.e.,*

$$\int_{t_1}^{t_2} y(t)\, d\eta \leq 0 \quad \forall\, y \in C(I) :\ y(t) \in T_G(x_0(t)) \quad \forall\, t \in I;$$

(d) *the Hamiltonian inclusion:*

$$(-\dot{\psi}(t), \dot{h}(t), \dot{x}_0(t)) \overset{\text{a.e.}}{\in} \partial H\left(x_0(t), t, \psi(t) + \int_{t_1}^{t} d\eta\right);$$

(e) *the transversality condition:*

$$\left(\psi(t_1), -\psi(t_2) - \int_{t_1}^{t_2} d\eta\right) \in \lambda_0 \hat{\partial} k_0(x_0(t_1), x_0(t_2)) + \hat{N}_{C_1 \cap G}(x_0(t_1)) \times \hat{N}_{C_2 \cap G}(x_0(t_2));$$

(f) *the nontriviality condition:*

$$\lambda_0 + \|\psi(t_1)\| + \|\eta\| \neq 0.$$

Here $\eta(t)$ is the atomic component of the measure η at the point t.

The main distinction of the theorem formulated above from the other variants of the maximum principle consists of condition (b) at the endtimes t_1 and t_2, which is

crucial to the nondegeneracy conditions we derive below. Conditions (a) and (b) are not completely independent. For any $t > t_1$, condition (b) says $h(t) = h(t-0)$, which follows from (a). When $t = t_1$, however, condition (b) places a restriction on the atom $\eta(t_1)$. So, we shall prove condition (a) and then condition (b) at the point t_1. The sense of these two conditions is to take into consideration the energy conservation law.

We should note that the nontraditional transversality condition (e), in which the right side contains normal cones to intersections of endpoint sets C_1 and C_2 with the state constraint set G, appears to be essential to the validity of the jump condition (b) at the endpoints. Example 2 below shows that Theorem 1 may fail if (e) is replaced by a traditional transversality condition [2], [4].

Further, using the Radon–Nikodým theorem [23], one can show that there exist a bounded nonnegative regular scalar Borel measure ν and a ν-integrable function ζ on I such that the measure η is absolutely continuous with respect to ν and

$$\frac{d\eta}{d\nu} = \zeta(t), \qquad \zeta(t) \in N_G(x_0(t)) \quad \nu \text{ — a.e. on } I. \tag{2.5}$$

The inclusions

$$\operatorname{supp}\eta \subset \{t \in I : \ x(t) \in \partial G\}; \qquad \eta(t) \in N_G(x_0(t)) \quad \forall\, t \in I$$

are obvious consequences of the measure sign condition. Here ∂G designates the boundary of the set G.

The following example illustrates the correlation of the new transversality condition (e) and the jump condition (b) at the endtimes t_1, t_2.

Example 2. Consider the problem

$$\begin{aligned}
&\dot{x} \in F(x) = \{y \in R^1 : \ |y| \le 1\};\\
&x(0) \in C_1 = \{y \in R^1 : \ y \le 0\}, \qquad x(1) \in C_2 = R^1;\\
&x(t) \ge 0 \quad \forall\, t \in I = [0,1];\\
&J(x(\cdot)) = x(0) + x(1) \to \min.
\end{aligned}$$

Obviously, $x_0(t) \equiv 0$ is an optimal trajectory. In this example $H(x,t,\psi) = |\psi|$, $G = \{y \in R^1 : \ y \ge 0\}$. Due to the Hamiltonian inclusion we have $|\psi(t) + \int_{t_1}^{t} d\eta| = 0$ a.e. on I. Hence, $\psi(t) \equiv \psi(t_1) = -\eta(t_1)$ and $\eta = \eta(t_1)\delta_{t_1} + \eta(t_2)\delta_{t_2}$.

Consider the transversality condition (e). It gives

$$(-\eta(t_1), -\eta(t_2)) \in \lambda_0(1,1) + (\xi_1, \xi_2),$$

where $\xi_1 \in N_{C_1 \cap G} \Rightarrow \xi_1 \in R^1$; $\xi_2 \in N_{C_2 \cap G} \Rightarrow \xi_2 \le 0$. Due to the measure sign condition $\eta(t_1) \le 0$ and $\eta(t_2) \le 0$. Thus, $\eta(t_1) = -\lambda_0 - \xi_1 \le 0$, $\eta(t_2) = -\lambda_0 - \xi_2 \le 0$. Putting $\lambda_0 = 1$, $\xi_1 = \xi_2 = -1$, we can satisfy the jump conditions at the endtimes t_1, t_2: $\eta(t_1) = 0$, $\eta(t_2) = 0$.

Consider now the standard transversality condition [2], [4]:

$$\left(\psi(t_1), -\psi(t_2) - \int_{t_1}^{t_2} d\eta\right) \in \lambda_0 \partial k_0(x(t_1), x(t_2)) + N_{C_1}(x(t_1)) \times N_{C_2}(x(t_2)).$$

It gives

$$(-\eta(t_1), -\eta(t_2)) \in \lambda_0(1,1) + (\xi_1, \xi_2),$$

where $\xi_1 \in N_{C_1}(0) \Rightarrow \xi_1 \geq 0$; $\xi_2 \in N_{C_2}(0) \Rightarrow \xi_2 = 0$. Hence, we can satisfy the jump conditions at the endtimes t_1, t_2 only by putting $\lambda_0 = \xi_1 = \xi_2 = 0$. But in this case $\lambda_0 = 0$, $\psi(t) \equiv 0$, $\eta = 0$, which contradicts the nontriviality condition (f). Thus, the jump condition (b) at the endtimes cannot be added to the standard maximum principles [1], [2], [3], [4], [5], [6], [7].

3. Proof of Theorem 1. Let x_0 be an optimal trajectory, $\rho(x) = \min_{y \in G} \|y - x\|$ and $\rho_i(x) = \int_{R^n} \rho(x+y) w_i(y)\, dy$. Here w_i is a smooth probabilistic density such that $\operatorname{supp} w_i \subset \{x : \|x\| \leq \frac{1}{2^i}\}$, $i = 1, 2, \ldots$.

Consider now a sequence of auxiliary problems:

$$\dot{x} \in F(x,t), \tag{3.1}$$

$$x(t_1) \in C_1 \cap G, \qquad x(t_2) \in C_2 \cap G, \tag{3.2}$$

$$\|x(t_1) - x_0(t_1)\| \leq 1, \tag{3.3}$$

$$J_i(x(\cdot)) = k_0(x(t_1), x(t_2)) + \int_{t_1}^{t_2} \varphi_i(x,t)\, dt \to \min. \tag{3.4}$$

Here $\varphi_i(x,t) = i\rho_i(x) + \|x - x_0(t)\|^2$, $i = 1, 2, \ldots$.

By virtue of Filippov's existence theorem, for any $i = 1, 2, \ldots$ there exists a solution x_i of auxiliary problem (3.1)–(3.4), and the sequence $\{x_i\}$ is relatively compact in $C(I)$. Furthermore, any limit point $x_* = \lim_{i \to \infty} x_i$ is a trajectory of the differential inclusion (3.1).

LEMMA. *Let $\{x_i\}$ be a sequence of solutions of* (3.1)–(3.4). *Then $x_i \rightrightarrows x_0$ on I, $\dot{x}_i \to \dot{x}_0$ weakly in $L_1(I)$, and $i\rho_i(x_i(t)) \to 0$ a.e. on I, $i \to \infty$.*

Proof. Let x_* be some limit point of $\{x_i\}$ and $x_i \to x_*$ in $C(I)$, $i \to \infty$. Then due to the optimality of x_i we have

$$J_i(x_i(\cdot)) \leq J_i(x_0(\cdot)) \leq k_0(x_0(t_1), x_0(t_2)) + \frac{i}{2^i}(t_2 - t_1). \tag{3.5}$$

Thus, $\int_{t_1}^{t_2} \rho_i(x_i(t))\, dt \to 0$ and $x_*(t) \in G\ \forall\, t \in I$. So the trajectory x_* satisfies the state constraints (2.3) and the endpoint constraints (2.2). Hence, due to the optimality of x_0 we have

$$k_0(x_0(t_1), x_0(t_2)) \leq k_0(x_*(t_1), x_*(t_2)).$$

Substituting this inequality into (3.5), we obtain

$$k_0(x_i(t_1), x_i(t_2)) + \int_{t_1}^{t_2} \varphi_i(x_i(t), t)\, dt \leq k_0(x_*(t_1), x_*(t_2)) + \frac{i}{2^i}(t_2 - t_1).$$

Passing to the limit in the last inequality, we get

$$\int_{t_1}^{t_2} i\rho_i(x_i(t))\, dt \to 0, \qquad \int_{t_1}^{t_2} \|x_*(t) - x_0(t)\|^2\, dt = 0.$$

Hence, $i\rho_i(x_i(t)) \to 0$ a.e. on I, $i \to \infty$ and $x_*(t) \equiv x_0(t)$. The weak convergence of $\{\dot{x}_i\}$ to $\dot{x}_0$ in $L_1(I)$ is a consequence of the uniform convergence of $\{x_i\}$ to x_0 and of the uniform boundness of $\{\|\dot{x}_i(t)\|\}$. □

Due to the lemma above, the inequality in endpoint constraint (3.3) holds strictly for all sufficiently large numbers i. Hence, it may be neglected in the transversality conditions of the maximum principle for the problem (3.1)–(3.4). Thus, due to the Hamiltonian necessary conditions of optimality for the problems without state constraints [2][2] with transversality conditions suggested by Mordukhovich [21] (see [5] for details) we obtain that there exist an absolutely continuous function ψ_i and a number $\lambda_i \geq 0$ such that

$$(-\dot{\psi}_i(t), \dot{h}_i(t), \dot{x}_i(t)) \overset{\text{a.e.}}{\in} \partial H(x_i(t), t, \psi_i(t)) \tag{3.6}$$

$$-\lambda_i \left(i\frac{\partial \rho_i(x_i(t))}{\partial x} + 2(x_i(t) - x_0(t)), -2\langle x_i(t) - x_0(t), \partial x_0(t)\rangle, 0 \right),$$

$$(\psi_i(t_1), -\psi_i(t_2)) \in \lambda_i \hat{\partial} k_0(x_i(t_1), x_i(t_2)) + \hat{N}_{C_1 \cap G}(x_i(t_1)) \times \hat{N}_{C_2 \cap G}(x_i(t_2)), \tag{3.7}$$

$$\lambda_i + \|\psi_i(t_1)\| \neq 0. \tag{3.8}$$

Here $h_i(t) = H(x_i(t), t, \psi_i(t)) - \lambda_i \varphi_i(x_i(t), t)$,

$$\langle x_i(t) - x_0(t), \partial x_0(t)\rangle = \bigcup_{\xi \in \partial x_0(t)} \langle x_i(t) - x_0(t), \xi\rangle.$$

Further, introducing the absolutely continuous function

$$\tilde{\psi}_i = \psi_i(t) - \int_{t_1}^{t} i\lambda_i \frac{\partial \rho_i}{\partial x}(x_i(s))\, ds,$$

the conditions (3.6)–(3.8) can be rewritten as

$$(-\dot{\tilde{\psi}}_i(t), \dot{h}_i(t), \dot{x}_i(t)) \overset{\text{a.e.}}{\in} \partial H\left(x_i(t), t, \tilde{\psi}_i(t) + i\lambda_i \int_{t_2}^{t} \frac{\partial \rho_i}{\partial x}(x_i(s))\, ds \right) \tag{3.9}$$

$$-\lambda_i (2(x_i(t) - x_0(t)), -2\langle x_i(t) - x_0(t), \partial x_0(t)\rangle, 0),$$

$$\left(\tilde{\psi}_i(t_1), -\tilde{\psi}_i(t_2) - \int_{t_1}^{t_2} i\lambda_i \frac{\partial \rho_i}{\partial x}(x_i(s))\, ds \right) \in \lambda_i \hat{\partial} k_0(x_i(t_1), x_i(t_2)) \tag{3.10}$$

$$+\hat{N}_{C_1 \cap G}(x_i(t_1)) \times \hat{N}_{C_2 \cap G}(x_i(t_2)),$$

$$\lambda_i + \|\tilde{\psi}_i(t_1)\| \neq 0. \tag{3.11}$$

Let us normalize the multipliers λ_i, $\tilde{\psi}_i$ as follows:

$$\lambda_i + \|\tilde{\psi}_i(t_1)\| + i\lambda_i \int_{t_1}^{t_2} \left\| \frac{\partial \rho_i}{\partial x}(x_i(s)) \right\| ds = 1. \tag{3.12}$$

[2]Using the smooth approximation procedure [17], [24], it is possible to derive our main result directly from Pontryagin's maximum principle [25].

Then, passing to a subsequence, we have $\lambda_i \to \lambda_0 \geq 0$, $\tilde{\psi}_i(t_1) \to \psi_1$, and by virtue of Helly's theorem [26] the sequence $\{i\lambda_i \frac{\partial \rho_i}{\partial x}(x_i(\cdot))\}$ converges weakly to a bounded regular Borel vector measure η on I.

Further, due to the Hamiltonian inclusion (3.9) and Proposition 3.2.4(e) in [2] we have

$$\|\dot{\tilde{\psi}}_i(t)\| \leq \kappa_1(\|\tilde{\psi}_i(t)\| + 1) \text{ a.e. on } I.$$

Here and in what follows $\kappa_1, \kappa_2, \ldots$ denote positive constants. Hence, due to the Bellman–Gronwall inequality and (3.12) we can suppose $\tilde{\psi}_i \rightrightarrows \psi$ on I and $\dot{\tilde{\psi}}_i \to \dot{\psi}$ weakly in $L_1(I)$, $i \to \infty$, where ψ is a Lipschitz continuous vector function. So, $\psi(t_1) = \psi_1$.

Now we prove the measure sign condition (c). For this we show that for any continuous vector function y on I such that $y(t) \in T_G(x_0(t))\ \forall\, t \in I$ the following inequality holds:

$$\langle \eta, y \rangle = \int_{t_1}^{t_2} y(t)\, d\eta \leq 0.$$

Using the assumption $\operatorname{int} T_G(x_0(t)) \neq \emptyset\ \forall\, t \in I$ and Michael's continuous selection theorem [27], it is not difficult to prove that there is a sequence of continuous functions $\{y_i\}$ such that $y_i \rightrightarrows y$, $i \to \infty$ and $y_i(t) \in \operatorname{int} T_G(x_0(t))\ \forall\, t \in I$.

Hence, without loss of generality we can assume that there is $\delta > 0$ such that

$$y(t) \in N_\delta^*(t) \quad \forall\, t \in I. \tag{3.13}$$

Here $N_\delta(t) = \{\lambda y : \ \|y - x\| \leq \delta,\ x \in N_G(x_0(t)),\ \|x\| = 1,\ \lambda \geq 0\}$ is the conic δ-neighborhood of the normal cone $N_G(x_0(t))$ and $N_\delta^*(t)$ its adjoint.

Let us take an arbitrary point $\tau \in I$ and prove that there is $\varepsilon(\tau) > 0$ such that

$$\frac{\partial \rho_i}{\partial x}(x_i(t)) \in N_{\delta/2}(\tau) \tag{3.14}$$

for all large enough numbers i and all t such that $|t - \tau| \leq \varepsilon(\tau)$. Indeed, from the definition of the function ρ_i we have

$$\frac{\partial \rho_i}{\partial x}(x_i(t)) = \int \frac{\partial \rho}{\partial x}(x_i(t) + y) w_i(y)\, dy. \tag{3.15}$$

If $x_0(\tau) \in \operatorname{int} G$, then $N_G(x_0(\tau)) = \{0\}$ and (3.14) holds, obviously. Suppose that $x_0(\tau) \in \partial G$ and (3.14) is violated. Then there is a sequence $t_i \to \tau$, $i \to \infty$ such that $\frac{\partial \rho_i}{\partial x}(x_i(t_i)) \notin N_{\delta/2}(\tau)$. Due to (3.15) there is a sequence $y_i \to 0$, $i \to \infty$ such that

$$v_i = \frac{\partial \rho}{\partial x}(x_i(t_i) + y_i) \notin N_{\delta/2}(\tau). \tag{3.16}$$

According to the properties of the distance function ρ we have $\|v_i\| = 1$ and $v_i \in N_G(z_i)$, where $z_i \in G$ is a closest point to $x_i(t_i) + y_i$ (see Proposition 2.5.4 in [2]). Since x_i converges to x_0, the sequence $\{z_i\}$ converges to $x_0(\tau)$, $i \to \infty$.

Passing to a subsequence we obtain

$$v_i \to v \quad \text{as} \quad i \to \infty, \qquad \|v\| = 1.$$

Due to the upper semicontinuity of Clarke's normal cone $N_G(x)$ in the case in which $\operatorname{int} T_G(x) \neq \emptyset$ we obtain $v \in N_G(x_0(\tau))$, which contradicts (3.16). Thus, the inclusion (3.14) is proved.

Decreasing the number $\varepsilon > 0$ if necessary, due to (3.13) we get

$$y(t) \in N^*_{\delta/2}(\tau) \quad \forall\, t:\ |t-\tau| \leq \varepsilon.$$

This inclusion and (3.14) imply

$$\left\langle i\lambda_i \frac{\partial \rho_i}{\partial x}(x_i(t)), y(t) \right\rangle \leq 0 \quad \forall\, t \in I:\ |t-\tau| \leq \varepsilon$$

for all large enough numbers i.

From this fact and the definition of the measure η, we conclude that for any point $\tau \in I$ there is $\varepsilon(\tau)$ such that

$$\int_{I \cap [\tau-\varepsilon(\tau), \tau+\varepsilon(\tau)]} y(t)\, d\eta \leq 0,$$

from where it is not difficult to derive the desired inequality

$$\int_{t_1}^{t_2} y(t)\, d\eta \leq 0.$$

Let us prove the nontriviality condition (f); i.e.,

$$\lambda_0 + \|\psi(t_1)\| + \|\eta\| \neq 0.$$

Indeed, let us assume $\lambda_0 = 0$, $\|\psi(t_1)\| = 0$, $\|\eta\| = 0$. Then from (3.12),

$$i\lambda_i \int_{t_1}^{t_2} \Big\| \frac{\partial \rho_i(x_i(t))}{\partial x} \Big\|\, dt \to 1, i \to \infty. \tag{3.17}$$

Due to the condition $\operatorname{int} T_G(x_0(t)) \neq \emptyset$ there is a continuous vector function g on I and a number $\delta > 0$ such that $\|g(t)\| = 1$, $\{y:\ \|y - g(t)\| \leq 2\delta\} \subset T_G(x_0(t))\ \forall\, t \in I$. Then, obviously,

$$\langle g(t), y \rangle \leq -2\delta \|y\| \quad \forall\, y \in N_G(x_0(t)) \quad \forall\, t \in I. \tag{3.18}$$

We show now that

$$\int_{t_1}^{t_2} g(s)\, d\eta \leq -\frac{\delta}{2} < 0.$$

Let $\delta > 0$; then from (3.14) for any point $\tau \in I$ there is $\varepsilon(\tau) > 0$ such that

$$\frac{\partial \rho_i}{\partial x}(x_i(t)) \in N_{\delta/2}(\tau) \quad \forall\, t \in I \cap [\tau - \varepsilon(\tau), \tau + \varepsilon(\tau)]$$

for all large enough numbers i. Hence, $\forall\, t \in I \cap [\tau - \varepsilon(\tau), \tau + \varepsilon(\tau)]$ there exist $z(t) \in N_G(x_0(\tau))$ with $\|z(t)\| = 1$ and $\xi(t)$ with $\|\xi(t)\| \leq \delta/2$ such that

$$\frac{\partial \rho_i}{\partial x}(x_i(t)) = (z(t) + \xi(t)) \Big\| \frac{\partial \rho_i}{\partial x}(x_i(t)) \Big\|. \tag{3.19}$$

Due to the continuity of g and (3.18) and decreasing $\varepsilon(\tau) > 0$ if necessary, we get

$$\langle g(t), z\rangle \le -\delta\|z\| \quad \forall\, z \in N_G(x_0(\tau)) \quad \forall\, t \in I \cap [\tau - \varepsilon(\tau), \tau + \varepsilon(\tau)].$$

Thus, by (3.19) we obtain

$$\int_{\tau-\varepsilon(\tau)}^{\tau+\varepsilon(\tau)} \left\langle g(s), \frac{\partial \rho_i}{\partial x}(x_i(s))\right\rangle ds \le -\frac{\delta}{2}\int_{\tau-\varepsilon(\tau)}^{\tau+\varepsilon(\tau)} \left\|\frac{\partial \rho_i}{\partial x}(x_i(s))\right\| ds$$

for all large enough numbers i.

Using the compactness of the time interval I, it is not difficult to show the existence of the finite system of disjoint half-open intervals $\{I_j\}$, $j = 1, \dots, N$ such that $I = \cup_{j=1}^{N} I_j$ and for all large enough numbers i

$$\int_{I_j} \left\langle g(s), \frac{\partial \rho_i}{\partial x}(x_i(s))\right\rangle ds \le -\frac{\delta}{2}\int_{I_j} \left\|\frac{\partial \rho_i}{\partial x}(x_i(s))\right\| ds.$$

Hence,

$$\int_{t_1}^{t_2} \left\langle g(s), \frac{\partial \rho_i}{\partial x}(x_i(s))\right\rangle ds \le -\frac{\delta}{2}\int_{t_1}^{t_2} \left\|\frac{\partial \rho_i}{\partial x}(x_i(s))\right\| ds$$

for all large enough numbers i. According to the definition of the measure η and (3.17) we have

$$\int_{t_1}^{t_2} g(s)\, d\eta \le -\frac{\delta}{2} < 0.$$

The nontriviality condition (f) is proved.

Let us denote

$$p_i(t) = H\left(x_i(t), t, \tilde{\psi}_i(t) + \int_{t_1}^{t} i\lambda_i \frac{\partial \rho_i}{\partial x}(x_i(s))\, ds\right),$$

$$g_i(t) = i\lambda_i \rho_i(x_i(t)) + \lambda_i \|x_i(t) - x_0(t)\|^2.$$

Then $h_i(t) = p_i(t) - g_i(t)$ is an absolutely continuous function $\forall\, i = 1, 2, \dots$. Due to the endpoint constraints (3.2) we have

$$p_i(t_1) = H(x_i(t_1), t_1, \tilde{\psi}_i(t_1)) \to H(x_0(t_1), t_1, \psi(t_1)),$$

$$|g_i(t_1)| \le \lambda_i \left(\frac{i}{2^i} + \|x_i(t_1) - x_0(t_1)\|^2\right) \to 0.$$

So, the sequence $\{|h_i(t_1)|\}$ is bounded. Further, due to the Hamiltonian inclusion (3.9) and (3.12) we have $|\dot{h}_i(t)| \le \kappa_2$ $\forall\, t \in I$. Therefore, without loss of generality we can suppose $h_i \rightrightarrows h$ on I, $\dot{h}_i \to \dot{h}$ weakly in $L_1(I)$ as $i \to \infty$, where h is a Lipschitz continuous function on I.

Let $t_* < t_2$. Since

$$\int_{t_1}^{t} d\eta = \lim_{\varepsilon \to 0+} \lim_{i \to \infty} \int_{t_1}^{t+\varepsilon} i\lambda_i \frac{\partial \rho_i}{\partial x}(x_i(s))\, ds \quad \forall\, t < t_2$$

and $g_i(t) \to 0$ a.e. on I, $i \to \infty$ (see the lemma above), one can choose a sequence $\tau_i \to t_* + 0$, $i \to \infty$ such that, passing to a subsequence, we get

$$p_i(\tau_i) = H\biggl(x_i(\tau_i), \tau_i, \tilde{\psi}_i(\tau_i) + \int_{t_1}^{\tau_i} i\lambda_i \frac{\partial \rho_i}{\partial x}(x_i(s))\, ds\biggr)$$

$$\to H\biggl(x_0(t_*), t_*, \psi(t_*) + \int_{t_1}^{t_*} d\eta\biggr), \qquad g_i(\tau_i) \to 0.$$

Then

$$h(t_*) = \lim_{i\to\infty} h_i(\tau_i) = H\biggl(x_0(t_*), t_*, \psi(t_*) + \int_{t_1}^{t_*} d\eta\biggr).$$

Now let us consider the point t_2. Due to the endpoint constraints (3.2), $g_i(t_2) \to 0$, and according to the definition of η

$$h(t_2) = \lim_{i\to\infty} h_i(t_2) = H\biggl(x_0(t_2), t_2, \psi(t_2) + \int_{t_1}^{t_2} d\eta\biggr).$$

Thus, $H(x_0(t), t, \psi(t) + \int_{t_1}^{t} d\eta) = h(t)\ \forall t \in I$. Hence, the function $H(x_0(t), t, \psi(t) + \int_{t_1}^{t} d\eta)$ is absolutely continuous on I.

The jump condition (b) is a consequence of the continuity of h for any $t > t_1$. Let us prove the jump condition (b) for $t = t_1$. We have

$$h_i(t_1) = H(x_i(t_1), t_1), \tilde{\psi}_i(t_1)) - i\lambda_i \rho_i(x_i(t_1))$$

$$-\lambda_i \|x_i(t_1) - x_0(t_1)\|^2 \to H(x_0(t_1), t_1, \psi(t_1)).$$

On the other hand $h_i(t_1) \to h(t_1) = H(x_0(t_1), t_1, \psi(t_1) + \eta(t_1))$, $i \to \infty$. Thus, the jump condition (b) holds for $t = t_1$.

Let us prove the Hamiltonian inclusion (d). As we have proved above, $x_i \rightrightarrows x_0$, $\tilde{\psi}_i \rightrightarrows \psi$, $h_i \rightrightarrows h$ on I, and $\dot{x}_i \to \dot{x}_0$, $\dot{\tilde{\psi}}_i \to \dot{\psi}_i$, $\dot{h}_i \to \dot{h}$ weakly in $L_1(I)$. Due to the definition of the measure η we have

$$\int_{t_1}^{t} d\eta = \lim_{i\to\infty} i\lambda_i \int_{t_1}^{t} \frac{\partial \rho_i}{\partial x}(x_i(s))\, ds \quad \text{a.e. on } I.$$

Hence, due to the upper semicontinuity of Clarke's generalized gradient [2] we obtain

$$\partial H\biggl(x_0(t), t, \psi(t) + \int_{t_1}^{t} d\eta\biggr)$$

$$\supseteq \limsup_{i\to\infty} \partial H\biggl(x_i(t), t, \psi_i(t) + i\lambda_i \int_{t_1}^{t} \frac{\partial \rho_i}{\partial x}(x_i(s))\, ds\biggr) \quad \text{a.e. on } I.$$

Hence, due to the inclusion (3.9) and Mazur's theorem [1]

$$(-\dot{\psi}(t), \dot{h}(t), \dot{x}_0(t)) \in \partial H\biggl(x_0(t), t, \psi(t) + \int_{t_1}^{t} d\eta\biggr) \quad \text{a.e. on } I.$$

The property (d) is proved.

The transversality condition (e) follows directly from (3.10) and upper semicontinuity of the cone of limiting normals and of the set of limiting subgradients [5], [21], [22]. □

4. The conditions of nondegeneracy and pointwise nontriviality. The following notion of the controllability of the reference trajectory at the endpoints plays the central role in our study of the degeneracy phenomenon.

DEFINITION 1. *A trajectory x_0 of the control system* (2.1) *is called controllable at the endpoints (with respect to the state and endpoints constraints) if*

$$H(x_0(t_i), t_i, (-1)^i g) > 0 \tag{4.1}$$

$$\forall\, g \in N_G(x_0(t_i)) \cap [-\hat{N}_{C_i \cap G}(x_0(t_i))], \quad g \neq 0, \quad i = 1, 2.$$

The controllability condition (4.1) is a very natural one. Indeed, let the state constraint set G be regular at the endpoints x_1, x_2. This means that Clarke's tangent cone $T_G(x_i)$ coincides with the contingent cone $K_G(x_i)$, $i = 1, 2$ [2]. Let us suppose that there exists a trajectory x_0 of the control system (2.1) satisfying the state constraints (2.3) and transferring the point x_1 to the point x_2. Then

$$\limsup_{t \xrightarrow{I} t_i} \frac{x(t) - x_i}{t - t_i} \subseteq F(x_i, t_i) \cap (-1)^{i+1} K_G(x_i), \qquad i = 1, 2.$$

Hence, $H(x_i, t_i, (-1)^i g) \geq 0\ \forall\, g \in N_G(x_i)$, $i = 1, 2$. Obviously, if the set G is regular at the endpoints x_i, $i = 1, 2$ then the controllability condition (4.1) is the generic one. This means not only that if there is at least one admissible trajectory x_0 transferring the point x_1 to the point x_2 then using small perturbation $F_\varepsilon(x, t) = F(x, t) + \varepsilon B$, $\varepsilon > 0$ of the right-hand side of the system we can satisfy (4.1) but also that this condition is stable under all small enough perturbations of the right-hand side of the system. Here B denotes the closed unit ball with center at the origin.

THEOREM 2. *Let the trajectory x_0 satisfying the maximum principle* (*Theorem* 1) *be controllable at the endpoints. Then*

$$\lambda_0 + \operatorname{meas}\left\{t :\ \psi(t) + \int_{t_1}^{t} d\eta \neq 0\right\} > 0. \tag{4.2}$$

Proof. Let us suppose that

$$\lambda_0 = 0 \quad \text{and} \quad \psi(t) + \int_{t_1}^{t} d\eta = 0 \quad \text{a.e. on } I.$$

According to the Hamiltonian inclusion we have

$$\|\dot{\psi}(t)\| \leq \kappa_2 \left\| \psi(t) + \int_{t_1}^{t} d\eta \right\| \quad \text{a.e. on } I.$$

Hence, $\psi(t) \equiv \psi(t_1)$, measure η is equal to 0 on (t_1, t_2), and $h_0(t) \equiv 0$. Further, due to the measure sign condition $\eta(t_1) \in N_G(x_1)$. Hence, $\psi(t_1) = -\eta(t_1)$ and due to the transversality condition $\eta(t_1) \in N_G(x_1) \cap [-\hat{N}_{C_1 \cap G}(x_1)]$, where $x_i = x_0(t_1)$, $i = 1, 2$.

It follows from the jump condition that $H(x_1, t_1, -\eta(t_1)) = 0$. We have $\eta(t_1) = 0$ from the previous equality and controllability of the trajectory x_0 at the left endpoint.

Let us consider now the right endpoint. Then we obtain similarly that $\eta(t_2) \in N_G(x_2) \cap (-\hat{N}_{C_2 \cap G}(x_2))$ and $H(x_2, t_2, \eta(t_2)) = 0$. Hence, from the controllability of the trajectory x_0 at the right endpoint we have $\eta(t_2) = 0$.

Thus, $\lambda_0 = 0$, $\psi(t) \equiv 0$, and $\eta = 0$. But this contradicts the nontriviality condition. Hence, the condition (4.2) is proved. □

It turns out that in the case in which the set G is regular [2] at the endpoints, the controllability condition (4.1) is the necessary and sufficient one for the maximum principle (Theorem 1) to be informative.

THEOREM 3. *Let the set G be regular at the endpoints x_1, x_2 of the trajectory x_0 satisfying the maximum principle. Then the nondegeneracy condition* (4.2) *holds for all collections of Lagrange multipliers λ_0, ψ, η if and only if the trajectory x_0 is controllable at the endpoints.*

Proof. Due to Theorem 2 we need to prove only the necessity.

Suppose that there exists a vector $g \in N_G(x_1) \cap [-\hat{N}_{C_1 \cap G}(x_1)]$, $g \neq 0$ such that $H(x_1, t_1, -g) = 0$ (the case of the right endpoint is completely analogous). Then we put $\lambda_0 = 0$, $\psi(t) \equiv g$ and $\eta = -g\delta_{t_1}$. Obviously, all conditions of Theorem 1 are satisfied and $\lambda_0 + \operatorname{meas}\{t : \psi(t) + \int_{t_1}^{t} d\eta \neq 0\} = 0$. □

THEOREM 4. *Let a trajectory x_0 satisfying the maximum principle* (*Theorem* 1) *be controllable at the endpoints and*

$$H(x_0(t), t, (-1)^i g) > 0 \quad \forall\, g \in N_G(x_0(t)) :\ g \neq 0, \tag{4.3}$$

$$\forall\, t \in (t_1, t_2), \qquad i = 1, 2.$$

Then

$$\lambda_0 + \left\| \psi(t) + \int_{t_1}^{t} d\eta \right\| \neq 0 \quad \forall\, t \in (t_1, t_2). \tag{4.4}$$

Proof. Let us consider the set

$$T = \left\{ t \in (t_1, t_2) :\ \psi(t) + \int_{t_1}^{t} d\eta = 0 \right\}.$$

Suppose that (4.4) is violated. Then $\lambda_0 = 0$ and $T \neq \emptyset$. We prove first that T is open. Let $\tau \in T$. We show that there exists a right-half neighborhood O_τ^+ of the point τ such that $O_\tau^+ \subset T$. Let us denote $\xi(t) = \| \int_\tau^t d\eta \|$. Suppose that $\xi(t) \not\equiv 0$ in some right-half neighborhood of τ.

Since $\tau \in T$ and the Hamiltonian inclusion holds, we have

$$\|\dot{\psi}(t)\| \leq \kappa_3(\|\psi(t) - \psi(\tau)\| + \xi(t)),$$

$$|\dot{h}(t)| \leq \kappa_3(\|\psi(t) - \psi(\tau)\| + \xi(t))$$

for all $t \geq \tau$. Therefore, Bellman–Gronwall's inequality leads to

$$\|\psi(t) - \psi(\tau)\| \leq \kappa_4 \int_\tau^t \xi(s)\, ds, \qquad t \geq \tau. \tag{4.5}$$

The bounded function ξ is nonnegative. Hence

$$\int_\tau^\sigma \int_\tau^t \xi(s)\, ds dt \leq \int_\tau^\sigma \xi(s) ds \qquad \forall \sigma \in [\tau, \tau + 1].$$

Due to the assumption $\tau \in T$ we have $h(\tau) = 0$. Therefore, due to the absolute continuity of h, (4.5), and the above inequality we obtain

$$|h(t)| \leq \kappa_5 \int_\tau^t \xi(s)\, ds \qquad \forall\, t \geq \tau. \tag{4.6}$$

Now Bellman–Gronwall's inequality leads to the existence of a sequence $\{t_i\}:\ t_i \to \tau + 0$ as $i \to \infty$ such that

$$\xi(t_i) \neq 0; \qquad \xi(t_i)^{-1} \int_\tau^{t_i} \xi(s)\, ds \to 0, \qquad i \to \infty.$$

Hence, due to (4.5), (4.6) we obtain

$$\xi(t_i)^{-1} h(t_i) \to 0, \qquad \xi(t_i)^{-1} \|\psi(t_i) - \psi(\tau)\| \to 0, \qquad i \to \infty. \tag{4.7}$$

Further, due to the upper semicontinuity of the normal cone $N_G(x)$ and (2.5) there exists $g \in N_G(x_0(\tau))$, $\|g\| = 1$ such that, passing to a subsequence, we obtain

$$\xi(t_i)^{-1} \int_\tau^{t_i} d\eta \to g, \qquad i \to \infty.$$

Due to the assumption $\tau \in T$ we have

$$h(t_i) = H\bigg(x_0(t_i), t_i, \psi(t_i) - \psi(\tau) + \int_\tau^{t_i} d\eta\bigg).$$

Dividing both parts of this equality by $\xi(t_i)$ and passing to the limit, due to (4.7) we obtain

$$H(x_0(\tau), \tau, g) = 0.$$

Thus, we come to a contradiction with assumption (4.3) for $i = 2$.

This contradiction proves that $\xi(t) \equiv 0\ \forall\, t \in O_\tau^+$ for some right-half neighborhood O_τ^+ of τ. Due to (4.4) we have

$$\psi(t) \equiv \psi(\tau) \quad \forall\, t \in O_\tau^+ \Rightarrow \psi(t) + \int_{t_1}^t d\eta \equiv 0 \quad \forall\, t \in O_\tau^+.$$

Hence, $O_\tau^+ \subset T$. Using similar arguments, one can prove the existence of some left-half neighborhood O_τ^- of the point τ such that $O_\tau^- \subset T$.

So, we have proved that the set T is open. Now let us prove the closedness of the set T with respect to interval (t_1, t_2).

Indeed, let a sequence $\{t_i\}$ converge to $\tau \in (t_1, t_2)$, $t_i \in T$, $i = 1, 2, \ldots$. Without loss of generality we can consider two cases: $t_i \to \tau + 0$ and $t_i \to \tau - 0$.

Let $t_i \to \tau + 0$. Then due to the regularity of the measure η we obtain that the function $\psi(t) + \int_{t_1}^t d\eta$ is continuous from the right. Hence,

$$\psi(\tau) + \int_{t_1}^\tau d\eta = \lim_{i \to \infty} \bigg(\psi(t_i) + \int_{t_1}^{t_i} d\eta\bigg) = 0.$$

Thus, $\tau \in T$.

Let $t_i \to \tau - 0$. Then

$$\begin{aligned} \eta(\tau) &= \eta(\tau) + \lim_{i \to \infty} \bigg(\psi(t_i) + \int_{t_1}^{t_i} d\eta\bigg) \\ &= \eta(\tau) + \psi(\tau) + \int_{[t_1, \tau)} d\eta = \psi(\tau) + \int_{t_1}^\tau d\eta. \end{aligned}$$

Due to the continuity of the Hamiltonian $h(\tau) = \lim_{i\to\infty} h(t_i) = 0$. Hence,

$$H(x_0(\tau), \tau, \eta(\tau)) = H\left(x_0(\tau), \tau, \psi(\tau) + \int_{t_1}^{\tau} d\eta\right) = 0.$$

Due to the measure sign condition $\eta(\tau) \in N_G(x_0(\tau))$. Therefore, the condition (4.3) with $i = 1$ gives $\eta(\tau) = 0$. So $\psi(\tau) + \int_{t_1}^{\tau} d\eta = 0$ and $\tau \in T$.

Thus, we have proved that the set T is closed with respect to the interval (t_1, t_2). We have proved above that T is open. The interval (t_1, t_2) is an arcwise connected set. Hence, $T = (t_1, t_2)$. But this contradicts the statement of Theorem 2. Hence, (4.4) is proved. □

We should note that in the problem considered above in Example 2 the state constraint set G is regular at any point and $H(x, g) > 0 \ \forall\, g \neq 0$. Hence, according to Theorem 4 the pointwise nontriviality condition (4.4) holds for any solution of this problem.

5. The classical optimal control problem. Now let us consider the classical optimal control problem for systems described by ordinary differential equations with control parameters. This problem is formulated as follows:

$$\dot{x} = f(x, t, u), \qquad u \in U; \tag{5.1}$$

$$x(t_1) = x_1, \qquad x(t_2) = x_2, \qquad p = (t_1, t_2, x_1, x_2); \tag{5.2}$$

$$k_1(p) \leq 0, \qquad k_2(p) = 0; \tag{5.3}$$

$$g(x(t)) \leq 0 \quad \forall\, t \in I = [t_1, t_2]; \tag{5.4}$$

$$J(u, p) = k_0(p) \to \min. \tag{5.5}$$

Here $x \in R^n$; U is a given closed subset of R^k; smooth vector functions k_1, k_2, g take values in spaces of dimensions d_1, d_2, and m, respectively; and the scalar function k_0 is also supposed to be smooth. As usual, we assume that the set of admissible controls consists of all bounded measurable functions u, each of them defined on its own time interval I such that $u(t) \in U$ a.e. on I.

Before formulating the maximum principle for the problem (5.1)–(5.5) let us introduce some notations and assumptions.

We define the small Lagrangian l and Hamilton–Pontryagin's function $\mathcal{H}$ as follows:

$$l(p, \lambda) = \lambda_0 k_0(p) + \langle \lambda_1, k_1(p) \rangle + \langle \lambda_2, k_2(p) \rangle,$$

$$\mathcal{H}(x, t, u, \psi) = \langle f(x, t, u), \psi \rangle.$$

Here $\lambda = (\lambda_0, \lambda_1, \lambda_2) \in R^{d_1+d_2+1}$, $p = (t_1, t_2, x_1, x_2) \in R^{2n+2}$, $\psi \in R^n$.

The Hamiltonian associated with the control system (5.1) is defined by

$$H(x, t, \psi) = \sup_{u\in U} \langle f(x, t, u), \psi \rangle.$$

Let a pair x_0, u_0 be a solution of the problem (5.1)–(5.5) and $I_0 = [t_{1,0}, t_{2,0}]$, $p_0 = (t_{1,0}, t_{2,0}, x_{1,0}, x_{2,0})$ correspond to it. We suppose that $t_{1,0} < t_{2,0}$ and the following assumptions are fulfilled:

(H1) The vectors $\left\{\frac{\partial g_j(x_0(t))}{\partial x}\right\}_{j:\, g_j(x_0(t))=0}$ are linearly independent $\forall\, t \in I_0$; $g = (g_1, \dots, g_m)$.

(H2) rank $\frac{\partial k_2}{\partial p}(p_0) = d_2$.

$$\exists\, y \in R^{d_2} :\ \frac{\partial k_2}{\partial p}(p_0)y = 0, \qquad \left\langle \frac{\partial k_{1,i}}{\partial p}(p_0), y \right\rangle < 0,$$

where $k_{1,i}(p_0)$ are all the coordinates of the vector $k_1(p_0)$ such that $k_{1,i}(p_0) = 0$.

(H3) The state constraints (5.4) are in accordance with the endpoint constraints (5.3) in some neighborhood of the ends of x_0. This means that there exists $\varepsilon > 0$ such that

$$\{p :\ \|p - p_0\| \le \varepsilon,\ k_1(p) \le 0,\ k_2(p) = 0,\ k_0(p) \le k_0(p_0)\}$$

$$\subset \{p :\ g(x_1) \le 0,\ g(x_2) \le 0\}.$$

(H4) The sets $F(x,t) = \{f(x,t,u) :\ u \in U\}$ are convex for all x, t.

The main distinction of the problem (5.1)–(5.5) from (2.1)–(2.4) consists of the smooth parametric form of the differential, endpoint, and state constraints.

It follows from (H1) that the state constraints set $G = \{x \in R^n :\ g(x) \le 0\}$ is regular at $x_0(t)$ $\forall\, t \in I_0$ and

$$N_G(x_0(t)) = \operatorname{cone}\left\{\frac{\partial g_j}{\partial x}(x_0(t))\right\}_{j:\, g_j(x_0(t))=0} \qquad \forall\, t \in I_0.$$

Analogously, (H2) provides the regularity of the endpoint constraint set $P = \{p \in R^{2n+2} :\ k_1(p) \le 0,\ k_2(p) = 0\}$ at p_0 and

$$N_P(p_0) = \Bigg\{z \in R^{2n+2} :\ z = \sum_{i=1}^{d_1} \alpha_i \frac{\partial k_{1,i}(p_0)}{\partial p}$$

$$+ \sum_{i=1}^{d_2} \beta_i \frac{\partial k_{2,i}(p_0)}{\partial p},\ \beta \in R^1,\ \alpha_i \ge 0,\ \sum_{i=1}^{d_1} \alpha_i k_{1,i}(p_0) = 0\Bigg\}.$$

Note that (H3) holds automatically, if some neighborhood of p_0 in P is contained in $\{z = (t_1, t_2, x_1, x_2) \in R^{2n+2} :\ x_i \in G,\ i = 1, 2\}$. In this case the normal cones to the sets P and $P_G = \{p \in P :\ x_i \in G,\ i = 1, 2\}$ are the same.

The following result is the modified version of Pontryagin's maximum principle.

THEOREM 5 (maximum principle). *Let the pair x_0, u_0 solve the problem* (5.1)–(5.5). *Then there exist a vector $\lambda = (\lambda_0, \lambda_1, \lambda_2)$: $\lambda_0 \in R^1$, $\lambda_i \in R^{d_i}$, $i = 1, 2$; a left-continuous vector function ψ of bounded variation; and a bounded nonnegative m-dimensional regular Borel measure η on I_0 such that following conditions hold:*

(a) $$\psi(t) = -\frac{\partial l}{\partial x_2}(p_0, \lambda) + \int_t^{t_{2,0}} \frac{\partial \mathcal{H}}{\partial x}(x_0(s), s, u_0(s), \psi(s))\, ds$$

$$- \int_t^{t_{2,0}} \frac{\partial g}{\partial x}(x_0(s))\, d\eta \quad \forall\, t \in I_0;$$

(b) $\mathcal{H}(x_0(t), t, u_0(t), \psi_0(t)) = H(x_0(t), t, \psi(t))$ *a.e. on* I_0;

(c) *the function* $h(t) = H(x_0(t), t, \psi(t))$ *is absolutely continuous on* I_0 *and*

$$\frac{dh(t)}{dt} = \frac{\partial \mathcal{H}}{\partial t}(x_0(t), t, u_0(t), \psi(t)) \quad \textit{a.e. on } I_0;$$

(d) $H(x_0(t), t, \psi(t)) = H\left(x_0(t), t, \psi(t) + \dfrac{\partial g}{\partial x}(x_0(t))\eta(t)\right)\ \forall\, t \in I_0$;

(e) $\psi(t_{1,0}) = \dfrac{\partial l}{\partial x_1}(p_0, \lambda)$,

$$-h(t_{1,0}) = \frac{\partial l}{\partial t_1}(p_0, \lambda), \qquad h(t_{2,0}) = \frac{\partial l}{\partial t_2}(p_0, \lambda);$$

(f) $\operatorname{supp}\eta_i \subset \{t \in I_0 :\ g_i(x_0(t)) = 0\}$, $i = 1, 2, \dots, m$;
(g) $\lambda_0 \geq 0$, $\lambda_1 \geq 0$, $\langle k_1(p_0), \lambda_1\rangle = 0$;
(h) $\|\lambda\| + \|\eta\| \neq 0$.

Note that if assumption (H3) is violated, then the endpoint constraint set P should be considered in intrinsic form. In this case the assertion of Theorem 5 is valid with the transversality condition analogous to the one in Theorem 1 (see also [19]). The smooth nonautonomous problem with state constraint $g(x, t) \leq 0$ can be reduced to the autonomous one standardly [25] by introducing an additional variable $\dot{x}^{n+1} = 1$.

DEFINITION 2. *A trajectory* x_0 *is called controllable at the endpoints*[3] *(with respect to the state and endpoint constraints) if*

$$H(x_0(t_{i,0}), t_{i,0}, (-1)^i z_i) > 0$$

$$\forall\, z_i \neq 0 : \exists z = (z_1, z_2) \in N \cap D, \qquad i = 1, 2;$$

$$N = N_G(x_0(t_{1,0})) \times N_G(x_0(t_{2,0})), \qquad D = \Big\{(z_1, z_2) \in R^{2n} : (z_1, z_2)$$

$$= -\frac{\partial}{\partial(x_1, x_2)}(\langle\lambda_1, k_1(p_0)\rangle + \langle\lambda_2, k_2(p_0)\rangle),\ \lambda_i \in R^{d_i},\ i = 1, 2;\ \lambda_1 \geq 0\Big\}.$$

THEOREM 6. *Let pair* x_0, u_0 *satisfy the maximum principle. Then for all collections of Lagrange multipliers*

$$\lambda_0 + \operatorname{meas}\{t \in I_0 :\ \psi(t) \neq 0\} > 0$$

if and only if the trajectory x_0 *is controllable at the endpoints.*

THEOREM 7. *Let pair* x_0, u_0 *satisfy the maximum principle and the trajectory* x_0 *be controllable at the endpoints. Let* $\forall\, t \in (t_{1,0}, t_{2,0})\ \exists\, u_i(t) \in U$:

$$(-1)^i \left\langle \frac{\partial g_j}{\partial x}(x_0(t)), f(x_0(t), t, u_i(t)) \right\rangle > 0$$

$$\forall\, j :\ g_j(x_0(t)) = 0, \qquad i = 1, 2.$$

[3]This notion of controllability at the endpoints is a little bit weaker than the one introduced in [9].

Then

$$\lambda_0 + \|\psi(t)\| \neq 0 \quad \forall\, t \in (t_{1,0}, t_{2,0}).$$

The proofs of Theorems 5–7 are analogous to the ones of Theorems 1–4 above. Some details of these proofs concerning time-transversality conditions and problems formulated in the class of generalized controls can be found in [14].

Finally, let us consider the case in which maximum principle degenerates. In this case under some controllability conditions it is possible to guarantee the existence of at least one nontrivial collection of Lagrange multipliers.

THEOREM 8. *Let the pair x_0, u_0 satisfy the maximum principle* (*Theorem* 5) *and* $N = N_G(x_0(t_{1,0})) \times N_G(x_0(t_{2,0}))$. *Suppose that* $\forall\, z = (z_1, z_2) \in N \cap \operatorname{Im} \left[\frac{\partial k_2(p_0)}{\partial(x_1, x_2)}\right]^*$, $z \neq 0$ $\exists i = 1, 2$ *such that the following condition holds:*

$$H(x_0(t_{i,0}), t_{i,0}, (-1)^i z_i) > 0. \tag{5.6}$$

Then there exists at least one nontrivial collection of Lagrange multipliers λ, ψ, η (*i.e., such that* $\lambda_0 + \operatorname{meas}\{t :\ \psi(t) \neq 0\} > 0$) *corresponding to* x_0.

Proof. Let $\lambda = (\lambda_0, \lambda_1, \lambda_2)$, ψ, η be a collection of Lagrange multipliers corresponding to x_0.

If (4.2) holds, then λ, ψ, η are desired Lagrange multipliers.

Suppose now that $\lambda_0 = 0$, $\psi(t) = 0$ $\forall\, t \in (t_{1,0}, t_{2,0})$. Then $\eta = \eta_1 \delta_{t_{1,0}} + \eta_2 \delta_{t_{2,0}}$ ($\eta_j = \eta(t_{j,0})$, $j = 1, 2$) and the conditions of the maximum principle lead to

$$\frac{\partial l}{\partial x_j}(p_0, \lambda) = -\frac{\partial g}{\partial x}(x_0(t_{j,0}))\eta_j, \qquad j = 1, 2; \tag{5.7}$$

$$H(x_0(t_{j,0}), t_{j,0}, (-1)^j \frac{\partial g}{\partial x}(x_0(t_{j,0}))\eta_j) = 0, \qquad j = 1, 2. \tag{5.8}$$

Consider the following auxiliary mathematical programming problem:

$$\chi(p) = \langle g(x_1), \eta_1 \rangle + \langle g(x_2), \eta_2 \rangle \to \max,$$

$$k_0(p) \leq k_0(p_0), \qquad k_1(p) \leq 0, \qquad k_2(p) = 0.$$

Due to (H3) p_0 is a local solution of this problem. Hence, by virtue of the Lagrange multiplier rule there exist $\gamma \geq 0$, $\beta = (\beta_0, \beta_1, \beta_2) \in R^{d_1 + d_2 + 1}$ such that

$$\gamma \frac{\partial g}{\partial x}(x_{j,0})\eta_j = \frac{\partial l}{\partial x_j}(p_0, \beta), \qquad j = 1, 2;$$

$$\beta_0 \geq 0, \quad \beta_1 \geq 0, \quad \beta_0 k_0(p_0) = 0; \quad \langle \beta_1, k_1(p_0) \rangle = 0; \quad \gamma + \|\beta\| \neq 0.$$

If $\gamma = 0$, then $\beta \neq 0$ and, hence, p_0 satisfies the Lagrange multiplier rule for the problem

$$k_0(p) \to \min, \tag{5.9}$$

$$k_1(p) \leq 0, \qquad k_2(p) = 0. \tag{5.10}$$

In this case by virtue of (H2) $\beta_0 > 0$. Hence, $\tilde{\lambda} = (\beta_0, \beta_1, \beta_2)$, $\tilde{\psi}(t) = 0$, $\tilde{\eta} = 0$ is a nontrivial collection of Lagrange multipliers corresponding to x_0.

Suppose that $\gamma \neq 0$. Then, setting $\gamma = 1$, due to (5.7) we obtain

$$\frac{\partial g}{\partial x}(x_{j,0})\eta_j = \frac{\partial l}{\partial x_j}(p_0, \beta) = -\frac{\partial l}{\partial x_j}(p_0, \lambda), \qquad j = 1, 2. \tag{5.11}$$

Hence,

$$\frac{\partial l}{\partial x_j}(p_0, \lambda + \beta) = 0, \qquad j = 1, 2.$$

If $\lambda + \beta \neq 0$, then p_0 satisfy the Lagrange multipliers rule for (5.9), (5.10). Hence, due to (H2), $\beta_0 > 0$ and $\tilde{\lambda} = (\beta_0, \lambda_1 + \beta_1, \lambda_2 + \beta_2)$, $\tilde{\psi}(t) \equiv 0$, $\tilde{\eta} = 0$ is a nontrivial collection of Lagrange multipliers corresponding to x_0.

If $\lambda + \beta = 0$, then $\beta_0 = 0$, $\lambda_1 = -\beta_1 = 0$, $\lambda_2 = -\beta_2 \neq 0$. Hence, by (5.11) and (H2),

$$z = (z_1, z_2) = \left(\frac{\partial g}{\partial x}(x_{1,0})\eta_1, \frac{\partial g}{\partial x}(x_{2,0})\eta_2\right) = \left[\frac{\partial k_2(p_0)}{\partial(x_1, x_2)}\right]^* \beta_2 \neq 0.$$

Due to (5.8) this equality contradicts assumptions (5.6) of the theorem. □

Let us consider now the problem of Example 1. In this case the optimal trajectory $x_0(t) \equiv 0$ is not controllable at the left endpoint and any corresponding collection of Lagrange multipliers is trivial. The question arises as how to modify it to exhibit the degeneracy phenomenon. In our opinion, if the optimal trajectory fails to be controllable at the endpoints the class of admissible Lagrange multipliers should be generalized. Some results in this direction can be found in [15], [17]. Here we restrict ourselves to an illustrative example.

Example 3. Following a suggestion of Maurer we introduce a small parameter $\alpha < 0$ in the problem of Example 1:

$$\begin{aligned}
&\dot{x}_1 = tu, \qquad |u| \leq 1,\\
&\dot{x}_2 = u;\\
&x_1(0) = 0, \qquad x_2(0) = 0, \qquad x_1(1) \geq \alpha;\\
&x_1(t) \geq \alpha \quad \forall\, t \in [0, 1];\\
&J(u(\cdot)) = x_2(1) \to \min.
\end{aligned}$$

Then the optimal pair x^α, u^α is the following:

$$x_1^\alpha(t) = \begin{cases} -\dfrac{t^2}{2}, & t \in [0, \sqrt{-2\alpha}];\\ -\alpha, & t \in [\sqrt{-2\alpha}, 1]; \end{cases} \qquad x_2^\alpha(t) = \begin{cases} -t, & t \in [0, \sqrt{-2\alpha}];\\ -\sqrt{-2\alpha}, & t \in [\sqrt{-2\alpha}, 1]; \end{cases}$$

$$u^\alpha(t) = \begin{cases} -1 & \text{a.e. on } [0, \sqrt{-2\alpha}];\\ 0 & \text{a.e. on } [\sqrt{-2\alpha}, 1]. \end{cases}$$

Let us apply Theorem 5. The easy calculations give the Lagrange multipliers λ_0^α, λ_1^α,

ψ^α, η^α corresponding to x^α, u^α:

$$\lambda_0^\alpha = 1, \qquad \lambda_1^\alpha = 1;$$

$$\psi_1^\alpha(t) = \begin{cases} \dfrac{1}{\sqrt{-2\alpha}}, & t \in [0, \sqrt{-2\alpha}]; \\ 1/t, & t \in [\sqrt{-2\alpha}, 1]; \end{cases} \qquad \psi_2^\alpha(t) = -1;$$

$$d\eta^\alpha = 1/t^2 dt, \qquad t \in (\sqrt{-2\alpha}, 1];$$

$$\operatorname{supp} \eta^\alpha = (\sqrt{-2\alpha}, 1].$$

In the limit as $\alpha \to -0$, we obtain

$$\lambda_0^\alpha \to \lambda_0 = 1, \qquad \lambda_1^\alpha \to \lambda_1 = 1;$$

$$\psi^\alpha(t) \to \psi(t) = \left(\frac{1}{t}, -1\right) \quad \forall\, t \in (0, 1];$$

$$\eta^\alpha \to \eta : \ \operatorname{supp} \eta = (0, 1], \qquad d\eta = \frac{1}{t^2} dt.$$

The obtained collection λ_0, λ_1, ψ, η satisfies all the conditions of Theorem 5 except the conditions for ψ to be a function of bounded variation and the measure η to be bounded.

6. Bibliographical notes. The first version of Pontryagin's maximum principle for problems with state constraints was obtained by Gamkrelidze in 1959 (see [25], [28]) under a priori assumptions on an optimal trajectory. (It is supposed that the reference trajectory consists of the finite number of boundary and interior arcs and corresponding control is regular.) This maximum principle contains no measures and is the nondegenerate one.

The present paper deals with the degeneracy phenomenon arising in a framework of the Dubovitskii–Milutin version of Pontryagin's maximum principle proven in 1963 [29] for trajectories without any a priori assumptions (see also [1], [3]).

The first conditions which guarantee the nondegeneracy of this maximum principle for classical optimal control problem with state constraints were obtained in [16], [18], and [9]. A maximum principle with pointwise nontriviality conditions of the type as in Theorem 7 above was obtained in [16], [18] under a controllability assumption along the reference trajectory. The maximum principle with zero atomic components of the measure at the endpoints was obtained in [9] under the controllability condition at the endpoints similar to (4.1). These results showed the importance of assumptions of controllability and accordance of state and endpoint constraints for strengthened nontriviality conditions. We should note that the assumption of accordance for state and endpoint constraints were used but not formulated explicitly in [16].

The results obtained in [9], [16] were generalized and strengthened in [8], [10], [11], [14].

In [12], [13] the maximum principle with new time-transversality conditions was obtained. It was proved that this maximum principle holds with nontriviality condition (4.2) under controllability conditions at the endpoints similar to (4.1).

The results obtained in [12], [13] are close to the ones obtained in [8], [9], [10], but the methods used are essentially different. As in present paper a perturbation

technique is used in [12], [13] to derive a new version of the maximum principle without any controllability assumptions and then controllability conditions are used to prove the strengthened nontriviality conditions, whereas in [8], [9], [10] the controllability conditions are used from the beginning to obtain the nondegenerate maximum principle.

Nondegenerate first-order necessary optimality conditions for differential inclusions problem were obtained in [15] in the case when the support function of the right-hand side is smooth in x (these results are analogous to [12], [13]). Moreover, [15] contains the maximum principle with unbounded measures for trajectories which are weakly controllable at the endpoints. These results were generalized to a more natural class of problems with locally Lipschitz support function in [17] (this paper contains main technical devices used in present investigation).

The present paper summarizes and synthesizes the essence of previous works [8], [9], [10], [11], [12], [13], [14], [15], [16], [17], [18]. The principal distinction of the result obtained here is the following. The maximum principle for problems with state constraints is obtained with additional jump condition at the endtimes without any a priori assumptions. This condition is a new one (even in the smooth case of classical optimal control problem) and enables one to describe the degeneracy phenomenon.

The paper [30] deals with the degeneracy phenomenon for the classical optimal control problem on the fixed time interval $I = [t_1, t_2]$. The state constraint set is given by the time-dependent inequality $g(x,t) \le 0$, the left endpoint x_1 is fixed on the state boundary and the right endpoint constraint is given by the inequality. The right-hand side f of the control system satisfies (H4) and is measurable in t.

The main result of [30, Proposition 2.1] asserts that if there is an admissible control u' such that

$$\lim_{\varepsilon\to 0+} \operatorname*{ess\,sup}_{t_1\le s<t_1+\varepsilon} \left\langle \frac{\partial g}{\partial x}(x_1,t_1), (f(x_1,s,u'(s)) - f(x_1,s,u_0(s))) \right\rangle < 0, \tag{6.1}$$

then the maximum principle [2] holds with the following nontriviality condition:

$$\lambda_0 + \|\lambda_1\| + \left| \int_{(t_1,t_2]} d\eta \right| \ne 0. \tag{6.2}$$

Here u_0 is the optimal control; λ_0, λ_1 are the Lagrange multipliers corresponding to the functional and right endpoint inequality constraint, respectively; η is the measure arising in the maximum principle.

This assertion follows from [12], [13] (see Proposition), and also from Theorem 5 above if (H3) is valid and the function f is smooth in t. To our knowledge it is a new one in the measurable in t case.

Consider condition (6.1). If the function f is continuous in t uniformly in u, the state constraint is autonomous, and the left endpoint x_1 is fixed, then (6.1) is more restrictive than (4.1) for $i = 1$ (or similar controllability condition at the left endpoint introduced earlier in [9]). Indeed, (4.1) follows from (6.1) in this case. On the other hand (6.1) is an a priori assumption on the behavior of an optimal control u_0 near t_1 while (4.1) is an assumption on the controllability of the system at the endpoints and is the checkable one. Condition (6.1) is equivalent to (4.1) under the additional assumption that the reference trajectory lies on the state boundary in some right neighborhood $O^+_{t_1}$ of t_1 [30] or that the corresponding control u_0 is continuous on $O^+_{t_1}$.

Some normality conditions (normality means $\lambda_0 > 0$) are also suggested in [30]. Note that a sufficient condition for $\lambda_0 > 0$ has been obtained earlier by Maurer [31] by expressing a regularity condition in optimization in terms of control problems.

Acknowledgment. We are very grateful to the anonymous referees for helpful suggestions.

REFERENCES

[1] A. D. Ioffe and V. M. Tikhomirov, *Theory of Extremal Problems*, North–Holland, Amsterdam, 1979.
[2] F. H. Clarke, *Optimization and Nonsmooth Analysis*, John Wiley, New York, 1983.
[3] A. Ya. Dubovitskii and A. A. Milutin, *Extremal problems with constraints*, Zh. Vychisl. Mat. i Mat. Fiz., 5 (1965), pp. 395–453 (in Russian). English translation in U.S.S.R. Comput. Math. and Math. Phys., 5 (1965), pp. 1–80.
[4] P. D. Loewen and R. T. Rockafellar, *The adjoint arc in nonsmooth optimization*, Trans. Amer. Math. Soc., 325 (1991), pp. 39–72.
[5] P. D. Loewen and R. T. Rockafellar, *Optimal control of unbounded differential inclusions*, SIAM J. Control Optim., 32 (1994), pp. 442–470.
[6] G. V. Smirnov, *Extremal problems for differential inclusions with phase constraints*, Dokl. Akad. Nauk SSSR, 302 (1988), pp. 541–544 (in Russian).
[7] A. P. Aphanasyev, V. V. Dikusar, A. A. Milutin, and C. A. Chukanov, *Necessary conditions in optimal control*, Nauka, Moscow, 1990 (in Russian).
[8] A. Ya. Dubovitskii and V. A. Dubovitskii, *The principle of the maximum in regular optimal control problems in which the endpoints of the phase trajectory lie on the boundary of the phase constraint*, Avtomat. i Telemekh., 12 (1987), pp. 25–33 (in Russian).
[9] A. Ya. Dubovitskii and V. A. Dubovitskii, *Necessary conditions for a strong minimum in optimal control problems with degenerate endpoint constraints and phase constraints*, Uspekhi Mat. Nauk, 40 (1985), pp. 175–176 (in Russian). English translation in Russian Math. Surveys, 40 (1985).
[10] A. Ya. Dubovitskii and V. A. Dubovitskii, *Conditions for pointwise nontriviality of the maximum principle in a regular optimal control problem*, in Proc. Inst. Prikl. Mat. Tbilis. Gos. Univ., 27, Tbilisi, 1988, pp. 60–88 (in Russian).
[11] A. Ya. Dubovitskii and V. A. Dubovitskii, *The maximum principle for trajectories whose endpoints lie on the phase boundary*, Chernogolovka, 1988, preprint, (in Russian).
[12] A. V. Arutyunov, *On the theory of the maximum principle in optimal control problems with phase constraints*, Dokl. Akad. Nauk SSSR, 304 (1989), pp. 11–14 (in Russian). English translation in Soviet Math. Dokl., 39 (1989), pp. 1–4.
[13] A. V. Arutyunov, *First-order necessary conditions in a problem of optimal control with phase constraints*, in Proc. Inst. Prikl. Mat. Tbilis. Gos. Univ., 27, Tbilisi, 1988, pp. 46–58 (in Russian).
[14] A. V. Arutyunov, *Perturbations of extremal problems with constraints and necessary optimality conditions*, J. Soviet Math., 54 (1991), pp. 1342–1400.
[15] A. V. Arutyunov and V. I. Blagodatskikh, *The maximum principle for differential inclusions with phase constraints*, Trudy Mat. Inst. Steklov, 200 (1991) (in Russian). English translation in Proc. of the Steklov Institute of Mathematics, 2 (1993), pp. 3–25.
[16] A. V. Arutyunov and N. T. Tynyanskii, *The maximum principle in a problem with phase constraints*, Izv. Akad. Nauk SSSR Tekhn. Kibern., 4 (1984), pp. 60–68 (in Russian). English translation in Soviet J. Comput. Systems Sci., 23 (1985), pp. 28–35.
[17] A. V. Arutyunov, S. M. Aseev, and V. I.Blagodatskikh, *First order necessary conditions for optimal control problem for differential inclusion with state constraints*, Mat. Sb., 184 (1993), pp. 3–32 (in Russian). English translation in Russian Acad. Sci. Sb. Math., 79 (1994), pp. 117–139.
[18] A. V. Arutyunov, *On necessary conditions of optimality for problem with phase constraints*. Dokl. Akad. Nauk SSSR, 280 (1985), pp. 1033–1037 (in Russian). English translation in Soviet Math. Dokl., 31 (1985), pp. 174–177.
[19] A. V. Arutyunov and S. M. Aseev, *The maximum principle for optimal control problems with phase constraints. Nondegeneracy and stability*, Dokl. Russian Akad. Nauk, 334 (1994), pp. 134–137 (in Russian). English translation in Russian Acad. Sci. Dokl. Math., 49 (1994), pp. 38–42.

[20] A. V. Arutyunov and S. M. Aseev, *State constraints in optimal control. The degeneracy phenomenon*, Systems Control Lett., 26 (1995), pp. 267–273.
[21] B. Sh. Mordukhovich, *The maximum principle in the problem of time-optimal control with nonsmooth constraints*, Prikl. Mat. Mekh., 40 (1976), pp. 1014–1023 (in Russian). English translation in J. Appl. Math. Mech., 40 (1976), pp. 960–969.
[22] B. Sh. Mordukhovich, *Approximation methods in problems of optimization and control*, Nauka, Moscow, 1988 (in Russian).
[23] J. Warga, *Optimal Control of Differential and Functional Equations*, Academic Press, New York, 1972.
[24] S. M. Aseev, *Smooth Approximation of Differential Inclusions and Time-Optimal Problem*, Technical Report CRM–1610, Centre de Recherches Mathematiques, Université de Montreal, 1989.
[25] L. S. Pontryagin, V. G. Boltyanskii, R. V. Gamkrelidze, and E. F. Mischenko, *The Mathematical Theory of Optimal Processes*, John Wiley, New York, 1962.
[26] R. E. Edwards, *Functional Analysis*, Holt, Rinehart and Winston, New York, 1965.
[27] E. Michael, *Continuous selections*. I, Ann. of Math. (2), 63 (1956), pp. 361–382.
[28] R. V. Gamkrelidze, *Time–optimal processes with phase constraints*, Dokl. Akad. Nauk SSSR, 125 (1959), pp. 475–478 (in Russian).
[29] A. Ya. Dubovitskii and A. A. Milutin, *Extremal problems with constraints*, Dokl. Akad. Nauk SSSR, 149 (1963), pp. 759–762 (in Russian).
[30] M. M. A. Ferreira and R. B. Vinter, *When is the maximum principle for state constrained problems nondegenerate?* J. Math. Anal. Appl., 187 (1994), pp. 438–467.
[31] H. Maurer, *Differential stability in optimal control problems*, Appl. Math. Optim., 5 (1979), pp. 283–295.

SIAM J. CONTROL OPTIM.
Vol. 35, No. 3, pp. 953–979, May 1997

010

GENERALIZED CONTROLLED INVARIANCE FOR NONLINEAR SYSTEMS *

H. J. C. HUIJBERTS†, C. H. MOOG‡, AND R. ANDIARTI§

Abstract. A general setting is developed which describes controlled invariance for nonlinear control systems and which incorporates the previous approaches dealing with controlled invariant (co -) distributions. A special class of controlled invariant subspaces, called controllability cospaces, is introduced. These geometric notions are shown to be useful for deriving a (geometric) solution to the dynamic disturbance decoupling problem and for characterizing the so-called fixed dynamics for noninteracting control. These fixed dynamics are a central issue in studying noninteracting control with stability. The class of quasi-static state feedbacks is used.

Key words. nonlinear systems, controlled invariance, quasi-static state feedback

AMS subject classifications. 93C10, 93B27, 93C60, 93C35

PII. S0363012994277190

1. Introduction.

During the last two decades, nonlinear control theory was developed thanks to the increasing number of researchers involved in this area. A main goal of the research in the 1980s was the generalization of the so-called geometric approach which proved to be particularly efficient for linear time-invariant systems (see [48], [4] for an overview). In this linear theory, controlled invariance plays a fundamental role in both static and dynamic feedback control problems. The goal of generalizing the linear approach to the nonlinear case was only *partially* reached: the situation is quite well understood when regular static feedback synthesis problems are considered; limits of the standard (geometric) notions became clear at the end of the 1980s in the study of such problems as

- control problems involving dynamic feedback,
- the inversion of a nonlinear system, the definition of its rank, and so on.

Alternative (algebraic) tools have been developed from 1985 on [19] and a definition of the rank of a system was provided by a differential algebraic theory [20].

The goal of this paper is to introduce a generalized notion of controlled invariance. The motivation is to clarify the geometric structure of nonlinear systems and to develop a geometric framework to tackle synthesis problems via dynamic feedback. In particular, we answer the two following questions:

Question 1. Does there exist any *geometric* solution to the dynamic disturbance decoupling problem (DDDP)?

*Received by the editors August 17, 1995; accepted for publication (in revised form) April 1, 1996. This research was performed in part under the auspices of the GR Automatique of CNRS, France.
http://www.siam.org/journals/sicon/35-3/27719.html

†Department of Mathematics and Computing Science, Eindhoven University of Technology, P.O. Box 513, 5600 MB Eindhoven, the Netherlands (hjch@win.tue.nl). The research of this author was done partly while visiting the Laboratoire d'Automatique de Nantes with the financial support of the Region "Pays de la Loire," France.

‡Laboratoire d'Automatique de Nantes (URA CNRS 823), Ecole Centrale de Nantes–Université de Nantes, 1 rue de la Noë, BP 92101, 44321 Nantes Cedex 3, France (moog@lan.ec-nantes.fr). This research was performed in part while this author was a visiting Professor at Eindhoven University of Technology, supported by the Dutch Systems and Control Theory Network.

§LAPAN, Jl. Pemuda Persil 1, P.O. Box 20/JAT, Jakarta, Indonesia. The research of this author was performed while she was with the Laboratoire d'Automatique de Nantes, supported by LAPAN, Indonesia.

TABLE 1
(Geometric) solution to the DDP.

	Static feedback	Dynamic feedback
Linear systems	$\mathcal{E} \subset \mathcal{V}^*$	
Nonlinear systems	$\mathcal{P} \subset \Delta^*$	?

TABLE 2
Decoupling zero structure.

Feedback	Invertible decoupling matrix	Noninvertible decoupling matrix
(Quasi-) Static	dim $(\mathcal{P}^*)$ Isidori and Grizzle [32]	?
Dynamic	dim (Δ_{mix}) Wagner [47]	dim $(\Delta_{mix}(\Sigma_p))$ Zhan, Tarn, and Isidori [50]

Question 2. Does there exist a *geometric* structure of nonlinear systems which displays the rank, the so-called decoupling zeros (under dynamic feedback), and the like?

The answer to such questions is of major importance since these questions motivate the search for geometric solutions to any other synthesis problem which involves dynamic feedback. Such solutions will contribute to the completion of the extension to nonlinear systems of the linear geometric theory [48], [4].

DDDP, considered in Question 1, is a special control problem involving dynamic feedback and was first stated and studied in [26], [25], [44], where an (algebraic) solution was provided based on the inversion algorithm. Also, a geometric interpretation was given by using (nonintrinsic) standard controlled invariant (co-)distributions on an extended state space and then projecting these (co-)distributions on the original state space to obtain intrinsic objects. Parallel results can be found in [42], [41], [27]. The generalization of controlled invariance which is introduced in this paper is shown to give a *natural* geometric solution to the DDDP, i.e., without taking recourse to nonintrinsic objects defined on an extended state space that are rendered intrinsic after projection. Of course, it goes without saying that the objects defined in this paper and the geometric objects defined in [26], [25], [44] carry the same information concerning the solvability of the DDDP. Recall that in the special case of linear systems, DDDP is equivalent to DDP (static feedback disturbance decoupling problem). The state of the art is summarized in Table 1, where notations are borrowed from [48], [29], [38].

One contribution of the paper is the completion of Table 1.

Question 2 originated in [30]. Standard controlled invariant distributions cannot be used to characterize the rank of a system in a straightforward manner. The rank was introduced in [19] based on a differential algebraic analysis. A geometric interpretation of the rank may be found in [45], based on controllability distributions defined on a certain extended state space. Contributions which parallel the geometric and algebraic approaches can be found in [49].

Generalized controlled invariance introduced in this paper is shown to give a natural and intrinsic geometric characterization of the rank. It further displays new (geometric) structures of a nonlinear system. We focus on the structure related to the so-called decoupling zeros (under quasi-static feedback [12], [13], [14]). We summarize once again the state of the art in Table 2, and we borrow the notations from the given references.

TABLE 3
Controlled invariance.

Feedback	References
$u = \alpha(x) + v$	Brockett
$u = \alpha(x) + \beta(x)v$	Isidori et al. Hirschorn Nijmeijer and van der Schaft
$u = \alpha(x, v, \dot{v}, \ldots, v^{(k)})$	?

In this paper, Table 2 is completed thanks to the controllability cospaces introduced in what follows. Moreover, throughout the text, the new geometric structures are compared with the standard ones. Both embody different and complementary properties.

The study of controlled invariance for nonlinear systems of the form

$$\dot{x} = f(x) + g(x)u, \tag{1}$$

where $x \in \mathbb{R}^n$, $u \in \mathbb{R}^m$ was initiated in [8]. In this paper invariants were sought under feedback transformations of the form

$$u = \alpha(x) + v. \tag{2}$$

Later on, controlled invariance was tackled by various authors [31], [24], [36], [37]. The group of feedback transformations acting on (1) was enlarged to transformations of the form

$$u = \alpha(x) + \beta(x)v, \tag{3}$$

where $\beta(x)$ is square and locally invertible. These works yielded the definition of a controlled invariant distribution. The key was found for the solution of synthesis problems, such as the disturbance decoupling problem and the noninteracting control problem, via regular (or invertible) static state feedback (see [29], [38] for an overview). The study of controlled invariance under the class of feedbacks (3) remains active; see [10], [22], [11], [43], [49] for recent contributions. A special class of controlled invariant distributions is given by controllability distributions [39], [33], [34]. They became a basic tool for solving the noninteracting control problem with or without stability. Indeed, the controllability distributions allow us to characterize the fixed dynamics of the decoupled system via static feedback [32].

In this paper, a generalized notion of controlled invariance is introduced by allowing an enlarged class of feedback transformations acting on (1), namely the class of quasi-static feedbacks $u = \alpha(x, v, \dot{v}, \ldots, v^{(k)})$. This class of feedbacks describes intrinsic properties of the system with respect to the solvability of synthesis problems via dynamic feedback as disturbance decoupling or noninteracting control. In this sense, quasi-static feedbacks are considered a mathematical tool rather than a new class of feedbacks to be used in practical applications. The various contributions to the study of controlled invariance are summarized in Table 3. This table will be completed in this paper. Preliminary results can be found in [28].

Quasi-static feedback has been used in [40] to derive canonical forms (see also [46]) and was formalized in [12], [13], [14], where the input-output decoupling problem under quasi-static state feedback was solved as well. Practical applications of quasi-static feedback can be found in [16].

In what follows we consider a nonlinear control system (1), where the entries of $f(x)$ and $g(x)$ are meromorphic functions from $\mathbb{R}^n$ to $\mathbb{R}$. Recall that a meromorphic function is the quotient of two analytic functions. This allows us to derive properties of the system under consideration on an open and dense subset of the state space. Different classes of systems can also be treated:

- C^∞ systems, where all results should be explicitly stated as local results, valid around a *regular* point, where regularity is to be defined in an appropriate way, depending on the problem under consideration;
- analytic systems, in which case the results are also valid on some open dense submanifold of the state space.

In the rest of this paper we use mainly a function field formalism. It is assumed that rank $g(x) = m$ and that $n \geq 1$.

The organization of the paper is as follows. In section 2 we define the generalized notion of invariance with respect to the dynamics (1). Section 3 is devoted to controlled invariance and related properties. A geometric necessary and sufficient condition for the existence of a solution to DDDP is obtained. Controllability cospaces and their applications as well as the fixed modes or decoupling zero dynamics under quasi-static feedback are treated in section 4.

2. Invariant subspaces. We follow the notations and setting of [18]. Let $\mathcal{K}$ denote the field of meromorphic functions of $\{x, u^{(k)}, k \geq 0\}$. $\mathcal{E}$ is the formal vector space spanned by $\{d\eta \mid \eta \in \mathcal{K}\}$ over $\mathcal{K}$. The notation dx stands for $\{dx_1, \ldots, dx_n\}$ and $du^{(k)}$ for $\{du_1^{(k)}, \ldots, du_m^{(k)}\}$. Let $\mathcal{X} := \text{span}_{\mathcal{K}}\{dx\}$ and $\mathcal{U} := \text{span}_{\mathcal{K}}\{du, d\dot{u}, \ldots, du^{(k)} \mid k \geq 0\}$.

Throughout this paper we employ the following terminology. A vector $\omega \subset \mathcal{E}$ is called *exact* if there exists a $\phi \in \mathcal{K}$ such that $\omega = d\phi$. A subspace $\Omega \subset \mathcal{E}$ of dimension r is called *exact* if there exist functions $\phi_1, \ldots, \phi_r \in \mathcal{K}$ such that $\Omega = \text{span}_{\mathcal{K}}\{d\phi_1, \ldots, d\phi_r\}$. Given subspaces $\Omega_1 \subset \Omega_2 \subset \mathcal{E}$, (Ω_2/Ω_1) is said to be *exact* if there exist functions $\phi_1, \ldots, \phi_d \in \mathcal{E}$, with $d = \dim(\Omega_2) - \dim(\Omega_1)$, such that $\Omega_2 = \Omega_1 \oplus \text{span}_{\mathcal{K}}\{d\phi_1, \ldots, d\phi_d\}$, or, in other words, (Ω_2/Ω_1) is isomorphic to an exact subspace of $\mathcal{E}$. Consider a subspace $\Omega \subset \mathcal{E}$. Then clearly $\{0\} \subset \Omega$ is exact. Furthermore, if $\Omega_1 \subset \Omega$, $\Omega_2 \subset \Omega$ are exact, then $\Omega_1 + \Omega_2 \subset \Omega$ is also exact. Hence there exists a unique maximal exact subspace in Ω.

Consider a subspace $\Omega \subset \mathcal{X}$. Define

$$\dot{\Omega} = \text{span}_{\mathcal{K}}\{\dot{\omega} \mid \omega \in \Omega\}, \tag{4}$$

where $\omega = \sum_{i=1}^n \omega_i(x, u, \dot{u}, \ldots, u^{(n-1)})dx_i$ and time derivation is defined by

$$\dot{\omega} = \sum_{i=1}^n (\omega_i d\dot{x}_i + \dot{\omega}_i dx_i).$$

Thus $\dot{\omega} \in \text{span}_{\mathcal{K}}\{dx, du\}$.

DEFINITION 2.1. *A subspace $\Omega \subset \mathcal{X}$ is said to be invariant with respect to* (1) *if*

$$\dot{\Omega} \subset \Omega + \text{span}_{\mathcal{K}}\{du\}. \tag{5}$$

Remark 2.2. Let $\mathcal{K}_k$ be the field of meromorphic functions of $x, u, \ldots, u^{(k)}$ and define

$$\mathcal{K}' = \bigcup_{k \in \mathbb{N}} \mathcal{K}_k.$$

Then (5) is equivalent to the statement that $(\Omega + \text{span}_{\mathcal{K}'}\{du^{(k)} \mid k \geq 0\})$ is a differential vector space, with the derivation defined above.

Example 2.3. Let Ω be an integrable invariant codistribution for (1) in the sense of, e.g., [29], [38], and let (x_1, x_2) be a local system of coordinates such that $\Omega = \text{span}\{dx_1\}$. Then in the coordinates (x_1, x_2), (1) takes the form (cf. [29], [38])

$$\begin{aligned} \dot{x}_1 &= f_1(x_1) + g_1(x_1)u, \\ \dot{x}_2 &= f_2(x_1, x_2) + g_2(x_1, x_2)u. \end{aligned} \tag{6}$$

Interpreting Ω as a subspace of $\text{span}_{\mathcal{K}}\{dx\}$, we then obtain

$$\dot{\Omega} = \text{span}_{\mathcal{K}}\{d\dot{x}_1\} = \text{span}_{\mathcal{K}}\{d(f_1(x_1) + g_1(x_1)u)\} \subset \Omega + \text{span}_{\mathcal{K}}\{du\}. \tag{7}$$

Hence Ω is invariant in the sense of Definition 2.1.

When a given subspace is not invariant, it is interesting to know whether or not there exists a feedback transformation that renders it invariant. This is the topic of the next section.

3. Controlled invariant subspaces. In this section we define and characterize the controlled invariance of subspaces $\Omega \subset \mathcal{X}$ under quasi-static state feedback. In subsection 3.1 we first define quasi-static state feedback, based on [12], [13], [14]. In subsection 3.2 we give a definition of controlled invariance under quasi-static state feedback. In subsection 3.3 some properties of controlled invariance under regular static state feedback (3) are given. Conditions for controlled invariance of subspaces $\Omega \subset \mathcal{X}$ under quasi-static state feedback are investigated in subsection 3.4. We make some remarks about the smallest controlled invariant subspace containing some given subspace in subsection 3.4.2. As shown in section 3.5, this subspace allows us to characterize the solvability conditions of the DDDP.

3.1. Quasi-static state feedback. Consider the nonlinear system (1). A *generalized static state feedback* for (1) is a feedback of the form

$$u = \phi(x, v, \ldots, v^{(r)}), \tag{8}$$

where $v \in \mathbb{R}^m$ denotes the new controls. Let $\mathcal{K}_v$ denote the field of meromorphic functions of $\{x, \{v^{(k)} \mid k \geq 0\}\}$, and define the formal vector space $\mathcal{E}_v := \text{span}_{\mathcal{K}_v}\{d\xi \mid \xi \in \mathcal{K}_v\}$. As in [12], [13], we define the following *filtrations* [3] of $\mathcal{E}_v$:

$$\begin{aligned} \mathcal{V}_{-1} &:= \text{span}_{\mathcal{K}_v}\{dx\}, \\ \mathcal{V}_k &:= \text{span}_{\mathcal{K}_v}\{dx, dv, \ldots, dv^{(k)}\} \quad (k \geq 0), \end{aligned} \tag{9}$$

$$\begin{aligned} \mathcal{U}_{-1} &:= \text{span}_{\mathcal{K}_v}\{dx\}, \\ \mathcal{U}_k &:= \text{span}_{\mathcal{K}_v}\{dx, d\phi, \ldots, d\phi^{(k)}\} \quad (k \geq 0). \end{aligned} \tag{10}$$

The filtrations $\mathcal{U}_k$ and $\mathcal{V}_k$ are said to have *bounded difference* [3] if there exists an $s \in \mathbb{N}$ such that for all $k \geq -1$

$$\begin{aligned} \mathcal{U}_k &\subset \mathcal{V}_{k+s}, \\ \mathcal{V}_k &\subset \mathcal{U}_{k+s}. \end{aligned} \tag{11}$$

DEFINITION 3.1 ([12], [13], [14]). *u given by* (8) *is said to be a* quasi-static state feedback *for* (1) *if the filtrations $\mathcal{U}_k$ and $\mathcal{V}_k$ have bounded difference.*

Remark 3.2. It is easily verified that a regular static state feedback (3) is a quasi-static state feedback.

The following result is also easily proven.

PROPOSITION 3.3. *Let u given by* (8) *be a quasi-static state feedback. Then there locally exists a function $\psi(x, u, \dots, u^{(r)})$ such that*

$$v = \psi(x, u, \dots, u^{(r)}). \qquad \square \tag{12}$$

Remark 3.4. In [17] a definition of quasi-static state feedback is given for generalized systems (systems of the form $\dot{x} = f(x, u, \dot{u}, \dots, u^{(s)})$, $y = h(x, u, \dot{u}, \dots, u^{(s)})$). This definition is the same as Definition 3.1 with the extra requirement that x is a state (in the sense of [17]) of the closed-loop system.

3.2. Controlled invariance. Consider the control system (1) together with a quasi-static state feedback (8) and define $\mathcal{V} := \operatorname{span}_{\mathcal{K}_v}\{dv^{(k)} \mid k \geq 0\}$. We denote by $\Theta^{(k)}$ the time derivative of order k of Θ along the trajectories of the system (1) and by $\Theta^{[k]}$ the time derivative of order k of Θ along the trajectories of the closed-loop system (1), (8). We will write $\dot{\Theta}$ for $\Theta^{(1)}$.

DEFINITION 3.5. *A subspace $\Omega \subset \mathcal{X}$ is said to be controlled invariant for* (1) *if there exists a quasi-static state feedback* (8) *such that for* (1), (8) *one has*

$$\Omega^{[1]} \subset \Omega + \mathcal{V}. \tag{13}$$

The definition of controlled invariance given in Definition 3.5 is in accordance with the well-known definition of a controlled invariant codistribution. Recall from [29], [38], e.g., that a codistribution Ω is controlled invariant if there exists a regular static state feedback (3) such that

$$\begin{aligned} \mathcal{L}_{f+g\alpha}\Omega &\subset \Omega, \\ \mathcal{L}_{(g\beta)_{*i}}\Omega &\subset \Omega \quad (i = 1, \dots, m). \end{aligned} \tag{14}$$

Let $\omega \in \Omega$. Then for (1), (3) we have

$$\omega^{[1]} = \mathcal{L}_{f+g\alpha}\omega + \sum_{i=1}^{m}(v_i \mathcal{L}_{(g\beta)_{*i}}\omega + \langle \omega, (g\beta)_{*i}\rangle dv_i) \in \Omega + \mathcal{V} \tag{15}$$

when we interpret Ω as a subspace of $\operatorname{span}_{\mathcal{K}}\{dx\}$.

Example 3.6. Consider a nonlinear system given by

$$\dot{x}_1 = u_1, \quad \dot{x}_2 = x_3 u_1 + x_2, \quad \dot{x}_3 = u_2.$$

Let $\Omega = \operatorname{span}_{\mathcal{K}}\{u_1 dx_3 + dx_2\}$ and

$$u_1 = v_1, \quad u_2 = (v_2 - x_3(\dot{v}_1 + v_1) - x_2)/v_1,$$

where $v = (v_1, v_2)^T$ is the new input. This is a quasi-static feedback since

$$v_1 = u_1, \quad v_2 = u_1 u_2 + x_3(\dot{u}_1 + u_1) + x_2.$$

This feedback renders Ω invariant, since we have

$$\Omega^{[1]} = \operatorname{span}_{\mathcal{K}_v}\{\dot{v}_1 dx_3 + v_1 d((v_2 - x_3(\dot{v}_1 + v_1) - x_2)/v_1) + d(x_3 v_1 + x_2)\} \subset \mathcal{V}.$$

The following theorem gives a necessary condition for controlled invariance. For (1), let $\mathcal{G}$ denote the distribution spanned by the input vector fields. Define the subspace $\mathcal{G}^\perp \subset \mathcal{X}$ by

$$\mathcal{G}^\perp = \{\omega \in \mathcal{X} \mid \langle \omega, g \rangle \equiv 0 \ \forall g \in \mathcal{G}\}. \tag{16}$$

THEOREM 3.7. *Let $\Omega \subset \mathcal{X}$. Then Ω is controlled invariant only if*

$$\widehat{(\dot{\Omega \cap \mathcal{G}^\perp})} \subset \Omega. \tag{17}$$

Proof. By definition of $\mathcal{G}^\perp$, $\widehat{(\dot{\Omega \cap \mathcal{G}^\perp})} \subset \mathcal{X}$. Controlled invariance of Ω then implies (17). □

Remark 3.8. Let Ω be an integrable codistribution. Using (15), it may be shown that (17) (with Ω interpreted as a subspace of $\mathcal{X}$) is equivalent to the well-known conditions $\mathcal{L}_f(\Omega \cap \mathcal{G}^\perp) \subset \Omega$, $\mathcal{L}_{g_i}(\Omega \cap \mathcal{G}^\perp) \subset \Omega \, (i = 1, \ldots, m)$ for controlled invariance of Ω (cf. [29], [38]).

3.3. Characterization of controlled invariant subspaces under regular static state feedback. In this subsection we investigate under what conditions a subspace $\Omega \subset \mathcal{X}$ is controlled invariant under regular static state feedback. Recall from subsection 3.2 that a regular static state feedback is a special sort of quasi-static state feedback. A first result is the following.

PROPOSITION 3.9. *Consider a d-dimensional subspace $\Omega \subset \mathcal{X}$. Assume that Ω is controlled invariant under a quasi-static state feedback of the form $u = \phi(x, v)$. Then Ω admits a basis $\omega_1, \ldots, \omega_d$ with*

$$\omega_i = \sum_{j=1}^{n} \omega_{ij}(x) dx_j. \tag{18}$$

Proof. Assume that $\Omega = \operatorname{span}_{\mathcal{K}}\{\tilde{\omega}_1, \ldots, \tilde{\omega}_d\}$, with

$$\tilde{\omega}_i = \sum_{j=1}^{n} \tilde{\omega}_{ij}(x, u) dx_j \qquad \square \tag{19}$$

Let $A(x, u)$ be the matrix with entries $\tilde{\omega}_{ij}$ $(i = 1, \ldots, d; j = 1, \ldots, n)$. Viewing Ω as a linear subspace (over $\mathcal{K}$) of $\mathcal{X} \oplus \operatorname{span}_{\mathcal{K}}\{du\}$, it may be characterized by

$$\Omega = \operatorname{rowspan}_{\mathcal{K}} \begin{pmatrix} A(x,u) & 0 \end{pmatrix}. \tag{20}$$

Similarly, $\Omega + \dot{\Omega}$ is characterized by

$$\Omega + \dot{\Omega} = \operatorname{rowspan}_{\mathcal{K}} \begin{pmatrix} A(x,u) & 0 \\ B(x,u,\dot{u}) & (Ag)(x,u) \end{pmatrix}, \tag{21}$$

where

$$B(x,u,\dot{u}) = \sum_{i=1}^{n} \frac{\partial A}{\partial x_i}(x,u)\dot{x}_i(x,u) + \sum_{j=1}^{m} \frac{\partial A}{\partial u_j}\dot{u}_j + A(x,u)\left(f_x(x) + \sum_{i=1}^{n} \frac{\partial g}{\partial x_i} u\right), \tag{22}$$

with f_x the Jacobian of f. Since Ω is rendered invariant via $u = \phi(x, v)$ there exist matrices $P(x, v, \dot{v})$ and $Q(x, v)$ such that

$$B(x, \phi, \dot{\phi})dx + (Ag)(x, \phi)d\phi = P(x, v, \dot{v})A(x, \phi)dx + Q(x, v)dv \tag{23}$$

or, equivalently,

$$\begin{aligned} B(x,\phi,\dot{\phi}) &= P(x,v,\dot{v})A(x,\phi) - (Ag)(x,\phi)\phi_x(x,v), \\ (Ag)(x,\phi)\phi_v(x,v) &= Q(x,v). \end{aligned} \tag{24}$$

Since $\phi_v(x,v)$ is invertible, this yields

$$B(x,\phi,\dot{\phi}) = P(x,v,\dot{v})A(x,\phi) - Q(x,v)\phi_v(x,v)^{-1}\phi_x(x,v). \tag{25}$$

Since $u = \phi(x,v)$ is a quasi-static state feedback, by Proposition 3.3 there locally exists a function $\psi(x,u)$ such that $\phi(x,\psi(x,u)) = u$. This yields in particular that

$$\psi_x(x,u) = -\phi_v(x,\psi(x,u))^{-1}\phi_x(x,\psi(x,u)).$$

Hence (25) yields

$$B(x,u,\dot{u}) = \tilde{P}(x,u,\dot{u})A(x,u) + \tilde{Q}(x,u)\psi_x(x,u), \tag{26}$$

where $\tilde{P}(x,u,\dot{u}) = P(x,\psi(x,u),\dot{\psi}(x,u,\dot{u}))$ and $\tilde{Q}(x,u) = Q(x,\psi(x,u))$. Taking partial derivatives with respect to $\dot{u}_i$, we obtain

$$\frac{\partial A}{\partial u_i} = \frac{\partial \tilde{P}}{\partial \dot{u}_i}A(x,u) \quad (i = 1,\dots,m). \tag{27}$$

Obviously,

$$\frac{\partial^2 \tilde{P}}{\partial \dot{u}_i \partial \dot{u}_j} = 0 \quad (i,j = 1,\dots,m).$$

Hence there exist matrices $R_i(x,u)$ $(i = 1,\dots,m)$ such that

$$\frac{\partial A}{\partial u_i} = R_i(x,u)A(x,u). \tag{28}$$

Using arguments from the theory of linear time-varying ordinary differential equations this yields that $A(x,u)$ is of the form

$$A(x,u) = \Phi(x,u)\Psi(x),$$

where $\Phi(x,u)$ is a square invertible matrix. Hence

$$\Omega = \mathrm{rowspan}_{\mathcal{K}}\,(A(x,u)\ \ 0) = \mathrm{rowspan}_{\mathcal{K}}\,(\Psi(x)\ \ 0)\,, \tag{29}$$

which establishes our claim. If $\Omega = \mathrm{rowspan}_{\mathcal{K}}\left(A(x,u,\dots,u^{(\ell)})\ \ 0\right)$ with $\ell > 1$, the claim is established by using the same arguments together with an induction argument. □

From the above proposition it follows that the set of subspaces $\Omega \subset \mathcal{X}$ that are controlled invariant under a quasi-static state feedback $u = \phi(x,v)$ may be identified with the set of "classical" controlled invariant codistributions. The following theorem gives a characterization of controlled invariance in our generalized framework.

THEOREM 3.10. *Let $\Omega \subset \mathcal{X}$ be a subspace such that*

$$(\Omega + \dot{\Omega})/\Omega \text{ is exact} \tag{30}$$

and admits a basis satisfying (18). *Then* Ω *is controlled invariant under a quasi-static state feedback* $u = \phi(x, v)$ *if and only if*

$$\widehat{(\dot{\Omega \cap \mathcal{G}^\perp})} \subset \Omega. \tag{31}$$

Moreover, if the conditions above are satisfied, then $\phi(x, v)$ *rendering* Ω *invariant may be chosen of the form* (3).

Proof. The necessity was proven in Theorem 3.7. To establish the sufficiency, assume that (31) holds. Note that $\Omega + \dot{\Omega} \subset \operatorname{span}_{\mathcal{K}}\{dx, du\}$. Let $\tilde{\Omega} \subset \mathcal{X}$ be such that $\Omega = (\Omega \cap \mathcal{G}^\perp) \oplus \tilde{\Omega}$. Assume that $\dot{\tilde{\Omega}} \cap \mathcal{X} \neq \{0\}$. This implies that there is an $\tilde{\omega} \in \tilde{\Omega}$, $\tilde{\omega} \neq 0$, such that $\dot{\tilde{\omega}} \in \mathcal{X}$ and hence $\tilde{\omega} \in (\Omega \cap \mathcal{G}^\perp)$, which gives a contradiction. Thus

$$\dot{\tilde{\Omega}} \cap \mathcal{X} = \{0\}. \tag{32}$$

By (30), there exists $v_1(x, u)$ such that

$$\Omega + \dot{\Omega} = \Omega \oplus \operatorname{span}_{\mathcal{K}}\{dv_1\}. \tag{33}$$

Since (31) and (32) hold, we must have that $(\partial v_1/\partial u)$ has full row rank. Then there exists a function $v_2(u)$ such that $(\partial v/\partial u)$ is square and invertible, where $v = (v_1^T\ v_2^T)^T$. By (33) we now have that

$$\Omega^{[1]} \subset \Omega + \mathcal{V}. \tag{34}$$

Moreover, since $(\partial v/\partial u)$ is invertible, there exists a $\psi(x, v)$ such that $u = \psi(x, v)$. Hence ψ defines a quasi-static state feedback and thus Ω can be rendered invariant via quasi-static state feedback. Since we are dealing with a control system (1) that is affine in u, it is easily seen that v can be taken affine in u and thus ψ can be taken affine in v. This implies that Ω can be rendered invariant via a static state feedback (3). □

Remark 3.11.

(i) If Ω is exact, then clearly also $(\Omega + \dot{\Omega})/\Omega$ is exact. Hence the set of subspaces $\Omega \subset \mathcal{X}$ such that $(\Omega + \dot{\Omega})/\Omega$ is exact incorporates the standard integrable codistributions.

(ii) The exactness of $(\Omega + \dot{\Omega})/\Omega$ is not necessary for controlled invariance. This can be seen from the following counter example. Take the system $\dot{x}_1 = u_1$, $\dot{x}_2 = u_2$, $\dot{x}_3 = 0$, and $\Omega = \operatorname{span}_{\mathcal{K}}\{dx_1 + x_2 dx_3\}$. It is straightforward to check that $\widehat{(\dot{\Omega \cap \mathcal{G}^\perp})} \subset \Omega$ and that $(\Omega + \dot{\Omega})/\Omega$ is not exact. However, with the regular static state feedback $u_1 = v_1 - x_3 v_2$, $u_2 = v_2$ we obtain

$$\dot{\Omega} = \operatorname{span}_{\mathcal{K}}\{dv_1 - x_3 dv_2\} \subset \Omega + \mathcal{V}$$

and hence Ω is controlled invariant.

3.4. Some characterizations of controlled invariance. In this subsection, conditions are derived for controlled invariance of a subspace under a quasi-static state feedback.

3.4.1. The general case: A sufficient condition. Let us consider a general subspace $\Omega \subset \mathcal{X}$. Define by induction

$$\begin{aligned} \hat{\Omega}_0 &:= 0, \\ \Omega_0 &:= \Omega, \\ \hat{\Omega}_{k+1} &:= \text{maximal exact subspace in } \frac{\Omega_k + \dot{\Omega}_k}{\Omega_k}, \\ \Omega_{k+1} &:= \Omega_k + \hat{\Omega}_{k+1}. \end{aligned}$$

Furthermore, define

$$k^* := \max\{k \geq 1 \mid \dim(\hat{\Omega}_k) > \dim(\hat{\Omega}_{k-1})\}.$$

THEOREM 3.12. *Let $\Omega \subset \mathcal{X}$. If*

(i) $(\widehat{\dot{\Omega \cap \mathcal{G}^\perp}}) \subset \Omega$,

(ii) $\dfrac{\Omega_{k^*-1} + \dot{\Omega}_{k^*-1}}{\Omega_{k^*-1}}$ *is exact,*

then Ω is controlled invariant for (1).

Proof. From the definition of k^* there exist vector-valued $dv_1, \ldots, dv_{k^*}$ in $\mathcal{E}$, where each dv_i is nonempty, such that

$$\begin{aligned}
\hat{\Omega}_1 &= \operatorname{span}_{\mathcal{K}}\{dv_1\} \subset \frac{\Omega_0 + \dot{\Omega}_0}{\Omega_0}, \\
\hat{\Omega}_2 &= \operatorname{span}_{\mathcal{K}}\{d\dot{v}_1, dv_2\} \subset \frac{\Omega_1 + \dot{\Omega}_1}{\Omega_1}, \\
&\vdots \\
\hat{\Omega}_{k^*} &= \operatorname{span}_{\mathcal{K}}\{dv_1^{(k^*-1)}, dv_2^{(k^*-2)}, \ldots, dv_{k^*}\} \subset \frac{\Omega_{k^*-1} + \dot{\Omega}_{k^*-1}}{\Omega_{k^*-1}}.
\end{aligned} \tag{35}$$

Note that from (ii) the last inclusion in (35) is in fact an equality. We now have

$$\begin{aligned}
\dot{\Omega} &\subset \Omega_0 + \hat{\Omega}_1 + \dot{\Omega}_0 + \dot{\hat{\Omega}}_1 = \Omega_1 + \dot{\Omega}_1 \subset \cdots \\
&\subset \Omega_{k^*-1} + \dot{\Omega}_{k^*-1} = \Omega_{k^*-1} + \operatorname{span}_{\mathcal{K}}\{dv_1^{(k^*-1)}, \ldots, dv_{k^*}\} \\
&\subset \Omega + \operatorname{span}_{\mathcal{K}}\{dv^{(k)} \mid k \geq 0\}.
\end{aligned} \tag{36}$$

It remains to be shown that v defines a quasi-static state feedback. From the above construction, one has

$$\begin{aligned}
v_1 &= \phi_1(x, u), \\
v_2 &= \phi_2(x, v_1, \dot{v}_1, u), \\
&\vdots \\
v_{k^*} &= \phi_{k^*}(x, \{v_i^{(j)} \mid 1 \leq i \leq k^* - 1, 0 \leq j \leq k^* - i\}, u).
\end{aligned} \tag{37}$$

From (i), $(\partial(\phi_1, \ldots, \phi_{k^*})/\partial u)$ has full row rank on an open and dense subset of $\mathbb{R}^n \times \mathbb{R}^{(k^*-1)(k^*-i+1)} \times \mathbb{R}^m$. By the implicit function theorem, for every point of this open and dense subset there exists a neighborhood of this point and a function ψ such that $u = \psi(x, v, \dot{v}, \ldots, v^{(k^*)})$. By applying this feedback, one has

$$\Omega^{[1]} \subset \Omega + \operatorname{span}_{\mathcal{K}}\{dv^{(k)} \mid k \geq 0\}. \qquad \square$$

Remark 3.13. Theorem 3.12 gives only sufficient conditions for the controlled invariance of a subspace $\Omega \subset \mathcal{X}$. In Theorem 3.7 it was shown that (i) is also a necessary condition. But the condition (ii) is not. This is shown by the following example.

Example 3.14 (see [30]). We consider a nonlinear system on $\mathbb{R}^4$ with three inputs u_1, u_2, u_3 given by

$$\dot{x}_1 = u_1, \quad \dot{x}_2 = x_4 + u_2, \quad \dot{x}_3 = x_3 u_1 + u_2, \quad \dot{x}_4 = u_3.$$

Let $\Omega = \text{span}_{\mathcal{K}}\{dx_1 - u_1 dx_3, dx_4\}$. Then Ω is not exact, and $\dot{\Omega}$ is given by

$$\dot{\Omega} = \text{span}_{\mathcal{K}}\{(1 - u_1 x_1)du_1 - u_1 du_2 - \dot{u}_1 dx_3 - u_1^2 dx_1, du_3\}.$$

Ω is rendered invariant by $u_1 = v_1$, $u_2 = -\frac{\dot{v}_1}{v_1^2}x_1 - v_1 x_1 + v_2$, and $u_3 = v_3$. One obtains $k^* = 1$, but $\frac{\Omega + \dot{\Omega}}{\Omega}$ is not exact.

3.4.2. The smallest controlled invariant subspace containing a given subspace. Given a subspace $\Pi \subset \mathcal{X}$, it is unclear whether (or under what conditions) there exists a smallest controlled invariant subspace containing Π. This is due to the fact that for two controlled invariant subspaces $\Omega_1, \Omega_2 \subset \mathcal{X}$, we do not necessarily have that $\Omega_1 \cap \Omega_2$ is controlled invariant, so that we cannot use the "standard" arguments (as in, e.g., [48], [29], [38]). In this subsection we will give some comments on this question.

We will use the following notation. Given a subspace $\Pi \subset \mathcal{X}$, we define

$$\Pi_* := \mathcal{X} \cap (\Pi + \Pi^{(1)} + \cdots + \Pi^{(n-1)}). \tag{38}$$

In what follows, we will need the following lemma.

LEMMA 3.15. *Consider a subspace $\Omega \subset \mathcal{X}$ satisfying $(\Omega \cap \mathcal{G}^{\perp}) = \{0\}$. Then we have for all $k \in \mathbb{N}$:*

$$\mathcal{X} \cap (\Omega^{(1)} + \cdots + \Omega^{(k)}) = \{0\}. \tag{39}$$

Proof. Let $d := \dim(\Omega)$, and let $\omega_1, \ldots, \omega_d$ be a basis of Ω, with

$$\omega_i = \sum_{j=1}^{n} \omega_{ij}(x, u, \ldots, u^{(r)})dx_j \quad (i = 1, \ldots, d). \tag{40}$$

Let $A(x, u, \ldots, u^{(r)})$ be the (d, n)-matrix with entries ω_{ij} $(i = 1, \ldots, d; j = 1, \ldots, n)$. Since $\omega_1, \ldots, \omega_d$ forms a basis of Ω, the matrix A has full row rank over $\mathcal{K}$. We may now characterize Ω by

$$\Omega = \text{rowspan}_{\mathcal{K}}(A(x, u, \ldots, u^{(r)})\ 0\ \cdots\ 0), \tag{41}$$

while $\Omega^{(k)}$ $(k = 1, 2, \ldots)$ may be characterized by

$$\Omega^{(k)} = \text{rowspan}_{\mathcal{K}}(X_{k0}\ X_{k1}\ \cdots\ X_{kk-1}\ (Ag)\ 0\ \cdots\ 0) \tag{42}$$

for certain matrices $X_{k0}, \ldots, X_{kk-1}$. Now assume that (Ag) is not right invertible over $\mathcal{K}$. This implies that there exists a nonzero row vector $\eta^T := (\eta_1 \cdots \eta_d)$ such that

$$\eta^T(Ag) = 0. \tag{43}$$

This gives that $\omega := \sum_{j=1}^{d} \eta_j \omega_j$ satisfies

$$\langle \omega, \tau \rangle = 0 \quad (\forall \tau \in \mathcal{G}), \tag{44}$$

which contradicts the fact that $(\Omega \cap \mathcal{G}^{\perp}) = \{0\}$. Hence we have that (Ag) is right invertible over $\mathcal{K}$. Next, let $\omega \in \mathcal{X} \cap (\Omega^{(1)} + \cdots + \Omega^{(k)})$ $(k \in \{1, 2, \ldots\})$. Since $\omega \in (\Omega^{(1)} + \cdots + \Omega^{(k)})$, we may represent ω by a row vector

$$(\eta_1^T\ \cdots\ \eta_k^T) \begin{pmatrix} X_{10} & (Ag) & 0 & \cdots & \cdots & 0 \\ X_{20} & X_{21} & (Ag) & \cdots & \cdots & 0 \\ \vdots & \vdots & \vdots & \ddots & & \vdots \\ \vdots & \vdots & \vdots & & \ddots & \vdots \\ X_{k0} & X_{k1} & X_{k2} & \cdots & X_{kk-1} & (Ag) \end{pmatrix}.$$

The fact that $\omega \in \mathcal{X}$ implies that necessarily

$$(\eta_1^T \ \cdots \ \eta_k^T) \begin{pmatrix} (Ag) & 0 & 0 & \cdots & \cdots & 0 \\ X_{21} & (Ag) & 0 & \cdots & \cdots & 0 \\ \vdots & \vdots & \vdots & \ddots & & \vdots \\ \vdots & \vdots & \vdots & & \ddots & \vdots \\ X_{k1} & X_{k2} & X_{k3} & \cdots & X_{kk-1} & (Ag) \end{pmatrix} = 0,$$

and thus

$$\eta_i^T (Ag) = 0,$$

which give $\eta_i^T = 0$, since (Ag) is right invertible. Thus $\omega = 0$, which establishes our claim. □

PROPOSITION 3.16. *Let $\Omega \subset \mathcal{X}$ be a subspace satisfying $(\widehat{\Omega \cap \mathcal{G}^\perp}) \subset \Omega$. Then*

$$\Omega_* = \Omega.$$

Proof. Let $\tilde{\Omega}$ be such that

$$\Omega = (\Omega \cap \mathcal{G}^\perp) \oplus \tilde{\Omega}. \tag{45}$$

By hypothesis we have

$$(\widehat{\Omega \cap \mathcal{G}^\perp}) \subset \Omega. \tag{46}$$

We now prove by induction that we have

$$(\Omega \cap \mathcal{G}^\perp)^{(k)} \subset \Omega + \tilde{\Omega}^{(1)} + \cdots + \tilde{\Omega}^{(k-1)} \quad (k = 1, 2, \ldots). \tag{47}$$

By (46), we have that (47) holds for $k = 1$. Next assume that (47) holds for $k = 1, \ldots, \ell - 1$. Then

$$\begin{aligned} (\Omega \cap \mathcal{G}^\perp)^{(\ell)} &= ((\Omega \cap \mathcal{G}^\perp)^{(\ell-1)})^{(1)} \overset{\text{IH}}{\subset} (\Omega + \tilde{\Omega}^{(1)} + \cdots + \tilde{\Omega}^{(\ell-2)})^{(1)} \\ &= (\Omega^{(1)} + \tilde{\Omega}^{(2)} + \cdots + \tilde{\Omega}^{(\ell-1)}) \overset{(45)}{=} ((\widehat{\Omega \cap \mathcal{G}^\perp}) + \tilde{\Omega}^{(1)} + \cdots + \tilde{\Omega}^{(\ell-1)}) \\ &\overset{(46)}{\subset} (\Omega + \tilde{\Omega}^{(1)} + \cdots + \tilde{\Omega}^{(\ell-1)}), \end{aligned}$$

which establishes (47). Using (47) and the modular distributive rule (see, e.g., [48, section 0.3]) we obtain

$$\begin{aligned} \Omega_* &= \mathcal{X} \cap (\Omega + \Omega^{(1)} + \cdots + \Omega^{(n-1)}) \\ &= \mathcal{X} \cap (\Omega + (\Omega \cap \mathcal{G}^\perp)^{(1)} + \tilde{\Omega}^{(1)} + \cdots + (\Omega \cap \mathcal{G}^\perp)^{(n-1)} + \tilde{\Omega}^{(n-1)}) \\ &\subset \mathcal{X} \cap (\Omega + \tilde{\Omega}^{(1)} + \cdots + \tilde{\Omega}^{(n-1)}) = \Omega + \mathcal{X} \cap (\tilde{\Omega}^{(1)} + \cdots + \tilde{\Omega}^{(n-1)}). \end{aligned} \tag{48}$$

Since by definition of $\tilde{\Omega}$ we have that $(\tilde{\Omega} \cap \mathcal{G}^\perp) = \{0\}$, we obtain from (48) and Lemma 3.15 that

$$\Omega_* \subset \Omega. \tag{49}$$

Furthermore, we have by definition of Ω_* that

$$\Omega \subset \Omega_*. \tag{50}$$

Hence we have that $\Omega_* = \Omega$, which establishes our claim. □

COROLLARY 3.17. *Consider a subspace $\Pi \subset \mathcal{X}$, and let $\Omega \subset \mathcal{X}$ be a controlled invariant subspace containing Π. Then $\Pi_* \subset \Omega$.*

Proof. Using the definition of Π_*, the fact that $\Pi \subset \Omega$ and combining the results of Theorem 3.7 and Proposition 3.16, we obtain

$$\Pi_* = \mathcal{X} \cap (\Pi + \Pi^{(1)} + \cdots + \Pi^{(n-1)}) \subset \mathcal{X} \cap (\Omega + \Omega^{(1)} + \cdots + \Omega^{(n-1)}) = \Omega_* = \Omega,$$

which establishes our claim. □

The subspace Π_* defined in (38) is, by Corollary 3.17, a candidate for being the smallest controlled invariant subspace containing Π. If Π is exact, it can be shown that indeed it *is*. This may be shown in the following way. Let $r = \dim \Pi$ and choose meromorphic functions $h_1(x), \ldots, h_r(x)$ such that $\Pi = \operatorname{span}_{\mathcal{K}}\{dh_1, \ldots, dh_r\}$. Next consider the system

$$\begin{aligned} \dot{x} &= f(x) + g(x)u, \\ y &= h(x). \end{aligned} \tag{51}$$

Then for this system, $\Pi_* = \mathcal{X} \cap \mathcal{Y}$, where $\mathcal{Y} = \operatorname{span}_{\mathcal{K}}\{dy, \ldots, dy^{(n-1)}\}$. (The subspace $\mathcal{X} \cap \mathcal{Y}$ was introduced in [9] for the study of the minimal order input-output decoupling problem.) If the system (51) is right invertible, one can construct a quasi-static state feedback which renders Π_* invariant by using the construction in [41]. If (51) is not right invertible, the same construction, together with Lemma 1 from [35], may be used to show that Π_* is controlled invariant. Summarizing, we have the following result.

THEOREM 3.18. *Consider a subspace $\Pi \subset \mathcal{X}$ which is exact. Then $\Pi_* := \mathcal{X} \cap (\Pi + \cdots + \Pi^{(n-1)})$ is the smallest controlled invariant subspace containing Π.* □

An application of the subspace $\Omega_* = \mathcal{X} \cap \mathcal{Y}$ is given in Section 3.5, where we consider the DDDP.

It has been shown that a "standard" controlled invariant codistribution is a controlled invariant subspace in the sense of Definition 3.5. If Δ^* denotes the largest controlled invariant distribution contained in $\ker dh$, then $\Delta^{*^\perp} \cap \mathcal{X}$ is a controlled invariant subspace containing the differential of the output. Since $\Omega_* = \mathcal{X} \cap \mathcal{Y}$ is the smallest controlled invariant subspace containing $\operatorname{span}_{\mathcal{K}}\{dy\}$, one has $\Delta^{*^\perp} \cap \mathcal{X} \supset \Omega_*$. These two different geometric structures are displayed in the lattice diagram in Figure 1.

In the special case of linear systems, this lattice diagram is simplified since $\Delta^{*^\perp} \cap \mathcal{X} = \Omega_*$.

3.4.3. A special case. Let us consider a subspace $\Omega \subset \mathcal{X}$ such that

$$\Omega = \Omega \cap \mathcal{G}^\perp + \Phi_*, \tag{52}$$

where Φ is an exact subspace of $\mathcal{X}$.

PROPOSITION 3.19. *Let $\Omega \subset \mathcal{X}$ satisfy* (52); *then Ω is controlled invariant if and only if*

$$\widehat{(\Omega \cap \mathcal{G}^\perp)^{\cdot}} \subset \Omega. \tag{53}$$

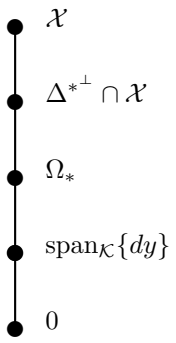


FIG. 1. *Lattice diagram: (geometric) structure of nonlinear systems.*

Proof. By Theorem 3.7 we only need to show the sufficiency. Clearly Φ_* is controlled invariant (see Theorem 3.18). Hence there exists a quasi-static feedback (8) such that

$$\Phi_*^{[1]} \subset \Phi_* + \mathcal{V}.$$

Now (53) implies that

$$\Omega^{[1]} \subset \Omega + \mathcal{V},$$

and hence Ω is controlled invariant. □

The following proposition gives conditions for the existence of a subspace $\Phi \subset \mathcal{X}$ such that (52) holds.

PROPOSITION 3.20. *Let $\Omega \subset \mathcal{X}$ be a subspace such that* (53) *holds. Then there exists an exact subspace $\Phi \subset \Omega$ satisfying* (52) *if and only if*

$$\Omega = \Omega \cap \mathcal{G}^\perp + \hat{\Phi}_*, \tag{54}$$

where $\hat{\Phi}$ is the largest exact subspace in Ω.

Proof. Assume that (54) holds. Taking $\Phi = \hat{\Phi}$, we then have (52). Conversely, assume that there exists an exact subspace $\Phi \subset \mathcal{X}$ such that (52) holds. Clearly $\Phi_* \subset \hat{\Phi}_*$. Now $\hat{\Phi} \subset \Omega$ implies by Proposition 3.16 that $\hat{\Phi}_* \subset \Omega$. Thus

$$\Omega = \Omega \cap \mathcal{G}^\perp + \Phi_* \subset \Omega \cap \mathcal{G}^\perp + \hat{\Phi}_* \subset \Omega.$$

Hence (54) is verified. □

3.5. Dynamic disturbance decoupling. A fundamental application of controlled invariance is disturbance decoupling [48], [29], [38]. In this section, generalized controlled invariance is shown to yield a geometric condition that characterizes the solvability of the dynamic feedback disturbance decoupling problem (DDDP). DDDP is stated as follows.

Consider a perturbed system Σ_q given by

$$\Sigma_q : \begin{cases} \dot{x} &= f(x) + g(x)u + p(x)q, \\ y &= h(x), \end{cases} \tag{55}$$

where q represents a disturbance. Find, if possible, a dynamic state feedback such that the disturbance q does not affect the output y.

Let $\mathcal{P}$ denote the distribution spanned by the disturbance vector fields. Define the subspace $\mathcal{P}^\perp$ by

$$\mathcal{P}^\perp = \{\omega \in \mathcal{X} \mid \langle \omega, p \rangle = 0 \ \forall p \in \mathcal{P}\}. \tag{56}$$

The following result gives a necessary and sufficient condition for DDDP to be solvable.

THEOREM 3.21. *DDDP is solvable if and only if there exists a controlled invariant subspace Ω such that*

$$\mathrm{span}_{\mathcal{K}}\{dy\} \subset \Omega \subset \mathcal{P}^\perp. \tag{57}$$

Proof. From Theorem 2.3 in [41] it follows that DDDP is solvable if and only if it is solvable by quasi-static state feedback. Thus, to prove Theorem 3.21, it suffices to show that the DDP is solvable by quasi-static state feedback.

Sufficiency. Controlled invariance of Ω implies that there exists a quasi-static state feedback $u = \phi(x, v, \ldots, v^{(r)})$ such that

$$\Omega^{[1]} \subset \Omega + \mathcal{V}. \tag{58}$$

By (57) and (58), one has

$$dy^{[k]} \subset \Omega + \mathcal{V} \quad \forall k \geq 0. \tag{59}$$

Thus in the closed loop system the output y is decoupled from the disturbance.

Necessity. Suppose that the quasi-static state feedback $u = \phi(x, v, \ldots, v^{(r)})$ solves the DDP. Then for the system Σ_q fed back with $u = \phi(x, v, \ldots, v^{(r)})$, one has

$$dy^{[k]} \subset \mathrm{span}_{\mathcal{K}_v}\{dx, dv, \ldots, dv^{(r+k-1)}\} \quad \forall k \geq 0. \tag{60}$$

Define the sequence Ω_μ as

$$\begin{aligned} \Omega_0 &= \mathcal{P}^\perp, \\ \Omega_{\mu+1} &= \{\omega \in \Omega_\mu \mid \omega^{[1]} \in \Omega_\mu + \mathcal{V}\} \quad \forall \mu \geq 1, \end{aligned} \tag{61}$$

and

$$\Omega = \lim_{\mu \to \infty} \Omega_\mu.$$

Obviously $\Omega^{[1]} \subset \Omega + \mathcal{V}$. Thus, Ω is a controlled invariant subspace. Since $\mathrm{span}_{\mathcal{K}}\{dy\} \subset \Omega$ and $\Omega \subset \mathcal{P}^\perp$, (57) also holds. □

Condition (57) in Theorem 3.21 is not constructive. The corresponding constructive condition is obtained when considering the smallest controlled invariant subspace containing the differential of the output Ω_*. From Theorem 3.18, Ω_* is given by $\mathcal{X} \cap \mathcal{Y}$. An immediate consequence of Theorem 3.21 is then as follows.

COROLLARY 3.22. *The DDDP is solvable if and only if*

$$\Omega_* \subset \mathcal{P}^\perp. \tag{62}$$ □

Remark 3.23. Theorem 3.21 gives the nonlinear feedback analogon of Theorem 4.2 in [48] for the linear (D)DDP. Also, it gives the dynamic feedback analogon of condition (3.1) in [29] and Proposition 7.8 in [38] for the nonlinear DDP. In this way

TABLE 4
(Geometric) solution to DDP (complete).

	Static feedback	Dynamic feedback
Linear systems	$\mathcal{P} \subset \mathcal{V}^*$	
Nonlinear systems	$\mathcal{P} \subset \Delta^*$	$\Omega_* \subset \mathcal{P}^\perp$

it is established that our generalized notion of controlled invariance is the natural generalization to the nonlinear dynamic feedback case of the linear notion of controlled invariance defined in [48]. It may be checked that condition (62) is equivalent to the geometric conditions (44) in [26] and (4.5) in [44]. Further, (62) is exactly the same as the condition for solvability of the DDDP derived in [41]. However, in [41] the concept of controlled invariance was missing.

Table 1, which displayed the various solutions of the DDP, is now completed in Table 4.

4. Controllability cospaces. In this section, we study controllability cospaces under quasi-static state feedback. These controllability cospaces form a special class of the controlled invariant subspaces defined previously. They parallel the dynamic controllability distributions [45]. In subsection 4.1 we first define controllability cospaces. An algorithm which characterizes these cospaces is then given in subsection 4.2, and some properties of these controllability cospaces are discussed. In subsection 4.3 we derive an algorithm computing the smallest controllability cospace containing a given exact subspace. Applications of controllability cospaces are treated in subsections 4.4 and 4.5. In particular, the fixed modes or decoupling zero dynamics under quasi-static feedback are characterized using controllability cospaces.

4.1. Definition of controllability cospaces. Controllability cospaces are vector spaces that are autonomous after having applied a certain quasi-static state feedback $u = \psi(x, v, \dots, v^{(r)})$ and zeroing certain input channels v_j, where $j \in \mathcal{J} \subset \{1, \dots, m\}$. Such nonregular transformations are not defined for every element in $\mathcal{K}_v$. One possibility to circumvent this problem is to consider the module $\mathrm{span}_{\mathcal{A}}\{dx\}$ over the ring of analytic functions rather than the linear space over the field of meromorphic functions. Another way is chosen here; it consists in taking a particular basis of a given subspace of $\mathrm{span}_{\mathcal{K}}\{dx\}$ so that its time derivative is well defined when applying nonregular feedback. Such a basis always exists. More precisely, let $\Theta \subset \mathcal{X}$ be a subspace which admits a basis $\theta_1, \dots, \theta_d$ with

$$\theta_i = \sum_{k=1}^{n} \frac{\alpha_{ik}(x, v, \dots, v^{(\nu)})}{\beta_{ik}(x, v, \dots, v^{(\nu)})} dx_i,$$

where α_{ik} and β_{ik} are in $\mathcal{A}$, the ring of analytic functions of $\{x, v^{(k)} \mid k \geq 0\}$. Obviously, we can choose another basis for Θ, $\tilde{\theta}_1, \dots, \tilde{\theta}_d$, in the module $\mathrm{span}_{\mathcal{A}}\{dx\}$ over the ring $\mathcal{A}$ by taking

$$\tilde{\theta}_i = \left(\prod_{k=1}^{n} \beta_{ik} \right) \theta_i.$$

DEFINITION 4.1. *A subspace $\mathcal{C} \subset \mathcal{X}$ is said to be a controllability cospace for* (1) *if there exist a quasi-static state feedback* (8) *and a set of integers $\mathcal{J} \subset \{1, \dots, m\}$ such that for* (1), (8) *one has*

$$\mathcal{C}^{[1]} \subset \mathcal{C} + \mathcal{V} \tag{63}$$

and

$$\mathcal{C} = \max\{\Theta \subset \mathcal{X} \mid \operatorname{span}_{\mathcal{K}}\{\tilde{\theta}_i^{[1]} \mid_{v_j=0, j\in\mathcal{J}}\} \subset \Theta\}, \tag{64}$$

where $\tilde{\theta}_i$ is defined as above.

This means that $\mathcal{C}$ is the largest autonomous subspace in $\mathcal{X}$ of the closed loop system. Moreover, by this definition, it is clear that a controllability cospace is controlled invariant. The following example illustrates the above definition.

Example 4.2. Consider again the nonlinear system given in Example 3.14. Let $\mathcal{C} = \operatorname{span}_{\mathcal{K}} \{dx_1, d(x_2 - x_3), dx_4 - u_1 dx_3\}$, and suppose that $u_1 = v_1 + c$, where c is a nonzero constant, $u_2 = v_2$ and $u_3 = v_3 + \dot{v}_1 x_3 + (v_1 + c)^2 x_3 + (v_1 + c)v_2$. This feedback is quasi-static since $v_1 = u_1 - c$ and $v_2 = u_2$ and $v_3 = u_3 - \dot{u}_1 x_3 - u_1^2 x_3 - u_1 u_2$.

From this, it is easy to show that

$$\mathcal{C}^{[1]} = \operatorname{span}_{\mathcal{K}}\{dv_1, dx_4 - u_1 dx_3, dv_3 + (x_3(v_1 + c) + v_2)dv_1 + x_3 d\dot{v}_1\} \subset \mathcal{C} + \mathcal{V}$$

and

$$\mathcal{C}^{[1]} \mid_{v_1=0, v_3=0} = \operatorname{span}_{\mathcal{K}}\{dx_4 - u_1 dx_3\} \subset \mathcal{C}.$$

Furthermore

$$\mathcal{C} = \max\{\Theta \subset \mathcal{X} \mid \Theta^{[1]} \mid_{v_1=0, v_3=0} \subset \Theta\}.$$

Hence $\mathcal{C}$ is a controllability cospace in the sense of Definition 4.1.

4.2. Controllability cospace algorithm. First of all, we give an algorithm characterizing controllability cospaces called *the controllability cospace algorithm*. Some properties of a general controllability cospace are then derived. Let $\mathcal{C}$ be a given subspace and define a sequence $\mathcal{S}_\mu$ according to

$$\begin{aligned} \mathcal{S}_0 &:= \mathcal{X}, \\ \mathcal{S}_{\mu+1} &:= \operatorname{span}_{\mathcal{K}}\{\omega \in \mathcal{S}_\mu \mid \dot{\omega} \in \mathcal{S}_\mu + \dot{\mathcal{C}}\} \quad (\mu \in \mathbb{N}). \end{aligned} \tag{65}$$

The sequence $\mathcal{S}_\mu$ is decreasing. Thus, there exists a $\mu^* \in \mathbb{N}$ such that $\mathcal{S}_{\mu^*} = \mathcal{S}_{\mu^*+k}$ for all $k \in \mathbb{N}$. Define $\mathcal{S}^* := \mathcal{S}_{\mu^*}$.

Algorithm (65) yields a necessary condition for a subspace $\mathcal{C}$ of $\mathcal{X}$ to be a controllability cospace. This is shown in the following lemma.

LEMMA 4.3. *Let $\mathcal{C} \subset \mathcal{X}$. If $\mathcal{C}$ is a controllability cospace, then $\mathcal{C} = \mathcal{S}^*$.*

Proof. Assume that $\mathcal{C}$ is a controllability cospace. Let $\{\tilde{\omega}_i\}$ be a basis for $\mathcal{C}$ in the module $\operatorname{span}_{\mathcal{A}}\{dx\}$ over the ring $\mathcal{A}$. Then by definition there exists a quasi-static state feedback (8) and a set of integers $\mathcal{J} \subset \{1, \ldots, m\}$ such that $\mathcal{C}^{[1]} \subset \mathcal{C} + \mathcal{V}$ and $\mathcal{C}^{[1]} = \operatorname{span}_{\mathcal{K}}\{\tilde{\omega}_i^{[1]} \mid_{v_j=0, j\in\mathcal{J}}\} \subset \mathcal{C}$. From (65), it follows that $\mathcal{S}^*$ satisfies

$$\mathcal{S}^* = \operatorname{span}_{\mathcal{K}}\{\omega \in \mathcal{X} \mid \dot{\omega} \in \mathcal{S}^* + \dot{\mathcal{C}}\}. \tag{66}$$

Let $\omega \in \mathcal{C}$. We have $\dot{\omega} \in \dot{\mathcal{C}}$ and hence $\omega \in \mathcal{S}^*$. This implies that $\mathcal{C} \subset \mathcal{S}^*$. Now, $\dot{\mathcal{S}}^* \subset \mathcal{S}^* + \dot{\mathcal{C}}$. By the feedback, which yields $\mathcal{C}^{[1]} \subset \mathcal{C}$, one has $\mathcal{S}^{*[1]} \subset \mathcal{S}^*$. Since $\mathcal{C}$ is the largest subspace in $\mathcal{X}$ such that $\mathcal{C}^{[1]} \subset \mathcal{C}$, one has $\mathcal{S}^* \subset \mathcal{C}$. □

In the next section, we give an algorithm computing the smallest controllability cospace containing a given subspace based on algorithm (65).

4.3. The smallest controllability cospace containing a given subspace. In general, the intersection of two controllability cospaces is not a controllability cospace. Thus it is unclear if there exists a smallest controllability cospace containing some given subspace. However, if an exact subspace $\Pi \subset \mathcal{X}$ is given, then there exists a smallest controllability cospace containing Π.

Consider a nonlinear system given by (1). By Theorem 3.18, Π_* is the smallest controlled invariant subspace containing Π. The next theorem will relate Π_* to the smallest controllability cospace containing Π.

THEOREM 4.4. *Define the sequence* $\mathcal{D}_\mu$ *by*

$$\begin{aligned} \mathcal{D}_0 &= \mathcal{X}, \\ \mathcal{D}_{\mu+1} &= \operatorname{span}_{\mathcal{K}}\{\omega \in \mathcal{D}_\mu \mid \dot{\omega} \in \mathcal{D}_\mu + \dot{\Pi}_*\} \quad (\mu \in \mathbb{N}). \end{aligned} \tag{67}$$

Then $\mathcal{D}_* = \lim_{\mu\to\infty} \mathcal{D}_\mu$ *is the smallest controllability cospace containing* Π.

Proof. Note that

$$\mathcal{D}_* = \operatorname{span}_{\mathcal{K}}\{\omega \in \mathcal{X} \mid \dot{\omega} \in \mathcal{D}_* + \dot{\Pi}_*\}. \tag{68}$$

Let $r = \dim\Pi$. The fact that Π is exact implies that there exist meromorphic functions $\varphi_1(x), \ldots, \varphi_r(x)$ such that $\Pi = \operatorname{span}_{\mathcal{K}}\{d\varphi_1, \ldots, d\varphi_r\}$. Consider the system (1) with a "dummy" output $\varphi = (\varphi_1, \ldots, \varphi_r)^T$. We decompose the output φ as $\varphi = (\tilde{\varphi}, \hat{\varphi})^T$ so that the system (1) with the output $\tilde{\varphi}$ is right invertible. Define $\rho := \dim(\tilde{\varphi})$.

Construct a quasi-static state feedback $u = \phi(x, v, \ldots, v^{(r)})$ by taking $v_i = \tilde{\varphi}_i^{(n_i')}$, where $\{n_i'\}$ is the set of orders of zeros at infinity [18] for $i = 1, \ldots, \rho$ and $v_i = w_i$ for $i = \rho + 1, \ldots, m$. This feedback always renders Π_* invariant. Thus, $\mathcal{D}_*$ is rendered invariant too; i.e., $\mathcal{D}_*^{[1]} \subset \mathcal{D}_* + \mathcal{V}$. Let now $\{\tilde{\omega}_i\}$ be a basis for $\mathcal{D}_*$ in the module $\operatorname{span}_{\mathcal{A}}\{dx\}$ over the ring $\mathcal{A}$. If we set $v_i = 0$ for $i = 1, \ldots, \rho$ one obtains

$$\mathcal{D}_*^{[1]} = \operatorname{span}_{\mathcal{K}}\{\tilde{\omega}_i^{[i]} \mid_{v_j=0, j=1,\cdots,\rho}\} \subset \mathcal{D}_*.$$

Hence $\mathcal{D}_*$ is a controllability cospace. In order to prove that $\mathcal{D}_*$ is the smallest controllability cospace containing Π, we consider another controllability cospace $\mathcal{D}$ such that $\mathcal{D} \supset \Pi$. By definition $\mathcal{D}$ is controlled invariant, and, according to Lemma 4.3, $\mathcal{D}$ satisfies

$$\mathcal{D} = \operatorname{span}_{\mathcal{K}}\{\omega \in \mathcal{X} \mid \dot{\omega} \in \mathcal{D} + \dot{\mathcal{D}}\}. \tag{69}$$

Since Π_* is the smallest controlled invariant subspace containing Π, this implies that $\mathcal{D} \supset \Pi_*$. From (68) and (69), one has $\mathcal{D}_* \subset \mathcal{D}$. ∎

COROLLARY 4.5. *Consider a nonlinear system of the form* (51). *Define the sequence* $\mathcal{C}_\mu$ *according to*

$$\begin{aligned} \mathcal{C}_0 &= \mathcal{X}, \\ \mathcal{C}_{\mu+1} &= \operatorname{span}_{\mathcal{K}}\{\omega \in \mathcal{C}_\mu \mid \dot{\omega} \in \mathcal{C}_\mu + \dot{\Omega}_*\} \quad (\mu \in \mathbb{N}). \end{aligned} \tag{70}$$

Then $\mathcal{C}_* = \lim_{\mu\to\infty} \mathcal{C}_\mu$ *is the smallest controllability cospace containing* $\operatorname{span}_{\mathcal{K}}\{dh(x)\}$.

Proof. Clearly, $\Omega_* = \mathcal{X} \cap \mathcal{Y}$ is the smallest controlled invariant subspace containing the differential of the outputs. The result then immediately follows from Theorem 4.4. ∎

Remark 4.6. When specialized to linear systems, the sequence $\mathcal{C}_\mu$ (70) turns out to be equal to the dual of the sequence $\mathcal{R}_\mu$ (the sequence computing the maximal

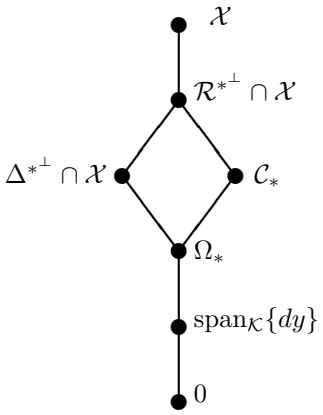


FIG. 2. *Lattice diagram: (Geometric) structure of nonlinear systems* (*continued; see Figure* 1).

controllability subspace in the kernel of the output mapping). A proof of this can be found in the appendix.

The geometric structure of a nonlinear system as presented in Figure 1 can now be completed. Let $\mathcal{R}^*$ be the largest controllability distribution contained in the kernel of the output. As an immediate consequence, $\mathcal{R}^{*^\perp} \cap \mathcal{X}$ is a controllability cospace in the sense of Definition 4.1. Figure 2 displays further geometric structures of nonlinear systems.

4.4. The block input-output decoupling problem. We now use the smallest controllability cospace $\mathcal{C}_*$, previously defined, to solve a quasi-static state feedback input-output decoupling problem. For this, we consider the system (1) together with the partitioned output blocks y_i for $i = 1, \ldots, k$, given by

$$y_i = h_i(x). \tag{71}$$

The problem can be stated as follows: find a quasi-static state feedback and a partition of the new control $v = (v_1^T, \ldots, v_k^T)^T$ such that the new input v_i affects only the output y_i.

Define $\mathcal{C}_{i*}$ and Ω_{i*} to be the smallest controllability cospace and the smallest controlled invariant subspace, respectively, both containing $\text{span}_{\mathcal{K}}\{dh_i(x)\}$.

First, let us give the following property which is derived from Theorem 5.1 in [42].

PROPERTY 4.7. *Consider system* (51)*, and assume that* $\dim(\mathcal{G}^\perp) = n - m$*. Let* ρ *be its differential output rank. Then*

$$\dim(\mathcal{G}^\perp + \Omega_*) = \dim(\mathcal{G}^\perp + \mathcal{C}_*) = (n - m + \rho). \tag{72}$$

Moreover, if the system (51) *is right invertible, then*

$$\dim(\mathcal{G}^\perp + \Omega_*) = \dim(\mathcal{G}^\perp + \mathcal{C}_*) = (n - m + p). \tag{73}$$

This property is a generalization of a known result on linear systems. It gives a geometric interpretation of the rank of a system. The property was also derived by Respondek in [45] using dynamic controllability distributions.

COROLLARY 4.8. *The block input-output decoupling problem via quasi-static (or dynamic) state feedback for the system* (1), (71) *is solvable if and only if*

$$\dim\left(\frac{\mathcal{G}^\perp + \mathcal{C}_*}{\mathcal{G}^\perp}\right) = \sum_{i=1}^{k} \dim\left(\frac{\mathcal{G}^\perp + \mathcal{C}_{i*}}{\mathcal{G}^\perp}\right). \tag{74}$$

Condition (74) coincides with the condition given by Di Benedetto, Grizzle, and Moog [18], in the case of the dynamic block decoupling problem. Indeed, if ρ denotes the rank of the system (1), (71) and ρ_i denotes the rank of the subsystem (1) with the output y_i, then by Property 4.7, (74) is equivalent to

$$\rho = \sum_{i=1}^{k} \rho_i. \tag{75}$$

By applying the structure algorithm to the system (1), (71), a quasi-static feedback which decouples the system is obtained [14].

Further, controllability cospaces also allow us to characterize the fixed dynamics with respect to any quasi-static feedback. This will be the topic of the next section.

4.5. Fixed modes by quasi-static state feedback. The problem of noninteraction with stability of nonlinear systems by means of static feedback has first been considered by Isidori and Grizzle [32]. They have shown that there exists a fixed internal dynamics called P^* *dynamics* whose stability is a necessary condition for solving the noninteracting control problem with stability via static feedback. In the case where the P^* dynamics are unstable, Wagner has shown [47] that there exists a well-defined dynamics called Δ_{mix} *dynamics* which cannot be eliminated by any regular dynamic feedback that renders the considered system noninteractive. The Δ_{mix} dynamics must then be asymptotically stable if noninteracting control with stability is to be achieved by means of dynamic state feedback. Glumineau, Moog, and Tarn [21] used a dynamic compensator to remove a one-dimensional interconnection zero dynamics and showed that such a compensator is able to cancel only the fixed dynamics which have a certain linearity property. A sufficient condition to solve the problem of noninteracting control with stability by means of dynamic state feedback was given in [5], [6], [7]. In these references, the problem of dynamic feedback noninteracting control with stability is solved if some regularity assumptions are satisfied, the Δ_{mix} dynamics are asymptotically stable and each decoupled subsystem is asymptotically stabilizable.

All results above are valid under the assumption that the decoupling matrix $A(x)$ is nonsingular. In the case where $A(x)$ is singular and the system is square and invertible, Zhan, Tarn, and Isidori [50] introduced the so-called canonical dynamic decoupling algorithm to construct a canonical dynamic extension (Σ_p). They have shown that the dynamically decoupled system is stable only if the Δ_{mix} dynamics of the canonical dynamic extension is stable, which is an intrinsic property of the given system. These different contributions are summarized in Table 2.

In this section, we investigate the case where the decoupling matrix is not necessarily invertible and study the noninteracting control problem with stability by means of quasi-static feedback. The goal is to show that the controllability cospaces introduced before are able to describe intrinsic geometric conditions with respect to quasi-static feedbacks, analogous to the above ones. Preliminary results may be found in [1].

Let us consider a square invertible nonlinear affine system (Σ) of the form

$$\Sigma : \begin{cases} \dot{x} &= f(x) + \displaystyle\sum_{i=1}^{m} g_i(x)u_i, \qquad x \in \mathbb{R}^n,\ u_i \in \mathbb{R}, \\ y_i &= h_i(x), \qquad i = 1, \ldots, m,\ y_i \in \mathbb{R}. \end{cases} \tag{76}$$

Let $\{n_i'\}$ be the set of orders of zeros at infinity [18], where $n_1' > n_2' > \cdots > n_m'$.

Permute if necessary y_i such that the corresponding order of zero at infinity is $n_i^{'}$. Let $\mathcal{C}_{i*}$ be the smallest controllability cospace containing $\text{span}_{\mathcal{K}}\{dh_i(x)\}$. A first result is the following.

LEMMA 4.9. *Suppose that the system* (76) *can be decoupled by a quasi-static state feedback* $u = \psi(x, v, \ldots, v^{(s)})$. *Then there always exist coordinates* $\xi = (\xi_0, \xi_1, \ldots, \xi_m, \hat{\xi})$ *such that the system* (76) *has the following form:*

$$
\begin{aligned}
\dot{\xi}_0 &= f_0(\xi_0), \\
\dot{\xi}_1 &= f_1(\xi_0, \xi_1, v_1), \\
&\vdots \\
\dot{\xi}_m &= f_m(\xi_0, \xi_m, v_m), \\
\dot{\hat{\xi}} &= \hat{f}(\xi, v, \dot{v}, \ldots, v^{(s)}), \\
y_i &= h_i(\xi_0, \xi_i). \qquad \square
\end{aligned} \tag{77}
$$

The system (77) will be referred to as a *standard decomposed system*, analogous to [23]. To prove Lemma 4.9, we first need the following property of $\mathcal{C}_{i*}$.

LEMMA 4.10. *For a scalar output* $y_i = h_i(x)$, $\mathcal{C}_{i*}$ *is an exact subspace.*

Proof. Let Ω_{i*} be the smallest controlled invariant subspace containing $\text{span}_{\mathcal{K}}\{dh_i\}$. If Δ_i^* is the maximal controlled invariant distribution in $\ker\{dh_i(x)\}$, we have $\Omega_{i*} = \Delta_i^{*\perp}$. Now let $\mathcal{R}_i^*$ be the maximal controllability distribution in $\ker\{dh_i(x)\}$. Clearly $\mathcal{R}_i^{*\perp}$ is a controllability cospace containing $\text{span}_{\mathcal{K}}\{dh_i(x)\}$, and thus $\mathcal{C}_{i*} \subset \mathcal{R}_i^{*\perp}$. From [29] we have

$$
\mathcal{R}_i^* = \Delta_i^* \cap \left([f, \mathcal{R}_i^*] + \sum_{j=1}^{m} [g_j, \mathcal{R}_i^*] + \mathcal{G} \right) \tag{78}
$$

and thus

$$
\begin{aligned}
\mathcal{R}_i^{*\perp} &= \Omega_{i*} + [f, \mathcal{R}_i^*]^\perp \cap \Big(\bigcap_{j=1}^{m} [g_j, \mathcal{R}_i^*]^\perp \Big) \cap \mathcal{G}^\perp \\
&= \{\omega \in \mathcal{X} \mid \exists \omega_1 \in \Omega_{i*}, \exists \omega_2 \in \mathcal{G}^\perp \text{ such that } \omega = \omega_1 + \omega_2 \\
&\qquad \text{and } (\forall \tau \in \mathcal{R}_i^*)(\forall \sigma \in \{f, g_1, \ldots, g_m\})(\langle [\sigma, \tau], \omega_2 \rangle = 0)\}.
\end{aligned} \tag{79}
$$

Let $\omega \in \mathcal{R}_i^{*\perp}$. Then there exist $\omega_1 \in \Omega_{i*}$ and $\omega_2 \in \mathcal{G}^\perp$ such that $\omega = \omega_1 + \omega_2$, and $\forall \tau \in \mathcal{R}_i^*, \forall \sigma \in \{f, g_1, \ldots, g_m\}$, one has $\langle [\sigma, \tau], \omega_2 \rangle = 0$. Compute

$$
\dot{\omega} = \dot{\omega}_1 + \dot{\omega}_2.
$$

Clearly $\dot{\omega}_1 \in \dot{\Omega}_{i*}$. Furthermore,

$$
\begin{aligned}
\dot{\omega}_2 &= \mathcal{L}_f \omega_2 + \sum_{j=1}^{m} (u_j \mathcal{L}_{g_j} \omega_2 + \langle \omega_2, g_j \rangle du_j) \\
&= \mathcal{L}_f \omega_2 + \sum_{j=1}^{m} u_j \mathcal{L}_{g_j} \omega_2.
\end{aligned} \tag{80}
$$

Now, let $\tau \in \mathcal{R}_i^*$ and $\sigma \in \{f, g_1, \ldots, g_m\}$. Then

$$
\begin{aligned}
\langle \tau, \mathcal{L}_\sigma \omega_2 \rangle &= \mathcal{L}_\sigma \langle \tau, \omega_2 \rangle - \langle [\sigma, \tau], \omega_2 \rangle \\
&= \mathcal{L}_\sigma \langle \tau, \omega_2 \rangle = \mathcal{L}_\sigma \langle \tau, (\omega - \omega_1) \rangle = 0,
\end{aligned} \tag{81}
$$

where the last equality follows from the fact that $\omega \in \mathcal{R}_i^{*\perp}$ and $\omega_1 \in \Omega_{i*} \subset \mathcal{R}_i^{*\perp}$. By (80), (81), we then have $\dot{\omega}_2 \in \mathcal{R}_i^{*\perp}$, and hence

$$\widehat{\dot{\mathcal{R}}_i^{*\perp}} \subset \dot{\Omega}_{i*} + \mathcal{R}_i^{*\perp}.$$

By construction, $\mathcal{C}_{i*}$ is the largest subspace in $\mathcal{X}$ which verifies $\dot{\mathcal{C}}_{i*} \subset \mathcal{C}_{i*} + \dot{\Omega}_{i*}$. This implies $\mathcal{R}_i^{*\perp} \subset \mathcal{C}_{i*}$. So $\mathcal{C}_{i*}$ is the annihilator of $\mathcal{R}_i^*$, which is defined to be involutive [39], [29]. Hence $\mathcal{C}_{i*}$ is exact, which establishes our claim. □

Proof of Lemma 4.9. By Lemma 4.10, $\mathcal{C}_{i*}$ is an exact subspace. Thus, $\dot{\mathcal{C}}_{i*}$ as well as $\sum_{j=0}^{n_i'-1} \mathcal{C}_{i*}^{(j)}$ is also exact. Let us define $\mathcal{C}_0$ as the uncontrollable subspace of (Σ) which is the subspace $\mathcal{H}_\infty$ introduced in [2]. It is obvious that for each $i = 1, \ldots, m$

$$\mathcal{C}_0 = \sum_{j=0}^{n_i'-1} \mathcal{C}_{i*}^{(j)} \cap \sum_{k=\neq i} \sum_{j=0}^{n_k'-1} \mathcal{C}_{k*}^{(j)}.$$

Let $\{d\xi_0\}$ be a basis of $\mathcal{C}_0$; thus $\dot{\xi}_0 = f_0(\xi_0)$. For an invertible system, we can construct a quasi-static state feedback which decouples system (Σ) by taking $v_i = y_i^{(n_i')}$. For $i = 1, \ldots, m$, choose $d\xi_i$ such that $\{d\xi_0, d\xi_i\}$ is a basis of $\sum_{j=0}^{n_i'-1} \mathcal{C}_{i*}^{(j)}$. Then one has

$$\dot{\xi}_i = f_i(\xi_0, \xi_i, v_i).$$

Complete the new coordinates by choosing $\hat{\xi}$ such that $\{d\xi_0, d\xi_1, \ldots, d\xi_m, d\hat{\xi}\}$ is linearly independent. Without loss of generality, $\hat{\xi}$ can be chosen so that $\operatorname{span}\{d\hat{\xi}\} \subset \mathcal{X}$. Thus, one has

$$\dot{\hat{\xi}} = \hat{f}(\xi, v, \dot{v}, \ldots, v^{(s)}),$$

and (77) is established. □

Now we may state the following theorem.

THEOREM 4.11. *For a square invertible nonlinear system, the dimension of the fixed dynamics with respect to any quasi-static state feedback is*

$$n - \dim \left(\mathcal{X} \cap \sum_{i=1}^{m} \sum_{j \geq 0} \mathcal{C}_{i*}^{(j)} \right). \tag{82}$$

Moreover, if the origin is an equilibrium point for Σ *and the quasi-static state feedback rendering* (76) *noninteractive preserves this equilibrium point, then the induced fixed dynamics are*

$$\dot{\hat{\xi}} = \hat{f}(0, \ldots, 0, \hat{\xi}, 0, \ldots, 0), \tag{83}$$

where $\hat{\xi}$ *is as defined in Lemma* 4.9.

Proof. From the proof of Lemma 4.9, the dimension of the fixed dynamics with respect to any quasi-static feedback which decouples the system is

$$n - \dim \left(\sum_{i=1}^{m} \sum_{j=0}^{n_i'-1} \mathcal{C}_{i*}^{(j)} \right). \tag{84}$$

From the definition of the structure at infinity, one gets

$$\dim\left(\sum_{i=1}^{m}\sum_{j=0}^{n_i'-1}\mathcal{C}_{i*}^{(j)}\right)=\dim\left(\mathcal{X}\cap\sum_{i=1}^{m}\sum_{j\geq 0}\mathcal{C}_{i*}^{(j)}\right). \tag{85}$$

Thus (82) is established. Once the system is in standard decomposed form (77), and analogously to [23], any decoupling quasi-static state feedback is of the form $v_i=\alpha_i(\xi_0,\xi_i,w_i,\ldots,w_i^{(\nu)})$. Hence, if the α_i's preserve the equilibrium, the second statement in Theorem 4.11 is immediate. □

The asymptotic stability of dynamics (83) is a necessary condition for noninteracting control with internal stability by quasi-static state feedback.

The next example illustrates Theorem 4.11.

Example 4.12. Let us consider a nonlinear system given by

$$\dot{x}_1=u_1,\quad \dot{x}_2=x_4+x_3u_1,\quad \dot{x}_3=x_3+x_4,\quad \dot{x}_4=u_2,\quad \dot{x}_5=x_1+x_2,$$
$$y_1=x_1,\qquad y_2=x_2.$$

We have $\{n_i'\}=\{2,1\}$. Permute then y_i, and thus $\mathcal{C}_{1*}=\{dx_2\}$ and $\mathcal{C}_{2*}=\{dx_1\}$. The quasi-static feedback which decouples the system is $u_1=v_1$ and $u_2=v_2-(x_3+x_4)v_1-x_3\dot{v}_1$, where (v_1,v_2) is a new input vector. It is clear that $\mathcal{C}_0=0$. We choose $d\xi_1=\{dx_2,d(x_4+x_3u_1)\}$ as a basis of $\{\mathcal{C}_{1*}+\dot{\mathcal{C}}_{1*}\}$, and thus

$$\dot{\xi}_1=\begin{pmatrix}\dot{\xi}_{11}\\ \dot{\xi}_{12}\end{pmatrix}=\begin{pmatrix}\xi_{12}\\ v_2\end{pmatrix}. \tag{86}$$

Now choose $\{d\xi_2\}=\{dx_1\}$ as a basis of $\mathcal{C}_{2*}$, and one has

$$\dot{\xi}_2=v_1. \tag{87}$$

We complete our coordinate transformation by taking

$$\hat{\xi}=\begin{pmatrix}\hat{\xi}_1\\ \hat{\xi}_2\end{pmatrix}=\begin{pmatrix}x_4\\ x_5\end{pmatrix}.$$

So in the new coordinates $(\xi_1,\xi_2,\hat{\xi})$, the considered system becomes

$$\begin{aligned}
\dot{\xi}_{11} &= \xi_{12},\\
\dot{\xi}_{12} &= v_2,\\
\dot{\xi}_2 &= v_1,\\
\dot{\hat{\xi}}_1 &= v_2-(\xi_{12}-\hat{\xi}_1)-\hat{\xi}_1v_1-(\xi_{12}-\hat{\xi}_1)\dot{v}_1/v_1,\\
\dot{\hat{\xi}}_2 &= \xi_2+\xi_{11},\\
y_1 &= \xi_2,\\
y_2 &= \xi_{11}.
\end{aligned} \tag{88}$$

Clearly, $\dim(\hat{\xi})=2=n-\dim(\mathcal{X}\cap(\sum_{i=1}^{m}\sum_{j\geq 0}\mathcal{C}_{i*}^{(j)}))=n-\dim(\mathrm{span}\{dx_1,dx_2,dx_4+u_1dx_3\})$. Thus, the dimension of the fixed dynamics equals two. Since the origin is an equilibrium point, the fixed dynamics are then

$$\begin{aligned}
\dot{\hat{\xi}}_1 &= \hat{\xi}_1,\\
\dot{\hat{\xi}}_2 &= 0.
\end{aligned} \tag{89}$$

TABLE 5
Decoupling zero structure (complete).

Feedback	$A(x)$ invertible	$A(x)$ noninvertible
(Quasi-) Static	dim $(\mathcal{P}^*)$ Isidori and Grizzle [32]	$n - \dim \left(\mathcal{X} \cap \left(\sum_{i=1}^{m} \sum_{j\geq 0} \mathcal{C}_i^{*(j)}\right)\right)$
Dynamic	dim (Δ_{mix}) Wagner [47]	dim $(\Delta_{mix}(\Sigma_p))$ Zhan, Tarn, and Isidori [50]

Similarly to Wagner's and Battilotti's results, in the case where no quasi-static state feedback can render the system simultaneously noninteractive and stable, a suitable dynamic feedback may still solve the problem. This reduces to the results in Zhan, Tarn, and Isidori [50].

Table 2, which displays the dimension of the various decoupling zero dynamics, is now completed in Table 5.

5. Conclusions. A generalized notion of controlled invariance under quasi-static state feedback for nonlinear systems was introduced. It was shown that this notion coincides with the standard notion of a controlled invariant distribution under regular static state feedback. Using the generalized notion of controlled invariance, a condition for the controlled invariance of not necessarily integrable codistributions was derived. For a subspace $\Omega \subset \mathcal{X}$, we gave sufficient conditions for controlled invariance under quasi-static state feedback. Furthermore, a necessary and sufficient condition for controlled invariance was also given for a special class of subspaces Ω. The generalized controlled invariance was applied to the DDP by dynamic feedback. A necessary and sufficient condition for solvability of this DDDP was obtained.

For a controllability cospace $\mathcal{C} \subset \mathcal{X}$, some properties were derived by means of the controllability cospace algorithm. Moreover, the smallest controllability cospace containing the differential of the output mapping allowed us to solve the block input-output decoupling problem. It also characterized the dimension of the fixed dynamics with respect to any quasi-static state feedback in the case of one to one decoupling.

This paper leaves some interesting open questions, which are topics for further research. A first question is related to necessary and sufficient conditions for controlled invariance for a general class of subspaces. A second question is whether (or under what conditions) there exists a smallest controlled invariant subspace containing some given subspace. It seems that for the answer to both questions a better understanding of quasi-static state feedback is needed.

Finally, let us remark that throughout the paper we have restricted ourselves to "Kalmanian" systems and to subspaces $\Omega \subset \mathcal{X}$. However, the definition of controlled invariance and the characterizations of controlled invariance in this paper can, mutatis mutandis, be translated to non-Kalmanian systems and subspaces $\Omega \subset \mathcal{X} \times \mathcal{U}$.

Appendix. According to Remark 4.6, we will prove that the sequence (70) computing $\mathcal{C}_*$ is the same as the one computing $\mathcal{R}^{*^\perp}$ (the dual of $\mathcal{R}^*$, the maximal controllability subspace in kernel of the output) for linear time-invariant systems. We proceed by induction. First, we recall some basic operations that we need.

Consider a linear system given by

$$\begin{aligned} \dot{x} &= Ax + Bu,\ x \in \mathbb{R}^n, u \in \mathbb{R}^m, \\ y &= Cx. \end{aligned} \tag{90}$$

Identify elements of $\mathbb{R}^n$ with column vectors while elements of $\mathbb{R}^{n^\perp}$, its dual, are identified with row vectors. Thus, $\omega = \sum_{i=1}^n \alpha_i dx_i \in \mathbb{R}^{n^\perp}$ is identified with the row vector $\alpha := (\alpha_1, \ldots, \alpha_n) \in \mathbb{R}^n$. With this notation,

$$\dot{\omega} = \alpha d\dot{x} = \alpha A dx + \alpha B du \in (\mathbb{R}^n \times \mathbb{R}^m)^\perp \tag{91}$$

is identified with the row vector $(\alpha A \ \alpha B)$.

Let a subspace $V \subset \mathbb{R}^n$ be given. Then

$$\begin{array}{rcl} (AV)^\perp & = & \{\omega \in \mathrm{span}\{dx\} \mid < \omega, Av >= 0 \, \forall v \in V\} \\ & = & \{\alpha \in \mathbb{R}^n \mid \alpha A v = 0, \forall v \in V\} = \{\alpha \in \mathbb{R}^n \mid \alpha A \in V^\perp\} \\ & =: & {}^{-1}AV^\perp \end{array} \tag{92}$$

if $\omega = \alpha dx \in (AV)^\perp \cap \mathcal{B}^\perp$, where $\mathcal{B} = \mathrm{Im}B$. Then

$$\dot{\omega} = \alpha A dx + \alpha B du = \alpha A dx \simeq \alpha A \in V^\perp. \tag{93}$$

The two sequences to be compared are

$$\left\{ \begin{array}{rcl} \mathcal{R}_0^\perp & := & \mathcal{X}, \\ \mathcal{R}_{\mu+1}^\perp & := & \mathcal{V}^{*^\perp} + {}^{-1}A\mathcal{R}_\mu^\perp \cap \mathcal{B}^\perp \quad (\mu \in \mathbb{N}) \end{array} \right. \tag{94}$$

and

$$\left\{ \begin{array}{rcl} \mathcal{C}_0 & := & \mathcal{X}, \\ \mathcal{C}_{\mu+1} & := & \{\omega \in \mathcal{C}_\mu \mid \dot{\omega} \in \mathcal{C}_\mu + \dot{\mathcal{V}}^{*^\perp}\} \ \ (\mu \in \mathbb{N}), \end{array} \right. \tag{95}$$

where $\mathcal{V}^*$ is the maximal controlled invariant subspace in $\mathrm{Ker}C$ for the system (90). For step 0, it is obvious that $\mathcal{R}_0^\perp = \mathcal{C}_0$. Suppose that $\mathcal{R}_\mu^\perp = \mathcal{C}_\mu$ for $\mu = 0, \ldots, \ell$. Let $\omega \in \mathcal{R}_{\ell+1}^\perp$, thus there exist $\omega_1 \in \mathcal{V}^{*^\perp}$ and $\omega_2 \in {}^{-1}A\mathcal{R}_\ell^\perp \cap \mathcal{B}^\perp$ such that $\omega = \omega_1 + \omega_2$. By (93), $\dot{\omega}_2 \in \mathcal{R}_\ell^\perp = \mathcal{C}_\ell$ and hence $\mathcal{R}_{\ell+1}^\perp \subset \mathcal{C}_{\ell+1}$. To show the other inclusion, let $\omega \in \mathcal{C}_{\ell+1}$; then

$$\dot{\omega} \in \mathcal{C}_\ell + \dot{\mathcal{V}}^{*^\perp} = \mathcal{R}_\ell^\perp + \dot{\mathcal{V}}^{*^\perp}. \tag{96}$$

Thus there exists $\omega_1 \in \mathcal{V}^{*^\perp}$ and $\omega_2 \in \mathcal{R}_\ell^\perp$ such that $\dot{\omega} = \dot{\omega}_1 + \omega_2$. Let now $\dot{\omega}_0 = \dot{\overbrace{\omega - \omega_1}} = \omega_2$. So $\dot{\omega}_0 \in \mathcal{R}_\ell^\perp$. This implies that

$$\begin{array}{rcl} \omega_0 & \in & \{\omega = \alpha dx \mid \dot{\omega} \in \mathcal{R}_\ell^\perp\} = \{\alpha dx \mid \alpha A dx + \alpha B du \in \mathcal{R}_\ell^\perp\} \\ & = & \{\alpha \mid \alpha A \in \mathcal{R}_\ell^\perp\} \cap \mathcal{B}^\perp = {}^{-1}A\mathcal{R}_\ell^\perp \cap \mathcal{B}^\perp. \end{array} \tag{97}$$

So $\omega = \omega_1 + \omega_0 \in \mathcal{R}_{\ell+1}^\perp$, which yields that $\mathcal{C}_{\ell+1} \subset \mathcal{R}_{\ell+1}^\perp$. Thus, we have that $\mathcal{C}_\mu = \mathcal{R}_\mu^\perp$ for all $\mu \in \mathbb{N}$, which establishes our claim. □

REFERENCES

[1] R. Andiarti, C.H. Moog, and H.J.C. Huijberts, *Fixed modes in quasi-static state feedback decoupling*, in Proc. 3rd International Federation on Automatic Control Conference on System Structure and Control, Nantes, France, 1995, pp. 132–136.

[2] E. Aranda-Bricaire, C.H. Moog, and J.B. Pomet, *A linear algebraic framework for dynamic feedback linearization*, IEEE Trans. Automat. Control, 40 (1995), pp. 127–132.

[3] M.F. Atiyah and I.G. MacDonald, *Introduction to Commutative Algebra*, Addison-Wesley, Reading, MA, 1969.
[4] G. Basile and G. Marro, *Controlled and Conditioned Invariants in Linear System Theory*, Prentice-Hall, Englewood Cliffs, NJ, 1992.
[5] S. Battilotti, *A sufficient condition for nonlinear noninteracting control with stability via dynamic state feedback*, IEEE Trans. Automat. Control, 36 (1991), pp. 1033–1045.
[6] S. Battilotti and W.P. Dayawansa, *Noninteracting control with stability for a class of nonlinear system*, Systems Control Lett., 17 (1991), pp. 327–338.
[7] S. Battilotti, *Noninteracting Control With Stability for Nonlinear Systems*, Lecture Notes in Control Information Sciences 196, Springer-Verlag, Berlin, 1994.
[8] R.W. Brockett, *Feedback invariants for nonlinear systems*, in Proc. International Federation on Automatic Control Congress, Helsinki, 1978, pp. 1115–1120.
[9] L. Cao and Y.F. Zheng, *On minimal compensators for decoupling control*, Systems Control Lett., 18 (1992), pp. 121–128.
[10] D. Cheng and T.J. Tarn, *New results on (f,g)-invariance*, Systems Control Lett., 12 (1989), pp. 319–326.
[11] W.P. Dayawansa, D. Cheng, W.M. Boothby, and T.J. Tarn, *Global (f, g)-invariance of nonlinear systems*, SIAM J. Control Optim., 25 (1988), pp. 1119–1132.
[12] E. Delaleau and M. Fliess, *An algebraic interpretation of the structure algorithm with an application to feedback decoupling*, in Proc. Nonlinear Control Symposium 1992, Bordeaux, France, pp. 489–494.
[13] E. Delaleau and M. Fliess, *Algorithme de structure, filtrations et découplage*, C. R. Acad. Sci. Paris Sér. I Math., 315 (1992), pp. 101–106.
[14] E. Delaleau, *Sur les derivées de l'entrée en représentation et commande des systèmes non linéaires*, Doctoral thesis in Science, Université Paris XI Orsay, 1993.
[15] E. Delaleau and P.S. Pereira da Silva, *Rank conditions for dynamic disturbance decoupling problem*, in Proc. 33rd IEEE Conference on Decision and Control, Lake Buena Vista, FL, 1994.
[16] E. Delaleau and J. Rudolph, *Some remarks on quasi-static feedback of generalized states*, in Proc. 1st International Federation on Automatic Control Workshop on New Trends in Design of Control Systems, Smolenice, Slovakia, 1994, pp. 47–52.
[17] E. Delaleau and J. Rudolph, *Decoupling and linearization by quasi-static feedback of generalized states*, in Proc. European Control Conference, Rome, 1995, pp. 1069–1074.
[18] M.D. Di Benedetto, J.W. Grizzle, and C.H. Moog, *Rank invariants of nonlinear systems*, SIAM J. Control Optim., 27 (1989), pp. 658–672.
[19] M. Fliess, *A new approach to the noninteracting control problem in nonlinear systems theory*, in Proc. 23rd IEEE Conference on Decision and Control, Allerton, Monticello, 1985, pp. 123–129.
[20] M. Fliess, *Automatique et corps différentiels*, Forum Math., 1 (1989), pp. 227–238.
[21] A. Glumineau, C.H. Moog, and T.J. Tarn, *Interconnected zero dynamics in nonlinear systems and their role in dynamic noninteracting control with stability*, in New Trends in Systems Theory, G. Conte, A.M. Perdon, and B. Wyman, eds., Birkhäuser Boston, Cambridge, MA, 1991, pp. 316–323.
[22] K.A. Grasse, *On controlled invariance for fully nonlinear systems*, Internat. J. Control, 56 (1992), pp. 1121–1137.
[23] I.J. Ha, *The standard decomposed system and noninteracting feedback control of nonlinear systems*, SIAM J. Control Optim., 26 (1988), pp. 1235–1249.
[24] R.M. Hirschorn, *(A, B)-invariant distributions and disturbance decoupling of nonlinear systems*, SIAM J. Control Optim., 19 (1981), pp. 1–19.
[25] H.J.C. Huijberts, H. Nijmeijer, and L.L.M. van der Wegen, *Dynamic disturbance decoupling for nonlinear systems: The nonsquare and noninvertible case*, in Analysis of Controlled Dynamical Systems, B. Bonnard, B. Bride, J.P. Gauthier, and I. Kupka, eds., Birkhäuser Boston, Cambridge, MA, 1991, pp. 243–252.
[26] H.J.C. Huijberts, H. Nijmeijer, and L.L.M. van der Wegen, *Dynamic disturbance decoupling for nonlinear systems*, SIAM J. Control Optim., 30 (1992), pp. 336–349.
[27] H.J.C. Huijberts, *Dynamic Feedback in Nonlinear Synthesis Problems*, CWI Tract 101, Centrum voor Wiskunde en Informatica (CWI), Amsterdam, 1994.
[28] H.J.C. Huijberts and C.H. Moog, *Controlled invariance of nonlinear systems: Nonexact forms speak louder than exact forms*, in Systems and Networks: Mathematical Theory and Applications, Vol. II, U. Helmke, R. Mennicken, and J. Saurer, eds., Akademie Verlag, Berlin, 1994, pp. 245–248.

[29] A. Isidori, *Nonlinear Control Systems* 2nd ed., Springer-Verlag, Berlin, 1989.
[30] A. Isidori, *Control of nonlinear systems via dynamic state feedback*, in Algebraic and Geometric Methods in Nonlinear Control Theory, M. Fliess and M. Hazewinkel, eds., D. Reidel, Dordrecht, The Netherlands, 1986.
[31] A. Isidori, A.J. Krener, C. Gori-Giorgi, and S. Monaco, *Nonlinear decoupling via feedback: A differential geometric approach*, IEEE Trans. Automat. Control, 26 (1981), pp. 331–345.
[32] A. Isidori and J.W. Grizzle, *Fixed modes and nonlinear noninteracting control with stability*, IEEE Trans. Automat. Control, 33 (1988), pp. 907–914.
[33] A.J. Krener, (Ad_f, g), (ad_f, g) *and locally* (ad_f, g) *invariant and controllability distributions*, SIAM J. Control Optim., 23 (1985), pp. 523–549.
[34] A.J. Krener and A. Isidori, (Ad_f, G) *invariant and controllability distributions*, in Feedback Control of Linear and Nonlinear Systems, Lecture Notes in Control and Information Science 39, D. Hinrichsen and A. Isidori, eds., Springer-Verlag, Berlin, 1982, pp. 157–164.
[35] C.H. Moog, *Nonlinear decoupling and structure at infinity*, Math. Control Signals Systems, 1 (1988), pp. 257–268.
[36] H. Nijmeijer, *Controlled invariance for affine control systems*, Internat. J. Control, 34 (1981), pp. 824–833.
[37] H. Nijmeijer and A.J. van der Schaft, *Controlled invariance for nonlinear systems*, IEEE Trans. Automat. Control, 27 (1982), pp. 904–914.
[38] H. Nijmeijer and A.J. van der Schaft, *Nonlinear Dynamical Control Systems*, Springer-Verlag, New York, 1990.
[39] H. Nijmeijer, *Controllability distributions for nonlinear control systems*, Systems Control Lett., 2 (1982), pp. 122–129.
[40] A.M. Perdon, G. Conte, and C.H. Moog, *Some canonical properties of nonlinear systems*, in Robust Control of Linear Systems and Nonlinear Control, M.A. Kaashoek, J.H. van Schuppen, and A.C.M. Ran, eds., Birkhäuser Boston, Cambridge, MA, 1990, pp. 89–96.
[41] A.M. Perdon, Y.F. Zheng, C.H. Moog, and G. Conte, *Disturbance decoupling for nonlinear systems: A unified approach*, Kybernetika, 29 (1993), pp. 479–489.
[42] P.S. Pereira da Silva and V.M. Pinto Leite, *Disturbance decoupling by regular dynamic feedback for affine nonlinear systems: A linear algebraic approach*, in Proc. International Federation on Automatic Control Control Conference, Sydney, 1993, pp. VI-387-VI-390.
[43] V. Ramakrishna, *Controlled invariance for singular distributions*, SIAM J. Control Optim., 32 (1994), pp. 790–807.
[44] W. Respondek, *Disturbance decoupling via dynamic feedback*, in Analysis of Controlled Dynamical Systems, B. Bonnard, B. Bride, J.P. Gauthier, and I. Kupka, eds., Birkhäuser Boston, Cambridge, MA, 1991, pp. 347–357.
[45] W. Respondek, *Dynamic controllability distributions in nonlinear systems*, in Proc. 2nd International Federation on Automatic Control Workshop on Systems Structure and Control, Prague, 1992, pp. 256–258.
[46] J. Rudolph, *A canonical form under quasistatic feedback*, C. R. Acad. Sci. Paris Sér. I Math., 316 (1993), pp. 1323–1328.
[47] K.G. Wagner, *Nonlinear noninteraction with stability by dynamic state feedback*, SIAM J. Control Optim., 29 (1991), pp. 609–622.
[48] W.M. Wonham, *Linear Multivariable Control: A Geometric Approach*, 3rd ed., Springer-Verlag, Berlin, 1985.
[49] X. Xia, *Parametrization of decoupling control laws for affine nonlinear systems*, IEEE Trans. Automat. Control, 38 (1993), pp. 916–928.
[50] W. Zhan, T.J. Tarnm, and A. Isidori, *A canonical dynamic extension for noninteraction with stability for affine nonlinear square systems*, Systems Control Lett., 17 (1991), pp. 177–184.

SIAM J. CONTROL OPTIM.
Vol. 35, No. 3, pp. 980–1001, May 1997

011

FINITE-DIMENSIONAL FILTERS. PART I: THE WEI–NORMAN TECHNIQUE*

M. COHEN DE LARA†

Abstract. This two-part paper deals with necessary or sufficient conditions for the existence of finite-dimensional filters. In this first part, we set the problem and propose a construction of such filters by the Wei–Norman technique. After having formulated the problem of finite-dimensional filters in terms of finite-dimensional realizations of input-output mappings, we specify the dependence with respect to the initial measure. We show how different notions of dependence imply different properties of the so-called estimation algebra $\mathcal{E}$: $\mathcal{E}$ is homomorphic to a Lie algebra of vector fields; $\mathcal{E}$ contains only operators of order less than or equal to two; $\mathcal{E}$ is finite dimensional and contains only operators of order less than or equal to two. These results depend on a precise definition of a finite-dimensional realization, especially on what concerns the domain of the output function. The last (and most stringent) condition on $\mathcal{E}$ will be shown to be almost sufficient to recover a family of finite-dimensional realizations thanks to the proof of a Baker–Campbell–Hausdorff formula which allows us to apply the Wei–Norman technique in a quite general setting.

Key words. finite-dimensional filter, estimation Lie algebra, bilinear stochastic partial differential equation, Baker–Campbell–Hausdorff formula.

AMS subject classifications. 93E11, 60G35, 22E15, 47D06

PII. S0363012994270904

1. Introduction.

The filtering problem for systems with correlated noises of the form (see [11, 48])

$$\begin{cases} dx_t = f(x_t)dt + g(x_t)dv_t + \tilde{g}(x_t)dy_t, & x_0 \rightsquigarrow \mu_0, \\ dy_t = h(x_t)dt + dw_t, & y_0 = 0, \end{cases} \tag{1}$$

where $x \in \mathbb{R}^n$, $f(x) \in \mathbb{R}^n$, $g(x) = (g_1(x), ..., g_m(x)) \in (\mathbb{R}^n)^m$, $\tilde{g}(x) = (\tilde{g}_1(x), ..., \tilde{g}_m(x)) \in (\mathbb{R}^n)^m$, $y \in \mathbb{R}^p$, $h(x) \in \mathbb{R}^p$ is the characterization of the conditional law Π_t of x_t knowing $\mathcal{Y}_t$, the σ-field generated by $\{y_s \,|\, 0 \le s \le t\}$, or at least of some statistics $\langle \Pi_t, \phi \rangle$ such as the conditional mean or variance.

Under quite general assumptions, the reference probability method allows us to define an unnormalized conditional law σ_t which satisfies a (generally) infinite-dimensional bilinear stochastic partial differential equation (PDE): the Duncan–Mortensen–Zakai equation (Zakai equation, for short) [47, 44, 15, 37]

$$\begin{aligned} d\sigma_t(\phi) = \sigma_t \left(\frac{1}{2} \sum_{k=1}^{m} L_{g_k}^2 \phi + L_f \phi - \frac{1}{2} \left(\|h\|^2 + \sum_{i=1}^{p} L_{\tilde{g}_i} h_i \right) \phi \right) dt \\ + \sum_{i=1}^{p} \sigma_t (L_{\tilde{g}_i} \phi + h_i \phi) \circ dy_t^i, \quad \sigma_0 = \mu_0. \end{aligned} \tag{2}$$

Remark. For any vector field X, we make use of the Lie derivative notation

$$L_X \phi(\xi) = X \cdot \phi(\xi) = \langle d\phi, X \rangle (\xi) = \lim_{t \to 0} \frac{\phi(\Phi_t^X(\xi)) - \phi(\xi)}{t}, \tag{3}$$

*Received by the editors July 8, 1994; accepted for publication (in revised form) April 8, 1996.
http://www.siam.org/journals/sicon/35-3/27090.html

†Centre d'Enseignement et de Recherche pour la Gestion des Ressources Naturelles et de l'Environnement, École Nationale des Ponts et Chaussées, 6 et 8 av. Blaise Pascal, Cité Descartes, 77455 Marne la Vallée Cédex 2, France (mcdl@cergrene.enpc.fr).

where Φ_t^X denotes the local flow generated by X. Of course, we have $L_X^k\phi = L_X(L_X^{k-1}\phi)\ \forall\ k \geq 1$ with the convention $L_X^0\phi = \phi$.

As was noticed by Brockett [9, 8], the solution of (2) usually defines an input output map

$$\mathcal{F}^{\mu_0} : \mathbb{R}_+ \times C^0(\mathbb{R}_+, \mathbb{R}^p) \to \mathbb{M}_+(\mathbb{R}^n), \tag{4}$$

where μ_0 is the initial law and $\mathbb{M}_+(\mathbb{R}^n)$ denotes the space of nonnegative bounded measures on $\mathbb{R}^n$. Then, in the spirit of system theory, we can define a finite-dimensional realization (FDR) of such an input-output map. This is just a specific way to formulate finite-dimensional filtering problems.

Such problems find their origin in the linear Gaussian case, where the solution σ_t of the Zakai equation evolves in the set of unnormalized Gaussian measures and where the corresponding finite number of parameters satisfies a stochastic differential equation, driven by the observations, on a finite-dimensional manifold. The extension of this property to nonlinear systems, or non-Gaussian initial laws, has given rise to the notion of finite-dimensional filters (FDFs). After the example of Beneš [3] a few others were discovered [5, 45, 39, 4, 28, 25, 14] and also [27, 13, 18] in a different setting. The estimation Lie algebra $\mathcal{E}$ associated with the Zakai equation [9, 10] plays a crucial role in the study, thus enhancing the existing links with geometric or algebraic methods [8, 10, 31, 29, 19]. Results of nonexistence are mostly obtained [34, 26], even in the case of the finite-dimensional computation of some statistics [11, 20, 21]. Nevertheless, whenever FDFs are known to exist, the estimation Lie algebra is a powerful tool in the computation of a solution [3, 34, 33, 45, 25, 46]. Other approaches use analytic methods [2, 35, 36], geometric tools [1, 26], classical probabilistic methods [28, 24], or Malliavin stochastic calculus of variations [30, 25, 32].

In this paper, we restrict ourselves to the problem of the existence of an FDF for the conditional law and more generally to the study of FDRs of bilinear stochastic PDEs very similar to equation (2) as in [26]. Our aim is to give tools for explicit computation of such realizations. We focus on the role of the initial measure to define specific families of FDRs $\{\mathcal{F}^{\mu_0} \mid \mu_0 \in \mathbb{M}_0\}$, parametrized by a set of initial measures $\mathbb{M}_0$. According to the more or less strong dependence of the elements of the FDR on $\mu_0 \in \mathbb{M}_0$, this yields a classification of estimation Lie algebras based on the following results.

1. A necessary condition for the existence of an FDR for one single initial law μ_0 is the existence of a nontrivial homomorphism from $\mathcal{E}$ to a Lie algebra of vector fields on a finite-dimensional manifold. This property, known as the conjecture of Brockett, has already been sketched out in the case of finite-dimensional statistics for the cubic sensor problem (section 3) [20, 21].

2. For bilinear stochastic bilinear PDEs admitting FDRs for every Dirac initial law with an additional regularity property (conic or regular FDR), we show in section 4 that $\mathcal{E}$ contains only operators of order less than or equal to two. The proof relies on a characterization of C^∞-differential operators satisfying the maximum principle to be found in [6] for instance. As was noticed in [26], $\mathcal{E}$ is generally infinite dimensional.

3. For bilinear stochastic bilinear PDEs admitting regular FDRs parametrized by Dirac measures and uniform with respect to the dynamics and initial condition (namely the same regular FDR holds for all Dirac initial law up to the output function), we show in section 5 that $\mathcal{E}$ is finite dimensional and has a basis consisting of one operator of order less than or equal to two and operators of order less than or equal to one. It turns out that such uniform realizations may generally be extended to many more initial laws than the Dirac ones.

4. In section 5, these results are used to compute FDRs uniform with respect to the dynamics and initial condition under additional regularity and algebraic assumptions. The key point is a extension of the Baker–Campbell–Hausdorff formula to linear partial differential operators, which is proved in the appendix.

In Part II, we shall propose another means of constructing FDRs by invariance group technique developed in [12], and we shall relate both methods.

2. Problem statement. Let $(y_t)_{t\geq 0}$ be a standard p-dimensional Brownian motion, and let $\circ dy_t$ denote its "Stratonovitch differential." For $\mu_0 \in \mathbb{M}_+(\mathbb{R}^n)$, the space of nonnegative bounded measures on $\mathbb{R}^n$, we consider the following stochastic bilinear PDE:

$$d\nu_t(\phi) = \nu_t(M_0\phi)dt + \sum_{i=1}^{p} \nu_t(M_i\phi) \circ dy_t^i \quad \forall\, \phi \in \mathcal{D}(\mathbb{R}^n), \quad \nu_0 = \mu_0. \tag{5}$$

Here $\mathcal{D}(\mathbb{R}^n)$ denotes the space of smooth functions on $\mathbb{R}^n$ with compact support, and $M_0, M_1, \ldots, M_p$ are the linear differential operators defined by

$$\begin{cases} M_0\phi = \mathcal{L}\phi + H\phi, \\ M_i\phi = L_{\tilde{g}_i}\phi + h_i\phi, \quad i = 1, \ldots, p \qquad \forall\, \phi \in \mathcal{D}(\mathbb{R}^n), \end{cases} \tag{6}$$

$\mathcal{L}$ being a smooth diffusion operator ($\mathcal{L}1 = 0$), $\tilde{g}_1, \ldots, \tilde{g}_p$ being smooth vector fields on $\mathbb{R}^n$, and $H, h_1, \ldots, h_p$ being smooth functions from $\mathbb{R}^n$ to $\mathbb{R}$.

Remark. In the filtering problem stated in the introduction, the Zakai equation is a particular case of equation (5) with

$$\mathcal{L}\phi = \frac{1}{2}\sum_{k=1}^{m} L_{g_k}^2\phi + L_f\phi \quad \text{and} \quad H = -\frac{1}{2}\left(\|h\|^2 + \sum_{i=1}^{p} L_{\tilde{g}_i}h_i\right). \tag{7}$$

Assumption 1. *For all $\mu_0 \in \mathbb{M}_+(\mathbb{R}^n)$, we assume that there exists an input-output map*

$$\mathcal{F}^{\mu_0} : \mathbb{R}_+ \times C^0(\mathbb{R}_+, \mathbb{R}^p) \to \mathbb{M}_+(\mathbb{R}^n) \tag{8}$$

which satisfies the following property: for all p-dimensional Brownian motion $(y_t)_{t\geq 0}$, equation (5) *has a unique solution $(\nu_t)_{t\geq 0}$ given by*

$$\text{with probability one } (w.p.1) \quad \forall\, t \geq 0, \quad \nu_t = \mathcal{F}^{\mu_0}(t; y_s, s \leq t). \tag{9}$$

In particular, uniqueness for the solution of equation (5) *is thus assumed.*

Remark. Other notions of input-output mappings can be considered, such as $\mathcal{F}^{\mu_0}(t; y_s, s \leq t) = \nu_t(\phi) \in \mathbb{R}$ (conditional statistics).

This input-output map $\mathcal{F}^{\mu_0}$ can be realized in finite dimension when it can be written $\nu_t = \theta(\xi_t)$, where ξ_t is driven by $(y_s)_{s\geq 0}$ and evolves on a finite-dimensional manifold M and θ is a nonnegative output function.

DEFINITION 2.1. *An* output function θ *on a finite-dimensional manifold M is a map*

$$\theta : \mathrm{Dom}(\theta) \to \mathcal{D}'(\mathbb{R}^n), \tag{10}$$

where $\mathrm{Dom}(\theta)$ *is an open set of M and $\mathcal{D}'(\mathbb{R}^n)$ is the space of distributions such that for every $\phi \in \mathcal{D}(\mathbb{R}^n)$ the function*

$$\begin{array}{rcl} \mathrm{Dom}(\theta) & \to & \mathbb{R}, \\ \xi & \mapsto & \langle \theta(\xi), \phi \rangle \end{array} \tag{11}$$

is of class C^∞ *(on the open set* $\mathrm{Dom}(\theta)$*) and has continuous partial derivatives of all order on* $\overline{\mathrm{Dom}(\theta)}$*. A* nonnegative output function *is an output function with values in* $\mathbb{M}_+(\mathbb{R}^n)$*.*

A well-known example of such an output function is given by the Gaussian family

$$\theta(m,\sigma) = \frac{1}{\sigma\sqrt{2\pi}} \exp\left(-\frac{(x-m)^2}{2\sigma^2}\right) dx \in \mathbb{M}_+(\mathbb{R}^n) \subset \mathcal{D}'(\mathbb{R})$$

with $\mathrm{Dom}(\theta) = \mathbb{R} \times (0,+\infty) \subset M = \mathbb{R}^2$ such that for every $\phi \in \mathcal{D}(\mathbb{R})$ the function

$$(m,\sigma) \mapsto \langle \theta(\xi),\, \phi\rangle = \frac{1}{\sqrt{2\pi}} \int_{-\infty}^{+\infty} \exp\left(-\frac{z^2}{2}\right) \phi(m+\sigma z)\, dz$$

is of class C^∞ and has continuous partial derivatives of all order on $\mathbb{R}^2$ (thus a fortiori on $\overline{\mathrm{Dom}(\theta)} = \mathbb{R} \times [0,+\infty)$).

DEFINITION 2.2. *An* FDR *of the input-output map* $\mathcal{F}^{\mu_0}$ *given by* (8) *consists of a collection* $(M,\, \xi_0,\, b_0,\, b_1,\, \ldots,\, b_p,\, \theta)$*, where*

1. M *is a smooth finite-dimensional manifold,*
2. $\xi_0 \in M$*,*
3. $b_0,\, b_1,\, \ldots,\, b_p$ *are smooth vector fields on* M*,*
4. θ *is a nonnegative output function on* M*,*

such that, if $(y_t)_{t\geq 0}$ *is a Brownian motion, the Stratonovitch stochastic differential equation*

$$d\xi_t = b_0(\xi_t)dt + \sum_{i=1}^{p} b_i(\xi_t) \circ dy_t^i, \quad \xi_{t\,|\,t=0} = \xi_0, \tag{12}$$

*is conservative (that is, w.p.*1 *has a solution for all time or w.p.*1 *has infinite explosion time) with a solution satisfying*

$$w.p.1 \quad \forall\, t > 0, \quad \xi_t \in \mathrm{Dom}(\theta)$$

and

$$w.p.1 \quad \forall\, t \geq 0, \quad \theta(\xi_t) = \mathcal{F}^{\mu_0}(t; y_s, s \leq t) = \nu_t. \tag{13}$$

We shall also say that system (1) *admits an* FDF *if the corresponding Zakai equation* (2) *admits an* FDR*.*

Remark. With the above definition, note that the deterministic point ξ_0 necessarily belongs to $\overline{\mathrm{Dom}(\theta)}$.

This definition can be extended to the case of a family of input-output maps, corresponding to the case where we want to specify the dependence with respect to the initial measure μ_0. Clearly, if $\{\mathcal{F}^{\mu_0} \mid \mu_0 \in \mathbb{M}_0\}$ is a given family of input-output maps, the corresponding family of FDRs, if it exists, is denoted by $(M^{\mu_0},\, \xi_0^{\mu_0},\, b_0^{\mu_0},\, b_1^{\mu_0},\, \ldots,\, b_p^{\mu_0},\, \theta^{\mu_0})$.

However, particular subclasses of families of stochastic realizations may be useful and will receive particular attention.

DEFINITION 2.3. *Given* $\mathbb{M}_0 \subset \mathbb{M}_+(\mathbb{R}^n)$ *and a family of input-output maps* $\{\mathcal{F}^{\mu_0} \mid \mu_0 \in \mathbb{M}_0\}$*, an FDR of* $\{\mathcal{F}^{\mu_0} \mid \mu_0 \in \mathbb{M}_0\}$*,* uniform with respect to the dynamics *(resp.,* to the dynamics and initial condition*), is a family of FDRs for which* $M^{\mu_0},\, b_0^{\mu_0},\, b_1^{\mu_0},\, \ldots,\, b_p^{\mu_0}$ *(resp.,* $M^{\mu_0},\, b_0^{\mu_0},\, b_1^{\mu_0},\, \ldots,\, b_p^{\mu_0}$*, and* $\xi_0^{\mu_0}$*) do not depend on* $\mu_0 \in \mathbb{M}_0$*.*

Of course, when a family of FDRs is uniform with respect to the dynamics and initial condition, then only the nonnegative output function θ may depend upon the initial law μ_0.

FDFs give rise to FDRs, and examples of FDRs, uniform with respect to the dynamics, are given by the Kalman–Bucy filter [23] and by Beneš [3], while examples of FDRs, uniform with respect to the dynamics and initial condition, are given by Beneš and Karatzas [5], Makowski [28], Wong [45], Lévine [25], and Yau [46].

An important tool in the study of FDRs is the following estimation Lie algebra.

DEFINITION 2.4. *The* estimation Lie algebra *of equation* (5)*, noted $\mathcal{E}$, is the* $\mathbb{R}$*-vector space of linear differential operators generated by $M_0, \ldots, M_p$ (defined by* (6)*) with respect to the Lie bracket defined by*

$$[N_1, N_2]\phi = N_1(N_2\phi) - N_2(N_1\phi) \quad \forall\, \phi \in \mathcal{D}(\mathbb{R}^n).$$

Let $x_0 \in \mathbb{R}^n$. If N is any smooth linear partial differential operator, $N(x_0)$ denotes the linear partial differential operator with constant coefficients, obtained by freezing at x_0 all the coefficients of N. $\mathcal{E}(x_0)$ denotes the vector space consisting of all partial differential operators with constant coefficients of the form $N(x_0)$ for $N \in \mathcal{E}$. If $N \in \mathcal{E}$, its dual operator $N^\star$ is given by $\langle \lambda, N\phi\rangle = \langle N^\star\lambda, \phi\rangle$ for all ϕ in $\mathcal{D}(\mathbb{R}^n)$ and $\lambda \in \mathcal{D}'(\mathbb{R}^n)$.

3. FDRs and the homomorphism property. The conjecture of Brockett is as follows: if the system (1) admits an FDF for the conditional law or for conditional statistics, there should exist a nontrivial Lie algebra homomorphism from the estimation Lie algebra to a Lie algebra of vector fields [8].

In this section, we assume that the input-output map $\mathcal{F}^{\mu_0}$ given by (8) admits an FDR given by Definition 2.2.

DEFINITION 3.1. *$\mathcal{B}$ denotes the Lie algebra of vector fields on M generated by the vector fields $b_0, \ldots, b_p$ given by Definition* 2.2*.*

Proofs of the conjecture of Brockett are discussed and outlined by Hazewinkel, Marcus, and Sussmann in [20] and also by Hijab [21] in the case of existence of FDFs for conditional statistics of the cubic sensor. It should be noted that the construction of the homomorphism is different according to both authors, especially because of the choice of various ideals of $\mathcal{B}$ used to quotient this latter algebra. However, both methods can be extended to other systems than the cubic sensor. The proof of the conjecture in [11] is quite different since it relies on a specific dependence of a family of FDRs upon Dirac initial measures.

In the spirit of these previous works, we shall exhibit an ideal $\mathcal{J}$ of $\mathcal{B}$ as well as a nontrivial Lie algebra homomorphism from $\mathcal{E}$ to a quotient $\mathcal{B} \setminus \mathcal{J}$ of $\mathcal{B}$. But here the definition of the set $\mathcal{J}$ is different from others and does not obviously imply that it is an ideal of $\mathcal{B}$. In fact, this latter property holds because we are dealing with *universal* FDRs (in the terminology of [11]), that is, the realizations of the process ν_t and not only of some statistics $\langle \nu_t, \phi\rangle$. This homomorphism will be useful in the next sections.

DEFINITION 3.2. *Let θ be an output function on a finite-dimensional manifold M and X be a smooth vector field on M. We define a new output function $L_X\theta$ by*

$$\langle L_X\theta(\xi)\,,\, \phi\rangle = L_X\left(\langle \theta(\cdot)\,,\, \phi\rangle\right)_{|\xi} \tag{14}$$

for all $\xi \in \mathrm{Dom}(\theta)$ and $\phi \in \mathcal{D}(\mathbb{R}^n)$.

Note that $L_X\theta$ indeed is an output function because $L_X\theta(\xi)$ is a weak limit of distributions.

PROPOSITION 3.3. *If the input-output map $\mathcal{F}^{\mu_0}$ given by* (8) *admits an FDR given by Definition* 2.2*, then the subset*

$$\mathcal{J} = \{\, j \in \mathcal{B} \mid w.p.1 \quad \forall\, t > 0, \quad L_j\theta(\xi_t) = 0 \,\} \tag{15}$$

is a Lie ideal of $\mathcal{B}$. Furthermore, there exists a surjective Lie algebras homomorphism from $\mathcal{E}$ to $\mathcal{B} \setminus \mathcal{J}$.

The proof relies on the following lemma which can be found in [26, proof of Theorem 3.1 and equations (12), (16), pp. 80–82].

LEMMA 3.4. *If the input-output map $\mathcal{F}^{\mu_0}$ given by* (8) *admits an FDR given by Definition* 2.2*, we have w.p.*1 *for all $t > 0$:*

$$\begin{aligned} M_i^\star\theta(\xi_t) &= L_{b_i}\theta(\xi_t), \quad i = 0, \ldots, p, \\ [M_j^\star, M_i^\star]\theta(\xi_t) &= L_{[b_i, b_j]}\theta(\xi_t), \quad i, j = 0, \ldots, p, \\ &\cdots = \cdots \end{aligned} \tag{16}$$

In these latter equalities,

- the left-hand side represents the differential operator (on $\mathbb{R}^n$) $M_i^\star$, or $[M_j^\star, M_i^\star]$, acting on the measure (on $\mathbb{R}^n$) $\theta(\xi_t)$;
- the right-hand side represents the derivative of the mapping $\xi \mapsto \theta(\xi)$ along the vector field b_i (see (14)) or along $[b_i, b_j]$, evaluated at the point $\xi = \xi_t$.

Since $\mathcal{B}$ is generated by linear combinations of iterated brackets of the vector fields $b_0, \ldots, b_p$ and $\mathcal{E}$ by linear combinations of iterated brackets of the operators $M_0, \ldots, M_p$, the following corollary of equations (16) is clear.

COROLLARY 3.5. *If the input-output map $\mathcal{F}^{\mu_0}$ given by* (8) *admits an FDR given by Definition* 2.2*, then for any $E \in \mathcal{E}$, there exists at least one $b \in \mathcal{B}$ such that*

$$w.p.1 \quad \forall\, t > 0, \quad L_b\theta(\xi_t) = E^\star\theta(\xi_t). \tag{17}$$

Conversely, for any $b \in \mathcal{B}$, there exists at least one $E \in \mathcal{E}$ such that equation (17) *is satisfied.*

The following remark will be quite useful in what follows.

Remark. By Definition 2.2 (and the remark following it), we know that the deterministic point ξ_0 belongs to $\overline{\mathrm{Dom}(\theta)}$.

Therefore, thanks to Definition 2.1 (the assumption that the mapping (11) has continuous partial derivatives of all order on $\overline{\mathrm{Dom}(\theta)}$), we can extend the equality (17) at $t = 0$. This provides the following deterministic relationship: if the input-output map $\mathcal{F}^{\mu_0}$ given by (8) admits an FDR given by Definition 2.2, then for any $E \in \mathcal{E}$, there exists at least one $b \in \mathcal{B}$ such that

$$L_b\theta(\xi_0) = E^\star\theta(\xi_0). \tag{18}$$

Conversely, for any $b \in \mathcal{B}$, there exists at least one $E \in \mathcal{E}$ such that equality (18) is satisfied.

Note that (18) is satisfied with $b = b_i$ and $E = M_i$ for $i = 0, \ldots, p$.

Now we turn to the proof of Proposition 3.3.

Proof of Proposition 3.3. Let $j \in \mathcal{J}$, $b \in \mathcal{B}$ be given. To prove that $\mathcal{J}$ is an ideal of $\mathcal{B}$, we shall show that

$$\text{w.p.1} \quad \forall\, t > 0, \quad L_{[b,\,j]}\theta(\xi_t) = L_bL_j\theta(\xi_t) \,-\, L_jL_b\theta(\xi_t) = 0. \tag{19}$$

First, we show that w.p.1 $\forall\, t > 0, L_bL_j\theta(\xi_t) = 0$. In [26], the Itô–Stratonovitch formula is applied to the semimartingale $L_j\theta(\xi_t)$ in the nuclear space of distributions

[42] as follows:

$$\text{w.p.1} \quad \forall\, t > 0, \quad dL_j\theta(\xi_t) = L_{b_0}L_j\theta(\xi_t)dt + \sum_{i=1}^{p} L_{b_i}L_j\theta(\xi_t) \circ dy_t^i \,. \tag{20}$$

But j belongs to $\mathcal{J}$, so that the semimartingale $L_j\theta(\xi_t)$ is zero. Therefore, so are the absolutely continous part and the martingale parts, and this gives

$$\text{w.p.1} \quad \forall\, t > 0, \quad L_{b_i}L_j\theta(\xi_t) = 0, \quad i = 0, \ldots, p \,. \tag{21}$$

Now, since $b \in \mathcal{B}$ can be written as a linear combination of iterated brackets of the vector fields $b_0, \ldots, b_p$, iterations of the Itô–Stratonovitch formula and proper linear combinations of the above equations provide the expected result:

$$\text{w.p.1} \quad \forall\, t > 0, \quad L_bL_j\theta(\xi_t) = 0 \,. \tag{22}$$

Second, we show that w.p.1 $\forall\, t > 0, L_jL_b\theta(\xi_t) = 0$, and this is not as straightforward here as above.

We know from Corollary 3.5 that there exists $E \in \mathcal{E}$ such that (17) is satisfied. Therefore, if we apply the Itô–Stratonovitch formula enough times to this latter equation, followed by appropriate linear combinations, we find that

$$\text{w.p.1} \quad \forall\, t > 0, \quad L_jL_b\theta(\xi_t) = L_jE^\star\theta(\xi_t).$$

We now show that the last term $L_jE^\star\theta(\xi_t)$ is zero.

On the one hand, by (14), we have the following for $\xi \in \mathrm{Dom}(\theta)$:

$$\begin{aligned} \forall\, \phi \in \mathcal{D}(\mathbb{R}^n), \quad \langle L_jE^\star\theta(\xi), \phi\rangle &= L_j\left(\langle E^\star\theta(.)\,,\, \phi\rangle\right)_{|\xi} \quad \text{by def. of } L_j \\ &= L_j\left(\langle \theta(.)\,,\, E\phi\rangle\right)_{|\xi} \quad \text{by duality} \\ &= \langle L_j\theta(\xi)\,,\, E\phi\rangle \quad \text{by def. of } L_j \\ &= \langle E^\star L_j\theta(\xi)\,,\, \phi\rangle \quad \text{by duality.} \end{aligned}$$

Thus, with $\xi = \xi_t$, we get

$$\text{w.p.1} \quad \forall\, t > 0, \quad L_jL_b\theta(\xi_t) = L_jE^\star\theta(\xi_t) = E^\star L_j\theta(\xi_t).$$

On the other hand, since $j \in \mathcal{J}$, we have

$$\text{w.p.1} \quad \forall\, t > 0, \quad L_j\theta(\xi_t) = 0,$$

so that for almost all $\omega \in \Omega$ and for all $t > 0$, the measure (on $\mathbb{R}^n$) $L_j\theta(\xi_t(\omega))$ is zero. Now, $E^\star$ is a differential operator on the state-space $\mathbb{R}^n$ (acting on measures on $\mathbb{R}^n$) and thus for almost all $\omega \in \Omega$ and for all $t > 0$ we have

$$E^\star L_j\theta(\xi_t) = E^\star 0 = 0.$$

Combining the last three equalities we get the expected result:

$$\text{w.p.1} \quad \forall\, t > 0, \quad L_jL_b\theta(\xi_t) = L_jE^\star\theta(\xi_t) = E^\star L_j\theta(\xi_t) = E^\star 0 = 0 \,.$$

To end the proof, we define the surjective homomorphism from $\mathcal{E}$ to $\mathcal{B} \setminus \mathcal{J}$ as follows:

$$\Upsilon(E) = \{\, b \in \mathcal{B} \,|\, \text{w.p.1} \;\; \forall\, t \;>\; 0, \; L_b\theta(\xi_t) = E^\star\theta(\xi_t) \,\}. \tag{23}$$

It is indeed easily seen that $\Upsilon(E)$ belongs to $\mathcal{B}\backslash\mathcal{J}$ and, by making use of Corollary 3.5, that Υ is a surjective homomorphism. □

COROLLARY 3.6. *If the input-output map $\mathcal{F}^{\mu_0}$ given by* (8) *admits an FDR given by Definition* 2.2 *such that $\theta(\xi_t)$ is not constant, then there exists a nontrivial Lie algebras homomorphism from the estimation Lie algebra to an algebra of vector fields on a finite-dimensional manifold.*

Proof. With the previous notations, we choose for manifold a leaf N of the foliation generated by the involutive distribution $\eta \to \mathcal{J}(\eta)$ (by the Frobenius theorem [22, p. 25], since this latter distribution is of fixed dimension if we restrict η to the open subset of M consisting of points η where the subspace $\mathcal{J}(\eta)$ of M has maximal dimension).

We now exhibit a Lie algebra homomorphism from $\mathcal{B} \setminus \mathcal{J}$ to the Lie algebra of vector fields on N.

Let $\eta_* \in N$ and $(\eta^1, \ldots, \eta^r)$ be a coordinate system on a neighborhood V of η_* such that $N = \{\, \eta \in V \mid \eta^1(\eta) = \eta^1(\eta_*), \ldots, \eta^k(\eta) = \eta^k(\eta_*) \,\}$. Thus, if j belongs to $\mathcal{J}$, we have

$$j = \sum_{i=1}^{k} j_i(\eta)\, \frac{\partial}{\partial \eta^i}.$$

Since $\mathcal{J}$ is an ideal, an easy computation provides for any $a \in \mathcal{B}$ [22, pp. 45–46]:

$$a = \sum_{i=1}^{k} a_i(\eta^1, \ldots, \eta^r)\, \frac{\partial}{\partial \eta^i} \; + \sum_{i=1}^{r-k} a_{k+i}(\eta^{k+1}, \ldots, \eta^r)\, \frac{\partial}{\partial \eta^{k+i}}.$$

Let C be a class of $\mathcal{B} \setminus \mathcal{J}$: there exists $c \in \mathcal{B}$ such that any $a \in C$ satisfies $c - a \in \mathcal{J}$. In coordinates, this means that there exist $r - k$ functions $c_{k+1}(\eta^{k+1}, \ldots, \eta^r)$, $\ldots$, $c_r(\eta^{k+1}, \ldots, \eta^r)$ on V such that any a in C can be written

$$a = \sum_{i=1}^{k} a_i(\eta^1, \ldots, \eta^r)\, \frac{\partial}{\partial \eta_i} \; + \sum_{i=1}^{r-k} c_{k+i}(\eta^{k+1}, \ldots, \eta^r)\, \frac{\partial}{\partial \eta^{k+i}}.$$

Thus $\overline{c} = \sum_{i=1}^{r-k} c_{k+i}(\eta^{k+1}, \ldots, \eta^r)\, \frac{\partial}{\partial \eta^{k+i}}$ is uniquely determined by the class C, and we can in this way assign $\overline{c}$ to C and thus define $\upsilon(C) = \overline{c}$.

It is easily seen that υ is a Lie algebras homomorphism from $\mathcal{B} \setminus \mathcal{J}$ to the Lie algebra of vector fields on the manifold N. On the other hand, υ is not trivial because $\upsilon = 0$ means that $\mathcal{B} \subset \mathcal{J}$, or equivalently that $\theta(\xi_t)$ is constant.

To end the proof, we note that since Υ in (23) is a surjective homomorphism from $\mathcal{E}$ to $\mathcal{B} \setminus \mathcal{J} \neq 0$, then $\Upsilon \circ \upsilon$ is a nontrivial homomorphism from $\mathcal{E}$ to the Lie algebra of vector fields on the manifold N. □

4. FDRs for Dirac initial laws. Although the estimation Lie algebra $\mathcal{E}$ of a system admitting an FDF is necessarily homomorphic to a Lie algebra of vector fields on a finite-dimensional manifold, it appears that, for all the known cases of systems admitting FDFs, the estimation Lie algebra $\mathcal{E}$ has much stronger properties: it is either both finite dimensional and made of operators of order less than or equal to two [23, 3, 5, 45, 4, 28, 14, 25, 41, 46] or made only of operators of order less than or equal to two [26].

We shall prove in this section and the following one that these last properties of $\mathcal{E}$ can be related to the existence of FDRs for the family $\{\mathcal{F}^{\mu_0} \mid \mu_0 \in \mathbb{D}_0\}$, where $\mathbb{D}_0$ is the set of Dirac measures.

First, a technical difficulty arises because Dirac measures are extremal measures in the cone of nonnegative measures. Indeed, we shall prove that, when an FDR exists, the parametrization of the nonnegative output function θ is almost always singular around any point ξ_0 such that $\theta(\xi_0) \in \mathbb{D}_0$.

LEMMA 4.1. *Let $x_0 \in \mathbb{R}^n$ be such that*

1. *the second-order part of the operator M_0 given in equation (6) is not zero at x_0, that is $M_0(x_0)$ is a constant linear differential operator of order equal to two (but may be degenerate),*
2. *$\mathcal{F}^{\delta_{x_0}}$, given by (8), admits an FDR.*

Then the initial point ξ_0 given in Definition 2.2 is necessarily a boundary point of the open domain $\mathrm{Dom}(\theta)$ of the nonnegative output function θ; that is, $\xi_0 \in \partial\mathrm{Dom}(\theta)$.

Proof. Let us note $M_0(x_0) = \sum_{i,j=1}^n s_{ij} \frac{\partial^2}{\partial x_i x_j} + \cdots$, where $S = (s_{ij})_{i,j=1,\ldots,n}$ is a nonnegative symmetric matrix. Let us choose $\phi \in \mathcal{D}(\mathbb{R}^n)$ nonnegative such that $\phi(x) = (x - x_0)' S (x - x_0)$ in a neighborhood of x_0. Since θ is a nonnegative output function and $\phi \geq 0$, we have $\langle \theta(\xi), \phi \rangle \geq 0$. On the other hand, we have $\langle \theta(\xi_0), \phi \rangle = \langle \delta_{x_0}, \phi \rangle = \phi(x_0) = 0$.

Now, assume that ξ_0 is not a boundary point of the open domain $\mathrm{Dom}(\theta)$. Therefore, all first-order partial derivatives of the mapping (11) are zero at ξ_0. In particular, we have $\langle L_{b_0} \theta(\xi_0), \phi \rangle = 0$, while on the other hand, we know from (16) that $M_0^\star \theta(\xi_0) = L_{b_0} \theta(\xi_0)$. Therefore,

$$(M_0\phi)(x_0) = \langle \delta_{x_0}, M_0\phi \rangle = \langle \theta(\xi_0), M_0\phi \rangle = \langle M_0^\star \theta(\xi_0), \phi \rangle = \langle L_{b_0}\theta(\xi_0), \phi \rangle = 0,$$

where ϕ is such that $(M_0\phi)(x_0) = \sum_{i,j=1}^n s_{ij}^2 > 0$, which contradicts the assumption on M_0. Thus, ξ_0 is a boundary point of the open domain $\mathrm{Dom}(\theta)$. □

In what follows, it appears that the shape of $\mathrm{Dom}(\theta)$ in the neighborhood of the boundary point ξ_0 is important to specify.

DEFINITION 4.2. *The* tangent cone *at ξ_0 is the cone of vectors v of the tangent space $T_{\xi_0}M$ at ξ_0 such that there exists a smooth path $\gamma : [0, T] \to M$ satisfying*

1. *$\gamma(0) = \xi_0$,*
2. *$T > 0$ and $\gamma(t) \in \mathrm{Dom}(\theta)$ for all $t \in (0, T]$,*
3. *$\gamma'(0) = v$.*

This indeed defines a cone (since v may be replaced by λv, $\lambda > 0$ by changing the path parametrization) that we note $K(\xi_0)$.

DEFINITION 4.3. *The FDR of Definition 2.2 is said to be*

- conic *if $b_0(\xi_0)$ belongs to the interior $\overset{\circ}{K}(\xi_0)$ of the tangent cone $K(\xi_0)$,*
- regular *if the tangent cone $K(\xi_0)$ contains an open half-space $T^+_{\xi_0}M$ and if $b_0(\xi_0)$ belongs to $T^+_{\xi_0}M$.*

With this definition, a regular FDR is conic.

The following proposition gives a necessary condition for $\mathcal{F}^{\delta_{x_0}}$ to have a conic or a regular FDR.

PROPOSITION 4.4. *Let $x_0 \in \mathbb{R}^n$ be such that the input-output map $\mathcal{F}^{\delta_{x_0}}$ given by (8) admits a conic FDR. Then $\mathcal{E}(x_0)$ contains no operators of order greater than two; that is,*

$$\mathcal{E}(x_0) \subset \mathbb{R} - \mathrm{span}\left\{ 1, \frac{\partial}{\partial x_i}, i = 1, \ldots, n, \frac{\partial^2}{\partial x_i \partial x_j}, i, j = 1, \ldots, n \right\}. \tag{24}$$

Moreover, if the FDR is not only conic but also regular, then

$$\mathcal{E}(x_0) \subset \mathbb{R}M_0(x_0) + \mathbb{R} - \mathrm{span}\left\{ 1, \frac{\partial}{\partial x_i}, i = 1, \ldots, n \right\}. \tag{25}$$

We need a lemma based on a characterization of the operators satisfying the maximum principle [6].

LEMMA 4.5. *Let R be a smooth differential operator on $\mathbb{R}^n$ such that for all smooth nonnegative function ϕ and for all $x_0 \in \mathbb{R}^n$, we have*

$$\phi(x_0) = 0 \Rightarrow (R\phi)(x_0) \geq 0 \,(resp., = 0). \tag{26}$$

Then R is a differential operator of order less than or equal to two (resp., of order less than or equal to one).

Proof. Let us write R under the form

$$R = \sum_{\alpha=(\alpha_1,\ldots,\alpha_n)\in\mathbb{N}^n} r_\alpha \frac{\partial^{\alpha_1+\cdots+\alpha_n}}{\partial x_1^{\alpha_1}\cdots\partial x_n^{\alpha_n}}. \tag{27}$$

Let $x_0 \in \mathbb{R}^n$ be given.

Let Q be a nonnegative symmetric matrix, let $\lambda \in \mathbb{R}$, and let $\beta = (\beta_1, \ldots, \beta_n) \in \mathbb{N}^n$ be such that $\mid \beta \mid = \beta_1 + \cdots + \beta_n > 2$. Then, let ϕ be a smooth nonnegative function such that in a neighborhood of x_0 we have

$$\phi(x) = \frac{(x-x_0)'Q(x-x_0)}{2} + \frac{\lambda}{\beta_1!\ldots\beta_n!}(x_1-(x_0)_1)^{\beta_1}\cdots(x_n-(x_0)_n)^{\beta_n}. \tag{28}$$

By (27) and (28) we have

$$(R\phi)(x_0) = \sum_{|\alpha|=2} r_\alpha(x_0)Q_\alpha + \lambda r_\beta(x_0).$$

In the notation Q_α, one has to understand the following: when $\mid \alpha \mid = 2$, there exists

- either $i \neq j$ such that $\alpha_i = \alpha_j = 1$ (and all others $\alpha_k = 0$) and then $Q_\alpha = Q_{i,j}$,
- or i such that $\alpha_i = 2$ (and all others $\alpha_k = 0$) and then $Q_\alpha = Q_{i,i}$.

Now, if $(R\phi)(x_0) \geq 0$, then

$$\sum_{|\alpha|=2} r_\alpha(x_0)Q_\alpha + \lambda r_\beta(x_0) \geq 0.$$

This inequality being true for all $\lambda \in \mathbb{R}$, we must have $r_\beta(x_0) = 0$ and this latter equality holds for all $\beta = (\beta_1, \ldots, \beta_n) \in \mathbb{N}^n$ such that $\mid \beta \mid = \beta_1 + \cdots + \beta_n > 2$. Thus, R is a differential operator of order less than or equal to two at x_0 (and then on all R^n since x_0 is arbitrary).

Now, if $(R\phi)(x_0) = 0$, we know that $r_\beta(x_0) = 0$ for all $\beta = (\beta_1, \ldots, \beta_n) \in \mathbb{N}^n$ such that $\mid \beta \mid = \beta_1 + \cdots + \beta_n > 2$. Thus, we simply have

$$\sum_{|\alpha|=2} r_\alpha(x_0)Q_\alpha = 0.$$

This equality being true for all Q nonnegative symmetric matrices, we must have $r_\alpha(x_0) = 0$ for all α such that $\mid \alpha \mid = 2$. Thus, R is a differential operator of order less than or equal to one at x_0 (and then on all R^n since x_0 is arbitrary). This ends the proof of Lemma 4.5. □

Proof of Proposition 4.4. Let $E \in \mathcal{E}$ be given and $b \in \mathcal{B}$ be such that (18) is satisfied by the remark following Corollary 3.5.

Assume that the FDR is conic. Since $b_0(\xi_0) \in \overset{\circ}{K}(\xi_0)$, there exists $\varepsilon \neq 0$ such that $b_0(\xi_0) + \varepsilon b(\xi_0) \in \overset{\circ}{K}(\xi_0)$. Let γ be a corresponding path in $\mathrm{Dom}(\theta)$ (apart from the starting point $\xi_0 \in \partial\mathrm{Dom}(\theta)$), with $\gamma'(0) = b_0(\xi_0) + \varepsilon b(\xi_0)$. We prove that

$$R = M_0 + \varepsilon E$$

and thus E itself is a differential operator of order less than or equal to two.

Let x_0 be given and let ϕ be a nonnegative smooth function such that $\phi(x_0) = 0$. By the remark following Corollary 3.5, we have both $(L_{b_0}\theta)(\xi_0) = M_0^\star\theta(\xi_0)$ and $(L_b\theta)(\xi_0) = E^\star\theta(\xi_0)$, so that

$$(R\phi)(x_0) = \langle L_{b_0+\varepsilon b}\theta(\xi_0), \phi\rangle\,.$$

We now prove that the above right-hand side is nonnegative.

The function

$$\varrho : t \in [0,T] \mapsto \langle \theta(\gamma(t)), \phi\rangle \in \mathbb{R}.$$

has the following properties:

1. It is smooth on $(0,T]$ with continous derivatives of all order on $[0,T]$.
2. It is nonnegative since ϕ is nonnegative.
3. It satisfies $\varrho(0) = 0$ since

$$\varrho(0) = \langle \theta(\gamma(0)), \phi\rangle = \langle \theta(\xi_0), \phi\rangle = \langle \delta_{x_0}, \phi\rangle = \phi(x_0) = 0.$$

By continuity of $\varrho'(t)$ at $t = 0$, we have

$$\langle L_{b_0+\varepsilon b}\theta(\xi_0), \phi\rangle \;=\; \left\langle L_{\gamma'(0)}\theta(\xi_0), \phi\right\rangle \;= \lim_{t\downarrow 0} \varrho'(t) = \lim_{t\downarrow 0} \frac{1}{t}\int_0^t \varrho'(s)ds.$$

We evaluate the last term sign by

$$\int_0^t \varrho'(s)ds = \lim_{\eta\downarrow 0}\int_\eta^t \varrho'(s)ds = \lim_{\eta\downarrow 0}(\varrho(t) - \varrho(\eta)) = \varrho(t) - 0 \geq 0.$$

Thus, we get $\langle L_{b_0+\varepsilon b}\theta(\xi_0), \phi\rangle \geq 0$.

We conclude that $R = M_0 + \varepsilon E$, and thus E itself is a differential operator of order less than or equal to two by Lemma 4.5.

Assume that the FDR is regular. Then there exists an open half-space $T^+_{\xi_0}M$ in the tangent space $T_{\xi_0}M$ of M at ξ_0 such that we can split

$$T_{\xi_0}M = T^+_{\xi_0}M \oplus T^0_{\xi_0}M \oplus T^-_{\xi_0}M,$$

where $T^+_{\xi_0}M = -T^-_{\xi_0}M$ and $T^0_{\xi_0}M = \partial T^+_{\xi_0}M = \overline{T^+_{\xi_0}M} \cap \overline{T^-_{\xi_0}M}$.

By symmetry ($a(\xi_0) \in T^+_{\xi_0}M \iff -a(\xi_0) \in T^-_{\xi_0}M$) and by continuity ($T^0_{\xi_0}M = \partial T^+_{\xi_0}M = \overline{T^+_{\xi_0}M} \cap \overline{T^-_{\xi_0}M}$), we get for all smooth function ϕ and for all $x_0 \in \mathbb{R}^n$ by the study here above,

$$\phi \geq \phi(x_0) = 0 \Rightarrow \begin{cases} \langle L_a\theta(\xi_0), \phi\rangle \;\geq\; 0 & \text{if } a(\xi_0) \in T^+_{\xi_0}M, \\ \langle L_a\theta(\xi_0), \phi\rangle \;\leq\; 0 & \text{if } a(\xi_0) \in T^-_{\xi_0}M, \\ \langle L_a\theta(\xi_0), \phi\rangle \;=\; 0 & \text{if } a(\xi_0) \in T^0_{\xi_0}M. \end{cases}$$

Now, since $b_0(\xi_0) \in T^+_{\xi_0}M$, there exists $\rho \in \mathbb{R}$ such that $\rho b_0(\xi_0) + b(\xi_0) \in T^0_{\xi_0}M$. Thus we have

$$\phi \geq \phi(x_0) = 0 \Rightarrow \langle L_{\rho b_0 + b}\theta(\xi_0), \phi\rangle \ = \ 0.$$

But by the remark following Corollary 3.5, we also have

$$((\rho M_0 + E)\phi)(x_0) = \langle L_{\rho b_0 + b}\theta(\xi_0), \phi\rangle .$$

We conclude that $R = \rho M_0 + E$ is a differential operator of order less than or equal to one by Lemma 4.5. Therefore, for any $E \in \mathcal{E}$ there exists a real ρ such that

$$E = \rho M_0 + \text{a first-order differential operator.}$$

This ends the proof of Proposition 4.4. □

The following proposition is an easy corollary.

PROPOSITION 4.6. *If the family of input-output maps* $\{\mathcal{F}^{\mu_0} \mid \mu_0 \in \mathbb{D}_0\}$*, given by* (8)*, admits a family of regular FDRs, then the estimation Lie algebra* $\mathcal{E}$ *contains only operators of order less than or equal to two. In particular, we have for all* x_0 *in* $\mathbb{R}^n$*,* $\dim \mathcal{E}(x_0) \leq \frac{(n+1)(n+2)}{2} < +\infty$.

Let us stress the fact that the assumption under which Proposition 4.6 holds differs from that of [11, p. 88], where it is assumed that the diffusion ξ_t depends on the point x_0 of $\mathbb{R}^n$ through a relationship of the form $\xi_0 = \psi(x_0)$. Moreover, this last result on $\dim \mathcal{E}(x_0)$ is to be compared with results in [11, 26] asserting that the estimation Lie algebra is finite dimensional when estimated on any point $x_0 \in \mathbb{R}^n$.

In [26], Lévine notices that estimation Lie algebras with this last property are not necessarily finite dimensional. The example given relies upon the use of the time-varying Kalman–Bucy filter, where the variable t is considered as a state variable (in fact, note that any estimation Lie algebra $\mathcal{E}$, where M_0 is a differential operator of order less than or equal to one is generally infinite dimensional, though finite dimensional when estimated on any point of $\mathbb{R}^n$ since it is made of operators of order less than or equal to one). Nevertheless this trick allows us to exhibit a class of systems (1) (and therefore of operators (6)) such that there exist FDFs for every Dirac initial law (and therefore FDRs for every Dirac measure) but that the estimation Lie algebra is not finite dimensional. Here is an example.

Consider the following system with state $z = (x, t) \in \mathbb{R}^2$ and observation $y \in \mathbb{R}$:

$$\left\{ \begin{array}{ll} dx_t \ = \ dv_t, & x_0 \,\text{deterministic,} \\ dt = dt, & t_0 \,\text{deterministic,} \\ dy_t \ = \ h(t)x_t dt \ + \ dw_t, & y_0 \ = \ 0. \end{array} \right. \tag{29}$$

It is clear that there exists for each Dirac initial law an FDF given by the (time-varying) Kalman–Bucy filter. On the other hand, the estimation Lie algebra is generally infinite dimensional. Indeed, we have

$$\left\{ \begin{array}{rcl} M_0 & = & \dfrac{1}{2}\dfrac{\partial^2}{\partial x^2} + \dfrac{\partial}{\partial t} - \dfrac{1}{2}h^2(t)x^2, \\ M_1 & = & h(t)x, \end{array} \right. \tag{30}$$

and it is easily seen that

$$[M_0, M_1] = h(t)\frac{\partial}{\partial x} + h'(t)x,$$

$$[[M_0, M_1], M_1] = h^2(t),$$

$$\mathrm{ad}_{M_0}([[M_0, M_1], M_1]) = [M_0, [[M_0, M_1], M_1]] = \frac{d}{dt}h^2(t),$$

$$\mathrm{ad}^2_{M_0}([[M_0, M_1], M_1]) = [M_0, \mathrm{ad}_{M_0}([[M_0, M_1], M_1])] = \frac{d^2}{dt^2}h^2(t).$$

Thus, a k-iterate Lie bracket as above is of the form

$$\mathrm{ad}^k_{M_0}([[M_0, M_1], M_1]) = [M_0, [M_0, [...[M_0, h^2(t)]...]] = \frac{d^k}{dt^k} h^2(t). \tag{31}$$

If we choose h such that all derivatives of h^2 are linearly independent, then the estimation Lie algebra is infinite dimensional.

Although such estimation Lie algebras have a certain finite-dimensional property, we shall see in the next chapter how to specify the dependency on the initial laws to have finite-dimensional estimation Lie algebras.

5. Uniform families of FDRs. In filtering theory it is remarkable that almost all the systems (except, for instance, those from [26] seen at the end of the previous section or the conditionally Gaussian processes of [18]) for which an FDF has been found have two particularities:

- they admit FDFs which depend on a broad class of initial measures only through the output function (any nonnegative measure in [28] or almost any nonnegative measure in [5] for linear systems, the domain of an operator of $L^1(\mathbb{R}^2)$ in [45], any probability measure in [25]),
- their estimation Lie algebra is finite dimensional (and contains only operators of order less than or equal to two).

It can be checked that all the FDFs in [45, 5, 28, 25] lead to families of FDRs for which the nonnegative output function θ^{μ_0} associated with the initial law μ_0 is of the form

$$\langle \theta^{\mu_0}, \phi \rangle = \int d\mu_0(x_0) \, \langle \theta^{\delta_{x_0}}, \phi \rangle, \tag{32}$$

where $\theta^{\delta_{x_0}}$ is the nonnegative output function of a family of FDRs, uniform with respect to the dynamics and initial condition. Indeed, in all these examples, the family of input-output maps $\{\mathcal{F}^{\mu_0} \mid \mu_0 \in \mathbb{M}_0\}$ is parametrized by a set $\mathbb{M}_0$ like $L^1(\mathbb{R}^n)$ which is too big to be parametrized by some finite-dimensional manifold (unlike the Dirac or the Gaussian measures), and then it is a reasonable choice to have the FDRs depend on μ through the nonnegative output function and not through the diffusion ξ_t by its starting point ξ_0 (like in the Kalman–Bucy filter or in [11]).

When $\mathbb{M}_0 = \mathbb{D}_0$, the set of Dirac measures, the following proposition yields the structure of the estimation Lie algebra.

PROPOSITION 5.1. *If the family of input-output maps $\{\mathcal{F}^{\mu_0} \mid \mu_0 \in \mathbb{D}_0\}$ given by* (8) *admits a regular FDR uniform with respect to the dynamics and initial condition, then $\mathcal{E}$ is finite dimensional and of the form*

$$\mathcal{E} = \mathbb{R}M_0 + \mathcal{R}, \tag{33}$$

where $\mathcal{R}$ is a finite-dimensional subalgebra of differential operators of order less than or equal to one.

Proof. Once we prove that $\mathcal{E}$ is finite dimensional, then it is clear that (33) holds by a straightforward application of Proposition 4.6.

Thus, to show that $\mathcal{E}$ is finite dimensional, we first exhibit a Lie ideal $\mathcal{J}$ of $\mathcal{B}$ such that $\mathcal{E}$ is isomorphic to $\mathcal{B} \setminus \mathcal{J}$, then show that this latter vector space is finite dimensional.

We introduce the following family of Lie ideals of $\mathcal{B}$ (Lie ideals by Proposition 3.3):

$$\mathcal{J}_{x_0} = \{\, j \in \mathcal{B} \,|\, \text{w.p.1} \;\; \forall\, t \geq 0, \; L_j \theta^{\delta_{x_0}}(\xi_t) = 0 \,\},$$

and we define the Lie ideal $\mathcal{J}$ of $\mathcal{B}$ by

$$\mathcal{J} = \bigcap_{x_0 \in \mathbb{R}^n} \mathcal{J}_{x_0}.$$

Now, we define a homomorphism Λ from $\mathcal{E}$ to $\mathcal{B} \setminus \mathcal{J}$ as follows. Let $E \in \mathcal{E}$. We can always write

$$E = \sum_{k \leq l} \sum_{i_1=0}^{p} \cdots \sum_{i_k=0}^{p} e_{i_1,\ldots,i_k} [\, M_{i_1}, [\ldots, M_{i_k}\,] \ldots]. \tag{34}$$

It is then clear by Lemma 3.4 that we have

$$\forall\, x_0 \in \mathbb{R}^n \quad \text{w.p.1} \quad \forall\, t \,>\, 0, \quad L_b \theta^{\delta_{x_0}}(\xi_t) = E^\star \theta^{\delta_{x_0}}(\xi_t), \tag{35}$$

where $b \in \mathcal{B}$ is defined by

$$b = \sum_{k \leq l} \sum_{i_1=0}^{p} \cdots \sum_{i_k=0}^{p} e_{i_1,\ldots,i_k} [\, b_{i_1}, [\ldots, b_{i_k}\,] \ldots]\,. \tag{36}$$

Thus, we can assign to $E \in \mathcal{E}$ a nonempty subset $\Lambda(E)$ of $\mathcal{B}$ by

$$\Lambda(E) = \{\, b \in \mathcal{B} \mid \text{w.p.1} \quad \forall\, t \,>\, 0, \quad L_b \theta^{\delta_{x_0}}(\xi_t) = E^\star \theta^{\delta_{x_0}}(\xi_t) \,\}\,. \tag{37}$$

What is more, it is easily seen that $\Lambda(E)$ belongs to $\mathcal{B} \setminus \mathcal{J}$.

The above-defined mapping Λ is one to one:

- Λ is injective because $\Lambda(E) = \mathcal{J}$ implies, by the definitions of $\Lambda(E)$ and $\mathcal{J}$, that for all $x_0 \in \mathbb{R}^n$, one has $E^\star \delta_{x_0} = 0$ and thus $E = 0$,
- Λ is onto, because for any $b \in \mathcal{B}$, we can always write it as in (36) and then $E \in \mathcal{E}$ given by (34) is such that $\Lambda(E)$ is the class of b.

Therefore, $\mathcal{E}$ is isomorphic to $\mathcal{B} \setminus \mathcal{J}$.

We now prove that $\mathcal{B} \setminus \mathcal{J}$ is finite dimensional. For this, we consider the following linear mapping ϖ:

$$\varpi \,:\, \mathcal{B} \to T_{\xi_0} M, \quad b \mapsto b(\xi_0)\,. \tag{38}$$

We study $\mathrm{Ker}\varpi$. Let $b \in \mathcal{B}$ such that $b(\xi_0) = 0$ and, since Λ is onto, let $E \in \mathcal{E}$ be such that $\Lambda(E)$ is the class of b. As a consequence of (37) for $t = 0$, E satisfies $L_b \theta^{\delta_{x_0}}(\xi_0) = E^\star \delta_{x_0}$ for all $x_0 \in \mathbb{R}^n$. Since $b(\xi_0) = 0$, then

$$\forall\, x_0 \in \mathbb{R}^n, \quad 0 = L_{b(\xi_0)} \theta^{\delta_{x_0}}(\xi_0) = L_b \theta^{\delta_{x_0}}(\xi_0) = E^\star \delta_{x_0}$$

and therefore $E = 0$. We have just shown that $\mathrm{Ker}\varpi \subset \Lambda(0) = \mathcal{J}$, and thus $\mathcal{B} \setminus \mathcal{J} \subset \mathcal{B} \setminus \mathrm{Ker}\varpi$.

Moreover, by (38), there exists an injection, derived from ϖ, from $\mathcal{B} \setminus \mathrm{Ker}\varpi$ to $T_{\xi_0} M$, and we obtain

$$\mathcal{E} \simeq \mathcal{B} \setminus \mathcal{J} \subset \mathcal{B} \setminus \mathrm{Ker}\varpi \hookrightarrow T_{\xi_0} M.$$

Since the last mapping is an injection, we finally get

$$\dim \mathcal{E} \leq \dim M \,<\, +\infty. \tag{39}$$

This ends the proof. □

Now we shall prove that the above necessary conditions of Proposition 5.1 for the existence of certain classes of FDRs are almost sufficient.

THEOREM 5.2. *Assume that*

1. $\mathcal{E}$ *is finite dimensional and given by* (33)*, where* $\mathcal{R}$ *is a finite-dimensional subalgebra of differential operators of order less than or equal to one,*

2. $\mathcal{E}$ *is a solvable Lie algebra, with a basis* $\{F_0, F_1, F_2, \ldots, F_q\}$ *such that* $F_0 = M_0$ *and* $\mathcal{F}_i = \mathbb{R} - \text{span}\{ F_i, \ldots, F_q \}$ *is a Lie ideal of* $\mathcal{F}_{i-1}$ *for* $i = 0, \ldots, q$*,*

3. *the first-order part of each operator* F_1*, ...,* F_q *defines a complete vector field on* $\mathbb{R}^n$*,*

4. *for all* $\phi \in \mathcal{D}(\mathbb{R}^n)$*, there exists a unique solution* $u \in C^\infty(]0, +\infty[\times\mathbb{R}^n) \cap C^0([0, +\infty[\times\mathbb{R}^n)$ *of the PDE*

$$\frac{\partial u}{\partial t} = M_0\, u, \quad u(0, x) = \phi(x). \tag{40}$$

Then the family of input-output maps $\{\mathcal{F}^{\mu_0} \mid \mu_0 \in \mathbb{D}_0\}$ *given by* (8) *admits a regular FDR, uniform with respect to the dynamics and initial condition.*

Remark. The Lie algebra $\mathcal{E}$ can be finite dimensional without being solvable, as is the case for $\mathcal{E} = \{\frac{d^2}{dx^2}, x^2\}_{\mathcal{L}.\mathcal{A}.}$. However, Tam, Wong, and Yau prove in [41] that finite-dimensional estimation algebra from filtering problems (without correlated noise) are solvable when the second-order part of the infinitesimal generator of the signal is the Laplacian on $\mathbb{R}^n$ and the drift is a gradient vector field.

First, we shall define the collection of output functions.

LEMMA 5.3. *Suppose that the assumptions of Theorem* 5.2 *are satisfied. Then there exists a family of nonnegative output functions* $\left(\theta^{\delta_{x_0}}\right)_{x_0 \in \mathbb{R}^n}$ *on* $\text{Dom}(\theta) =]0, +\infty[\times\mathbb{R}^q$ *which satisfy*

$$\left\langle \frac{\partial \theta^{\delta_{x_0}}}{\partial \xi^i}(\xi), \phi \right\rangle = \left\langle \theta^{\delta_{x_0}}(\xi), e^{-\xi^0\, ad_{F_0}} \cdots e^{-\xi^{i-1}\, ad_{F_{i-1}}} F_i \phi \right\rangle, \tag{41}$$

where $F_0 = M_0$*. Here,* $ad_M(N) = [M, N]$ *and* $e^{t\, ad_M}$ *is the linear operator* $e^{t\, ad_M} = \sum_{j=0}^{+\infty} \frac{t^j\, ad_M^j}{j!}$ *of the finite-dimensional Lie algebra* $\mathcal{E}$*.*

Proof. If $F_1 = X_1 + c_1, \ldots, F_q = X_q + c_q$ is a basis of $\mathcal{R}$, where $X_1, \ldots, X_q$ are (operators of order one identified with) complete vector fields and c_1, ..., c_q are smooth functions, then the evolution equations

$$\frac{\partial u}{\partial t} = F_i\, u = (X_i + c_i)u, \quad u(0, x) = \phi(x), \tag{42}$$

generate positive transition groups $(P_t^1)_{t\in\mathbb{R}}$, ..., $(P_t^q)_{t\in\mathbb{R}}$ on $C^\infty(\mathbb{R}^n)$ given by

$$P_t^i\, \phi\,(x) \,=\, \exp\left(\int_0^t c_i(\Phi_s^{X_i}(x))\, ds \right)\, \phi\,(\Phi_t^{X_i}(x))\,. \tag{43}$$

Moreover, if $(P_t^0)_{t\geq 0}$ is the transition semigroup on $\mathcal{D}(\mathbb{R}^n)$ generated by the evolution equation (40), let us define the map $\theta^{\delta_{x_0}}$ from $]0, +\infty[\times\mathbb{R}^q$ to $\mathbb{M}_+(\mathbb{R}^n)$ by

$$\left\langle \theta^{\delta_{x_0}}(\xi),\, \phi \right\rangle \,=\, \left(P_{\xi^q}^q \,\cdots\, (P_{\xi^0}^0\, \phi) \right)\, (x_0) \,=\, P_{\xi^q}^q \,\cdots\, P_{\xi^0}^0\, \phi\,(x_0). \tag{44}$$

The function $\left\langle \theta^{\delta_{x_0}}(\cdot),\, \phi \right\rangle$ is clearly continuous on $[0, +\infty[\times\mathbb{R}^q$ and smooth on $]0, +\infty[\times\mathbb{R}^q$ since it can be expressed by the formula

$$\begin{aligned}
&\left\langle \theta^{\delta_{x_0}}(\xi),\, \phi \right\rangle \\
&\quad = \exp\left(\int_0^{\xi^q} c_q(\Phi_s^{X_q}(x_0))\, ds \right) \times \cdots \times \exp\left(\int_0^{\xi^1} c_1(\Phi_s^{X_1} \circ \Phi_{\xi^2}^{X_2} \circ \cdots \circ \Phi_{\xi^q}^{X_q}(x_0))\, ds \right) \\
&\qquad \times\, (P_{\xi^0}^0 \phi)(\Phi_{\xi^1}^{X_1} \circ \cdots \circ \Phi_{\xi^q}^{X_q}(x_0))\,.
\end{aligned}$$

To prove (41), we put

$$\phi_0 = \phi \in \mathcal{D}(\mathbb{R}^n), \phi_1 = P^0_{\xi^0}\,\phi_0 \in C^\infty(\mathbb{R}^n), \ldots, \phi_q = P^q_{\xi^q}\,\phi_{q-1} \in C^\infty(\mathbb{R}^n)$$

so that if $\varepsilon_0, \ldots, \varepsilon_q$ is the canonical basis of $\mathbb{R}^{q+1}$, we have

$$\frac{\left\langle \theta^{\delta_{x_0}}(\xi),\, \phi \right\rangle - \left\langle \theta^{\delta_{x_0}}(\xi + t\varepsilon_i),\, \phi \right\rangle}{t} = \left(P^q_{\xi^q} \cdots P^i_{\xi^i}\, \frac{P^i_t\,\phi_i - \phi_i}{t} \right)(x_0).$$

When t tends to zero, this last term converges to $(P^q_{\xi^q} \cdots P^i_{\xi^i}\, F_i\phi_i)(x_0)$ since the operators P^1_t, …, P^q_t are easily seen to be continuous on $C^0(\mathbb{R}^n)$ for the pointwise convergence. Therefore, we have for $i = 0, \ldots, q$:

$$\left\langle \frac{\partial \theta^{\delta_{x_0}}}{\partial \xi^i}(\xi), \phi \right\rangle = \frac{\partial}{\partial \xi^i} \left\langle \theta^{\delta_{x_0}}(\cdot),\, \phi \right\rangle_{|\xi} = P^q_{\xi^q} \cdots P^i_{\xi^i}\, F_i\, \phi_i\,(x_0)\,. \tag{45}$$

By Theorem A.1 in the Appendix, the Baker–Campbell–Hausdorff formula (A.57) applies to $F_i\,\phi_i$ for $i = 1, \ldots, q$. Indeed, we have

$$\begin{aligned} F_1\,\phi_1 &= F_1\,P^0_{\xi^0}\,\phi_0 \\ &= P^0_{\xi^0}\,(e^{-\xi^0\,\mathrm{ad}_{F_0}} F_1\,\phi_0) \end{aligned}$$

since $\phi_0 = \phi$, $F_0\phi_0$, …, $F_q\phi_0$ all belong to $\mathcal{D}(\mathbb{R}^n)$ on which $P^0_{\xi^0}$ is defined (see Theorem A.1). Now, with $\phi_1, \ldots, \phi_q \in C^\infty(\mathbb{R}^n)$, we show in the same way for $i = 1, \ldots, q$,

$$\begin{aligned} F_i\,\phi_i &= F_i\,P^{i-1}_{\xi^{i-1}}\,\phi_{i-1} \\ &= P^{i-1}_{\xi^{i-1}}\,e^{-\xi^{i-1}\,\mathrm{ad}_{F_{i-1}}} F_i\,\phi_{i-1} \\ &= \cdots \\ &= P^{i-1}_{\xi^{i-1}} \cdots P^0_{\xi^0}\,e^{-\xi^0\,\mathrm{ad}_{F_0}} \cdots e^{-\xi^{i-1}\,\mathrm{ad}_{F_{i-1}}}\, F_i\phi\,. \end{aligned}$$

After reporting this last expression in (45), we find (41).

Now, there remains to prove that all partial derivatives of $\left\langle \theta^{\delta_{x_0}}(\xi),\, \phi \right\rangle$ have limits on $[0, +\infty[\times\mathbb{R}^q$ (see Definition 2.1). For this, we can notice that in (41) the term $e^{-\xi^0 \mathrm{ad}_{F_0}} \cdots e^{-\xi^{i-1} \mathrm{ad}_{F_{i-1}}} F_i$ is an analytic expression in $(\xi^0, \ldots, \xi^{i-1})$ since $\mathcal{E}$ is finite dimensional. Therefore, we can write

$$e^{-\xi^0\,\mathrm{ad}_{F_0}} \cdots e^{-\xi^{i-1}\,\mathrm{ad}_{F_{i-1}}}\, F_i = \sum_{j=0}^{q} \alpha_{ij}(\xi^0, \ldots, \xi^{i-1})\, F_j, \tag{46}$$

where α_{i0}, …, α_{iq} are analytic functions. As a consequence of (41), we have

$$\left\langle \frac{\partial \theta^{\delta_{x_0}}}{\partial \xi^i}(\xi),\, \phi \right\rangle = \sum_{j=0}^{q} \alpha_{ij}(\xi^0, \ldots, \xi^{i-1}) \left\langle \theta^{\delta_{x_0}}(\xi),\, F_j\,\phi \right\rangle$$

so that the first partial derivatives of $\left\langle \theta^{\delta_{x_0}}(\cdot)\,,\, \phi \right\rangle$ have limits on $[0, +\infty[\times\mathbb{R}^q$ since $\left\langle \theta^{\delta_{x_0}}(\cdot)\,,\, \phi \right\rangle$ is continuous on $[0, +\infty[\times\mathbb{R}^q$. If we iterate the procedure, we prove that this is the case for all partial derivatives, so that $\theta^{\delta_{x_0}}$ is an output function on $]0, +\infty[\times\mathbb{R}^q$. □

Proof of Theorem 5.2. If we show that there exist $p+1$ vector fields b_0, b_1, …, b_p on $\mathbb{R}^q$ such that equation (12) is conservative on $]0,+\infty[\times\mathbb{R}^q$ (with $\xi_{t\,|t=0}=0$) and that

$$\left\langle L_{b_i}\theta^{\delta_{x_0}}(\xi),\,\phi\right\rangle \,=\, \left\langle \theta^{\delta_{x_0}}(\xi),\,M_i\phi\right\rangle, \quad i=0,\dots,p, \tag{47}$$

then $(\mathbb{R}^{q+1},0,b_0,b_1,\dots,b_p,\theta^{\delta_{x_0}})_{x_0\in\mathbb{R}^n}$ is a family of regular FDRs, uniform with respect to the dynamics and initial condition, for the family of input-output maps $\{\mathcal{F}^{\mu_0}\mid\mu_0\in\mathbb{D}_0\}$ given by (8).

Indeed, it would only remain to show that $\left\langle\theta^{\delta_{x_0}}(\xi_t),\,\phi\right\rangle$ satisfies equation (5). But (47) and (12) provide the result thanks to the (ordinary) Itô–Stratonovitch formula since for all $\phi\in\mathcal{D}(\mathbb{R}^n)$, w.p.1, we have

$$\begin{aligned} d\left\langle\theta^{\delta_{x_0}}(\xi_t),\,\phi\right\rangle &= L_{b_0}\left\langle\theta^{\delta_{x_0}},\phi\right\rangle_{|\xi_t}dt \,+\, \sum_{i=1}^{p} L_{b_i}\left\langle\theta^{\delta_{x_0}},\phi\right\rangle_{|\xi_t}\circ dy_t^i \\ &= \left\langle L_{b_0}\theta^{\delta_{x_0}}(\xi_t),\phi\right\rangle dt \,+\, \sum_{i=1}^{p}\left\langle L_{b_i}\theta^{\delta_{x_0}}(\xi_t),\phi\right\rangle\circ dy_t^i \\ &= \left\langle\theta^{\delta_{x_0}}(\xi_t),M_0\phi\right\rangle dt \,+\, \sum_{i=1}^{p}\left\langle\theta^{\delta_{x_0}}(\xi_t),M_i\phi\right\rangle\circ dy_t^i. \end{aligned}$$

We define the vector fields b_0, …, b_p on $\mathbb{R}^{q+1}$ as images of left-invariant vector fields on a Lie group as follows.

Since the Lie algebra $\mathcal{E}$ is finite dimensional, then it is the Lie algebra of a simply connected Lie group G [17]. Moreover, if $\mathcal{G}$ denotes the Lie algebra of left-invariant vector fields of the Lie group G and T_eG the tangent space to G at the neutral element e of G, there is a series of isomorphisms

$$\mathcal{E}\simeq\mathcal{G}\simeq T_eG. \tag{48}$$

Let us denote by $\mathbf{g}_0,\dots,\mathbf{g}_q$ the images of F_0, …, F_q in $\mathcal{G}$. The exponential mapping from T_eG to G satisfies $\Phi_t^{\mathbf{g}}(\tau)=\tau\,e^{t\mathbf{g}(e)}$ for all $\tau\in G$ and $\mathbf{g}\in\mathcal{G}$. Denoting $g_i=\mathbf{g}_i(e)\in T_eG$, the following mapping Ψ from $\mathbb{R}^{q+1}$ to G is analytic:

$$\Psi(\xi^0,\dots,\xi^q)=e^{\xi^q g_q}\,\cdots\,e^{\xi^0 g_0}=\Phi_{\xi^q}^{\mathbf{g}_q}\cdots\Phi_{\xi^0}^{\mathbf{g}_0}(e)\,. \tag{49}$$

Moreover, since $\mathcal{E}$ is solvable, then G is also solvable and, thanks to the choice of the basis $\{F_0,F_1,\dots,F_q\}$, Ψ is shown to be one to one with analytic inverse [7, p. 240]. Then, to the vector fields $\mathbf{g}_0$, …, $\mathbf{g}_q$ we associate the following vector fields on $\mathbb{R}^{q+1}$: $a_0=\Psi_\star^{-1}(\mathbf{g}_0)$, …, $a_q=\Psi_\star^{-1}(\mathbf{g}_q)$. They satisfy

$$\left\langle L_{a_i}\theta^{\delta_{x_0}}(\xi),\,\phi\right\rangle \,=\, \left\langle\theta^{\delta_{x_0}}(\xi),\,F_i\,\phi\right\rangle \quad\forall\,\xi\in\,[0,+\infty[\times\mathbb{R}^q\,. \tag{50}$$

Indeed, for ξ in $]0,+\infty[\times\mathbb{R}^q$ and $i=0,\dots,p$, we have

$$\begin{aligned} \left\langle L_{a_i}\theta^{\delta_{x_0}}(\xi),\,\phi\right\rangle &= L_{a_i}\left\langle\theta^{\delta_{x_0}}(\cdot)\,,\,\phi\right\rangle_{|\xi} \\ &= L_{\Psi_\star^{-1}(\mathbf{g}_i)}\left\langle\theta^{\delta_{x_0}}(\cdot)\,,\,\phi\right\rangle_{|\xi} \\ &= L_{\mathbf{g}_i}\left\langle\theta^{\delta_{x_0}}\left(\Psi^{-1}(\cdot)\right),\,\phi\right\rangle_{|\Psi(\xi)} \\ &= \tfrac{d}{dt}_{|t=0}\left\langle\theta^{\delta_{x_0}}\left(\Psi^{-1}(\Phi_t^{\mathbf{g}_i}(\Psi(\xi)))\right),\,\phi\right\rangle. \end{aligned} \tag{51}$$

Now, since Ψ is an analytic diffeomorphism of $\mathbb{R}^{q+1}$, there exist analytic functions $r_0, \ldots, r_q$ (which depend on $\xi^0, \ldots, \xi^q$) such that for all t

$$\Phi_t^{\mathbf{g}_i}(\Psi(\xi)) = e^{\xi^q g_q} \cdots e^{\xi^0 g_0} e^{t g_i} = e^{(\xi^q + r_q(t)) g_q} \cdots e^{(\xi^0 + r_0(t)) g_0}. \tag{52}$$

Therefore, by (51) we have

$$\begin{aligned} \left\langle L_{a_i} \theta^{\delta_{x_0}}(\xi),\, \phi \right\rangle &= \frac{d}{dt}_{|t=0} \left\langle \theta^{\delta_{x_0}} \left(\xi^0 + r_0(t), \ldots, \xi^q + r_q(t)\right),\, \phi \right\rangle \\ &= \sum_{j=0}^{q} \frac{dr_j}{dt}(0) \left\langle \frac{\partial \theta^{\delta_{x_0}}}{\partial \xi^j}(\xi),\, \phi \right\rangle \\ &= \left\langle \theta^{\delta_{x_0}}(\xi),\, \widetilde{F}_i\, \phi \right\rangle, \end{aligned}$$

where

$$\widetilde{F}_i = \sum_{j=0}^{q} \frac{dr_j}{dt}(0)\, e^{-\xi^0 \operatorname{ad}_{F_0}} \cdots e^{-\xi^{j-1} \operatorname{ad}_{F_{j-1}}} F_j. \tag{53}$$

We now prove that $\widetilde{F}_i = F_i$ for $i = 0, \ldots, q$, and for this purpose we focus on equation (52), which defines the functions $r_0, \ldots, r_q$. Denoting $\tau = e^{\xi^q g_q} \cdots e^{\xi^0 g_0}$, we obtain the following equality:

$$\tau\, g_i = \sum_{j=0}^{q} e^{\xi^q g_q} \cdots e^{\xi^j g_j}\, \frac{dr_j}{dt}(0)\, g_j\, e^{\xi^{j-1} g_{j-1}} \cdots e^{\xi^0 g_0}$$

in the tangent space $T_\tau(G)$ after differentiation in t. Here, $g\, g_i$ denotes the tangent mapping at e, estimated on the tangent vector g_i, of the left action on the Lie group G: $h \mapsto g\, h$ (this notation is convenient because it allows easy and theoretically justified manipulations). Multiplying this last equation by $e^{-\xi^q g_q} \cdots e^{-\xi^1 g_1}$, we finally get

$$g_i = \sum_{j=0}^{q} \frac{dr_j}{dt}(0)\, e^{-\xi^0 g_0} \cdots e^{-\xi^{j-1} g_{j-1}}\, g_j\, e^{\xi^{j-1} g_{j-1}} \cdots e^{\xi^0 g_0},$$

where, as in any Lie group, $e^{-\xi^{j-1} g_{j-1}}\, g_j\, e^{\xi^{j-1} g_{j-1}} = e^{-\xi^{j-1} \operatorname{ad}_{g_{j-1}}}\, g_j$ [16, p. 175], so that

$$g_i = \sum_{j=0}^{q} \frac{dr_j}{dt}(0)\, e^{-\xi^0 \operatorname{ad}_{g_0}} \cdots e^{-\xi^{j-1} \operatorname{ad}_{g_{j-1}}} (g_j).$$

Therefore, by the Lie isomorphisms (48) and by comparing this latter equality with (53), this proves that $\widetilde{F}_i = F_i$.

Now, if $M_i = \sum_{j=0}^{q} f_{ij}\, F_j$, we define $p+1$ vector fields $b_0, \ldots, b_p$ on $\mathbb{R}^{q+1}$ by $b_i = \sum_{j=0}^{q} f_{ij}\, a_j$. By (50), the vector fields $b_0, \ldots, b_p$ satisfy (47).

Since $\mathbf{g}_0, \ldots, \mathbf{g}_q$ are left-invariant vector fields on the Lie group G, then the stochastic differential equation

$$dg_t = \left(\sum_{j=0}^{q} f_{0j} \mathbf{g}_j \right)(g_t)\, dt + \sum_{i=1}^{p} \left(\sum_{j=0}^{q} f_{ij} \mathbf{g}_j \right)(g_t)\, dy_t^i, \quad g_0 = e \tag{54}$$

is conservative [40]. It is easily seen that $\xi_t = \Psi^{-1}(g_t)$ is a solution of the stochastic differential equation (12), thus proving that it is conservative. Moreover, the solution ξ_t is of the form (t, η_t). Indeed, (49) implies that

$$\begin{aligned}\frac{\partial \Psi}{\partial \xi^j} &= e^{\xi^q g_q} \cdots e^{\xi^j g_j}\, g_j\, e^{\xi^{j-1} g_{j-1}} \cdots e^{\xi^0 g_0} \\ &= e^{\xi^q g_q} \cdots e^{\xi^0 g_0} e^{-\xi^0 g_0} \cdots e^{-\xi^{j-1} g_{j-1}}\, g_j\, e^{\xi^{j-1} g_{j-1}} \cdots e^{\xi^0 g_0} \\ &= e^{\xi^q g_q} \cdots e^{\xi^0 g_0} e^{-\xi^0 \operatorname{ad}_{g_0}} \cdots e^{-\xi^{j-1} \operatorname{ad}_{g_{j-1}}}(g_j) \\ &= \Psi(\xi^0, \ldots, \xi^q) e^{-\xi^0 \operatorname{ad}_{g_0}} \cdots e^{-\xi^{j-1} \operatorname{ad}_{g_{j-1}}}(g_j);\end{aligned}$$

that is,

$$\begin{aligned}\frac{\partial}{\partial \xi^0} &= \Psi_\star^{-1}(\mathbf{g}_0), \\ \frac{\partial}{\partial \xi^1} &= \Psi_\star^{-1}\left(e^{-\xi^0 \operatorname{ad}_{\mathbf{g}_0}}\, \mathbf{g}_1\right), \\ \cdots &= \cdots \\ \frac{\partial}{\partial \xi^q} &= \Psi_\star^{-1}\left(e^{-\xi^0 \operatorname{ad}_{\mathbf{g}_0}} \cdots e^{-\xi^{q-2} \operatorname{ad}_{\mathbf{g}_{q-2}}}\, \mathbf{g}_q\right).\end{aligned}$$

By the property of the basis $\{F_0, F_1, F_2, \ldots, F_q\}$ of the solvable Lie algebra $\mathcal{E}$ (see assumption 2 of Theorem 5.2), we have

$$b_i(\xi) \in \mathbb{R} - \operatorname{span}\left\{\frac{\partial}{\partial \xi^1}, \ldots, \frac{\partial}{\partial \xi^q}\right\} \quad \forall\, i = 1, \ldots, q. \tag{55}$$

Since $b_0 = a_0 = \frac{\partial}{\partial \xi^0}$, it is clear that ξ_t is of the form (t, η_t) so that the stochastic differential equation (12) is conservative on $]0, +\infty[\times\mathbb{R}^q$. □

Remarks.

• This uniform FDR is minimal in the sense that the dimension q of the manifold is minimal by the inequality (39).

• Under the assumptions of Theorem 5.2, let μ_0 in $\mathbb{M}_+(\mathbb{R}^n)$ be such that for all $\xi \in]0, +\infty[\times\mathbb{R}^q$ there exists a neighborhood V such that $x_0 \mapsto \sup_{\xi\in V} |\langle \theta^{\delta_{x_0}}(\xi),\, \phi\rangle|$ belongs to $L^1(\mu_0)$ for any ϕ in $\mathcal{D}(\mathbb{R}^n)$ (in order to apply Lebesgue's dominated convergence theorem). Then an easy extension of the proof provides an FDR for $\mathcal{F}^{\mu_0}$, where the nonnegative output function θ^{μ_0} is defined by

$$\langle \theta^{\mu_0}(\xi),\, \phi\rangle \;=\; \int_{\mathbb{R}^n} d\mu_0(x_0)\, \left\langle \theta^{\delta_{x_0}}(\xi),\, \phi\right\rangle.$$

• If M_0 is a first-order differential operator, everything remains valid with P_t^0 defined like $P_t^1, \ldots, P_t^q$.

Appendix A. A Baker–Campbell–Hausdorff formula. We are providing a variation of the Baker–Campbell–Hausdorff formula which gives, under proper assumptions, the exponential expression of a product of exponentials in a Lie group [7, sections 6, 7]. Our formulation is close to that of Wei and Norman [43] but in the field of semigroups of operators, namely a sort of exponential of differential operator. The proof of the formula does not appeal to functional spaces theory as in [33, 45] for the particular cases treated but does appeal to partial differential calculations as in [38].

THEOREM A.1. *Let Y and Z be two smooth linear differential operators on $\mathbb{R}^n$. Let D be a set of smooth functions ϕ such that the PDE*

$$\frac{\partial u}{\partial t} \;=\; Y\, u, \quad u(0, x) \;=\; \phi(x) \tag{A.56}$$

has a unique solution $u(t,x)$ *in* $C^\infty(I\times\mathbb{R}^n)$*, where* $I=]0,+\infty[$ *or* $I=\mathbb{R}$*. In such a case, we denote* $u(t,x)$ *by* $P_t\,\phi\,(x)$*.*

If the Lie algebra $\mathcal{O}$ *of differential operators generated by* Y *and* Z *is finite dimensional and if* $\mathcal{O}\,D\,\subset\,D$*, then we have the following relation:*

$$\forall\, t\in I,\quad \forall\,\phi\in D,\quad Z\,(P_t\,\phi)\;=\;P_t\,(e^{-t\,ad_Y}\,Z\,\phi). \tag{A.57}$$

Here, $ad_Y(Z)=[Y,Z]$ *and* $e^{t\,ad_Y}$ *is the linear operator* $e^{t\,ad_Y}=\sum_{j=0}^{+\infty}\frac{t^j\,ad_Y^j}{j!}$ *of the finite-dimensional Lie algebra* $\mathcal{O}$*.*

Proof. Suppose that for any $X\in\mathcal{O}$ we show that

$$\forall\, t\in I,\quad \forall\,\phi\,\in\,D,\quad e^{t\,\mathrm{ad}_Y}\,X\,(P_t\,\phi)\;=\;P_t\,(X\,\phi)\,. \tag{A.58}$$

Then, if we replace X by $e^{-t\,\mathrm{ad}_Y}Z$, the theorem is proved.

With the definition of D, equation (A.58) means that the function u defined by $u(t,x)=e^{t\,\mathrm{ad}_Y}X\,(P_t\,\phi)(x)$ satisfies (A.56) with initial condition $u(0,x)=X\phi$. Since this last property of u is straightforward, let us compute $\frac{\partial u}{\partial t}$. If $\{W_1,\,\ldots,\,W_q\}$ is a basis of $\mathcal{O}$, then $e^{-t\,\mathrm{ad}_Y}X$ can be written as $e^{t\,\mathrm{ad}_Y}X=\sum_{j=1}^q\,z_j(t)\,W_j$, where $z_1,\,\ldots,\,z_q$ are smooth functions of t and therefore

$$u(t,x)=\sum_{j=1}^{q}\,z_j(t)\,W_j(P_t\,\phi)(x). \tag{A.59}$$

The function u is smooth on $I\times\mathbb{R}^n$ and we have

$$\frac{\partial u}{\partial t}=\sum_{j=1}^{q}\,\frac{\partial}{\partial t}(z_j(t)\,W_j\,(P_t\,\phi))=\sum_{j=1}^{q}\,\frac{dz_j(t)}{dt}W_j\,(P_t\,\phi)\,+\,\sum_{j=1}^{q}\,z_j(t)\,\frac{\partial}{\partial t}(W_j\,P_t\,\phi).$$

Now, $P_t\,\phi\,(x)$ is smooth in both variables (t,x) so that we can swap the differentiations $\frac{\partial}{\partial t}$ in the variable t and W_j in the variable x to get

$$\frac{\partial u}{\partial t}=\sum_{j=1}^{q}\,\frac{dz_j(t)}{dt}W_j\,(P_t\,\phi)\,+\,\sum_{j=1}^{q}\,z_j(t)\,W_j\,\frac{\partial}{\partial t}(P_t\,\phi)\,.$$

To compute $\frac{dz_j(t)}{dt}$, note that $Z_t=e^{t\,\mathrm{ad}_Y}X$ is the solution of the linear differential equation $\frac{dZ_t}{dt}=[Y,\,Z_t]$, $Z_0=Z$ in the finite-dimensional real vector space $\mathcal{O}$, so that $\frac{dZ_t}{dt}=[Y,\,Z_t]=\sum_{j=1}^q\,z_j(t)\,[Y,\,W_j]$. Since $\frac{dZ_t}{dt}=\sum_{j=1}^q\,\frac{dz_j(t)}{dt}\,W_i$, we have

$$\sum_{j=1}^{q}\,\frac{dz_j(t)}{dt}\,W_j=\sum_{j=1}^{q}\,z_j(t)\,[Y,\,W_j].$$

Therefore,

$$\begin{aligned}\frac{\partial u}{\partial t}\;&=\;\sum_{j=1}^{q}\,z_j(t)\,[Y,\,W_j]\,(P_t\,\phi)\,+\,\sum_{j=1}^{q}\,z_j(t)\,W_j\,Y\,(P_t\,\phi)\\&=\;\sum_{j=1}^{q}\,z_j(t)\,Y\,W_j\,(P_t\,\phi)\\&=\;Y\,\left(\sum_{j=1}^{q}\,z_j(t)\,W_j\right)\,(P_t\,\phi)\\&=\;Y\,e^{t\,\mathrm{ad}_Y}X\,(P_t\,\phi)\;=\;Y\,u.\end{aligned}$$

This ends the proof. □

Acknowledgment. The author is indebted to Professor J. Lévine for fruitful discussions.

REFERENCES

[1] J. Baras, *Group invariance methods in nonlinear filtering of diffusion processes*, in Stochastic Systems: The Mathematics of Filtering and Identification and Applications, M. Hazewinkel and J. Willems, eds., D. Reidel, Dordrecht, the Netherlands, 1981, pp. 565–572.

[2] J. Baras, G. Blankenship, and W. Hopkins Jr, *Existence, uniqueness and asymptotic behaviour of solutions to a class of Zakai equations with unbounded coefficients*, IEEE Trans. Automat. Control, 28 (1983), pp. 203–214.

[3] V. Beneš, *Exact finite-dimensional filters for certain diffusions with nonlinear drift*, Stochastics, 5 (1982), pp. 65–92.

[4] V. Beneš, *New exact nonlinear filters with large Lie algebras*, Systems Control Lett., 5 (1985), pp. 217–221.

[5] V. Beneš and I. Karatzas, *Estimation and control for linear, partially observable systems with non-gaussian initial distribution*, Stochastic Process. Appl., 14 (1983), pp. 233–248.

[6] J. Bony, P. Courrège, and P. Priouret, *Semi groupes de Feller sur une variété à bord compacte et problèmes aux limites intégro-différentiels donnant lieu au principe du maximum*, Ann. Inst. Fourier (Grenoble), 18 (1968), pp. 369–521.

[7] N. Bourbaki, *Groupes et Algèbres de Lie*, Hermann, Paris, 1972, Chapters 2 and 3.

[8] R. Brockett, *Remarks on finite dimensional nonlinear estimation*, Astérisque, 764 (1980), pp. 47–55.

[9] R. Brockett, *Nonlinear systems and nonlinear estimation theory*, in Stochastic Systems: The Mathematics of Filtering and Identification and Applications, M. Hazewinkel and J. Willems, eds., D. Reidel, Dordrecht, the Netherlands, 1981, pp. 442–477.

[10] R. Brockett and J. Clark, *The geometry of the conditional density equation*, in Analysis and Optimization of Stochastic Systems, O. Jacobs et al., eds., Academic Press, New York, 1980, pp. 299–310.

[11] M. Chaleyat-Maurel and D. Michel, *Des résultats de non existence de filtre de dimension finie*, Stochastics, 13 (1984), pp. 83–102.

[12] M. Cohen de Lara, *Application of symmetry semi-groups to discrete and continuous time filtering problems*, in Analysis of Controlled Dynamical Systems, B. Bonnard, B. Bride, J. Gauthier, and I. Kupka, eds., Birkhäuser Boston, Cambridge, MA, 1991, pp. 146–155.

[13] F. Daum, *Exact finite-dimensional nonlinear filters*, IEEE Trans. Automat. Control, 31 (1986), pp. 616–622.

[14] F. Daum, *Solution of the Zakai equation by separation of variables*, IEEE Trans. Automat. Control, 32 (1986), pp. 941–943.

[15] M. Davis and S. Marcus, *An introduction to nonlinear filtering*, in Stochastic Systems: The Mathematics of Filtering and Identification and Applications, M. Hazewinkel and J. Willems, eds., D. Reidel, Dordrecht, the Netherlands, 1981, pp. 53–75.

[16] J. Dieudonné, *Éléments d'Analyse*, tome 4, Gauthier-Villars, Paris, 1975.

[17] J. Dieudonné, *Éléments d'Analyse*, tome 5, Gauthier-Villars, Paris, 1977.

[18] U. Haussmann and E. Pardoux, *A conditionally almost linear filtering problem with non-gaussian initial condition*, Stochastics, 23 (1988), pp. 241–275.

[19] M. Hazewinkel and S. Marcus, *Some results and speculations on the role of Lie algebras in filtering*, in Stochastic Systems: The Mathematics of Filtering and Identification and Applications, M. Hazewinkel and J. Willems, eds., D. Reidel, Dordrecht, the Netherlands, 1981, pp. 591–604.

[20] M. Hazewinkel, S. Marcus, and H. Sussmann, *Nonexistence of exact finite dimensional filters for conditional statistics of the cubic sensor problem*, Systems Control Lett., 3 (1983), pp. 331–340.

[21] O. Hijab, *A realization theory for nonlinear stochastic systems*, in Proc. IEEE Conference on Decision and Control, San Antonio, TX, 1983, pp. 98–903.

[22] A. Isidori, *Nonlinear Control Systems: An Introduction*, 2nd ed., Springer-Verlag, Berlin, 1989.

[23] R. Kalman and R. Bucy, *New results in linear filtering and prediction theory*, Trans. ASME Ser. D. J. Basic Engrg., 83 (1961), pp. 95–108.

[24] T. Kurtz and D. Ocone, *Unique characterization of conditional distributions in nonlinear filtering*, Ann. Probab., 16 (1988), pp. 80–107.

[25] J. Lévine, *Finite dimensional filters for a class of nonlinear systems and immersion in a linear system*, SIAM J. Control Optim., 25 (1987), pp. 1430–1439.

[26] J. Lévine, *Finite dimensional realizations of stochastic p.d.e.'s and application to filtering*, Stochastics, 37 (1991), pp. 75–103.
[27] J. Lévine and G. Pignié, *Exact finite dimensional filters for a class of nonlinear discrete-time systems*, Stochastics, 10 (1986), pp. 97–132.
[28] A. Makowski, *Filtering formulae for partially observed linear systems with non-gaussian initial conditions*, Stochastics, 13 (1986), pp. 1–14.
[29] S. Marcus, *Algebraic and geometric methods in nonlinear filtering*, SIAM J. Control Optim., 22 (1984), pp. 817–844.
[30] D. Michel, *Régularité des lois conditionnelles en théorie du filtrage non-linéaire et calcul des variations stochastique*, J. Funct. Anal., 41 (1981), pp. 8–36.
[31] S. Mitter, *On the analogy between mathematical problems of non-linear filtering and quantum physics*, Ricerche Automat., 10 (1979), pp. 163–216.
[32] D. Ocone, *Stochastic Calculus of Variations for Stochastic Partial Differential Equations*, preprint.
[33] D. Ocone, *Topics in Nonlinear Filtering Theory*, Ph.D. thesis, Massachusetts Institute of Technology, Cambridge, MA, 1980.
[34] D. Ocone, *Finite dimensional Lie algebras in nonlinear filtering*, in Stochastic Systems: The Mathematics of Filtering and Identification and Applications, M. Hazewinkel and J. Willems, eds., D. Reidel, Dordrecht, the Netherlands, 1981, pp. 629–636.
[35] E. Pardoux, *Stochastic partial differential equations and filtering of diffusion processes*, Stochastics, 3 (1979), pp. 127–167.
[36] E. Pardoux, *Nonlinear filtering, prediction and smoothing*, in Stochastic Systems: The Mathematics of Filtering and Identification and Applications, M. Hazewinkel and J. Willems, eds., D. Reidel, Dordrecht, the Netherlands, 1981, pp. 529–558.
[37] E. Pardoux, *Filtrage non linéaire et équations aux dérivées partielles stochastiques associées*, École d'été de Probabilités de Saint-Flour, 1989.
[38] S. Rosencrans, *Perturbation algebra of an elliptic operator*, J. Math. Anal. Appl., 56 (1976), pp. 317–329.
[39] Z. Roth and K. Loparo, *Optimal filter realization for a class of nonlinear systems with finite dimensional estimation algebra*, Systems Control Lett., 4 (1984), pp. 23–26.
[40] I. Shigekawa, *Transformations of the brownian motion on the riemannian symmetric space*, Z. Wahrsch. Ver. Geb., 65 (1984), pp. 493–522.
[41] L. Tam, W. Wong, and S. Yau, *On a necessary and sufficient condition for finite dimensionality of estimation algebras*, SIAM J. Control Optim., 28 (1990), pp. 173–185.
[42] A. Ustunel, *Some applications of stochastic integration in infinite dimensions*, Stochastics, 7 (1982), pp. 255–268.
[43] J. Wei and E. Norman, *On the global representation of the solutions of linear differential equations as a product of exponentials*, Proc. Amer. Math. Soc., 15 (1964), pp. 327–334.
[44] E. Wong, *Stochatic Processes in Information and Dynamical Systems*, McGraw-Hill, New York, 1971.
[45] W. Wong, *New classes of finite dimensional filters*, Systems Control Lett., 3 (1983), pp. 155–164.
[46] S. S.-T. Yau, *Finite dimensional filters with nonlinear drift* I*: A class of filters including both Kalman-Bucy filters and Beneš filters*, J. Math. Systems, Estim. Control, 4 (1994), pp. 181–203.
[47] M. Zakai, *On the optimal filtering of diffusion processes*, Z. Wahrsch. Ver. Geb., 11 (1969), pp. 230–243.
[48] O. Zeitouni, *On some finite dimensional nonlinear filters for certain diffusions observed in correlated noise*, Systems Control Lett., 7 (1986), pp. 61–63.

SIAM J. CONTROL OPTIM.
Vol. 35, No. 3, pp. 1002–1029, May 1997

012

FINITE-DIMENSIONAL FILTERS. PART II: INVARIANCE GROUP TECHNIQUES*

M. COHEN DE LARA†

Abstract. This two-part paper deals with necessary or sufficient conditions for the existence of finite-dimensional filters. In the first part, we set the problem and proposed a construction of finite-dimensional filters by the Wei–Norman technique. In this second part, we show how geometric-methods offer another approach that is more powerful, as we shall see. The invariance group of a parabolic equation is introduced and its action on initial data enhanced. This is applied to the problem of finite-dimensional realization of bilinear stochastic PDEs and further simplified by the introduction of a Riemannian framework. We end by an analysis of partially observed systems having finite-dimensional filters, with emphasis on the case of systems with correlated noise.

Key words. finite-dimensional filter, estimation Lie algebra, bilinear stochastic partial differential equation, invariance group, Riemannian geometry

AMS subject classifications. 93E11, 60G35, 58D19, 35K10

PII. S0363012994270916

Notation.

$A(F)$: subgroup of the group $G(F)$ of (global) diffeomorphisms of $F = \mathbb{R} \times \mathbb{R}^n \times \mathbb{R}$, consisting of diffeomorphisms g of the form (26); that is, $g(s,x,u) = (\alpha(s), \beta(s,x), \delta(s,x)u)$.

$SA(F)$: semigroup of $A(F)$ consisting of diffeomorphisms g such that $g^{-1}(\mathbb{R}_+ \times \mathbb{R}^n \times (0,+\infty)) \subset \mathbb{R}_+ \times \mathbb{R}^n \times (0,+\infty)\}$; that is, $\alpha^{-1}(\mathbb{R}_+) \subset \mathbb{R}_+$ and $\delta > 0$.

$\mathcal{U}_P$: subset of $C^\infty(\mathbb{R}_+ \times \mathbb{R}^n)$, consisting of functions $u(t,x)$, solutions of $\frac{\partial u}{\partial t} = M_0^\star u$.

$SA_P(F)$: symmetry semigroup of $\mathcal{U}_P$, consisting of diffeomorphisms $g \in SA(F)$ such that for all $u \in \mathcal{U}_P$, $g \cdot u \in \mathcal{U}_P$:

$$SA_P(F) \subset SA(F) \subset A(F).$$

$\mathcal{A}^c(F)$: set of *complete* smooth vector fields Z on F of the form (37); that is, $Z = \zeta^0(s)\frac{\partial}{\partial s} + \sum_{i=1}^n \zeta^i(s,x)\frac{\partial}{\partial x_i} + \zeta^{n+1}(s,x)u\frac{\partial}{\partial u}$.

$\mathcal{SA}_P^c(F)$: set of complete infinitesimal symmetries of $\mathcal{U}_P$; that is, set of vector fields Z in $\mathcal{A}^c(F)$ such that Φ_r^Z belongs to $SA_P(F)$, for all $r \in \mathbb{R}$,

$$\mathcal{A}^c(F) \subset \mathcal{SA}_P^c(F) \subset \mathcal{X}^c(F) \subset \mathcal{X}(F).$$

1. Introduction.

In this second part we present geometric methods which are powerful to analyze a stochastic partial bilinear partial differential equation (PDE), such as the Zakai equation, and, should the occasion occur, to build a finite-dimensional realization (FDR). The range of application of such methods will be shown to be broader than the Wei–Norman techniques developed in Part I.

In section 2, we motivate the introduction of invariance groups by treating the case of the one-dimensional heat equation and, after stating the problem in section 3, we formalize in section 4 the notions behind the example developed above and introduce basic definitions, notations, and properties for the next sections. In section 5, we show how such invariance group techniques may bring an answer to the FDR problem and state our main result. In section 6, we compare main theorems from Parts I and II. A Riemannian geometric framework is introduced in section 7, which greatly simplifies

*Received by the editors July 8, 1994; accepted for publication (in revised form) April 8, 1996.
http://www.siam.org/journals/sicon/35-3/27091.html

†Centre d'Enseignement et de Recherche pour la Gestion des Ressources Naturelles et de l'Environnement, École Nationale des Ponts et Chaussées, 6 et 8 av. Blaise Pascal, Cité Descartes, 77455 Marne la Vallée Cédex 2, France (mcdl@cergrene.enpc.fr).

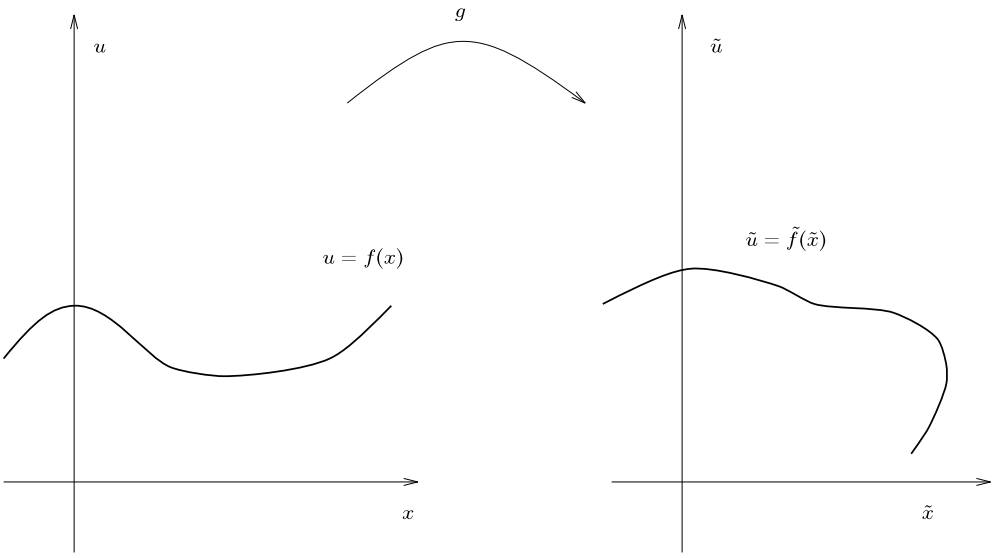


FIG. 1. *Action of a group transformation on a function.*

the analysis of the FDR problem. In section 8, a systematic treatment of the finite dimensional filter (FDF) problem by this latter technique is presented.

2. Motivation. The set of planar motions which keep a geometric figure invariant forms a group, the *symmetry group* of the figure (square, triangle, circle, etc.). In the case of an algebraic equation, a symmetry group or *invariance group* consists of transformations of the base space which permute solutions. (This is one of the basic concepts of Galois's theory.) In some cases, the knowledge of such a group may help solving the equation as in the classical example of the "bisquared" equation:

$$x^4 + bx^2 + c = 0 \iff z = x^2 \quad \text{and} \quad z^2 + bz + c = 0\,.$$

In the case of ordinary differential equations (ODEs), it was pointed out by Lie that all the special techniques to solve certain classes of ODEs had their origin in a general method related to the existence of a continuous invariance group for these ODEs (see the introduction of [16]). Basically, this (local) group consists of geometric transformations of the product space "independent variables" × "dependent variables," and its action on functions consists in transforming their graph as in Figure 1, these transformed graphs being graphs of solutions of the original ODE.

Continuous groups are interesting because they can be found by some calculation algorithms. It is indeed a crucial fact of Lie theory that "the nonlinear conditions expressing the invariance of a system of differential equations under the action of a group of transformations may, in the case of continuous groups, be replaced by *linear* but simpler equivalent conditions" [16]. The latter represent the infinitesimal invariance of the system under the action of the infinitesimal generators of the group.

All the Lie theory can be extended to PDEs [17, 16, 4]. The case of evolution equations, where the variable t plays a specific role, is particularly interesting to us.

Indeed, Rosencrans in [20] utilizes the invariance group of a linear evolution equation of the form $u_t = Au$, where A is a differential operator "acting on the space variables," not to find new solutions from one of them, but to exhibit solutions of

other evolution equations of the form $w_t = (A + P)w$. The set of possible operators P is shown to form a Lie algebra, called the *perturbation algebra*. We illustrate this in the case of the one-dimensional heat equation.

PROPOSITION 2.1. *Let $(g_{r,r'})_{(r,r')\in\mathbb{R}^2}$ be the family of two-parameter diffeomorphisms of $\mathbb{R}\times\mathbb{R}\times\mathbb{R}$ defined by*

$$g_{r,r'}(t,x,u) = (e^{-2r}t - r', e^{-r}x, u)\,. \tag{1}$$

The action of the diffeomorphism $g_{r,r'}$ on a function $u(t,x)$ defines a new function $g_{r,r'}\cdot u$ by the relation

$$g_{r,r'}(\mathit{graph}(u)) \supset \mathit{graph}(g_{r,r'}\cdot u) \tag{2}$$

(see Figure 1*), and its expression is given by*

$$(g_{r,r'}\cdot u)(t',x') = u(e^{2r}(t'+r'), e^r x')\,. \tag{3}$$

The family $(g_{r,r'})_{(r,r')\in\mathbb{R}^2}$ is actually a two-parameter group *of diffeomorphisms, and the mapping*

$$g_{r,r'} \mapsto g_{r,r'}\cdot u \tag{4}$$

defines a group action. Moreover, $(g_{r,r'})$ is an invariance group of the one-dimensional heat equation $u_t = u_{xx}$.

If u is a solution of the heat equation with initial data ϕ, then for $r\in\mathbb{R}$, $r'\geq 0$, $g_{r,r'}\cdot u$ is also a solution of the heat equation but with initial data $(g_{r,r'}\cdot u)(0,x') = u(e^{2r}r', e^r x')$. What is more, the group action (4) *of $g_{r,r'}$ for $r\in\mathbb{R}$, $r'\in\mathbb{R}_+$ induces a group action*

$$g_{r,r'} \mapsto \widehat{g}_{r,r'}\phi \tag{5}$$

by mapping the initial data ϕ to the latter initial data:

$$(\widehat{g}_{r,r'}\phi)(x') = (g_{r,r'}\cdot u)(0,x') = u(e^{2r}r', e^r x')\,. \tag{6}$$

To this group action corresponds a Lie infinitesimal action. The infinitesimal generators of the two-parameter group of diffeomorphisms $(g_{r,r'})_{(r,r')\in\mathbb{R}^2}$ form a Lie algebra generated by the vector fields $\frac{\partial}{\partial t}$ and $2t\frac{\partial}{\partial t} + x\frac{\partial}{\partial x}$. The one-parameter group of diffeomorphisms $\Phi_s^Z(t,x,u)$ generated by the vector field

$$Z = -a\frac{\partial}{\partial t} - b\left(2t\frac{\partial}{\partial t} + x\frac{\partial}{\partial x}\right) \tag{7}$$

belongs to the two-parameter group of diffeomorphisms $(g_{r,r'})_{(r,r')\in\mathbb{R}^2}$ and may be written

$$\Phi_s^Z(t,x,u) = \left(te^{-2bs} - a\frac{1-e^{-2bs}}{2b}, e^{-bs}x, u\right) = g_{r(s),r'(s)}(t,x,u)\,, \tag{8}$$

where $r(s) = bs$, $r'(s) = a(1-e^{-2bs})/2b$. Then, for $a\geq 0$, $(\widehat{g}_{r(s),r'(s)})_{s\geq 0}$ defines a semigroup of operators with infinitesimal generator

$$a\frac{d^2}{dx^2} + bx\frac{d}{dx}\,. \tag{9}$$

In particular, if $u(t,x)$ is the solution of the heat equation with initial data ϕ, then a solution to the Cauchy problem

$$\frac{\partial w}{\partial s} = a\frac{\partial^2 w}{\partial x^2} + bx\frac{\partial w}{\partial x}\,, \quad w(0,x) = \phi(x) \tag{10}$$

is given by

$$w(s,x) = u\left(a\frac{e^{2bs}-1}{2b}, e^{-bs}x\right)\,. \tag{11}$$

Proof. For a given function $u(t,x)$, we have

$$\begin{aligned} g_{r,r'}(\mathrm{graph}(u)) &= \{g_{r,r'}(t,x,u(t,x)) \mid (t,x) \in \mathrm{Dom}(u)\} \\ &= \{(e^{-2r}t - r', e^{-r}x, u(t,x)) \mid (t,x) \in \mathrm{Dom}(u)\} \\ &= \{(t',x',u(e^{2r}(t'+r'), e^r x')) \mid (t',x') \in g_{r,r'}\,(\mathrm{Dom}(u))\}. \end{aligned}$$

This justifies the expression of $(g_{r,r'}\cdot u)(t',x')$ in (3).

Note that the family $(g_{r,r'})_{r,r'\in\mathbb{R}^2}$ is generated by the subfamilies

$$g_{r,0}(t,x,u) = (e^{-2r}t, e^{-r}x, u) \quad \text{and} \quad g_{0,r'}(t,x,u) = (t-r',x,u) \tag{12}$$

since we have the relations

$$\begin{cases} g_{r,0}\circ g_{0,r'}(t,x,u) &= (e^{-2r}(t-r'), e^{-r}x, u) = g_{r,r'}(t,x,u), \\ g_{0,r'}\circ g_{r,0}(t,x,u) &= (e^{-2r}t - r', e^{-r}x, u) = g_{r,e^{2r}r'}(t,x,u)\,. \end{cases} \tag{13}$$

It is thus a two-parameter group of diffeomorphisms. The mapping (4) does define a group action since, on the one hand, we have

$$g_{q,q'}\,(g_{r,r'}(\mathrm{graph}(u))) \supset g_{q,q'}\,(\mathrm{graph}(g_{r,r'}\cdot u)) \supset \mathrm{graph}(g_{q,q'}\cdot(g_{r,r'}\cdot u))$$

and, on the other hand,

$$g_{q,q'}\,(g_{r,r'}(\mathrm{graph}(u))) = g_{q,q'}\circ g_{r,r'}\,(\mathrm{graph}(u)) \supset \mathrm{graph}\,((g_{q,q'}\circ g_{r,r'})\cdot u))$$

so that

$$g_{q,q'}\cdot(g_{r,r'}\cdot u) = (g_{q,q'}\circ g_{r,r'})\cdot u. \tag{14}$$

If $u_t = u_{xx}$, we deduce from (3) that

$$\frac{\partial g_{r,r'}\cdot u}{\partial t'}(t',x') = e^{2r}\frac{\partial^2 u}{\partial x^2}(e^{2r}(t'+r'), e^r x') = \frac{\partial^2 g_{r,r'}\cdot u}{\partial x'^2}(t',x')\,.$$

Therefore, the two-parameter group of diffeomorphisms $(g_{r,r'})_{r,r'\in\mathbb{R}^2}$ is an *invariance group* of the one-dimensional heat equation $u_t = u_{xx}$.

Let u be a function with domain $\mathrm{Dom}(u) = [0,+\infty)\times\mathbb{R}$. Its image $g_{r,r'}\cdot u$ by the diffeomorphism $g_{r,r'}$ has the domain $\mathrm{Dom}(g_{r,r'}\cdot u) = [-r',+\infty)\times\mathbb{R}$ by the formula (3). If u is solution of the heat equation with initial data $u(0,\cdot) = \phi$, then for $r\in\mathbb{R}$ and $r'\geq 0$, $g_{r,r'}\cdot u$ is also a solution of the heat equation, but for other initial data $(g_{r,r'}\cdot u)(0,x') = u(e^{2r}r', e^r x')$. The group action (4) of $g_{r,r'}$ for $r\in\mathbb{R}$, $r'\geq 0$, clearly induces a group action on the initial data.

To this group action corresponds a Lie infinitesimal action as follows. It is well known from elementary Lie group theory [16] that the infinitesimal generators of the two-parameter group of diffeomorphisms $(g_{r,r'})_{r,r'\in\mathbb{R}^2}$ form a Lie algebra generated by the vector fields

$$\left\{\begin{array}{rcl} \frac{\partial}{\partial r}_{|r=0} g_{r,0}(t,x,u) & = & (-2t,-x,0) = -2t\frac{\partial}{\partial t} - x\frac{\partial}{\partial x}, \\ \frac{\partial}{\partial r'}_{|r'=0} g_{0,r'}(t,x,u) & = & (-1,0,0) = -\frac{\partial}{\partial t}\,. \end{array}\right.$$

The one-parameter group of diffeomorphisms $\Phi_s^Z(t,x,u)$ generated by the vector field $Z = -a\frac{\partial}{\partial t} - b(2t\frac{\partial}{\partial t} + x\frac{\partial}{\partial x})$ is the flow of the differential equation

$$\frac{dt}{ds} = -a - 2bt\,, \quad \frac{dx}{ds} = -bx\,, \quad \frac{du}{ds} = 0\,. \tag{15}$$

One finds $t(s) = t(0)e^{-2bs} - a(1-e^{-2bs})/2b$, $x(s) = e^{-bs}x(0)$, $u(s) = u(0)$, hence the expression of $\Phi_s^Z(t,x,u)$ in (8). The comparison of (8) and (1) yields the expressions $r(s) = bs$ and $r'(s) = a(1-e^{-2bs})/2b$.

For $a \geq 0$ and $s \geq 0$, we have $r'(s) \geq 0$, and $(\widehat{g}_{r(s),r'(s)})_{s\geq 0}$ does define a semigroup of operators on the initial data since (5) is a group action ($(g_{r(s),r'(s)})_{s\geq 0}$ extends to the initial data its action on the solutions of the heat equation). If u is the solution of the heat equation with initial data ϕ, then the infinitesimal generator of this semigroup may be computed by

$$\begin{aligned} \frac{d}{ds}_{|s=0} \widehat{g}_{r(s),r'(s)}\phi(x') &= \frac{d}{ds}_{|s=0} u(e^{2r(s)}r'(s), e^{r(s)}x') \\ &= \frac{d}{ds}_{|s=0} u\left(ae^{2bs}\frac{1-e^{-2bs}}{2b}, e^{bs}x'\right) \\ &= \frac{d}{ds}_{|s=0} u\left(a\frac{e^{2bs}-1}{2b}, e^{bs}x'\right) \\ &= a\frac{\partial u}{\partial t}(0,x') + bx'\frac{\partial u}{\partial x}(0,x') \\ &= a\frac{\partial^2 u}{\partial x^2}(0,x') + bx'\frac{\partial u}{\partial x}(0,x') \\ &= a\frac{\partial^2 \phi}{\partial x^2}(x') + bx'\frac{\partial \phi}{\partial x}(x')\,. \end{aligned}$$

In particular, a solution of the Cauchy problem (10) is given by

$$\begin{aligned} w(s,x') &= \widehat{g}_{r(s),r'(s)}\phi(x') \\ &= g_{r(s),r'(s)}\cdot u(0,x) \\ &= u(e^{2r(s)}r'(s), e^{r(s)}x') \\ &= u\left(ae^{2bs}\frac{1-e^{-2bs}}{2b}, e^{bs}x'\right). \end{aligned}$$ □

The link between these techniques and the filtering problem appears when one writes the Zakai equation for the density as $dp_t = (M_0^\star + \sum_{i=1}^p M_i^\star \dot{y}_i(t))p_t dt$. Baras in [2] utilizes this form to define a notion of equivalence between two filtering problems. He gives general conditions, related to the existence of a common invariance group, under which the solution of a filtering problem may be computed from the solution of the other problem.

Here, we extend the method to the resolution of valued functions stochastic evolution equations. We first consider the unperturbed equation $dp_t = M_0^\star p_t dt$ (when $y_t \equiv 0$) and its invariance group. The generators of this latter group may be used to describe the deformations of solutions of the unperturbed equation for each trajectory of the process (y_t) and yield an accurate description of the structure of the resulting finite realization.

3. Problem statement. Let $(y_t)_{t\geq 0}$ be a standard p-dimensional Brownian motion and let $\circ dy_t$ denote its "Stratonovitch differential." For $p_0 \in C^\infty(\mathbb{R}^n)$, we consider the following stochastic bilinear PDE:

$$\text{with probability one (w.p.1) } \forall\, t \in \mathbb{R}_+\,, \quad dp_t = M_0^\star p_t dt + \sum_{i=1}^{p} M_i^\star p_t \circ dy_t^i\,, \quad p_{t|t=0} = p_0. \tag{16}$$

Here $\mathcal{D}(\mathbb{R}^n)$ denotes the space of smooth functions on $\mathbb{R}^n$ with compact support and M_0, M_1, ..., M_p are the linear differential operators defined by

$$\begin{cases} M_0\phi &= \mathcal{L}\phi + H\phi, \\ M_i\phi &= L_{\tilde{g}_i}\phi + h_i\phi\,, \quad i = 1,\ldots,p, \qquad \forall\, \phi \in \mathcal{D}(\mathbb{R}^n)\,, \end{cases} \tag{17}$$

$\mathcal{L}$ being a smooth diffusion operator ($\mathcal{L}1 = 0$), $\tilde{g}_1$, ..., $\tilde{g}_p$ smooth vector fields on $\mathbb{R}^n$ and H, h_1, ..., h_p smooth functions from $\mathbb{R}^n$ to $\mathbb{R}$.

Assumption 1. M_0 (or $\mathcal{L}$) is nondegenerate elliptic; that is, if

$$M_0 = \frac{1}{2}\sum_{i,j=1}^{n} a^{ij}(x)\frac{\partial^2}{\partial x_i \partial x_j} + \sum_{i=1}^{n} b^i(x)\frac{\partial}{\partial x_i} + H(x), \tag{18}$$

the symmetric matrix $(a^{ij}(x))_{i,j=1,\ldots,n}$ is positive definite for all $x \in \mathbb{R}^n$.

Remark. In the filtering problem stated in the introduction of Part I, the Zakai equation is a particular case of equation (16) with

$$\mathcal{L}\phi = \frac{1}{2}\sum_{k=1}^{m} L_{g_k}^2\phi + L_f\phi \quad \text{and} \quad H = -\frac{1}{2}\left(\|h\|^2 + \sum_{i=1}^{p} L_{\tilde{g}_i}h_i\right). \tag{19}$$

DEFINITION 3.1. *Let* Π *consist of all functions* $p_0 \in C^\infty(\mathbb{R}^n)$ *such that there exists an input-output map*

$$\mathcal{F}^{p_0} : \mathbb{R}_+ \times C^0(\mathbb{R}_+, \mathbb{R}^p) \to C^\infty(\mathbb{R}^n) \tag{20}$$

which satisfies the following property: for all $p_0 \in \Pi$ *and* p*-dimensional Brownian motion* $(y_t)_{t\in\mathbb{R}_+}$*, the stochastic bilinear PDE* (16) *has a unique solution* $(p_t)_{t\in\mathbb{R}_+}$ *in* $C^\infty(\mathbb{R}^n)$ *given by*

$$w.p.1 \quad \forall\, t \in \mathbb{R}_+\,, \quad p_t = \mathcal{F}^{p_0}(t; y_s, s \leq t)\,. \tag{21}$$

The definition of an FDR for the input-output map $\mathcal{F}^{p_0}$ is the same as in Part I except for the output function θ. Indeed, θ must be a smooth map from $\mathrm{Dom}(\theta) \subset M$ to $C^\infty(\mathbb{R}^n)$; that is, the following function $(\xi, x) \in \mathrm{Dom}(\theta) \times \mathbb{R}^n \mapsto \theta(\xi)(x) \in \mathbb{R}$ must be smooth.

DEFINITION 3.2. *An* FDR *of the input-output map* $\mathcal{F}^{p_0}$ *given by* (20) *consists of a collection* $(M, \xi_0, b_0, b_1, \ldots, b_p, \theta)$*, where*

1. *M is a smooth finite-dimensional manifold,*
2. *$\xi_0 \in M$,*
3. *b_0, b_1, ..., b_p are smooth vector fields on M,*
4. *θ is a smooth map from $\mathrm{Dom}(\theta) \subset M$ to $C^\infty(\mathbb{R}^n)$; that is, the following function $(\xi, x) \in \mathrm{Dom}(\theta) \times \mathbb{R}^n \mapsto \theta(\xi)(x) \in \mathbb{R}$ must be smooth,*

such that if $(y_t)_{t\geq 0}$ is a Brownian motion, the Stratonovitch stochastic differential equation (SDE)

$$d\xi_t = b_0(\xi_t)dt + \sum_{i=1}^{p} b_i(\xi_t) \circ dy_t^i\,, \quad \xi_{t\,|t=0} = \xi_0 \tag{22}$$

is conservative (that is, w.p.1 has a solution for all time or w.p.1 has infinite explosion time) with solution satisfying

$$w.p.1 \quad \forall\, t > 0\,, \quad \xi_t \in \mathrm{Dom}(\theta)$$

and

$$w.p.1 \quad \forall\, t \geq 0\,, \quad \theta(\xi_t) = \mathcal{F}^{p_0}(t; y_s, s \leq t) = \nu_t\,. \tag{23}$$

4. Symmetry semigroup and infinitesimal symmetries. In this section, we formalize the notions behind the example developed above and introduce basic definitions, notations, and properties for the next sections (see also [7]). The symmetry semigroup consists of a class of point transformations of

$$F = \mathbb{R} \times \mathbb{R}^n \times \mathbb{R} \tag{24}$$

which permute graphs of the solutions of the parabolic equation $u_t = M_0^\star u$. If symmetry groups commonly used in the study of symmetries of differential equations [17, 16, 4, 20] are *local*, we must consider here *global* groups because the Zakai equation is defined on all $\mathbb{R}^n$.

DEFINITION 4.1. *Let $\mathcal{D}(M_0^\star)$ be the subspace of $C^\infty(\mathbb{R}^n)$ consisting of functions p_0 such that the following PDE*

$$\frac{\partial u}{\partial t} = M_0^\star u\,, \quad u(0,\cdot) = p_0 \tag{25}$$

has a unique solution $u(t,x) \in C^\infty(\mathbb{R}_+ \times \mathbb{R}^n)$. Let $(P_t)_{t\in\mathbb{R}_+}$ be the nonnegative (since $M_0^\star$ is a second-order differential operator) semigroup on $\mathcal{D}(M_0^\star)$ generated by equation (25).

For $p_0 \in \mathcal{D}(M_0^\star)$, Pp_0 denotes the function $Pp_0(t,x) = (P_t p_0)(x)$ defined on $\mathbb{R}_+ \times \mathbb{R}^n$. The set of solutions of equation (25) is $\mathcal{U}_P \subset C^\infty(\mathbb{R}_+ \times \mathbb{R}^n)$, consisting of functions of the form Pp_0 for $p_0 \in \mathcal{D}(M_0^\star)$.

We now introduce groups and semigroups (that is, only stable by composition of diffeomorphisms, not by inversion) of F as follows.

DEFINITION 4.2. *Let $G(F)$ denote the group of (global) diffeomorphisms of $F = \mathbb{R} \times \mathbb{R}^n \times \mathbb{R}$, and let $A(F)$ be the subgroup of diffeomorphisms g of the form*

$$\begin{array}{rccc} g: & \mathbb{R} \times \mathbb{R}^n \times \mathbb{R} & \to & \mathbb{R} \times \mathbb{R}^n \times \mathbb{R}, \\ & (s,x,u) & \mapsto & (\alpha(s), \beta(s,x), \delta(s,x)u)\,. \end{array} \tag{26}$$

Let $SA(F) \subset A(F)$ be the semigroup defined by

$$SA(F) = \{g \in A(F) \mid g^{-1}(\mathbb{R}_+ \times \mathbb{R}^n \times (0,+\infty)) \subset \mathbb{R}_+ \times \mathbb{R}^n \times (0,+\infty)\}\,. \tag{27}$$

If g is given by (26)*, this can be rewritten as*

$$g \in SA(F) \iff g \in A(F) \quad \textit{and} \quad \alpha^{-1}(\mathbb{R}_+) \subset \mathbb{R}_+ \quad \textit{and} \quad \delta > 0\,. \tag{28}$$

We recall that a natural action of $SA(F)$ on $C^\infty(\mathbb{R}_+ \times \mathbb{R}^n)$ is given as in [16, 17]. For any function h, graph(h) will denote its graph.

PROPOSITION 4.3. *For any $u \in C^\infty(\mathbb{R}_+ \times \mathbb{R}^n)$ and $g \in SA(F)$, there exists a unique function $g{\cdot}u \in C^\infty(\mathbb{R}_+ \times \mathbb{R}^n)$ such that $g\,(graph(u)) \supset graph(g{\cdot}u)$. Moreover, if g_1 and g_2 belong to $SA(F)$, we have*

$$(g_1 \circ g_2){\cdot}u = g_1{\cdot}(g_2{\cdot}u). \tag{29}$$

If g is given by (26) *and m defined by $m(s,x) = (\alpha(s), \beta(s,x))$, we have*

$$g{\cdot}u\,(s,x) = (\delta \times u)(m^{-1}(s,x)) \quad \forall\ (s,x) \in \mathbb{R}_+ \times \mathbb{R}^n\,. \tag{30}$$

The symmetry semigroup $SA_P(F)$ *of $\mathcal{U}_P$ is defined by*

$$SA_P(F) = \{\, g \in SA(F) \mid \forall\, u \in \mathcal{U}_P\,,\ g{\cdot}u \in \mathcal{U}_P \,\}\,. \tag{31}$$

Example. For $r \in \mathbb{R}_+$, $\Phi_{-r}^{\frac{\partial}{\partial s}} \in SA(F)$ and we have

$$\Phi_{-r}^{\frac{\partial}{\partial s}}{\cdot}u\,(s,x) = u(s+r,x). \tag{32}$$

What is more, $\Phi_{-r}^{\frac{\partial}{\partial s}} \in SA_P(F)$ since, if u satisfies (25) with initial data $u(0,\cdot) = p_0$, then $\Phi_{-r}^{\frac{\partial}{\partial s}}{\cdot}u$ satisfies (25) with initial data $u(r,\cdot) = P_r p_0$.

Remark. In this formal definition of a symmetry semigroup, we are interested in *global* transformations of the state-space. For the *local* definition and results, we send the reader to [8]. (Some results are recalled in the appendix.)

It is easy to see that $SA_P(F)$ is a semigroup by equation (29). Moreover, it is noticed in [20] that since any function in $\mathcal{U}_P$ is characterized by its initial data $u(0,\cdot) \in \mathcal{D}(M_0^\star)$, then $SA_P(F)$ induces an action on $\mathcal{D}(M_0^\star)$ as follows.

PROPOSITION 4.4. *For any $p_0 \in \mathcal{D}(M_0^\star)$ and $g \in SA(F)$, let us define*

$$\widehat{g}{\cdot}p_0 = (g{\cdot}Pp_0)(0,\cdot) \in \mathcal{D}(M_0^\star). \tag{33}$$

If $g \in SA_P(F)$, then $\widehat{g}$ is a nonnegative linear endomorphism of $\mathcal{D}(M_0^\star)$. Moreover, if g_1 and g_2 belong to $SA_P(F)$, we have for all $p_0 \in \mathcal{D}(M_0^\star)$ the following homomorphism property:

$$\widehat{g_1 \circ g_2}{\cdot}p_0 = \widehat{g}_1{\cdot}(\widehat{g}_2{\cdot}p_0)\,. \tag{34}$$

Proof. If $g \in SA_P(F)$ and $p_0 \in \mathcal{D}(M_0^\star)$, then $g{\cdot}Pp_0$ belongs to $\mathcal{U}_P$; that is, $g{\cdot}Pp_0 = P\tilde{p}_0$ for some $\tilde{p}_0 \in \mathcal{D}(M_0^\star)$. But $\tilde{p}_0(x) = P\tilde{p}_0(0,x) = g{\cdot}Pp_0(0,x) = (\widehat{g}{\cdot}p_0)\,(x)$, so that $\widehat{g}{\cdot}p_0$ belongs to $\mathcal{D}(M_0^\star)$ and thus

$$\forall\, p_0 \in \mathcal{D}(M_0^\star)\,, \quad \forall\, g \in SA_P(F) \quad g{\cdot}(Pp_0) = P(\widehat{g}{\cdot}p_0)\,. \tag{35}$$

By (33), it is clear that $\widehat{g}{\cdot}p_0$ is linear in p_0, since each P_t is a linear operator, and nonnegative as soon as p_0 is nonnegative, since each P_t is a nonnegative linear operator (see Definition 4.1).

For g_1, $g_2 \in SA_P(F)$, we have by equation (29)

$$\widehat{g_1 \circ g_2}{\cdot}p_0 = ((g_1 \circ g_2){\cdot}Pp_0)\,(0,\cdot) = (g_1{\cdot}(g_2{\cdot}Pp_0))\,(0,\cdot)\,,$$

where $g_2 \cdot Pp_0 = P(\widehat{g_2} \cdot p_0)$ by (35), so that

$$\widehat{g_1 \circ g_2} \cdot p_0 = (g_1 \cdot P(\widehat{g_2} \cdot p_0))\,(0, \cdot) = \widehat{g_1} \cdot (\widehat{g_2} \cdot p_0). \qquad \square$$

Example. By the example following Proposition 4.3 and by (32), we have for $r \in \mathbb{R}_+$:

$$\Phi_{-r}^{\frac{\partial}{\partial s}} \in SA_P(F) \quad \text{and} \quad \forall\, p_0 \in \mathcal{D}(M_0^\star)\,, \widehat{\Phi_{-r}^{\frac{\partial}{\partial s}}} \cdot p_0 = P_r p_0. \tag{36}$$

Now we turn to infinitesimal aspects. Since we are interested in *global diffeomorphisms*, not all infinitesimal generators are to be considered but only those which are *complete* vector fields.

Let $\mathcal{X}(F)$ be the Lie algebra of smooth vector fields on $F = \mathbb{R} \times \mathbb{R}^n \times \mathbb{R}$ and $\mathcal{X}^c(F)$ be the set of complete smooth vector fields on F. We recall that a complete vector field is a vector field whose flow can be defined for all time (no explosion in finite time): if $Z \in \mathcal{X}^c(F)$, its global flow on F will be denoted $(\Phi_r^Z)_{r \in \mathbb{R}}$. The set $\mathcal{X}^c(F)$ is stable neither by addition, nor by Lie bracketing.

DEFINITION 4.5. *$\mathcal{A}^c(F) \subset \mathcal{X}^c(F)$ is the set of* complete *smooth vector fields Z on F of the form*

$$Z = \zeta^0(s)\frac{\partial}{\partial s} + \sum_{i=1}^{n} \zeta^i(s,x)\frac{\partial}{\partial x_i} + \zeta^{n+1}(s,x)u\frac{\partial}{\partial u}. \tag{37}$$

The vector field Z in $\mathcal{A}^c(F)$ of global flow $(\Phi_r^Z)_{r \in \mathbb{R}}$ is said to be a complete infinitesimal symmetry of $\mathcal{U}_P$ if Φ_r^Z belongs to $SA_P(F)$ for all $r \in \mathbb{R}$. We denote by $\mathcal{SA}_P^c(F)$ the set of complete infinitesimal symmetries of $\mathcal{U}_P$.

It should be noted that when $Z \in \mathcal{SA}_P^c(F)$, Φ_r^Z belongs to $SA_P(F)$ for all $r \in \mathbb{R}$ and not only for all $r \in \mathbb{R}_+$.

PROPOSITION 4.6. *Let Z be a complete infinitesimal symmetry of $\mathcal{U}_P$. If Z is given by* (37)*, then $\zeta^0(0) = 0$, and for any $p_0 \in \mathcal{D}(M_0^\star)$ and $x \in \mathbb{R}^n$ we have*

$$\frac{d}{dr}_{|r=0} \left(\widehat{\Phi_r^Z} \cdot p_0\right)(x) = -\sum_{i=1}^{n} \zeta^i(0,x)\frac{\partial p_0}{\partial x_i}(x) + \zeta^{n+1}(0,x)p_0(x) \stackrel{\text{def}}{=} \left(\widehat{Z}_0 p_0\right)(x). \tag{38}$$

Proof. If $\Phi_r^Z(s,x,u) = (\alpha_r(s), \beta_r(s,x), \delta_r(s,x)u)$, we have $\alpha_r^{-1}(\mathbb{R}_+) \subset \mathbb{R}_+$ and $\alpha_{-r}^{-1}(\mathbb{R}_+) \subset \mathbb{R}_+$ by (28) applied to Φ_r^Z and to Φ_{-r}^Z. Since $\alpha_{-r}^{-1} = \alpha_r$ by the group property, this implies that $\alpha_r(\mathbb{R}_+) = \mathbb{R}_+$. But α_r is a global diffeomorphism of $\mathbb{R}$, hence it is monotonous and therefore it is a strictly increasing function. Then, the relation $\alpha_r(\mathbb{R}_+) = \mathbb{R}_+$ implies that 0 is a fixed point of α_r, so that $\zeta^0(0) = \frac{d}{dr}_{|r=0}\alpha_r(0) = 0$. Now equation (38) comes from the classic computation (see [16, 17, 20])

$$\begin{aligned}
\left(\widehat{\Phi_r^Z} \cdot p_0\right)(x) &= \left(\Phi_r^Z \cdot p_0\right)(0,x) \\
&= (\delta_r \times Pp_0)\left(\alpha_r^{-1}(0), \beta_r(\alpha_r^{-1}(0), \cdot)^{-1}(x)\right) \quad \text{by (30)} \\
&= (\delta_r \times Pp_0)\,(0, \beta_{-r}(0,x)) \quad \text{since} \quad \alpha_{-r}(0) = 0 \\
&= \delta_r\,(0, \beta_{-r}(0,x)) \times p_0\,(\beta_{-r}(0,x)) \\
&= \left(1 + r\zeta^{n+1}(0,x) + o(r)\right) \\
&\quad \times p_0\left(x_1 - r\zeta^1(0,x) + o(r), \ldots, x_n - r\zeta^n(0,x) + o(r)\right) \\
&= p_0(x) + r\left(\zeta^{n+1}(0,x)p_0(x) - \sum_{i=1}^{n} \zeta^i(0,x) \cdot \frac{\partial p_0}{\partial x_i}(x)\right) + o(r). \qquad \square
\end{aligned}$$

5. Application to FDR. Unlike Part I, we are concerned here with FDRs of *mappings having values in a set of smooth functions* (densities) rather than measures (laws). We keep the same notations, however.

A general means of constructing FDRs is given by the following proposition (which will be further simplified in section 7).

PROPOSITION 5.1. *Assume that*

1. *there exist p complete infinitesimal symmetries* $Z^1, \ldots, Z^p$ *of* $\mathcal{U}_P$ *such that the first-order differential operators associated by* (38) *satisfy*

$$\forall\, p_0 \in \mathcal{D}(M_0^\star)\,, \quad \widehat{Z^1}_0 p_0 = M_1^\star p_0\,, \ldots,\, \widehat{Z^p}_0 p_0 = M_p^\star p_0; \tag{39}$$

2. *the subalgebra of* $\mathcal{X}(F)$ *generated by the vector fields* $Z^1, \ldots, Z^p$ *and* $\frac{\partial}{\partial s}$ *is finite dimensional.*

Then there exists an FDR, uniform with respect to the dynamics and initial condition, of the family of input-output maps $\{\mathcal{F}^{p_0} \mid p_0 \in \mathcal{D}(M_0^\star) \cap \Pi\}$.

Example. For the one-dimensional linear system ($dx_t = dv_t$, $dy_t = x_t + dw_t$) we have

$$M_0 = \frac{1}{2}\frac{d^2}{dx^2} - \frac{1}{2}x^2 (= M_0^\star) \quad \text{and} \quad M_1 = x (= M_1^\star).$$

If $Z = -s\frac{\partial}{\partial x} + xu\frac{\partial}{\partial u}$ we have

$$\Phi_r^Z(s,x,u) = \left(s, x - sr, u\exp(xr - sr^2/2)\right),$$

$$\Phi_r^Z \cdot u(s,x) = \exp(xr + sr^2/2) u(s, x + sr).$$

1. If u is solution of $\partial_t u - M_0^\star u = 0$, an easy computation shows that for all r, $v = \Phi_r^Z \cdot u$ is a solution of $\partial_t v - M_0^\star v = 0$; thus, Z is a complete infinitesimal symmetry of the PDE, $\partial_t u - M_0^\star u = 0$.

By (38), we have $\widehat{Z}_0 p_0 = x p_0 = M_1^\star p_0$.

2. Z and $\frac{\partial}{\partial s}$ generate a four-dimensional Lie algebra with basis: Z, $\frac{\partial}{\partial s}$, $\frac{\partial}{\partial x}$, $u\frac{\partial}{\partial u}$.

Proof of Proposition 5.1. First, let us define the manifold M of the FDR.

The manifold M is the subgroup of $G(F)$ generated by the global flows $(\Phi_r^{Z^1})_{r\in\mathbb{R}}$, $\ldots$, $(\Phi_r^{Z^p})_{r\in\mathbb{R}}$, and $(\Phi_r^{\frac{\partial}{\partial s}})_{r\in\mathbb{R}}$. By [18, pp. 99–105, Theorem VII], it can indeed be equipped with a unique manifold structure that makes it a Lie group since the vector fields $Z^1, \ldots, Z^p$ and $\frac{\partial}{\partial s}$ of $\mathcal{X}(F)$ are complete (by Assumption 1) and generate a finite-dimensional Lie algebra (by Assumption 2).

We denote by $\mathfrak{M}$ the finite-dimensional subalgebra of vector fields on F generated by the vector fields $Z^1, \ldots, Z^p$ and $\frac{\partial}{\partial s}$. The Lie group M is a Lie transformation group (in the terminology of [18, p. 99, Definition IV]), and thus there exists an isomomorphism $Z \mapsto Z^+$ from $\mathfrak{M}$ to the Lie algebra of right-invariant vector fields on the Lie group M: this isomomorphism is such that if $Z \in \mathfrak{M}$, then $\Phi_r^Z \in G(F)$ coincides with the exponential $\exp(rZ^+) \in M$. In particular, the following mapping is smooth:

$$r \in \mathbb{R} \mapsto \Phi_r^Z = \exp(rZ^+) \in M\,. \tag{40}$$

Second, let us define the output functions $(\theta^{p_0}, \mathrm{Dom}(\theta^{p_0}))$ of the FDR. We take

$$\mathrm{Dom}(\theta^{p_0}) = M \cap SA(F);$$

that is, all $g \in M$ such that $g^{-1}(\mathbb{R}_+ \times \mathbb{R}^n \times (0,+\infty)) \subset \mathbb{R}_+ \times \mathbb{R}^n \times (0,+\infty)$ (see Definition 4.2). The map θ^{p_0} is defined from $\mathcal{D}(M_0^\star)$ to $\mathcal{D}(M_0^\star)$ as follows (see (33)):

$$\forall\, p_0 \in \mathcal{D}(M_0^\star)\,, \quad \forall\, g \in \mathrm{Dom}(\theta^{p_0})\,, \quad \theta^{p_0}(g) = \widehat{g}\cdot p_0\,. \tag{41}$$

This map is smooth because by (33) and (30) we have, if g is given by (26),

$$(\widehat{g}\cdot p_0)(x) = (g\cdot Pp_0)(0,x) = (\delta \times Pp_0)(\alpha^{-1}(0), \beta(\alpha^{-1}(0),\cdot)^{-1}(x)),$$

where Pp_0 is smooth by assumption and where the following mapping is smooth by [18, p. 99, Definition IV]:

$$\begin{array}{rcl} M \times \mathbb{R} \times \mathbb{R}^n \times \mathbb{R} & \to & \mathbb{R} \times \mathbb{R}^n \times \mathbb{R}, \\ (g,s,x,u) & \mapsto & g(s,x,u) = (\alpha(s), \beta(s,x), \delta(s,x)u). \end{array} \tag{42}$$

Third, let us exhibit a conservative stochastic differential equation on M having values in $\mathrm{Dom}(\theta)$.

Let the vector fields b_0, b_1, …, b_p on M be given by

$$b_0 = \frac{\partial}{\partial s}^+, b_1 = Z^{1+}, \ldots, b_p = Z^{p+}\,.$$

The SDE

$$dg_t = \frac{\partial}{\partial s}^+ \cdot g_t\, dt + \sum_{i=1}^p Z^{i+}\cdot g_t \circ dy_t^i\,, \quad g_0 = \mathrm{Id}_F, \tag{43}$$

is conservative on M, since b_0, b_1, …, b_p are right-invariant vector fields on the Lie group M by [22, Lemma 2.2].

We now prove that the solution g_t belongs to $\mathrm{Dom}(\theta)$, that is, belongs to $SA(F)$. For this, we write $g_t(s,x,u) = (\alpha_t(s), \beta_t(s,x), \delta_t(s,x)u)$ and, by (28), we just have to show that $\alpha_t^{-1}(\mathbb{R}_+) \subset \mathbb{R}_+$ and that $\delta > 0$. It is clear that $\delta_t(s,x) > 0$ by continuity since $\delta_0(s,x) = 1$ and since $\delta_t(s,x) \neq 0$ (g_t belongs to a group, hence is invertible). Thus, we must study $\alpha_t(s)$.

Since the mapping (42) is smooth, so is the map $f_z(g) = g(z)$ for $z \in F$ and $g \in M$. By (40), we have for any $Z \in \mathfrak{M}$:

$$(Z^+ f_z)(g) = \frac{d}{dr}_{|r=0} f_z(\exp(rZ^+)\circ g) = \frac{d}{dr}_{|r=0} \Phi_t^Z(g(z)) = Z(g(z))\,.$$

Therefore, for any $z = (s,x,u) \in F$, we can apply the Itô–Stratonovitch formula to $f_z(g_t)$ to get w.p.1 for all $t \geq 0$:

$$\begin{aligned} dg_t(s,x,u) &= \left(\frac{\partial}{\partial s}^+ f_z\right)(g_t)\, dt + \sum_{i=1}^p (Z^{i+} f_z)(g_t) \circ dy_t^i \\ &= \frac{\partial}{\partial s}(g_t(s,x,u))\, dt + \sum_{i=1}^p Z^i(g_t(s,x,u)) \circ dy_t^i\,. \end{aligned}$$

Denoting, for $i = 1, \ldots, p$, $Z^j = \zeta_j^0(s)\frac{\partial}{\partial s} + \sum_{i=1}^n \zeta_j^i(s,x)\frac{\partial}{\partial x_i} + \zeta_j^{n+1}(s,x)u\frac{\partial}{\partial u}$, then $\alpha_t(s)$ satisfies

$$\begin{aligned} d\alpha_t(s) &= dt + \sum_{i=1}^p \zeta_i^0(\alpha_t(s)) \circ dy_t^i \\ &= dt + \frac{1}{2}\sum_{i=1}^p ((\zeta_i^0)'\zeta_i^0)(\alpha_t(s))dt + \sum_{i=1}^p \zeta_i^0(\alpha_t(s))\, dy_t^i\,, \quad \alpha_0(s) = s\,. \end{aligned}$$

By the diffeomorphism theorem [19, p. 139], the (Itô) SDE

$$d\alpha_t = dt + \frac{1}{2}\sum_{i=1}^{p}((\zeta_i^0)'\zeta_i^0)(\alpha_t)dt + \sum_{i=1}^{p}\zeta_i^0(\alpha_t)\,dy_t^i$$

generates a stochastic flow of diffeomorphisms and thus, in particular, we have $\alpha_t(\mathbb{R}) = \mathbb{R}$. On the other hand, we know from Proposition 4.6 that $\zeta_j^0(0) = 0$: therefore, when $\alpha_0(s) = s \geq 0$ (resp. $\alpha_0(s) = s \leq 0$), $\alpha_t(s)$ cannot escape from $\mathbb{R}_+$ (resp. $\mathbb{R}_-$) [11, p. 149]. Thus $\alpha_t(\mathbb{R}_+) \subset \mathbb{R}_+$ and $\alpha_t(\mathbb{R}_-) \subset \mathbb{R}_-$, so that $\alpha_t^{-1}(\mathbb{R}_+) \subset \mathbb{R}_+$.

Fourth, we prove that (23) is satisfied. For this, we start by showing that

$$\mathrm{Dom}(\theta^{p_0}) = M \cap SA(F) \subset SA_P(F),$$

which will allow us to use the homomorphism property (34). By Assumption 1 and the Definition 4.5 of complete infinitesimal symmetries of $\mathcal{U}_P$, the flows $(\Phi_r^{Z^1})_{r\in\mathbb{R}}$, …, $(\Phi_r^{Z^p})_{r\in\mathbb{R}}$ all belong to $SA_P(F)$. On the other hand, the whole flow $(\Phi_r^{\frac{\partial}{\partial s}})_{r\in\mathbb{R}}$ cannot belong to $SA_P(F)$ because for $r \in \mathbb{R}_+$, $\Phi_r^{\frac{\partial}{\partial s}} \notin SA(F)$. However, for $r \in \mathbb{R}_+$ we know by (36) that $\Phi_{-r}^{\frac{\partial}{\partial s}} \in SA_P(F)$. Therefore, $M \cap SA(F) \subset SA_P(F)$.

Now, we apply the Itô–Stratonovitch formula to $\theta^{p_0}(g_t)(x)$ (for notation reasons, we shall forget the term x in what follows):

$$\text{w.p.1} \quad \forall\, t \in \mathbb{R}_+\,, \quad d\theta^{p_0}(g_t) = \left(\frac{\partial}{\partial s}\cdot\theta^{p_0}\right)(g_t)\,dt + \sum_{i=1}^{p}\left(Z^i\cdot\theta^{p_0}\right)(g_t)\circ dy_t^i\,.$$

For $i = 1, \ldots, p$ we have by (34)

$$\theta^{p_0}(\Phi_r^{Z^i}(g)) = \theta^{p_0}(\Phi_r^{Z^i}\circ g) = \widehat{\Phi_r^{Z^i}\circ g}\cdot p_0 = \widehat{\Phi_r^{Z^i}}\cdot(\widehat{g}\cdot p_0) = \widehat{\Phi_r^{Z^i}}\cdot(\theta^{p_0}(g))$$

so that

$$\left(Z^i\cdot\theta^{p_0}\right)(g) = \frac{d}{dr}_{|r=0}\widehat{\Phi_r^{Z^i}}\cdot(\theta^{p_0}(g)) = \widehat{Z^i}_0(\theta^{p_0}(g)) = M_i^\star\theta^{p_0}(g)\,.$$

In the same way, (36) provides

$$\left(\frac{\partial}{\partial s}\cdot\theta^{p_0}\right)(g) = \frac{d}{dr}_{|r=0+}\widehat{\Phi_{-r}^{\frac{\partial}{\partial s}}}\cdot(\theta^{p_0}(g)) = \frac{d}{dr}_{|r=0+}P_r(\theta^{p_0}(g)) = M_0^\star\theta^{p_0}(g)$$

and therefore

$$\text{w.p.1} \quad \forall\, t \in \mathbb{R}_+\,, \quad d\theta^{p_0}(g_t) = M_0^\star\theta^{p_0}(g_t)dt + \sum_{i=1}^{p}M_i^\star\theta^{p_0}(g_t)\circ dy_t^i\,.$$

We conclude by uniqueness of smooth solutions of equation (16) for $p_0 \in \Pi$. □

The following theorem is the main result of Part II. (It will be compared with the main theorem of Part I.)

THEOREM 5.2. *Assume that*

1. *the estimation algebra is finite dimensional and can be written*

$$\mathcal{E} = \mathbb{R}M_0 + \mathcal{Q}, \tag{44}$$

where $\mathcal{Q}$ is a (finite-dimensional) subalgebra of differential operators on $\mathbb{R}^n$ of order less than or equal to one;

2. *the vector fields* $Z^1, \ldots, Z^p$ *on* $\mathbb{R} \times \mathbb{R}^n \times \mathbb{R}$ *defined hereafter*

$$Z^j = \zeta_j^0(s)\frac{\partial}{\partial s} + \sum_{i=1}^{n} \zeta_j^i(s,x)\frac{\partial}{\partial x_i} + \zeta_j^{n+1}(s,x)u\frac{\partial}{\partial u} \tag{45}$$

are complete. For each $j = 1, \ldots, p$ *the coefficients* $\zeta_j^0(s)$, $\zeta_j^1(s,x)$, $\ldots$, $\zeta_j^{n+1}(s,x)$ *are given by*

$$\exp(sad_{M_0})(M_j) = -\zeta_j^0(s)M_0 - \sum_{i=1}^{n}\zeta_j^i(s,x)\frac{\partial}{\partial x_i} + \zeta_j^{n+1}(s,x). \tag{46}$$

Then, there exists an FDR of $\{\mathcal{F}^{p_0} \mid p_0 \in \mathcal{D}(M_0^\star) \cap \Pi\}$*, uniform with respect to the dynamics and initial condition.*

Proof. It suffices to prove that the assumptions of Proposition 5.1 are satisfied. To begin with, note that the assumptions of the theorem may be rewritten as follows by taking the dual expressions of all equalities.

1. The estimation algebra is finite dimensional, and we have

$$\mathcal{E}^\star = \mathbb{R}M_0^\star + \mathcal{R}, \tag{47}$$

where $\mathcal{R}(= \mathcal{Q}^\star)$ is a (finite-dimensional) subalgebra of differential operators on $\mathbb{R}^n$ of order less than or equal to one.

2. The vector fields $T^1, \ldots, T^p$ on $\mathbb{R} \times \mathbb{R}^n \times \mathbb{R}$ defined hereafter

$$T^j = \zeta_j^0(s)\frac{\partial}{\partial s} - \sum_{i=1}^{n}\zeta_j^i(s,x)\frac{\partial}{\partial x_i} + \left(\zeta_j^{n+1}(s,x) + \sum_{i=1}^{n}\frac{\partial \zeta_j^i}{\partial x_i}(s,x)\right)u\frac{\partial}{\partial u} \tag{48}$$

are complete. For each $j = 1, \ldots, p$ the coefficients $\zeta_j^0(s)$, $\zeta_j^1(s,x)$, $\ldots$, $\zeta_j^{n+1}(s,x)$ are given by

$$\exp(s\mathrm{ad}_{M_0^\star})(M_j^\star) = -\zeta_j^0(s)M_0^\star + \sum_{i=1}^{n}\zeta_j^i(s,x)\frac{\partial}{\partial x_i} + \sum_{i=1}^{n}\frac{\partial \zeta_j^i}{\partial x_i}(s,x) + \zeta_j^{n+1}(s,x). \tag{49}$$

We shall point out why the vector fields $T^1, \ldots, T^p$ are complete. The equations of the flow of T^j are

$$\left\{\begin{array}{rcl} \dot{s} &=& \zeta_j^0(s), \\ \dot{x} &=& -\displaystyle\sum_{i=1}^{n}\zeta_j^i(s,x)\frac{\partial}{\partial x_i}, \\ \dot{u} &=& \left(\zeta_j^{n+1}(s,x) + \sum_{i=1}^{n}\frac{\partial \zeta_j^i}{\partial x_i}(s,x)\right)u, \end{array}\right.$$

and they generate a solution defined for all time, since this is the case for the following equations of the flow of Z^j (s is not changed, just invert time for x, u satisfies a linear equation):

$$\left\{\begin{array}{rcl} \dot{s} &=& \zeta_j^0(s), \\ \dot{x} &=& \displaystyle\sum_{i=1}^{n}\zeta_j^i(s,x)\frac{\partial}{\partial x_i}, \\ \dot{u} &=& \zeta_j^{n+1}(s,x)u. \end{array}\right.$$

For the proof, we then proceed as follows. We shall make use of some notions recalled in Appendix B, particularly the notion of *perturbation algebra* $\mathcal{P}_{M_0^\star}$ of $M_0^\star$ presented in Proposition B.1 (where A has to be replaced by $M_0^\star$).

1. We prove that the vector fields $T^1, \ldots, T^p$ are complete infinitesimal symmetries of $\mathcal{U}_P$, such that the first-order differential operators associated by (38) satisfy

$$\forall\, p_0 \in \mathcal{D}(M_0^\star)\,, \quad \widehat{T^1}_0 p_0 = M_1^\star p_0\,, \ldots, \widehat{T^p}_0 p_0 = M_p^\star p_0. \tag{50}$$

By (49) evaluated at $s = 0$, we have

$$\begin{aligned} M_j^\star &= \exp(0 \times \mathrm{ad}_{M_0^\star})(M_j^\star) \\ &= -\zeta_j^0(0) M_0^\star + \sum_{i=1}^n \zeta_j^i(0,x)\frac{\partial}{\partial x_i} + \sum_{i=1}^n \frac{\partial \zeta_j^i}{\partial x_i}(0,x) + \zeta_j^{n+1}(0,x) \\ &= \widehat{T^j}_0 \quad \text{by (38) and (48).} \end{aligned}$$

Now, we show that the vector fields $T^1, \ldots, T^p$ are complete infinitesimal symmetries of $\mathcal{U}_P$.

By Assumption 1 for each $i = 1, \ldots, p$ the Lie algebra generated by $M_0^\star$ and $M_i^\star$ is finite dimensional. Thus, by the characterization of the perturbation algebra in Proposition B.1 in the appendix, we see that $M_1^\star, \ldots, M_p^\star$ belong to $\mathcal{P}_{M_0^\star}$. Thus, by the above equalities we have

$$\widehat{T^1}_0 = M_1^\star \in \mathcal{P}_{M_0^\star}, \ldots, \widehat{T^p}_0 = M_p^\star \in \mathcal{P}_{M_0^\star}.$$

On the other hand, the equalities (48) and (49) may be rewritten, with the notation of (B.86),

$$\widehat{T^1}_t = \exp(t\mathrm{ad}_{M_0^\star})(M_1^\star), \ldots, \widehat{T^p}_t = \exp(t\mathrm{ad}_{M_0^\star})(M_p^\star).$$

These two latter equations precisely mean, by Proposition B.1 and especially (B.87), that the vector fields $T^1, \ldots, T^p$ are (local) infinitesimal symmetries of $\mathcal{U}_P$. Since, by completeness, $\Phi_r^{T^i}$ is a global diffeomorphism of $\mathbb{R}^n$ for all $r \in \mathbb{R}$, this implies that for all $u \in \mathcal{U}_P$, $\Phi_r^{T^i} \cdot u \in \mathcal{U}_P$. This means that $T^1, \ldots, T^p$ belong to $\mathcal{SA}_P^c(F)$. The first assumption of Proposition 5.1 is thus satisfied.

2. It remains to prove the second assumption of Proposition 5.1, namely that the subalgebra of $\mathcal{X}(F)$ generated by the vector fields $T^1, \ldots, T^p$ and $\frac{\partial}{\partial s}$ is finite dimensional. But $\frac{\partial}{\partial s}$ is an infinitesimal symmetry of $\mathcal{U}_P$ since $\partial_t - M_0^\star$ is invariant by time translations and the result then follows from Theorem 2.1 in [8]. (This result is recalled in Proposition B.1 in the appendix, where A has to be replaced by $M_0^\star$ and states that (nontrivial) infinitesimal symmetries form a finite-dimensional Lie algebra.)

The assumptions of Proposition 5.1 are thus satisfied and this completes the proof. □

6. Comparison between Part I and Part II main theorems. *Under the common assumption that M_0 is nondegenerate elliptic, Theorem* 5.2 *is stronger than the main theorem of Part* I *recalled here below.*

THEOREM 6.1. *Assume that*

1. *$\mathcal{E}$ is finite dimensional and given by*

$$\mathcal{E} = \mathbb{R}M_0 + \mathcal{R},$$

where $\mathcal{R}$ is a finite-dimensional subalgebra of differential operators of order less than or equal to one;

2. *$\mathcal{E}$ is a solvable Lie algebra, with a basis $\{F_0,F_1,F_2,\ldots,F_q\}$ such that $F_0 = M_0$ and $\mathcal{F}_i = \mathbb{R} - \mathrm{span}\{F_i,\ldots,F_q\}$ is a Lie ideal of $\mathcal{F}_{i-1}$ for $i = 0,\ldots,q$,*

3. *the first-order part of each operator F_1, ..., F_q defines a complete vector field on $\mathbb{R}^n$,*

4. *for all $\phi \in \mathcal{D}(\mathbb{R}^n)$, there exists a unique solution $u \in C^\infty(]0,+\infty[\times\mathbb{R}^n) \cap C^0([0,+\infty[\times\mathbb{R}^n)$ of the PDE*

$$\frac{\partial u}{\partial t} = M_0\, u\,, \quad u(0,x) = \phi(x)\,. \tag{51}$$

Then the family of input-output maps $\{\mathcal{F}^{\mu_0} \mid \mu_0 \in \mathbb{D}_0\}$ given by (20) admits a regular FDR, uniform with respect to the dynamics and initial condition.

If we suppose that the assumptions 1, 2, 3, and 4 of this latter theorem are satisfied, then the assumptions of Theorem 5.2 are automatically satisfied:

1. the estimation algebra $\mathcal{E}$ can be written as in (44), with $\mathcal{Q} = \mathcal{R}^\star$ by 1,
2. we have by 2 the following expression:

$$\exp(s\mathrm{ad}_{M_0})(M_j) = \sum_{i=1}^{q} \alpha_j^i(s) F_i = \sum_{i=1}^{q} \alpha_j^i(s) X_i + \sum_{i=1}^{q} \alpha_j^i(s) c_i \tag{52}$$

so that the vector fields Z^1, ..., Z^p are given by

$$Z^j = -\sum_{i=1}^{q} \alpha_j^i(s) X_i + \left(\sum_{i=1}^{q} \alpha_j^i(s) c_i(x)\right) u \frac{\partial}{\partial u}\,. \tag{53}$$

The equations of the flow of Z^j are therefore

$$\left\{\begin{array}{rcl} \dot{s} &=& 0, \\ \dot{x} &=& -\displaystyle\sum_{i=1}^{q} \alpha_j^i(s) X_i(x), \\ \dot{u} &=& \left(\displaystyle\sum_{i=1}^{q} \alpha_j^i(s) c_i(x)\right) u, \end{array}\right. \tag{54}$$

so that if $s(0) = s_0$, Z^j is complete on $\mathbb{R}\times\mathbb{R}^n\times\mathbb{R}$ as soon as the vector field $X = \sum_{i=1}^q \alpha_j^i(s_0) X_i(x)$ is complete on $\mathbb{R}^n$. By a theorem of Palais [18, Theorem III, pp. 91–97], X appears to be complete as the sum of complete vector fields generating a finite-dimensional Lie algebra of vector fields.

7. A Riemannian geometric point of view. In this section, we take advantage of the Riemannian structure induced by the operator M_0.

Thanks to the assumption that M_0 is nondegenerate elliptic, it is well known that we can introduce a Riemannian metric g on $\mathbb{R}^n$ as follows (see [17, 13, 15, 9]).

LEMMA 7.1. *If $(a_{ij}(x))_{i,j=1,\ldots,n}$ denotes the inverse matrix of $(a^{ij}(x))_{i,j=1,\ldots,n}$ in (18), then*

$$g = \sum_{i,j=1}^{n} a_{ij}(x) dx_i dx_j \tag{55}$$

defines a Riemannian metric g on $\mathbb{R}^n$, that we shall note $g = met(M_0)$. Moreover, if Δ_g is the Laplace–Beltrami operator (Laplacian) on the Riemannian manifold $(\mathbb{R}^n, g)$, then we can write

$$M_0 = \frac{1}{2}\Delta_g + B + H, \tag{56}$$

where B is a smooth vector field on $\mathbb{R}^n$ (which depends not only on $b^1, \ldots, b^n$ in (18) *but also on a^{ij}, $i, j = 1, \ldots, n$).*

The first assumption of Theorem 5.2 may be replaced by the equivalent formulation that $M_1, \ldots, M_p$ belong to the *perturbation algebra* $\mathcal{P}_{M_0}$ of M_0 (see Appendix B, where A has to be replaced by M_0). This is interesting since geometric characterizations of $\mathcal{P}_{M_0}$ may be found in [9]. This is the point of view that we develop here.

We utilize the geometric objects recalled in Appendix A (see also [9]). We identify a smooth differential operator of order less than or equal to one with the sum of a smooth vector field and of a smooth function.

PROPOSITION 7.2. *Assume that $(\mathbb{R}^n, g)$ is complete. Let us write $M_1 = X^1 + m^1$, $\ldots$, $M_p = X^p + m^p$. If*

$$X_0^1 = X^1 \in \mathcal{H}_g(\mathbb{R}^n), \ldots, X_0^p = X^p \in \mathcal{H}_g(\mathbb{R}^n) \tag{57}$$

and if there exist p sequences $(X_i^1)_{i\geq 1}$, $\ldots$, $(X_i^p)_{i\geq 1}$ in $\mathcal{I}_g(\mathbb{R}^n)$ such that

1. *for $j = 1, \ldots, p$,*

$$\begin{aligned} X_1^j &= K_{M_0} X_0^j + \nabla_g (m^j - g(X_0^j, B)) \\ (or &= -\eta_g(X_0^j) B + [B, X_0^j] + \nabla_g m^j), \end{aligned} \tag{58}$$

2. *for $j = 1, \ldots, p$,*

$$\begin{aligned} X_2^j &= K_{M_0} X_1^j + \frac{1}{2} \nabla_g (L_{X_0^j} H_{M_0} + \eta_g(X_0^j) H_{M_0}) \\ (or &= [B, X_1^j] + \nabla_g (g(X_1^j, B) + \tfrac{1}{2} L_{X_0^j} H_{M_0} + \tfrac{1}{2} \eta_g(X_0^j) H_{M_0})), \end{aligned} \tag{59}$$

3. *for $j = 1, \ldots, p$ and $i \geq 1$,*

$$\begin{aligned} X_{i+2}^j &= K_{M_0} X_{i+1}^j + \frac{1}{2} \nabla_g (L_{X_i^j} H_{M_0}) \\ (or &= [B, X_{i+1}^j] + \nabla_g (g(X_{i+1}^j, B) + \tfrac{1}{2} L_{X_i^j} H_{M_0})). \end{aligned} \tag{60}$$

Then, there exists an FDR of $\{\mathcal{F}^{p_0} \mid p_0 \in \mathcal{D}(M_0^\star) \cap \Pi\}$, uniform with respect to the dynamics and initial condition.

Proof. By Proposition 3.2 in [8] (recalled in Theorem B.3 in the appendix, where A has to be replaced by M_0), the above assumptions imply that $M_1, \ldots, M_p$ belong to the perturbation algebra $\mathcal{P}_{M_0}$ of M_0. Thus, the first assumption of Theorem 5.2 is satisfied by Proposition 3.1 in [8] (recalled in Proposition B.1).

By (B.97), the sequence $(X_i^j)_{i\geq 0}$ is such that

$$\mathrm{ad}_{M_0}^{i+1}(M^j) = \eta_g(X_i^j) M_0 + X_i^j + \text{function}.$$

Since $X_i^j \in \mathcal{I}_g(\mathbb{R}^n)$ for $i \geq 1$ and thus $\eta_g(X_i^j) = 0$, this implies that

$$\begin{aligned} \exp(s\mathrm{ad}_{M_0})(M_j) &= M_j + \sum_{i=0}^{+\infty} \frac{s^{i+1}}{(i+1)!} \mathrm{ad}_{M_0}^{i+1}(M^j) \\ &= s\eta_g(X_0^j) M_0 + X_j + \sum_{i=1}^{+\infty} \frac{s^{i+1}}{(i+1)!} X_i^j + \text{function}, \end{aligned}$$

so that by comparing with (46), we see that necessarily $\zeta_j^0(s,x) = s\eta_g(X_0^j) = s\lambda$. The flow of the vector field Z^j is thus given by

$$\left\{\begin{array}{rcl} \dot{s} &=& \lambda s, \\ \dot{x}_1 &=& \zeta^1(s,x), \\ & \cdots & \\ \dot{x}_n &=& \zeta^n(s,x), \\ \dot{u} &=& \zeta^{n+1}(s,x)u. \end{array}\right.$$

Thus, if $s(0) = s_0$, the vector field Z^j on $\mathbb{R} \times \mathbb{R}^n \times \mathbb{R}$ is complete as soon as the time-varying vector field $X = \sum_{k=1}^n \zeta^k(e^{\lambda t}s_0, x)\frac{\partial}{\partial x_k}$ on $\mathbb{R} \times \mathbb{R}^n$ is complete. But for each fixed t, X belongs to $\mathcal{H}_g(\mathbb{R}^n)$ and the result follows from Lemma C.1 in the appendix. □

COROLLARY 7.3. *Assume that $(\mathbb{R}^n, g)$ is complete and that $M_1 = h_1, \ldots, M_p = h_p$. If*

$$X_1^1 = \nabla_g h_1 \in \mathcal{P}_g(\mathbb{R}^n), \ldots, X_1^p = \nabla_g h_p \in \mathcal{P}_g(\mathbb{R}^n) \tag{61}$$

and if there exists p sequences $(X_i^1)_{i\geq 2}, \ldots, (X_i^p)_{i\geq 2}$ in $\mathcal{I}_g(\mathbb{R}^n)$ such that

1. *for $j = 1, \ldots, p$,*

$$\begin{array}{rcl} X_2^j &=& K_{M_0} X_1^j \\ (or &=& [B, \nabla_g h_j] + \nabla_g(L_B h_j)\,), \end{array} \tag{62}$$

2. *for $j = 1, \ldots, p$ and $i \geq 1$,*

$$\begin{array}{rcl} X_{i+2}^j &=& K_{M_0} X_{i+1}^j + \frac{1}{2}\nabla_g(L_{X_i^j} H_{M_0}) \\ (or &=& [B, X_{i+1}^j] + \nabla_g(g(X_{i+1}^j, B) + \frac{1}{2} L_{X_i^j} H_{M_0}\,). \end{array} \tag{63}$$

Then, there exists an FDR of $\{\mathcal{F}^{p_0} \mid p_0 \in \mathcal{D}(M_0^\star) \cap \Pi\}$, uniform with respect to the dynamics and initial condition.

8. Application to filtering. For the relation between FDFs and FDRs, we follow Part I. By Proposition 7.2 and Corollary 7.3, we see that the construction of FDRs by the methods developed above imposes strong constraints on the geometry of the Riemannian space $(\mathbb{R}^n, g)$, particularly on the Lie algebras $\mathcal{P}_g(\mathbb{R}^n)$ of *parallel vector fields* and $\mathcal{H}_g(\mathbb{R}^n)$ of *infinitesimal homothetic transformations*. For the sake of simplicity, we shall assume in this section that the Riemannian space $(\mathbb{R}^n, g)$ is complete.

Remark. Saying that the Riemannian space $(\mathbb{R}^n, g)$ is complete amounts to saying that $\mathbb{R}^n$ equipped with the distance

$$d(x,y) = \inf\left\{\int_0^1 \sqrt{g(\dot{\gamma}(s), \dot{\gamma}(s))}ds \mid \gamma : [0,1] \to \mathbb{R}^n\,,\ \gamma(0) = x\,,\ \gamma(1) = y\right\} \tag{64}$$

is complete or that any bounded closed subset of $(\mathbb{R}^n, d)$ is compact [14]. This latter property is satisfied as soon as there exists $k > 0$ such that

$$\forall\, z \in \mathbb{R}^n\,,\ \forall\, x \in \mathbb{R}^n\,, \quad \sum_{i,j=1}^n a_{ij}(x) z_i z_j \geq k\|z\|^2\,. \tag{65}$$

Indeed, for any path γ such that $\gamma(0) = x$ and $\gamma(1) = y$, we have

$$k\|x-y\| \leq \int_0^1 \|\dot{\gamma}(s)\| ds \leq \int_0^1 \sqrt{\sum_{i,j=1}^n a_{ij}(\gamma(s))\dot{\gamma}_i(s)\dot{\gamma}_j(s)} ds$$

so that, by taking the infimum on all such paths, we get

$$k\|x-y\| \leq d(x,y)\,. \tag{66}$$

Let K be a bounded closed subset of $(\mathbb{R}^n, d)$. The topologies of $(\mathbb{R}^n, d)$ and $(\mathbb{R}^n, \| \|)$ being the same [14], K also is a closed subset of $(\mathbb{R}^n, \| \|)$. By (66), K is a bounded closed subset of $(\mathbb{R}^n, \| \|)$ and thus a compact of $(\mathbb{R}^n, \| \|)$. It is also a compact of $(\mathbb{R}^n, d)$ since the topologies are the same.

It may be noted that the inequality (65) is "opposite" to the one expressing the uniform ellipticity of M_0 since (65) may be written as

$$\forall\, z \in \mathbb{R}^n\,,\ \forall\, x \in \mathbb{R}^n\,, \quad \sum_{i,j=1}^n a^{ij}(x) z_i z_j \leq k\|z\|^2\,. \tag{67}$$

8.1. The case without correlated noises. In this case, the operators M_1, $\ldots$, M_p are functions $h_1, \ldots, h_p$. By Corollary 7.3, the construction of FDRs by the methods developed above necessitates that their gradients should belong to the Lie algebra $\mathcal{P}_g(\mathbb{R}^n)$ of parallel vector fields of $(\mathbb{R}^n, g)$. Necessarily the Lie algebra $\mathcal{P}_g(\mathbb{R}^n)$ of parallel vector fields of $(\mathbb{R}^n, g)$ is not reduced to zero, and this is a strong condition as can be seen from Proposition B.4.

8.1.1. Flat metrics on $\mathbb{R}^n$ and gradient drifts. This case is particularly studied in [23, 10]. We shall omit the reference to the metrics g and note $\nabla_g = \nabla$ and $\Delta_g = \Delta$. Here, we have

$$\mathcal{P}_g(\mathbb{R}^n) = \mathbb{R} - \text{span}\{\nabla x_1, \ldots, \nabla x_n\},$$

that is, all the "constant" vector fields on $\mathbb{R}^n$. By Definition B.2, we have

$$\left\{ \begin{array}{rcl} M_0 & = & \frac{1}{2}\Delta + \nabla\varphi - \frac{1}{2}\sum_{i=1}^p h_i^2, \\ K_{M_0} & = & 0, \\ H_{M_0} & = & \Delta\varphi + \|\nabla\varphi\|^2 + \sum_{i=1}^p h_i^2. \end{array} \right.$$

The following class of systems having FDFs is already known [3, 23, 10].

PROPOSITION 8.1. *If $\Delta\varphi + \|\nabla\varphi\|^2$ is quadratic, then there exists an FDF to compute the unnormalized conditional density associated with a system of the form*

$$\left\{ \begin{array}{rcl} dx_t & = & \nabla\varphi(x_t)dt + dv_t\,, \quad x_0 \rightsquigarrow p_0(x)dx, \\ dy_t & = & (Cx_t + D)dt + dw_t \end{array} \right. \tag{68}$$

when the density p_0 of x_0 belongs to $\mathcal{D}(M_0^\star)$.

Proof. This is a straightforward application of Corollary 7.3 with

$$X_2^j = 0 \quad \text{and} \quad X_{i+2}^j = \frac{1}{2}\nabla_g(L_{X_i^j} H_{M_0}) \in \mathcal{P}_g(\mathbb{R}^n)$$

since H_{M_0} is quadratic and X_i^j is a constant vector field (that is, belongs to $\mathcal{P}_g(\mathbb{R}^n)$) as can be seen by induction. □

8.1.2. Flat metrics on $\mathbb{R}^n$ and other drifts. Here, we still have $\mathcal{P}_g(\mathbb{R}^n) = \mathbb{R} - \mathrm{span}\{\nabla x_1, \dots, \nabla x_n\}$, but we no longer have $K_{M_0} = 0$.

PROPOSITION 8.2. *Let K be a skew-symmetric matrix. If $\Delta\varphi + \|\nabla\varphi + Kx\|^2$ is quadratic, then there exists an FDF to compute the unnormalized conditional density associated with a system of the form*

$$\left\{ \begin{array}{rcl} dx_t &=& (\nabla^0\varphi(x_t) + Kx_t)dt + dv_t\,, \quad x_0 \rightsquigarrow p_0(x)dx, \\ dy_t &=& (Cx_t + D)dt + dw_t \end{array} \right. \tag{69}$$

when the density p_0 of x_0 belongs to $\mathcal{D}(M_0^\star)$.

Proof. We have by Definition B.2

$$\left\{ \begin{array}{rcl} M_0 &=& \frac{1}{2}\Delta + (Kx + \nabla\varphi) - \frac{1}{2}\sum_{i=1}^p h_i^2, \\ K_{M_0} &=& 2K, \\ H_{M_0} &=& \Delta\varphi + \|Kx + \nabla\varphi\|^2 + \sum_{i=1}^p h_i^2. \end{array} \right.$$

The proof is a straightforward application of Corollary 7.3 with

$$X_2^j = 2K\nabla h_j \in \mathcal{P}_g(\mathbb{R}^n) \quad \text{and} \quad X_{i+2}^j = 2KX_{i+1}^j + \frac{1}{2}\nabla_g(L_{X_i^j}H_{M_0}) \in \mathcal{P}_g(\mathbb{R}^n)$$

since H_{M_0} is quadratic and X_i^j is a constant vector field (that is, belongs to $\mathcal{P}_g(\mathbb{R}^n)$) as can be seen by induction. □

In the one-dimensional case, we do not find new finite filters (since all vector fields are gradients and $K = 0$). In the multidimensional case, the systems we find have already been studied by Yau in [25, Theorem 7] and by Haussmann and Pardoux in a more general setting with stochastic coefficients [12].

8.1.3. Other metrics. Here, we no longer have $\mathcal{P}_g(\mathbb{R}^n) = \mathbb{R} - \mathrm{span}\{\nabla x_1, \dots, \nabla x_n\}$. If $\mathcal{P}_g(\mathbb{R}^n)$ is not zero, it can be written as $\mathbb{R} - \mathrm{span}\{\nabla_g z_1, \dots, \nabla_g z_r\}$ by Proposition B.4.

PROPOSITION 8.3. *Assume that*

1. *$h_1, \dots, h_p$ belong to $\mathbb{R} - \mathrm{span}\{1, z_1, \dots, z_r\}$,*
2. *$\Delta\varphi + \|\nabla\varphi\|^2$ is quadratic in the variables $z_1, \dots, z_r$;*

then there exists an FDF to compute the unnormalized conditional density associated with a system of the form

$$\left\{ \begin{array}{rcl} x_t & \text{diffusion process with generator } \mathcal{L} = \frac{1}{2}\Delta_g + \nabla_g\varphi\,, & x_0 \rightsquigarrow p_0(x)dx, \\ dy_t = & h(x_t)dt + dw_t \end{array} \right. \tag{70}$$

when the density p_0 of x_0 belongs to $\mathcal{D}(M_0^\star)$.

Proof. We have by Definition B.2,

$$\left\{ \begin{array}{rcl} M_0 &=& \frac{1}{2}\Delta_g + \nabla_g\varphi - \frac{1}{2}\sum_{i=1}^p h_i^2, \\ K_{M_0} &=& 0, \\ H_{M_0} &=& \Delta_g\varphi + \|\nabla_g\varphi\|^2 + \sum_{i=1}^p h_i^2 \quad \text{is quadratic in} \quad z_1, \dots, z_r. \end{array} \right.$$

The proof follows that of Proposition 8.1. □

8.2. The case with correlated noises. Here, the operators $M_1 = L_{\tilde{g}_1} + h_1$, $\dots$, $M_p = L_{\tilde{g}_p} + h_p$ are not all functions, as was the case with uncorrelated noises, but some of them are differential operators of order equal to one. By Proposition 7.2,

the construction of FDRs by the methods developed here above necessitates that $\tilde{g}_1$, ..., $\tilde{g}_p$ should belong to the Lie algebra $\mathcal{H}_g(\mathbb{R}^n)$ of infinitesimal homothetic transformations of $(\mathbb{R}^n, g)$. The conditions for having $\mathcal{H}_g(\mathbb{R}^n) \neq 0$ are less restrictive than those for having $\mathcal{P}_g(\mathbb{R}^n) \neq 0$ since $\mathcal{P}_g(\mathbb{R}^n) \subset \mathcal{H}_g(\mathbb{R}^n)$. This is why this case is more favorable for the existence of FDFs.

8.2.1. Flat metrics on $\mathbb{R}^n$ and gradient drifts. Here, $\mathcal{H}(\mathbb{R}^n)$ consists of vector fields T of the form

$$T = \mu \sum_{i=1}^{n} \tilde{x}^i \frac{\partial}{\partial \tilde{x}^i} + \sum_{i=1}^{n} \left(\sum_{j=1}^{n} \alpha_{ij} \tilde{x}^j \right) \frac{\partial}{\partial \tilde{x}^i} + \sum_{i=1}^{n} \beta_i \frac{\partial}{\partial \tilde{x}^i},$$

where μ, β_1, ..., β_n belong to $\mathbb{R}$ and α is any skew-symmetric matrix. Moreover, we have by Definition B.2,

$$\left\{ \begin{array}{rcl} M_0 & = & \frac{1}{2}\Delta + \nabla\varphi - \frac{1}{2}\sum_{i=1}^{p} \left(h_i^2 + L_{\tilde{g}_i} h_i \right), \\ K_{M_0} & = & 0, \\ H_{M_0} & = & \Delta\varphi + \|\nabla\varphi\|^2 + \sum_{i=1}^{p} \left(h_i^2 + L_{\tilde{g}_i} h_i \right). \end{array} \right.$$

PROPOSITION 8.4. *Let*

1. *$\tilde{g}_1$, ..., $\tilde{g}_p$ be vector fields of the form*

$$\tilde{g}_j(x) = K_j x + G_j\,,\ K_j' = -K_j\,, \quad j = 1, \ldots, p\,, \tag{71}$$

2. *h_1, ..., h_p and φ be such that*

$$\Delta\varphi + \|\nabla\varphi\|^2 + \sum_{i=1}^{p} \left(h_i^2 + L_{\tilde{g}_i} h_i \right) \tag{72}$$

and $h_1(x) - L_{\tilde{g}_1}\varphi(x)$, ..., $h_p(x) - L_{\tilde{g}_p}\varphi(x)$ are all affine functions.
Then there exists an FDF to compute the unnormalized conditional density associated with a system of the form

$$\left\{ \begin{array}{rcl} dx_t & = & \nabla\varphi(x_t)dt + dv_t + \tilde{g}(x_t) \circ dy_t\,, \quad x_0 \rightsquigarrow p_0(x)dx, \\ dy_t & = & h(x_t)dt + dw_t \end{array} \right. \tag{73}$$

when the density p_0 of x_0 belongs to $\mathcal{D}(M_0^\star)$.

Proof. The proof is a straightforward application of Proposition 7.2 with $X_0^j = \tilde{g}_j \in \mathcal{I}_g(\mathbb{R}^n)$, $X_1^j = \nabla(h_i - L_{\tilde{g}_i}\varphi) \in \mathcal{P}_g(\mathbb{R}^n)$ and $X_{i+1}^j = \frac{1}{2}\nabla(L_{X_i^j} H_{M_0}) \in \mathcal{P}_g(\mathbb{R}^n)$ since H_{M_0} is affine and X_i^j is a linear vector field (as all infinitesimal homothetic transformations of $\mathbb{R}^n$). □

Remark. If $\tilde{g}_1(x) = G_1$, ..., $\tilde{g}_p(x) = G_p$, the previous result still holds when (72) is not affine but even quadratic.

8.2.2. Flat metrics on $\mathbb{R}^n$ and other drifts. Here, we no longer have $K_{M_0} = 0$.

PROPOSITION 8.5. *Let*

1. *$\tilde{g}_1$, ..., $\tilde{g}_p$ be constant vector fields,*
2. *K be a skew-symmetric matrix, h_1, ..., h_p and φ be such that*

$$\Delta\varphi + \|Kx + \nabla\varphi\|^2 + \sum_{i=1}^{p} \left(h_i^2 + L_{\tilde{g}_i} h_i \right) \tag{74}$$

is quadratic and that for all $i = 1$, ..., p, $h_i(x) - L_{\tilde{g}_i}\varphi(x)$ is an affine function.

Then there exists an FDF to compute the unnormalized conditional density associated with a system of the form

$$\left\{ \begin{array}{lcl} dx_t &=& (\nabla\varphi(x_t) + Kx_t)dt \,+\, dv_t \,+\, \tilde{g} \circ dy_t \,, \quad x_0 \rightsquigarrow p_0(x)dx, \\ dy_t &=& h(x_t)dt \,+\, dw_t \end{array} \right. \tag{75}$$

when the density p_0 of x_0 belongs to $\mathcal{D}(M_0^\star)$.

Proof. The proof is a straightforward application of Proposition 7.2 with $X_0^j = \tilde{g}_j \in \mathcal{P}_g(\mathbb{R}^n)$, $X_1^j = \nabla(h_i - L_{\tilde{g}_i}\varphi) \in \mathcal{P}_g(\mathbb{R}^n)$, and $X_{i+1}^j = \frac{1}{2}\nabla_g(L_{X_i^j}H_{M_0}) \in \mathcal{P}_g(\mathbb{R}^n)$ since H_{M_0} is quadratic and X_i^j is a constant vector field (that is, belongs to $\mathcal{P}_g(\mathbb{R}^n)$), as can be seen by induction. □

8.2.3. Other metrics. We shall not treat this case, which does not lead to amenable formulations for the results.

8.3. Examples of FDFs. The following examples are straightforward applications of Proposition 8.4 (and the remark following, for the first of them).

PROPOSITION 8.6. *If there exist δ, ϵ such that the function*

$$f'(x) + f^2(x) + (\delta x + \epsilon)f(x) \tag{76}$$

is quadratic, then there exists a finite filter for the system

$$\left\{ \begin{array}{ll} dx_t \,=\, f(x_t)dt \,+\, dv_t \,+\, dy_t \,, & x_0 \rightsquigarrow p_0(x)dx, \\ dy_t \,=\, (f(x_t) + \delta x_t + \epsilon)\, dt \,+\, dw_t \,, & y_0 \,=\, 0\,. \end{array} \right. \tag{77}$$

We do not detail this case since it can more or less be found in [26].

Remark. It can be easily seen that there is no quadratic f satisfying the above assumptions but that any affine function does. Setting $f(x) = ax + b + 1/u(x)$, we find a family of solutions

$$f(x) = ax + b + \frac{\exp(-(a+\epsilon/2)x^2 + (2b+\delta)x)}{c + \displaystyle\int_0^x \exp(-(a+\epsilon/2)z^2 + (2b+\delta)z)dz}\,.$$

For instance, there exists an FDF for the system

$$\left\{ \begin{array}{ll} dx_t \,=\, \dfrac{1}{e^{-x_t}+1}dt \,+\, dv_t \,+\, dy_t \,, & x_0 \rightsquigarrow p_0(x)dx \\ dy_t \,=\, -\dfrac{e^{-x_t}}{e^{-x_t}+1}dt \,+\, dw_t \,, & y_0 \,=\, 0\,. \end{array} \right. \tag{78}$$

PROPOSITION 8.7. *If there exist δ, ϵ such that the function*

$$(1+x^2)(f'(x) + f^2(x)) + x\,(2(\delta x + \epsilon) + 1)\, f(x) + \delta^2 x^2 \tag{79}$$

is an affine function, then there exists a finite filter for the system

$$\left\{ \begin{array}{ll} dx_t \,=\, f(x_t)dt \,+\, dv_t \,+\, x_t \circ dy_t \,, & x_0 \rightsquigarrow p_0(x)dx, \\ dy_t \,=\, (x_t f(x_t) + \delta x_t + \epsilon)\, dt \,+\, dw_t \,, & y_0 \,=\, 0\,. \end{array} \right. \tag{80}$$

Remark. Setting $f(x) = b + 1/u(x)$, it can be shown by computation that any $f(x)$ of the form

$$f(x) = -\delta + \frac{(1+x^2)^{-\epsilon-\frac{1}{2}} \exp(2\delta \arctan x)}{a + \displaystyle\int_0^x (1+z^2)^{-\epsilon-\frac{1}{2}} \exp(2\delta \arctan z) dz} \tag{81}$$

is such that (79) is quadratic.

For instance, there exists an FDF for the system

$$\begin{cases} dx_t = \dfrac{1}{(1+x_t^2)(\arctan x_t + a)} dt + dv_t + x_t \circ dy_t\,, & x_0 \rightsquigarrow p_0(x)dx, \\ dy_t = \left(\dfrac{x_t}{(1+x_t^2)(\arctan x_t + a)} - \dfrac{3}{2} \right) dt + dw_t\,, & y_0 = 0, \end{cases} \tag{82}$$

where $a \notin]-\pi/2, \pi/2[$, when $p_0 \in \mathcal{D}(M_0^\star)$.

Appendix A. On the Riemannian geometric framework. Here, we review the necessary mathematical background (our references are [14, 1]). We assume from now on that $\mathbb{R}^n$ is equipped with a Riemannian metric g.

DEFINITION A.1. *Let T be an r-form ($(0,r)$ tensor field), X, X_1, ..., X_r be vector fields and f a smooth function on $\mathbb{R}^n$. The* Lie derivation *(of tensors) L_X is characterized by the following relations.*

$$\begin{cases} L_X f = Xf = \langle df, X \rangle\,, \\ L_X X_1 = [X, X_1], \\ L_X(T(X_1, \ldots, X_r)) = (L_X T)(X_1, \ldots, X_r) + \displaystyle\sum_{i=1}^r T(X_1, \ldots, L_X X_i, \ldots, X_r). \end{cases}$$

The inner product *i_X is defined by*

$$(i_X T)(X_2, \ldots, X_r) = T(X, X_2, \ldots, X_r).$$

DEFINITION A.2. *Let X, Y, and Z be vector fields ($(1,0)$ tensor fields), ω be a one-form ($(0,1)$ tensor field), and f be a smooth function on $\mathbb{R}^n$.*

1. *D_Z denotes the covariant derivation and A_Z the derivation $A_Z = L_Z - D_Z$.*
2. *A_Z induces a $(1,1)$ tensor field by $A_Z X = -D_X Z$ whose adjoint $A_Z^\star$ is defined by $g(A_Z^\star X, Y) = g(X, A_Z Y)$.*
3. *Ω_g is the* volume form *on $(\mathbb{R}^n, g)$.*
4. *The* divergence *of X is the function $\mathrm{div}_g X$ which satisfies $L_X \Omega_g = \mathrm{div}_g X\, \Omega_g$.*
5. *$Z^\flat$ is the one-form defined by duality by $Z^\flat(X) = g(Z, X)$.*
6. *$\omega^\sharp$ is the vector field defined by duality by $g(\omega^\sharp, X) = \omega(X)$.*
7. *The* gradient *of f is the vector field $\nabla_g f = (df)^\sharp$ such that $g(\nabla_g f, X) = Xf$.*
8. *The* Laplacian (*or Laplace–Beltrami operator*) *Δ_g is given by $\Delta_g f = \mathrm{div}_g(\nabla_g f)$.*

DEFINITION A.3. *We note by $\mathcal{P}_g(\mathbb{R}^n)$ the Lie algebra of* parallel vector fields *of $(\mathbb{R}^n, g)$, namely,*

$$\begin{aligned} \mathcal{P}_g(\mathbb{R}^n) &= \{X \in \mathcal{X}(\mathbb{R}^n) \mid L_X g = 0 \quad \text{and} \quad dX^\flat = 0\} \\ &= \{X \in \mathcal{X}(\mathbb{R}^n) \mid A_X = 0\}. \end{aligned} \tag{A.83}$$

We note by $\mathcal{I}_g(\mathbb{R}^n)$ the Lie algebra of infinitesimal isometries *of $(\mathbb{R}^n, g)$, namely,*

$$\mathcal{I}_g(\mathbb{R}^n) = \{X \in \mathcal{X}(\mathbb{R}^n) \mid L_X g = 0\}. \tag{A.84}$$

We note by $\mathcal{H}_g(\mathbb{R}^n)$ *the Lie algebra of* infinitesimal homothetic transformations *of* $(\mathbb{R}^n, g)$*, namely,*

$$\mathcal{H}_g(\mathbb{R}^n) = \{X \in \mathcal{X}(\mathbb{R}^n) \mid \exists \lambda \in \mathbb{R}\,, \quad L_X g = \lambda g\}. \tag{A.85}$$

If $L_X g = \rho g$*, we note* $\rho = \eta_g(X)$*. It is clear that* $\mathcal{P}_g(\mathbb{R}^n) \subset \mathcal{I}_g(\mathbb{R}^n) \subset \mathcal{H}_g(\mathbb{R}^n)$*.*

Appendix B. Recalls on the perturbation algebra of a parabolic operator. In this section, we make a brief recall of the main results in [8], and we broaden their validity by relaxing an analyticity assumption in [8].

Let A be an elliptic operator, assumed to have smooth coefficients, and be non-degenerate elliptic:

$$A = \sum_{i,j=1}^{n} a^{ij}(x) \frac{\partial^2}{\partial x_i \partial x_j} + \sum_{i=1}^{n} b^i(x) \frac{\partial}{\partial x_i} + c(x).$$

Above in the paper, we make use of the following results with either $A = M_0$ or $A = M_0^\star$.

PROPOSITION B.1 (see [8]). *Let* $Z \in \mathfrak{X}(F)$ *be of the form* (37) *and let* $\widehat{Z}_t$ *be defined, for all* t*, as a smooth differential operator on* $\mathbb{R}^n$*: for all* $p \in C^\infty(\mathbb{R}^n)$*,*

$$\left(\widehat{Z}_t p\right)(x) = -Z^0(t) A p(x) - \sum_{i=1}^{n} Z^i(t,x) \frac{\partial p}{\partial x_i}(x) + Z^{n+1}(t,x) p(x). \tag{B.86}$$

Then, there exists a finite-dimensional Lie algebra $\mathcal{P}_A$*, the* perturbation algebra *of the parabolic operator* $\partial_t - A$*, of linear partial differential operators of order less than or equal to one on* $\mathbb{R}^n$ *such that* Z *is a (local) infinitesimal symmetry of the parabolic PDE* $\partial_t u - Au = 0$ *if and only if*

$$\widehat{Z}_0 \in \mathbb{R}A \oplus \mathcal{P}_A \quad \textit{and} \quad \forall\, t \in \mathbb{R}\,, \quad \widehat{Z}_t = \exp(t\, ad_A)(\widehat{Z}_0)\,. \tag{B.87}$$

Moreover, $\Lambda(Z) = \widehat{Z}_0$ *is an anti-isomorphism from the set of such (local) infinitesimal symmetries to* $\mathbb{R}A \oplus \mathcal{P}_A$*. In particular, (local) infinitesimal symmetries of the form* (37) *form a finite-dimensional Lie algebra.*

Let O *be linear partial differential operators of order less than or equal to one on* $\mathbb{R}^n$*. Then* $O \in \mathcal{P}_A$ *if and only if the Lie algebra* $\{A, O\}_{\mathcal{L}.\mathcal{A}.}$ *of smooth differential operators on* $\mathbb{R}^n$ *generated by* A *and* O *is of the form* $\{A, O\}_{\mathcal{L}.\mathcal{A}.} = \mathbb{R}A \oplus \mathcal{Q}$*, where* $\mathcal{Q}$ *is a finite-dimensional Lie algebra consisting of linear partial differential operators of order less than or equal to one on* $\mathbb{R}^n$*.*

Proof. We simply prove here that the assumption on the analyticity of the coefficients of A in [8] may be relaxed and replaced by the assumption that the coefficients of A are smooth.

This analyticity condition was used in the proof of Proposition 2.2 in [8], and we follow the notations of this proof.

Let $U^{(2)}$ be the second jet space of $\mathbb{R} \times \mathbb{R}^n$, with coordinates denoted by

$$u^{(2)} = (u, (u_t, u_{x_i},\ 1 \leq i \leq n), (u_{tt}, u_{tx_j},\ 1 \leq j \leq n, u_{x_i x_j},\ 1 \leq i \leq j \leq n)).$$

If $(t_0, x_0) \in \mathbb{R} \times \mathbb{R}^n$ and v is a smooth function on a neighborhood of (t_0, x_0), the two-jet of v (or second prolongation of v) at (t_0, x_0) is the collection $\mathrm{pr}^{(2)}[v](t_0, x_0)$ of partial derivatives of v at (t_0, x_0) up to order two.

Let Γ be the smooth map on $\mathbb{R} \times \mathbb{R}^n \times U^{(2)}$ given by

$$\Gamma(t,x,u^{(2)}) = u_t - \sum_{i,j=1}^{n} a^{ij}(x) u_{x_i x_j} - \sum_{i=1}^{n} b^i(x) u_{x_i} - c(x)u. \tag{B.88}$$

We have

$$\Gamma(t,x,\mathrm{pr}^{(2)}[v](t,x)) = (\partial_t - A)v(t,x).$$

Γ is of maximal rank and, to prove that it is *nondegenerate* in the sense of [16, Definition 2.70], it remains to be proven that it is *locally solvable* in the sense of [16, Definition 2.70]. This latter condition is satisfied if, for any $(t_0, x_0, u_0^{(2)}) \in \mathbb{R} \times \mathbb{R}^n \times U^{(2)}$ such that $\Gamma(t_0, x_0, u_0^{(2)}) = 0$, there exists a smooth function v on a neighborhood of (t_0, x_0) such that

$$\begin{cases} \Gamma(t,x,\mathrm{pr}^{(2)}[v](t,x)) &= 0 \text{ for all } (t,x) \text{ in a neighborhood of } (t_0,x_0) \\ & \quad (v \text{ is a solution of } (\partial_t - A)v = 0), \\ \Gamma(t_0,x_0,\mathrm{pr}^{(2)}[v](t_0,x_0)) &= u_0^{(2)}. \end{cases}$$

Let $(t_0, x_0, u_0^{(2)}) \in \mathbb{R} \times \mathbb{R}^n \times U^{(2)}$ be given and let us exhibit such a function v.

First, since the symmetric matrix $(a^{ij}(x_0))_{i,j=1,\ldots,n}$ is not degenerate, it is easily seen that there exists $f \in C^\infty(\mathbb{R}^n)$ whose derivatives up to four satisfy

$$\begin{cases} \mathrm{pr}^{(2)}[f](x_0) &= \left(u^0, (u^0_{x_i},\ 1\le i\le n), (u^0_{x_i x_j},\ 1\le i\le j\le n)\right), \\ (Af)(x_0) &= u^0_t, \\ (A^2 f)(x_0) &= u^0_{tt}, \\ (\partial_{x_j} Af)(x_0) &= u^0_{tx_j}, \quad j = 1,\ldots,n. \end{cases} \tag{B.89}$$

Second, given a smooth function ψ on $\mathbb{R}^n$ with compact support included in an open ball $\Omega = B(x_0, r)$ and having constant value one in a neighborhood of x_0, we exhibit (for $T > t_0$) a smooth function v in $C^\infty([t_0, T] \times \Omega)$ such that

$$(\partial_t - A)v(t,x) = 0 \quad \forall\ (t,x) \in]t_0, T[\times B(x_0, r), \tag{B.90}$$

$$v(t,x) = 0 \quad \forall\ (t,x) \in]t_0, T[\times S(x_0, r) =]t_0, T[\times \partial B(x_0, r), \tag{B.91}$$

$$v(t_0, x) = \psi(x) f(x) \quad \forall\ x \in B(x_0, r). \tag{B.92}$$

Indeed, the above problem has a unique solution $v \in C^\infty([t_0, T] \times \Omega)$ since the assumptions of [5, Theorem X.10] are satisfied:

- the smooth elliptic differential operator A is strictly elliptic on the bounded open set Ω: $\sum_{i,j=1}^n a_{ij}(x)\zeta_i\zeta_j \ge \alpha\|\zeta\|^2, x \in \Omega,\ \zeta \in \mathbb{R}^n,\ \alpha > 0$, because A is nondegenerate elliptic with smooth coefficients,
- the open set Ω is bounded and has a smooth frontier $\partial\Omega = S(x_0, r)$,
- a^{ij}, b^i, c, and $v(t_0, \cdot) = \psi f$ belong to $C^\infty(\overline{\Omega})$,
- the compatibility relations, $\forall\ k \in \mathbb{N}, A^k v(t_0, \cdot) = 0$ on $\partial\Omega$, are satisfied since $v(t_0, \cdot) = \psi f$ has compact support included in that of ψ, therefore in the open ball Ω, and is thus zero in a neighborhood of $\partial\Omega$.

Third, the function v answers the question since by (B.90) (extended down to $t = t_0$, since $v \in C^\infty([t_0, T] \times \Omega)$), by (B.89), and by (B.92) it satisfies

$$\left\{\begin{array}{rcl} v(t_0,x_0) & = & f(x_0) = u^0, \\ \partial_t v(t_0,x_0) & = & Av(t_0,x_0) = (Af)(x_0) = u^0_t, \\ \partial_{x_j} v(t_0,x_0) & = & \partial_{x_j} f(x_0) = u^0_{x_i}, \\ \partial^2_{tt} v(t_0,x_0) & = & \partial_t Av(t_0,x_0) = A\partial_t v(t_0,x_0) = (A^2 f)(x_0) = u^0_{tt}, \\ \partial^2_{tx_j} v(t_0,x_0) & = & \partial_{x_j}\partial_t v(t_0,x_0) = \partial_{x_j}(Af)(x_0) = u^0_{tx_j}, \\ \partial^2_{x_i x_j} v(t_0,x_0) & = & \partial^2_{x_i x_j} f(x_0) = u^0_{x_i x_j}\,. \end{array}\right.$$

This ends the proof by relaxing the analyticity assumption. □

We can characterize the perturbation algebra in geometric terms when A has the form

$$A = \frac{1}{2}\Delta_g + U + c, \tag{B.93}$$

where U is a smooth vector field on $\mathbb{R}^n$ and where c is a smooth function.

DEFINITION B.2. *If A is given by* (B.93)*, the skew-symmetric* $(1,1)$ *tensor field K_A and the function H_A are defined by*

$$K_A = A_U - A^\star_U \quad \textit{and} \quad H_A = \mathrm{div}_g U + g(U,U) - 2c. \tag{B.94}$$

If $U = \nabla_g \varphi$, then $K_A = 0$ and $H_A = \Delta_g \varphi + \|\nabla_g \varphi\|^2 - 2c$.

Remark. The case $U = \nabla_g \varphi$ is related to the so-called *exact estimation algebra* extensively studied in [24, 23, 10, 6].

We identify a smooth differential operator of order less than or equal to one with the sum of a smooth vector field and of a smooth function.

THEOREM B.3 (see [9]). *The operator $X + m \in \mathcal{X}(\mathbb{R}^n) \oplus C^\infty(\mathbb{R}^n)$ belongs to $\mathcal{P}_A$ if and only if there exists a sequence $(X_i)_{i\in\mathbb{N}}$ in $\mathcal{H}_g(\mathbb{R}^n)$ which satisfies one of the equivalent following inductions:*

$$\left\{\begin{array}{rcl} X_0 & = & X, \\ X_1 & = & K_A X_0 + \nabla_g(m - g(X_0,U)) \\ (or & = & -\eta_g(X_0)U + [U,X_0] + \nabla_g m\ ), \\ X_{i+2} & = & K_A X_{i+1} + \frac{1}{2}\nabla_g(L_{X_i} H_A + \eta_g(X_i) H_A) \\ (or & = & -\eta_g(X_{i+1})U + [U,X_{i+1}] + \nabla_g(g(X_{i+1},U) + \frac{1}{2}L_{X_i}H_A + \frac{1}{2}\eta_g(X_i)H_A)\ ) \end{array}\right. \tag{B.95}$$

or

$$\left\{\begin{array}{rcl} X_0 & = & X, \\ X_1^\flat & = & -i_{X_0}(dU^\flat) + d(m - g(X_0,U)), \\ X_{i+2}^\flat & = & -i_{X_{i+1}}(dU^\flat) + \frac{1}{2}d(L_{X_i}H_A + \eta_g(X_i)H_A), \end{array}\right. \tag{B.96}$$

where the skew-symmetric $(1,1)$ *tensor field K_A and the function H_A are defined in* (B.94)*. Moreover, we have*

$$ad_A^{k+1}(X+m) = \eta_g(X_k)A + X_{k+1} + \textit{function}. \tag{B.97}$$

PROPOSITION B.4. *Assume that $(\mathbb{R}^n, g)$ is complete. If $\mathcal{P}_A$ is not reduced to constants, then $\mathcal{H}_g(\mathbb{R}^n)$ is not trivial. If $\mathcal{P}_A$ contains a nonconstant function, then $\mathcal{P}_g(\mathbb{R}^n)$ is not trivial.*

If $\dim\mathcal{P}_g(\mathbb{R}^n) = r \neq 0$, there exist $\tilde{x}^1, \ldots, \tilde{x}^r$ smooth functions such that $\nabla\tilde{x}^1, \ldots, \nabla\tilde{x}^r$ is an orthonormal basis of $\mathcal{P}_g(\mathbb{R}^n)$. Let $h \in \mathcal{P}_A$.

1. *If $0 < r < n$, h necessarily is a linear combination of $1, \tilde{x}^1, \ldots, \tilde{x}^r$.*
2. *If $r = n$, h necessarily is a linear combination of $1, \tilde{x}^1, \ldots, \tilde{x}^n$ and $\|\tilde{x}\|^2$.*

Proof. If $\mathcal{P}_A$ is not reduced to constants, then either X_0 or X_1 in (B.95) is not zero, so that $\mathcal{H}_g(\mathbb{R}^n)$ is not trivial (this holds true even if $(\mathbb{R}^n, g)$ is not complete).

By Theorem B.3, we know that if $h \in \mathcal{P}_A$, then $X_0 = \nabla_g h \in \mathcal{H}_g(\mathbb{R}^n)$. Then, either $\mathcal{I}_g(\mathbb{R}^n) \subsetneq \mathcal{H}_g(\mathbb{R}^n)$ and then $\mathcal{P}_g(\mathbb{R}^n) \neq 0$ (since $(\mathbb{R}^n, g)$ is complete and by Lemma A.3 in [8]) or $\mathcal{I}_g(\mathbb{R}^n) = \mathcal{H}_g(\mathbb{R}^n)$ and then $\nabla_g h \in \mathcal{P}_g(\mathbb{R}^n)$ (since any gradient vector field T in $\mathcal{I}_g(\mathbb{R}^n)$ is in fact in $\mathcal{P}_g(\mathbb{R}^n)$ because we have both $A_T = A_T^\star$ and $A_T + A_T^\star = 0$ [8, 9]).

If $\dim\mathcal{P}_g(\mathbb{R}^n) = r \neq 0$, such $\tilde{x}^1, \ldots, \tilde{x}^r$ exist by Lemma A.3 in [8]. If $r < n$, then $\nabla_g h \in \mathcal{H}_g(\mathbb{R}^n) \iff \nabla_g h \in \mathcal{P}_g(\mathbb{R}^n)$ by the same argument as above, so that the result is clear. If $r = n$, then by Lemma A.3 in [8] $(\mathbb{R}^n, g)$ is isometric to $\mathbb{R}^n$ flat so that $\mathcal{H}_g(\mathbb{R}^n) = \mathcal{I}_g(\mathbb{R}^n) \oplus \mathbb{R}\nabla\|\tilde{x}\|^2$. This ends the proof. ∎

Appendix C. A lemma on completeness of time-varying vector fields. The following lemma is used in the proof of Proposition 7.2.

LEMMA C.1. *Assume that $(\mathbb{R}^n, g)$ is complete. Let $H_1, \ldots, H_k$ be a basis of $\mathcal{H}_g(\mathbb{R}^n)$ and $\alpha_1(t), \ldots, \alpha_k(t)$ be piecewise continuous functions. Then the ODE*

$$\dot{x}(t) = \sum_{l=1}^{k} \alpha_l(t) H_l(x(t)), \quad x(0) = x_0 \tag{C.98}$$

has a solution defined for all time t.

Proof. The sketch of the proof is as follows.

Let $T > 0$ and $C(T) = \sup_{l=1,\ldots,k} \sup_{0\leq t\leq T} |\alpha_l(t)|$.

1. When $\alpha_1(t), \ldots, \alpha_k(t)$ are constant functions, we know that (C.98) has a solution defined for all time t. We shall show that, for $0 \leq t \leq T$,

$$d(x_0, x(t)) \leq t e^{\gamma C(T) t} C(T) \phi(x_0), \tag{C.99}$$

where γ depends only on $H_1, \ldots, H_k$ and $\phi(x)$ depends only on the Riemannian structure.

2. We shall also prove that there exists $\lambda \geq 0$ (depending only on the structural constants of the Lie algebra $\mathcal{H}_g(\mathbb{R}^n)$) such that, for $0 \leq t \leq T$,

$$\phi(x(t)) \leq e^{t(\lambda+\gamma)C(T)} \phi(x_0). \tag{C.100}$$

3. When $\alpha_1(t), \ldots, \alpha_k(t)$ are piecewise constant functions, we know that (C.98) has a solution defined for all time t. We shall show that, for $0 \leq t \leq T$,

$$d(x_0, x(t)) \leq C(T) t e^{t(2\gamma+\lambda)C(T)} \phi(x_0)$$

4. When $\alpha_1(t), \ldots, \alpha_k(t)$ are piecewise continuous functions, we shall show that (C.98) has a solution defined at least up to time T.

Since $(\mathbb{R}^n, g)$ is complete, $\mathcal{H}_g(\mathbb{R}^n)$ consists of complete vector fields [14], and therefore (C.98) has a solution defined for all time t when $\alpha_1(t), \ldots, \alpha_k(t)$ are constant functions. What is more, we have

$$d(x_0, x(t)) \leq \int_0^t \sqrt{g(\dot{x}(s), \dot{x}(s))} ds,$$

where $g(\dot{x}(s), \dot{x}(s)) = g(\sum_{l=1}^k \alpha_l H_l(x(s)), \sum_{l=1}^k \alpha_l H_l(x(s))$ satisfies a linear differential equation in the variable s since $\sum_{l=1}^k \alpha_l H_l$ is an infinitesimal homothetic transformation. Thus

$$\begin{aligned} d(x_0, x(t)) &\leq \sqrt{g\left(\sum_{l=1}^k \alpha_l H_l(x_0), \sum_{l=1}^k \alpha_l H_l(x_0)\right)} \int_0^t e^{s\eta_g(\sum_{l=1}^k \alpha_l H_l)} ds \\ &\leq te^{t\gamma C(T)} C(T) \sqrt{\sum_{l,m=1}^k g(H_l, H_m)(x_0)} \end{aligned}$$

and the first point is proved with $\gamma = \sum_{l=1}^k |\eta_g(H_l)|$ and $\phi(x) = \sqrt{\sum_{l,m=1}^k g(H_l, H_m)(x)}$.

It is easy to see that the $g(H_l, H_m)(x(t))$ satisfy a linear differential system in the variable t whose coefficients depend linearly on $\eta_g(\sum_{l=1}^k \alpha_l H_l)$, on $\alpha_1, \ldots, \alpha_k$, and the structural constants of the Lie algebra $\mathcal{H}_g(\mathbb{R}^n)$. Thus, there exists $\lambda \geq 0$, depending only upon these structural constants, such that

$$\sum_{l,m=1}^k g(H_l, H_m)(x(t)) \leq e^{2t(\lambda+\gamma)C(T)} \sum_{l,m=1}^k g(H_l, H_m)(x_0).$$

This leads to $\phi(x(t)) \leq e^{t(\lambda+\gamma)C(T)} \phi(x_0)$ and the second point is proved.

Now, when $\alpha_1(t), \ldots, \alpha_k(t)$ are piecewise constant functions, we can piece together the previous inequalities as (C.99) obtained for $d(x(t_q), x(t_{q+1}))$ on each interval $[t_q, t_{q+1}]$, where $\alpha_1(t), \ldots, \alpha_k(t)$ are constant. This gives, for $0 \leq t \leq T$,

$$d(x_0, x(t)) \leq C(T) \sum_q (t_{q+1} - t_q) e^{\gamma C(T)(t_{q+1} - t_q)} \phi(x(t_q)).$$

Since by (C.100) an easy induction yields

$$\phi(x(t_q)) \leq e^{(\lambda+\gamma)C(T)(t_q - t_{q-1})} \phi(x(t_{q-1})) \leq e^{(\lambda+\gamma)C(T)t_q} \phi(x_0),$$

we obtain the following estimate for $0 \leq t \leq T$:

$$d(x_0, x(t)) \leq C(T) t e^{t(2\gamma+\lambda)C(T)} \phi(x_0).$$

The third point being proved, we turn to the last one. When $\alpha_1(t), \ldots, \alpha_k(t)$ are piecewise continuous functions, we know that (C.98) has a solution defined at least up to a positive time. Let $\tau > 0$ be the first time when $x(t)$ leaves the nonempty (when $C(T) > 0$) bounded open set

$$\Omega = \{x \in \mathbb{R}^n, d(x_0, x) < 2C(T) T e^{T(2\gamma+\lambda)C(T)} \phi(x_0)\}.$$

We shall prove that $\tau \geq T$. Indeed, by [21, p. 106], since Ω is included in a bounded closed set (hence compact on the complete Riemannian manifold $(\mathbb{R}^n, g)$ [14]), the whole trajectory $(x(t), 0 \leq t \leq \tau \wedge T)$ may be approximated by trajectories of (C.98) with $\alpha_1(t), \ldots, \alpha_k(t)$ being piecewise constant functions. Thus, for $0 \leq t \leq \tau \wedge T$, we have by continuity

$$d(x_0, x(t)) \leq C(T) t e^{t(2\gamma+\lambda)C(T)} \phi(x_0) \leq C(T) T e^{T(2\gamma+\lambda)C(T)} \phi(x_0).$$

If $\tau < T$, this leads to a contradiction with the definition of τ.

We have shown that, for any time $T > 0$, the ODE (C.98) has a solution defined at least up to time T. This proves the assertion of the lemma. □

REFERENCES

[1] R. Abraham, J. Marsden, and T. Ratiu, *Manifolds, Tensor Analysis, and Applications*, Springer-Verlag, New York, 1988.

[2] J. Baras, *Group invariance methods in nonlinear filtering of diffusion processes*, in Stochastic Systems: The Mathematics of Filtering and Identification and Applications, M. Hazewinkel and J. Willems, eds., D. Reidel, Dordrecht, the Netherlands, 1981, pp. 565–572.

[3] V. Beneš, *Exact finite-dimensional filters for certain diffusions with nonlinear drift*, Stochastics, 5 (1982), pp. 65–92.

[4] G. Bluman and S. Kumei, *Symmetries and Differential Equations*, Springer-Verlag, New York, 1989.

[5] H. Brezis, *Analyse Fonctionnelle*, Masson, Paris, 1982.

[6] W.-L. Chiou and S.-T. Yau, *Finite-dimensional filters with nonlinear drift* II*: Brockett's problem on classification of finite-dimensional estimation algebras*, SIAM J. Control Optim., 32 (1994), pp. 297–310.

[7] M. Cohen de Lara, *Application of symmetry semi-groups to discrete and continuous time filtering problems*, in Analysis of Controlled Dynamical Systems, B. Bonnard, B. Bride, J. Gauthier, and I. Kupka, eds., Birkhäuser Boston, Cambridge, MA, 1991, pp. 146–155.

[8] M. Cohen de Lara, *A note on the symmetry group and perturbation algebra of a parabolic partial differential operator*, J. Math. Phys., 32 (1991), pp. 1444–1449.

[9] M. Cohen de Lara, *Geometric and symmetry properties of a non degenerate diffusion process*, Ann. Probab., 23 (1995), pp. 1557–1604.

[10] R. Dong, L. Tam, W. Wong, and S. Yau, *Structure and classification theorems of finite-dimensional exact estimation algebras*, SIAM J. Control Optim., 29 (1991), pp. 866–877.

[11] I. Gihman and A. Skorohod, *Stochastic Differential Equations*, Springer-Verlag, Berlin, 1972.

[12] U. Haussmann and E. Pardoux, *A conditionally almost linear filtering problem with non-gaussian initial condition*, Stochastics, 23 (1988), pp. 241–275.

[13] N. Ikeda and S. Watanabe, *Stochastic Differential Equations and Diffusion Processes*, North–Holland, Amsterdam, 1989.

[14] S. Kobayashi and K. Nomizu, *Foundations of Differential Geometry*, Vol. 1, John Wiley, New York, 1963.

[15] M. Liao, *Symmetry groups of Markov processes*, Ann. Probab., 20 (1992), pp. 563–578.

[16] P. Olver, *Applications of Lie Groups to Differential Equations*, Springer-Verlag, New York, 1986.

[17] L. Ovsjannikov, *Group Analysis of Differential Equations*, Academic Press, New York, 1982.

[18] R. Palais, *A global formulation of the Lie theory of transformation groups*, Mem. Amer. Math. Soc., 22 (1957), pp. 1–123.

[19] L. Rogers and D. Williams, *Diffusions, Markov Processes, and Martingales*, Vol. 2, John Wiley, New York, 1987.

[20] S. Rosencrans, *Perturbation algebra of an elliptic operator*, J. Math. Anal. Appl., 56 (1976), pp. 317–329.

[21] N. Rouche and J. Mawhin, *Équations différentielles ordinaires*, Vol. 1, Masson, Paris, 1973.

[22] I. Shigekawa, *Transformations of the brownian motion on the riemannian symmetric space*, Z. Wahrsch. Ver. Geb., 65 (1984), pp. 493–522.

[23] L. Tam, W. Wong, and S. Yau, *On a necessary and sufficient condition for finite dimensionality of estimation algebras*, SIAM J. Control Optim., 28 (1990), pp. 173–185.

[24] W. Wong, *Theorems on the structure of finite dimensional estimation algebras*, Systems Control Lett., 9 (1987), pp. 117–124.

[25] S.-T. Yau, *Finite dimensional filters with nonlinear drift* I*: A class of filters including both Kalman-Bucy filters and Beneš filters*, J. Math. Systems, Estimation, Control, 4 (1994), pp. 181–203.

[26] O. Zeitouni, *On some finite dimensional nonlinear filters for certain diffusions observed in correlated noise*, Systems Control Lett., 7 (1986), pp. 61–63.

SIAM J. CONTROL OPTIM.
Vol. 35, No. 3, pp. 1030–1052, May 1997

013

OPTIMIZATION OF OBSERVATIONS: A STOCHASTIC CONTROL APPROACH*

BORIS M. MILLER† AND WOLFGANG J. RUNGGALDIER‡

Abstract. We study a stochastic control problem for the optimization of observations in a partially observable stochastic system. Using a method of discontinuous time transformation, we associate with the original problem with unbounded controls a problem that has bounded controls. This latter problem allows us to construct nearly optimal nonanticipative Lipschitz Markov controls with finite observation power for the original problem. Since the controlled observation equation may degenerate, we also derive a corresponding filtering result and show a separation property of the optimal controls.

Key words. partially observed stochastic systems, observation control, nonlinear filtering, separation principle, discontinuous time transformation

AMS subject classifications. Primary, 93E03, 93E20; Secondary, 93E11, 60G35, 49N25

PII. S0363012995287878

Introduction. The most common way to formulate a control problem is to let the control affect only the evolution of the state while possible partial observations of the state are supposed to be continuously available.

Many practical situations, however, lead to the possibility that the observations also can be controlled in a way that affects both their timing as well as their quality. This then leads to a control problem where one tries to choose the control in a way to maximize the information content of the observations regarding the state while at the same time also taking into account a possible penalization of the control effort. Since the information content of the observations can be measured by the state estimation covariance, maximization of the information content can be obtained by minimizing this estimation covariance.

Problems of optimization of observations were mainly studied in the East (see, e.g., [1], [2], [9], [11]); a first study in the West appears in [4]. In a stochastic context only the linear case has been studied so far. More precisely, in [9] the authors consider the following linear model:

$$dx_t = a_t x_t\,dt + b_t\,dw_t^{(1)}, \tag{0.1a}$$

$$dy_t = A_t(u_t)x_t\,dt + B_t\,dw_t^{(2)}, \tag{0.1b}$$

where u_t is the observation control and $(w_t^{(1)})$ and $(w_t^{(2)})$ are independent standard Wiener processes. In this linear case the estimation covariance γ does not depend on the observations and so the optimal control becomes a deterministic time function.

The purpose of this paper is an attempt to extend the investigations to the nonlinear case. As a first step in this direction we consider a model related to (0.1), where

*Received by the editors June 19, 1995; accepted for publication (in revised form) April 8, 1996. The results of this paper were announced at the ICCI'95 Conference in Hong Kong, and a preliminary version of this paper appeared in the proceedings. This research was supported by the Italian National Research Council (GNAFA visiting appointment) and INTAS contract 94-697.

http://www.siam.org/journals/sicon/35-3/28787.html

†Institute for Information Transmission Problems, Russian Academy of Sciences, Ermolovoj Str. 19, 101447 GSP-4 Moscow, Russia (bmiller@ippi.ac.msk.su).

‡Dipartimento di Matematica Pura e Applicata, Università di Padova, Via Belzoni 7, I-35131 Padova, Italy (runggal@math.unipd.it).

the coefficients depend on an unknown parameter and the observation noise is more realistically considered as an endogenous noise induced by the observations themselves; the observation power is restricted to be finite at all times with bounded total observation energy. Contrary to the linear case, the observation covariance here depends on the observations and the control problem itself becomes a stochastic control problem.

The structure of the paper is as follows: in section 1 we present our control model and study the associated filter problem which, due to the possibility that the (controlled) observation equation may degenerate, cannot be approached directly by standard techniques. In section 2 we then formulate the full control problem and show a separation property, namely, that among the optimal controls there is one depending on the observations through the filter values. This control problem is a nonstandard nonlinear problem with finite but unbounded controls. In section 3, using a stochastic version of the so-called method of discontinuous time transformation (see [10] for a deterministic context), we therefore derive an auxiliary problem with bounded controls and study the relationship between the original and the auxiliary control problems. While the auxiliary problem can be shown to admit an optimal solution, for the original problem there may not exist an optimal nonanticipative solution. On the other hand, the auxiliary problem gives also the possibility to derive a nearly optimal nonanticipative Lipschitz Markov (feedback) control for the original problem. Finally, in the concluding remarks we recall some of the delicate points of our approach.

1. The model and the associated filter process.

1.1. The model. On a given finite time interval $[0, T]$ consider a partially observed process (x_t, y_t) that satisfies the following linear system, parametrized by an unknown parameter θ and with a control in the observations:

$$x_t = x_0 + \int_0^t a_s(\theta) x_s \, ds + \int_0^t b_s(\theta) \, dw_s^{(1)}, \tag{1.1a}$$

$$y_t = \int_0^t A_s(\theta) x_s \, dv_s + \eta_t. \tag{1.1b}$$

In this system, where for simplicity of presentation we consider all processes to be scalar valued, $(w_t^{(1)})$ is a standard Wiener process with respect to a given filtration $(\mathcal{F}_t)$ with $\mathcal{F}_t \supseteq \mathcal{F}_t^y := \sigma\{y_s,\, s \le t\}$; x_0 is $\mathcal{F}_0$-measurable with distribution $\mathcal{N}(m_0, \sigma_0)$ and independent of $(w_t^{(1)})$. The observation control process (v_t) is an $\mathcal{F}_t^y$-adapted absolutely continuous and almost surely (a.s.) nondecreasing process with $v_0 = 0$ that thus has almost everywhere (a.e.) a derivative $u_t = \dot{v}_t$ that we assume satisfies the restrictions

$$0 \le u_t < +\infty \qquad \text{(finite observation power)}, \tag{1.2a}$$

$$\int_0^T u_t dt = v_T \le M < +\infty \qquad \text{(finite observation energy)}. \tag{1.2b}$$

The additive observation disturbance consists of an endogenously induced noise, due to the observation itself and represented by the (conditionally) Gaussian $\mathcal{F}_t^y$-martingale (η_t), whose quadratic variation satisfies the compatibility condition

$$\langle \eta \rangle_t = B^2 \, v_t \qquad (B \neq 0) \tag{1.3}$$

and is independent of $(w_t^{(1)})$ and x_0. Although other reasonable models could possibly be posited, in this first approach to the control of the observations in the nonlinear stochastic case we assume a compatibility condition in the form of (1.3), which implicitly states that drift and noise in the observation equation are both linear in the control (see the ensuing equivalent representation of model (1.1) in (1.6) below). By considering more complex situations, the control may enter the diffusion term in a nonmultiplicative way so that the absolute continuity of the drift with respect to the quadratic variation of the noise in the observation equation may be lost with ensuing additional problems for the filter. Note also that we may let the constant B in (1.3) be substituted by a time function B_t; dividing the observations by the (known) function B_t then reduces this more general case to the one treated here.

As a consequence of (1.3), there exists an $\mathcal{F}_t$-standard Wiener process $(w_t^{(2)})$, independent of $(w_t^{(1)})$ and x_0, so that the following representation holds:

$$\eta_t = B \int_0^t (u_s)^{1/2} dw_s^{(2)}. \tag{1.4}$$

For this representation (1.4) and analogous ones later, as is usually done we implicitly assume that, where necessary, the underlying probability space is sufficiently enlarged to support all required Wiener processes. (For an explicit construction of such an enlargement see, e.g., section 1.4.4 in [5].)

The dependence of u_t on the observation history implies that (1.1b) is actually an equation in (y_t). To ensure that (1.1b) is well defined, we shall thus assume that u_t as a function of the observation history $y_0^t := \{y_s, s \le t\}$ is such that it satisfies a Lipschitz property in the sense that for a nondecreasing and right continuous function $K(t)$, $0 \le K(t) \le 1$, and some nonnegative constants L_1, L_2 we have for all $t \ge 0$

$$|u_t(y_0^t) - u_t(\tilde{y}_0^t)\,|^2 \le L_1 \int_0^t |y_s - \tilde{y}_s|^2 dK(s) + L_2 |y_t - \tilde{y}_t|^2. \tag{1.5}$$

Furthermore, taking the Bayesian point of view, the unknown parameter θ is considered an $\mathcal{F}_0$-measurable random variable, independent of $x_0, (w_t^{(1)})$, and (η_t) and taking a finite number of possible values $\theta^i\,(i = 1, \dots, k)$ with prior probabilities $p_i = P(\theta = \theta^i)$. Finally, $a_t(\theta), b_t(\theta)$, and $A_t(\theta)$ are continuous and bounded functions of t for all θ.

In the setting just described, system (1.1) can equivalently be represented as

$$dx_t = a_t(\theta) x_t\, dt + b_t(\theta)\, dw_t^{(1)}, \tag{1.6a}$$

$$dy_t = A_t(\theta) x_t u_t\, dt + B \cdot (u_t)^{1/2} dw_t^{(2)}, \quad y_0 = 0, \tag{1.6b}$$

$$d\theta = 0. \tag{1.6c}$$

Remark 1.1. Since the value of u_t corresponds to the power applied to the observation of the signal x_t, conditions (1.2) imply that we require this power to be finite at each t with bounded total observation energy, which indeed corresponds to the actual physical situation. The fact that the additive observation noise in (1.1b), or equivalently in (1.6b), is given by an endogenously generated noise due to the observation itself justifies the assumption of a compatibility condition such as (1.3). The Gaussian assumption for this noise can be justified for those cases when the power u_t,

applied for observing x_t, is sufficiently large; take, e.g., an optical noise that is Poisson with intensity proportional to the power of the observation so that for large values of this power it can be approximated by a Gaussian. In a sense, the observation noise in (1.1b) or (1.6b) is thus a minimum-level noise, to which some independent exogenous Gaussian noise could possibly be added as well. Note finally that the observation inaccuracy is due not only to the additive observation noise but also to the averaging of the signal x_t as implied by the first term on the right of the observation equation (1.1b) or (1.6b). The averaging due to the choice of v_t (equivalently of u_t) thus affects both the timing and the quality of the observations: in the limit, when the observation power tends to infinity, u_t tends to a δ-function thus determining only the timing of the observations with no averaging of the signal. Note also that the observation power may tend to infinity, while the total observation energy remains still bounded by M.

Before going on to describe the full control problem, we study the filter process associated with the given partially observed control model.

1.2. The filter process. For $i = 1, \dots, k$ consider

$$m_t^i := E\{x_t | \mathcal{F}_t^y, \theta^i\}, \tag{1.7a}$$

$$\gamma_t^i := E\{(x_t - m_t^i)^2 | \mathcal{F}_t^y, \theta^i\}, \tag{1.7b}$$

$$\pi_t^i := P\{\theta = \theta^i | \mathcal{F}_t^y\}, \tag{1.7c}$$

and let the "filter process" X_t be given by the following set of triplets:

$$X_t := \{m_t^i, \gamma_t^i, \pi_t^i\}_{i=1,\dots,k}. \tag{1.8}$$

The main purposes of this subsection are to derive a stochastic differential equation for X_t and to show that, under the assumptions of subsection 1.1, it has a unique solution. We point out that these results will not simply be a direct application of known filtering results since, due to the possibility that u_t may be equal to zero on intervals of positive length, the observation equation may degenerate. The main result of this section is the following theorem.

THEOREM 1.2. *The filter process* (X_t) *in* (1.8) *satisfies, for a given control* v_t *(equivalently* u_t*) and* $i = 1, \dots, k$,

$$dm_t^i = a_t(\theta^i) m_t^i dt + B^{-2} A_t(\theta^i) \gamma_t^i [dy_t - A_t(\theta^i) m_t^i dv_t], \quad m_0^i = m_0 = E(x_0), \tag{1.9a}$$

$$d\gamma_t^i = 2a_t(\theta^i)\gamma_t^i dt + b_t^2(\theta^i)dt - B^{-2}[A_t(\theta^i)\gamma_t^i]^2 dv_t, \quad \gamma_0^i = \sigma_0 = \mathrm{Cov}(x_0), \tag{1.9b}$$

$$d\pi_t^i = \pi_t^i \left[A_t(\theta^i) m_t^i - \sum_{j=1}^k \pi_t^j A_t(\theta^j) m_t^j \right] B^{-2} d\xi_t, \quad \pi_0^i = p_i, \tag{1.9c}$$

where

$$\xi_t := \int_0^t \left[dy_s - \sum_{j=1}^k \pi_s^j A_s(\theta^j) m_s^j dv_s \right] \tag{1.10}$$

is an $\mathcal{F}_t^y$-conditionally Gaussian martingale with quadratic variation

$$\langle \xi \rangle_t = v_t(y_0^t). \tag{1.11}$$

Furthermore, the system (1.9) *can be represented in compact form as*

$$dX_t = F_t(X_t)dt + B_t(X_t)u_t dt + G_t(X_t)u_t^{1/2} dw_t^X \tag{1.12}$$

for suitable functions F, B, and G related to the coefficients in (1.9) *and where (w_t^X) is an $\mathcal{F}_t$-standard Wiener process. Finally, for any given control (v_t) or (u_t), the solution of* (1.9) (*or* (1.12)) *is unique.*

To prove this theorem (the proof will be given below) we shall need an intermediate result for an auxiliary filtering problem that will allow us to cope with the possible degeneracy of the original (controlled) filtering model. To derive the auxiliary problem we use an absolutely continuous time transformation. More precisely, for a given control v_t (recall that $v_0 = 0$) let

$$\Gamma_t := v_t + \int_0^t I\{s : \dot{v}_s = 0\}\, ds, \tag{1.13}$$

which is an $\mathcal{F}_t^y$-adapted absolutely continuous process with strictly positive derivative and satisfying $\Gamma_T \le M + T$ (see (1.2)). On the interval $[0, \Gamma_T]$ it admits thus the inverse function

$$\nu_s = \inf\{\tau : \Gamma_\tau > s\} = \inf\{\tau : \Gamma_\tau = s\}, \tag{1.14}$$

which satisfies $0 \le \nu_s \le T$ and is absolutely continuous with

$$\dot{\nu}_s = \frac{1}{\dot{\Gamma}_{t\,|t=\nu_s}} = \frac{1}{(\dot{v}_t + I\{t : \dot{v}_t = 0\})_{|t=\nu_s}} > 0. \tag{1.15}$$

Furthermore, for each $s \in [0, T]$, ν_s is an $\mathcal{F}_t^y$-stopping time and, as a process, (ν_s) is adapted to $\mathcal{F}_{\nu_s}^y = \sigma\{y_\tau : 0 \le \tau \le \nu_s\}$. Consider now the time-transformed observation process

$$\bar{y}_s = y_{\nu_s}, \tag{1.16}$$

which by (1.1) satisfies

$$d\bar{y}_s = A_{\nu_s}(\theta)x_{\nu_s} dv_{\nu_s} + d\eta_{\nu_s}. \tag{1.17}$$

Note that (η_{ν_s}) is a continuous, $\mathcal{F}_{\nu_s}^y$–conditionally Gaussian martingale with quadratic variation $\langle \eta \rangle_{\nu_s} = B^2 v_{\nu_s}$ such that

$$\frac{d}{ds}\langle \eta \rangle_{\nu_s} = \frac{B^2 \dot{v}_{\nu_s}}{\dot{v}_{\nu_s} + I\{s : \dot{v}_{\nu_s} = 0\}} = B^2 I\{s : \dot{v}_{\nu_s} \ne 0\}. \tag{1.18}$$

The process (η_{ν_s}) may thus degenerate and so, using a regularization procedure, we define

$$\tilde{\eta}_s := \eta_{\nu_s} + B\int_0^s I\{\tau : \dot{v}_{\nu_\tau} = 0\}\, dw_\tau, \tag{1.19}$$

where (w_t) is an $\mathcal{F}_t$-standard Wiener process, independent of $(w_t^{(1)})$ and (η_t). This process $(\tilde{\eta}_t)$ is thus a conditionally Gaussian martingale with respect to $\mathcal{F}_{\nu_s}^y = \mathcal{F}_s^{\tilde{y}}$, that has continuous trajectories and nondegenerate quadratic variation

$$\begin{aligned}\langle\tilde{\eta}\rangle_s &= \langle\eta\rangle_{\nu_s} + B^2\int_0^s I\{\tau : \dot{v}_{\nu_\tau} = 0\}\,d\tau \\ &= B^2\int_0^s I\{\tau : \dot{v}_{\nu_\tau} \neq 0\}\,d\tau + B^2\int_0^s I\{\tau : \dot{v}_{\nu_\tau} = 0\}\,d\tau = B^2 s.\end{aligned} \tag{1.20}$$

Since $(\tilde{\eta}_s)$ is independent of $(w_s^{(1)})$, in what follows we shall consider an $\mathcal{F}_t$-Wiener process $(\tilde{w}_s^{(2)})$, independent of $(w_s^{(1)})$, and represent $(\tilde{\eta}_s)$ as $\tilde{\eta}_s = B\,\tilde{w}_s^{(2)}$. On the other hand, since for the process $(w_{\nu_s}^{(1)})$ we have $\langle w^{(1)}\rangle_{\nu_s} = \nu_s$, we may also consider an $\mathcal{F}_s$-standard Wiener process $(\tilde{w}_s^{(1)})$, independent of $(\tilde{w}_s^{(2)})$, and obtain $(w_{\nu_s}^{(1)})$ as

$$w_{\nu_s}^{(1)} = \int_0^s (\dot{\nu}_\tau)^{1/2} d\tilde{w}_\tau^{(1)}. \tag{1.21}$$

Defining finally for $s \le \Gamma_T$ (see (1.15))

$$\tilde{a}_s(\theta) := \frac{a_{\nu_s}(\theta)}{(\dot{v}_t + I\{t : \dot{v}_t = 0\})_{|t=\nu_s}} = a_{\nu_s}(\theta)\,\dot{\nu}_s, \tag{1.22a}$$

$$\tilde{b}_s(\theta) := b_{\nu_s}(\theta), \tag{1.22b}$$

$$\tilde{A}_s(\theta) := \frac{A_{\nu_s}(\theta)\,\dot{v}_{t\,|t=\nu_s}}{(\dot{v}_t + I\{t : \dot{v}_t = 0\})_{|t=\nu_s}} = A_{\nu_s}(\theta)\,I\{s : \dot{v}_{\nu_s} \neq 0\}, \tag{1.22c}$$

consider on $[0, T+M]$ the process-pair $(\tilde{x}_s, \tilde{y}_s)$ defined, for $0 \le s \le \Gamma_T$, by

$$d\tilde{x}_s = \tilde{a}_s(\theta)\tilde{x}_s ds + \tilde{b}_s(\theta)(\dot{\nu}_s)^{1/2} d\tilde{w}_s^{(1)}, \quad \tilde{x}_0 = x_0, \tag{1.23a}$$

$$d\tilde{y}_s = \tilde{A}_s(\theta)\tilde{x}_s ds + B\,d\tilde{w}_s^{(2)}, \quad \tilde{y}_0 = 0, \tag{1.23b}$$

$$d\theta = 0 \tag{1.23c}$$

and by putting $d\tilde{x}_s = d\tilde{y}_s = d\theta = 0$ for $\Gamma_T < s < T+M$. Note that from the foregoing we immediately have

$$\tilde{x}_s := x_{\nu_s}, \tag{1.24}$$

$$\bar{y}_s = \int_0^s I\{\tau : \dot{v}_{\nu_\tau} \neq 0\}\,d\tilde{y}_\tau \tag{1.25}$$

so that $\bar{y}_s$ defined in (1.16) is $\mathcal{F}_s^{\tilde{y}}$-measurable and therefore

$$\mathcal{F}_{\nu_s}^y = \mathcal{F}_s^{\bar{y}} \subseteq \mathcal{F}_s^{\tilde{y}}. \tag{1.26}$$

Analogously to (1.7) now consider

$$\tilde{m}_s^i := E\{\tilde{x}_s |\, \mathcal{F}_s^{\tilde{y}}, \theta^i\}, \tag{1.27a}$$

$$\tilde{\gamma}_s^i := E\{(\tilde{x}_s - \tilde{m}_s^i)^2 |\, \mathcal{F}_s^{\tilde{y}}, \theta^i\}, \tag{1.27b}$$

$$\tilde{\pi}_s^i := P\{\theta = \theta^i |\, \mathcal{F}_s^{\tilde{y}}\}. \tag{1.27c}$$

We have the following proposition.

PROPOSITION 1.3. *The process* $\tilde{X}_s := \{\tilde{m}_s^i, \tilde{\gamma}_s^i, \tilde{\pi}_s^i\}_{i=1,\dots,k}$ *satisfies on* $[0,T]$ *(for a given control)*

$$d\tilde{m}_s^i = \tilde{a}_s(\theta)\tilde{m}_s^i ds + B^{-2}\tilde{A}_s(\theta^i)\tilde{\gamma}_s^i\,[d\tilde{y}_s - \tilde{A}_s(\theta^i)\tilde{m}_s^i ds], \quad \tilde{m}_0^i = m_0^i = m_0, \tag{1.28a}$$

$$d\tilde{\gamma}_s^i = 2\tilde{a}_s(\theta^i)\tilde{\gamma}_s^i ds + \tilde{b}_s^2(\theta^i)\,\dot{\nu}_s ds - B^{-2}[\tilde{A}_s(\theta^i)\tilde{\gamma}_s^i]^2 ds, \quad \tilde{\gamma}_0^i = \gamma_0^i\sigma_0, \tag{1.28b}$$

$$d\tilde{\pi}_s^i = \pi_s^i \left[\tilde{A}_s(\theta^i)\tilde{m}_s^i - \sum_{j=1}^k \tilde{\pi}_s^j \tilde{A}_s(\theta^j)\tilde{m}_s^j\right] B^{-2} d\tilde{\xi}_t, \quad \tilde{\pi}_0^i = \pi_0^i = p_i, \tag{1.28c}$$

where

$$\tilde{\xi}_s := \int_0^s \left[dy_\tau - \sum_{j=1}^k \tilde{\pi}_\tau^j \tilde{A}_\tau(\theta^j)\tilde{m}_\tau^j d\tau\right]. \tag{1.29}$$

Remark 1.4. From (1.22c) and (1.25) we have

$$\int_0^s \tilde{A}_\tau(\theta^i)\tilde{m}_\tau^i d\tilde{y}_\tau = \int_0^s A_{\nu_\tau}(\theta^i)\tilde{m}_\tau^i d\bar{y}_\tau, \tag{1.30}$$

$$\int_0^s \tilde{A}_\tau(\theta^i)\tilde{\gamma}_\tau^i d\tilde{y}_\tau = \int_0^s A_{\nu_\tau}(\theta^i)\tilde{\gamma}_\tau^i d\bar{y}_\tau \tag{1.31}$$

so that, besides being $\mathcal{F}_s^{\tilde{y}}$-adapted, the process $\tilde{X}_s$ in Proposition 1.3 can also be considered $\mathcal{F}_s^{\bar{y}}$-adapted. Since

$$E\left\{\tilde{x}_s | \mathcal{F}_s^{\tilde{y}}\right\} = E\left\{E\left\{\tilde{x}_s | \mathcal{F}_s^{\tilde{y}}, \theta\right\} | \mathcal{F}_s^{\tilde{y}}\right\} = \sum_{i=1}^k \tilde{m}_s^i \tilde{\pi}_s^i, \tag{1.32}$$

it thus follows that (see (1.26))

$$E\left\{\tilde{x}_s | \mathcal{F}_s^{\bar{y}}\right\} = E\left\{E\left\{\tilde{x}_s | \mathcal{F}_s^{\tilde{y}}\right\} | \mathcal{F}_s^{\bar{y}}\right\} = E\left\{\tilde{x}_s | \mathcal{F}_s^{\tilde{y}}\right\}. \tag{1.33}$$

Proof of Proposition 1.3. Note that the partially observed system $(\tilde{x}_s, \tilde{y}_s)$ defined in (1.23) is nondegenerate and corresponds to the so-called "conditionally Gaussian" case. For the first two sets of components in (1.28) we may thus make use of Theorem 12.1 in [8], whose assumptions can easily be seen to be satisfied; in fact (see (1.22a))

$$\begin{aligned}\int_0^{T+M} |\tilde{a}_s(\theta)|\, I\{s : s \le \Gamma_T\} ds &= \int_0^{T+M} |a_{\nu_s}(\theta)|\, \dot{\nu}_s\, I\{s : s \le \Gamma_T\} ds \\ &= \int_0^T |a_s(\theta)| ds < +\infty,\end{aligned} \tag{1.34}$$

and analogously (see (1.22c))

$$\int_0^{T+M} \tilde{A}_s^2(\theta)\, I\{s : s \le \Gamma_T\} ds = \int_0^{T+M} A_{\nu_s}^2(\theta)\, I\{s : s \le \Gamma_T\} dv_{\nu_s} = \int_0^T A_s^2(\theta) dv_s < +\infty. \tag{1.35}$$

Furthermore, the function $\tilde{b}_s(\theta)$ is continuous and $(\dot{\nu}_s)^{1/2}$ is integrable.

For the components $(\tilde{\pi}_t^i)$ in (1.28) we make use of the general (innovations form) nonlinear filtering equation of Theorem 8.1 in [8], putting, for a generic $i \le k$ and all $t \ge 0$,

$$h_s = h_s(\theta) := I\{\theta = \theta^i\}. \tag{1.36}$$

The assumptions of Theorem 8.1 in [8] are satisfied, and equation (8.10) in [8] with $H = D = 0$ then leads to

$$\begin{aligned} d\tilde{\pi}_s^i = B^{-2} &\left[\pi_s\left(I\{\theta = \theta^i\}\tilde{A}_s(\theta)\tilde{x}_s\right) - \pi_s\left(I\{\theta = \theta^i\}\right)\pi_s\left(\tilde{A}_s(\theta)\tilde{x}_s\right)\right] \\ &\times \left[d\tilde{y}_s - \pi_s\left(\tilde{A}_s(\theta)\tilde{x}_s\right) ds\right], \end{aligned} \tag{1.37}$$

where $\pi_s(Z) := E(Z|\mathcal{F}_s^{\tilde{y}})$. Noting now that

$$\begin{aligned} &\pi_s\left(I\{\theta = \theta^i\}\tilde{A}_s(\theta)\tilde{x}_s\right) \\ &\quad = E\left\{E\left\{I\{\theta = \theta^i\}\tilde{A}_s(\theta)\tilde{x}_s|\mathcal{F}_s^{\tilde{y}}, \theta\right\}|\mathcal{F}_s^{\tilde{y}}\right\} = \tilde{A}_s(\theta^i)\tilde{m}_s^i\,\tilde{\pi}_s^i \end{aligned} \tag{1.38}$$

and analogously

$$\pi_s\left(\tilde{A}_s(\theta)\tilde{x}_s\right) = \sum_{j=1}^k \tilde{A}_s(\theta^j)\tilde{m}_s^j\,\tilde{\pi}_s^j, \tag{1.39}$$

it follows that (1.37) is exactly (1.28c). We are now in a position to come to the proof of Theorem 1.2.

Proof of Theorem 1.2. Note first that, by (1.16), (1.24), and (1.33),

$$m_t^i = E\left\{\tilde{x}_{\Gamma(t)}|\mathcal{F}_{\Gamma(t)}^{\bar{y}}, \theta^i\right\} = E\left\{\tilde{x}_{\Gamma(t)}|\mathcal{F}_{\Gamma(t)}^{\tilde{y}}, \theta^i\right\} = \tilde{m}_{\Gamma(t)}^i. \tag{1.40}$$

Analogously,

$$\gamma_t^i = \tilde{\gamma}_{\Gamma(t)}^i, \qquad \pi_t^i = \tilde{\pi}_{\Gamma(t)}^i. \tag{1.41}$$

From (1.40) and (1.28a) we obtain

$$m_t^i = \tilde{m}_{\Gamma(t)}^i = \tilde{m}_0^i + \int_0^{\Gamma(t)} \tilde{a}_s(\theta^i)\tilde{m}_s^i\, ds + B^{-2}\int_0^{\Gamma(t)} \tilde{A}_s(\theta^i)\tilde{\gamma}_s^i\,[d\tilde{y}_s - \tilde{A}_s(\theta^i)\tilde{m}_s^i\, ds]. \tag{1.42}$$

We now evaluate the integrals in this last expression, namely (see (1.22) and (1.16) with (1.25)),

$$\int_0^{\Gamma(t)} \tilde{a}_s(\theta^i)\tilde{m}_s^i\, ds = \int_0^{\Gamma(t)} a_{\nu_s}(\theta^i)\tilde{m}_s^i\,\dot{\nu}_s ds = \int_0^t a_\tau(\theta^i)m_\tau^i\, d\tau, \tag{1.43}$$

$$
\begin{aligned}
&\int_0^{\Gamma(t)} \tilde{A}_s(\theta^i)\tilde{\gamma}_s^i\,[d\tilde{y}_s - \tilde{A}_s(\theta^i)\tilde{m}_s^i\,ds] \\
&\qquad = \int_0^{\Gamma(t)} A_{\nu_s}(\theta^i)\tilde{\gamma}_s^i\,[A_{\nu_s}(\theta)\dot{v}_{\nu_s}\dot{\nu}_s\tilde{x}_s\,ds + d\eta_{\nu_s} - A_{\nu_s}(\theta^i)\dot{v}_{\nu_s}\dot{\nu}_s\tilde{m}_s^i\,ds] \\
&\qquad = \int_0^t A_\tau(\theta^i)\gamma_\tau^i\,[A_\tau(\theta)x_\tau\,dv_\tau + d\eta_\tau - A_\tau(\theta^i)m_\tau^i\,dv_\tau] \\
&\qquad = \int_0^t A_\tau(\theta^i)\gamma_\tau^i\,[dy - A_\tau(\theta^i)m_\tau^i\,dv_\tau].
\end{aligned} \tag{1.44}
$$

Substituting (1.43) and (1.44) into (1.42) we obtain (1.9a). The remaining equations for (γ_t^i) and (π_t^i) in (1.9) follow analogously.

Coming to the statement of Theorem 1.2 concerning the process (ξ_t) note that, according to Theorem 7.12 in [8], the process $(\tilde{y}_s)$ defined in (1.23) admits the representation

$$
\tilde{y}_s = \int_0^s E\left\{\tilde{A}_\tau(\theta)\tilde{x}_\tau|\mathcal{F}_\tau^{\tilde{y}}\right\}d\tau + \tilde{w}_s = \int_0^s \sum_{j=1}^k \tilde{A}_\tau(\theta^j)\,\tilde{m}_\tau^j\tilde{\pi}_\tau^j\,d\tau + \tilde{w}_s, \tag{1.45}
$$

where $(\tilde{w}_s)$ is an $\mathcal{F}_s^{\tilde{y}}$-standard Wiener process and the second equality follows from (1.39). As a consequence

$$
\tilde{w}_s = \int_0^s \left[d\tilde{y}_s - \sum_{j=1}^k \tilde{A}_\tau(\theta^j)\,\tilde{m}_\tau^j\tilde{\pi}_\tau^j\,d\tau\right]. \tag{1.46}
$$

On the other hand, from the definition of (ξ_t) in (1.10) and from (1.16), (1.15), (1.22), and (1.25), it then follows that

$$
\begin{aligned}
\xi_t &= \int_0^t \left[dy_s - \sum_{j=1}^k A_s(\theta^j)\,m_s^j\pi_s^j\,dv_s\right] \\
&= \int_0^{\Gamma(t)} \left[d\bar{y}_s - \sum_{j=1}^k \tilde{A}_{\nu_s}(\theta^j)\,\tilde{m}_s^j\tilde{\pi}_s^j\,I\{s:\dot{v}_{\nu_s}\neq 0\}ds\right] \\
&= \int_0^{\Gamma(t)} I\{s:\dot{v}_{\nu_s}\neq 0\}\left[d\tilde{y}_s - \sum_{j=1}^k \tilde{A}_{\nu_s}(\theta^j)\,\tilde{m}_s^j\tilde{\pi}_s^j\,ds\right] \\
&= \int_0^{\Gamma(t)} I\{s:\dot{v}_{\nu_s}\neq 0\}\,d\tilde{w}_s,
\end{aligned} \tag{1.47}
$$

from which

$$
\langle\xi\rangle_t = \int_0^{\Gamma(t)} I\{s:\dot{v}_{\nu_s}\neq 0\}ds = \int_0^{\Gamma(t)} dv_{\nu_s} = v_{\nu_{\Gamma(t)}} = v(t). \tag{1.48}
$$

It follows that there exists an $\mathcal{F}_t$-Wiener process (w_t^x) such that

$$
d\xi_t = u_t^{1/2}dw_t^x. \tag{1.49}
$$

On the other hand, the driving random process in (1.9a) can, using (1.10) and (1.49), be expressed as

$$
\begin{aligned}
&[dy_t - A_t(\theta^i) m_t^i dv_t] \\
&\quad = \left[dy_t - \sum_{j=1}^k \pi_t^j A_t(\theta^j) m_t^j dv_t \right] + \left[\sum_{j=1}^k \pi_t^j A_t(\theta^j) m_t^j - A_t(\theta^i) m_t^i \right] dv_t \\
&\quad = u_t^{1/2} dw_t^x + \left[\sum_{j=1}^k \pi_t^j A_t(\theta^j) m_t^j - A_t(\theta^i) m_t^i \right] u_t dt.
\end{aligned} \tag{1.50}
$$

Equation (1.12) now follows from (1.9), thus concluding the first part of the proof of the theorem.

Concerning the uniqueness, we start from the equation (1.9b) for (γ_t^i). Its solution is uniformly bounded, implying the local Lipschitzianity (with integrable Lipschitz constant) of the right-hand side in (1.9b). Coming to (1.9a) for (m_t^i), its uniqueness follows from the linearity in m_t^i of the right-hand side. Finally concerning the equation for (π_t^i), consider the auxiliary process $(\pi_{t\wedge t_n}^i)$, where

$$
t_n := \inf\{t : \max_i |m_t^i| = n\} \wedge T, \tag{1.51}
$$

for which it is easily seen that $t_n \uparrow T$. Due to the linearity of its right-hand side, the equation for $(\pi_{t\wedge t_n}^i)$ now admits a unique solution that coincides with (π_t^i) for $t \le t_n$. If, besides (π_t^i), there is also a solution $(\hat\pi_t^i)$, then $\pi_{t\wedge t_n}^i = \hat\pi_{t\wedge t_n}^i$ so that, by $t_n \uparrow T$ and the continuity of π_t^i, it follows that $\pi_t^i = \hat\pi_t^i \ \ \forall t \in [0,T]$.

2. The control problem.

2.1. Formulation of the control problem. The purpose of the control problem is to choose the control v_t, or equivalently u_t, in (1.1) to maximize the information content of the observations regarding the state. This information content can be measured by the precision of the estimation of x_t on the basis of the observation history $y_0^t := \{y_s;\, 0 \le s \le t\}$, which is given by the inverse of the conditional estimation covariance

$$
\begin{aligned}
\gamma_t = \mathrm{Cov}(x_t |\, \mathcal{F}_t^y) &= E\left\{ \left(x_t - \sum_{i=1}^k \pi_t^i m_t^i \right)^2 \mid \mathcal{F}_t^y \right\} \\
&= \sum_{i=1}^k \pi_t^i \left(\gamma_t^i + (m_t^i)^2 \right) - \left(\sum_{i=1}^k \pi_t^i m_t^i \right)^2 .
\end{aligned} \tag{2.1}
$$

The control objective can therefore be seen as minimizing γ_t for each t. More generally, taking into consideration also a possible penalization of the control effort, we shall consider as control objective the minimization of the following (finite-horizon) functional:

$$
J(u) = E\left\{ \int_0^T \left[f_t^0(\gamma_t) + f_t^1(\gamma_t) u_t \right] dt + \phi^0(\gamma_T) \right\}, \tag{2.2}
$$

where f^0, f^1, and ϕ^0 are continuous functions of polynomial growth in γ. Note that the filter components γ_t^i and π_t^i are uniformly bounded; on the other hand, by (1.10), equation (1.9a) can (see also the proof of Theorem 1.2) be rewritten as

(2.3)
$$dm_t^i = a_t(\theta^i)\, m_t^i\, dt + B^{-2} A_t(\theta^i)\gamma_t^i \left[\sum_{j=1}^k \pi_t^j A_t(\theta^j) m_t^j - A_t(\theta^i) m_t^i\right] u_t dt + B^{-2} A_t(\theta^i)\gamma_t^i d\xi_t,$$

from which it follows that, for a given control u_t, m_t^i possesses uniformly bounded moments of all orders. By (2.1) and the polynomial growth property of f^0, we thus have the existence of the expectation

(2.4)
$$E\{f_t^0(\gamma_t)\} = E\{E\{f_t^0(\gamma_t)|\mathcal{F}_t^y\}\} = E\left\{f_t^0\left(\sum_{i=1}^k \pi_t^i\,(\gamma_t^i + (m_t^i)^2) - \left(\sum_{i=1}^k \pi_t^i m_t^i\right)^2\right)\right\} := E\{F_t^0(X_t)\},$$

with X_t as in (1.8). Analogously, for the remaining two terms in (2.2) we have

$$E\left\{f_t^1(\gamma_t)u_t\right\} = E\left\{F_t^1(X_t)\, u_t\right\}, \tag{2.5}$$

$$E\left\{\phi^0(\gamma_T)\right\} = E\left\{\Phi^0(X_T)\right\}. \tag{2.6}$$

From (2.4)–(2.6), which implicitly define the functions F^0, F^1, and Φ^0, we have that the criterion function in (2.2) is well defined for any control (u_t) satisfying (1.2) and that $J(u)$ can equivalently be represented as

$$J'(u) = E\left\{\int_0^T \left[F_t^0(X_t) + F_t^1(X_t)u_t\right]\, dt + \Phi^0(X_T)\right\} \tag{2.7}$$

for suitable functions F^0, F^1, and Φ^0 that inherit the polynomial growth property of f^0, f^1, and ϕ^0 (the prime distinguishing the representation (2.3) from that in (2.2)).

2.2. The separation property. So far the admissible controls were assumed to be $\mathcal{F}_t^y$-adapted. In line with stochastic control under partial state observation one may investigate whether among the possible $\mathcal{F}_t^y$-adapted optimal controls there is one that is $\mathcal{F}_t^X$- adapted, namely, a function of the observations through the filter values. This is in fact so, and for this purpose consider the two classes of controls

(2.8)
$$\mathcal{L}_0 := \{u : u_t \text{ is } \mathcal{F}_t^y\text{-adapted, is Lipschitz in the sense of (1.5), and satisfies (1.2)}\},$$

$$\mathcal{L}_1 := \{u \in \mathcal{L}_0 : u_t \text{ is in particular } \mathcal{F}_t^X\text{-adapted}\}. \tag{2.9}$$

Given these two classes of controls and recalling (2.2) and (2.7), we have the following *separation theorem*, which allows us to consider, instead of the original control system (1.1) with admissible controls in $\mathcal{L}_0$ and criterion functional $J(u)$ according to (2.2),

the equivalent problem for the filter system (1.12), admissible controls in $\mathcal{L}_1$, and criterion functional $J'(u)$ (see (2.7).

THEOREM 2.1. *Let* $\gamma_0 = \mathrm{Cov}(x_0) > 0$ *and* $f_t^1(\gamma_t) \geq 0$. *Then the strong principle of separation holds, namely,*

$$\inf_{u\in\mathcal{L}_1} J'(u) = \inf_{u\in\mathcal{L}_0} J(u). \tag{2.10}$$

Proof. Since $\mathcal{L}_1 \subseteq \mathcal{L}_0$, we immediately have

$$\inf_{u\in\mathcal{L}_1} J'(u) \geq \inf_{u\in\mathcal{L}_0} J(u), \tag{2.11}$$

so we need to show only the opposite inequality. For this purpose, defining

$$N_A := \left\{ t : \sup_{i\leq k} |A_t(\theta^i)| = 0 \right\}, \tag{2.12}$$

consider the subclasses of controls

$$\bar{\mathcal{L}}_0 = \{u \in \mathcal{L}_0 : u_t = 0 \quad \text{for} \quad t \in N_A\}, \tag{2.13a}$$

$$\bar{\mathcal{L}}_1 = \{u \in \mathcal{L}_1 : u_t = 0 \quad \text{for} \quad t \in N_A\}. \tag{2.13b}$$

For $u \in \mathcal{L}_0$ let

$$\bar{u}_t := u_t\, I\{t \notin N_A\} \in \bar{\mathcal{L}}_0, \tag{2.14}$$

and we have

$$J(\bar{u}) \leq J(u). \tag{2.15}$$

Analogously for $u \in \mathcal{L}_1$. In fact, it is easily seen from (1.9) that $u(\cdot)$ and $\bar{u}(\cdot)$ generate the same process $X_t = \left\{m_t^i, \gamma_t^i, \pi_t^i\right\}_{i=1,\dots,k}$, while due to the nonnegativity of $f_t^1(\gamma)$ and of u_t, one has

$$\int_0^T f_t^1(\gamma_t)\bar{u}_t\, dt \leq \int_0^T f_t^1(\gamma_t)u_t\, dt. \tag{2.16}$$

It follows that

$$\inf_{u\in\bar{\mathcal{L}}_0} J(u) \leq \inf_{u\in\mathcal{L}_0} J(u), \tag{2.17}$$

and, since $\bar{\mathcal{L}}_0 \subseteq \mathcal{L}_0$, we have

$$\inf_{u\in\bar{\mathcal{L}}_0} J(u) = \inf_{u\in\mathcal{L}_0} J(u) \tag{2.18}$$

and, analogously,

$$\inf_{u\in\bar{\mathcal{L}}_1} J'(u) = \inf_{u\in\mathcal{L}_1} J'(u). \tag{2.19}$$

Now let $\bar{u}$ denote a control with $\bar{u}_t = 0$ for $t \in N_A$; in particular, we may think of $\bar{u} \in \bar{\mathcal{L}}_0$. Corresponding to any such given control $\bar{u}$, the σ-algebra $\mathcal{F}_t^y$ generated by

y_s for $0 \le s \le t$ is contained in that $\mathcal{F}_t^X$ generated by $X_s = \{m_s^i, \gamma_s^i, \pi_s^i\}_{i=1,\dots,k}$ for $0 \le s \le t$. In fact, since $\gamma_0 > 0$, passing to its inverse, we have from (1.9b) that $\gamma_t > 0$ for all $t \in [0,T]$. Furthermore, on the complement of N_A there exists at least one value of $i \in \{1,\dots,k\}$ for which $A_t(\theta^i)\gamma_t^i \neq 0$; consequently, taking into account the continuity of $A_t(\theta^i)\gamma_t^i$, one can choose a measurable function $i(t)$ with values in $\{1,\dots,k\}$ so that

$$A_t(\theta^{i(t)})\gamma_t^{i(t)} \neq 0 \qquad \text{for} \quad t \in \bar{N}_A, \tag{2.20}$$

where $\bar{N}_A$ denotes the complement of N_A. This set $\bar{N}_A$ can then be represented in the form $\bar{N}_A = \cup_{i=1}^k \bar{N}_i$, where

$$\bar{N}_i := \{t : i(t) = i\}, \quad i = 1,\dots,k, \tag{2.21}$$

with $A_t(\theta^i)\gamma_t^i \neq 0$ for $t \in \bar{N}_i$, and we have

$$I\{t \notin N_A\} = \sum_i I\{t \notin N_i\}. \tag{2.22}$$

Recalling then that $\bar{u}_t = 0$ for $t \in N_A$, for the observation process we have

$$y_t = \int_0^t I\{s \notin N_A\}\, dy_s = \sum_{i=1}^k \int_0^t I\{s \notin N_i\}\, dy_s. \tag{2.23}$$

On the other hand, from (1.9a) we obtain

$$\begin{aligned} &\int_0^t B^{-2} A_s(\theta^i)\gamma_s^i\, dy_s \\ &\qquad = \int_0^t \left(dm_s^i - a_s(\theta^i) m_s^i\, ds\right) + \int_0^t B^{-2} A_s^2(\theta^i)\gamma_s^i\, m_s^i\, \bar{u}_s\, ds. \end{aligned} \tag{2.24}$$

Multiplying the integrands by $\left(B^{-2}A_s(\theta^i)\gamma_s^i\right)^+ I\{s \notin N_i\}$, where $(\cdot)^+$ denotes the generalized inverse, it follows that

$$\begin{aligned} \int_0^t I\{s \notin N_i\}\, dy_s = &\int_0^t \left(B^{-2}A_s(\theta^i)\gamma_s^i\right)^+ I\{s \notin N_i\} \left[dm_s^i - a_s(\theta^i) m_s^i ds\right] \\ &+ \int_0^t I\{s \notin N_i\} A_s(\theta^i)\, m_s^i\, \bar{u}_s\, ds, \end{aligned} \tag{2.25}$$

from which, taking (2.23) into account, we obtain for y_t the following representation:

$$\begin{aligned} y_t = \sum_{i=1}^k \int_0^t I\{s \notin N_i\} \Big\{ &\left(B^{-2}A_s(\theta^i)\gamma_s^i\right)^+ \left[dm_s^i - a_s(\theta^i) m_s^i ds\right] \\ &+ A_s(\theta^i)\, m_s^i\, \bar{u}_s\, ds \Big\}. \end{aligned} \tag{2.26}$$

This representation shows that, given a control $\bar{u}$ with $\bar{u}_t = 0$ for $t \in N_A$, the process (y_t) is also $\mathcal{F}_t^X$-adapted and so $\mathcal{F}_t^y \subseteq \mathcal{F}_t^X$. As a consequence, we have that $\bar{\mathcal{L}}_0 \subseteq \bar{\mathcal{L}}_1$ so that by (2.18) and (2.19)

$$\inf_{u \in \mathcal{L}_1} J'(u) = \inf_{u \in \bar{\mathcal{L}}_1} J'(u) \le \inf_{u \in \bar{\mathcal{L}}_0} J(u) = \inf_{u \in \mathcal{L}_0} J(u), \tag{2.27}$$

which is the desired opposite inequality.

3. The auxiliary control problem and nearly optimal Lipschitz Markov controls. Our original control problem now consists of controlling the filter process

$$X_t = \left\{m_t^i, \gamma_t^i, \pi_t^i\right\}_{i=1,\dots,k}$$

evolving according to (1.12) in order to minimize (see (2.7))

$$J'(u) = E\left\{\int_0^T \left[F_t^0(X_t) + F_t^1(X_t)u_t\right]\, dt + \Phi^0(X_T)\right\}. \tag{3.1}$$

As follows from Theorem 2.1, we may limit ourselves to considering controls from the class $\mathcal{L}_1$ so that they satisfy also the constraints (1.2), i.e.,

$$\int_0^T u_t dt \le M < +\infty, \quad 0 \le u_t < +\infty. \tag{3.2}$$

It is a nonlinear control problem with unbounded controls, so an optimal solution may not exist. Using the so-called method of discontinuous time transformation (see [10] in a deterministic context, where it is used for the representation of generalized—in particular, discontinuous—solutions in problems with impulse control) next we transform this original problem into an auxiliary problem with bounded controls for which an optimal solution can be shown to exist.

3.1. Method of discontinuous time transformation. To describe the method, let $u \in \mathcal{L}_1$ and consider similarly to section 1.2 the function

$$\Gamma_t := t + \int_0^t u_s ds = t + v_t \tag{3.3}$$

as well as its inverse

$$\nu_s = \Gamma_s^{-1} = \inf\{\tau : \Gamma_\tau > s\}, \tag{3.4}$$

which is an $\mathcal{F}_{\nu_s}^X$-adapted process defined on $[0, \Gamma_T]$, where, due to (3.2), $\Gamma_T \le T + M$. Furthermore, it is absolutely continuous since, for any s_1, s_2 with $s_1 \le s_2$, we have

$$0 < \nu_{s_2} - \nu_{s_1} \le s_2 - s_1, \tag{3.5}$$

so it is almost everywhere differentiable on $[0, \Gamma_T]$ with derivative (see (3.3))

$$\dot\nu_s = \left[\dot\Gamma_t\right]^{-1}_{|t=\nu_s} = (1 + u_t)^{-1}_{|t=\nu_s}, \tag{3.6}$$

which is $\mathcal{F}_{\nu_s}^X$-adapted and satisfies

$$0 < \dot\nu_s \le 1. \tag{3.7}$$

Then considering the process-pair

$$Z_s := X_{\nu_s}, \qquad \mu_s := v_{\nu_s}, \tag{3.8}$$

where (see section 1.1)

$$v_t = \int_0^t u_\tau d\tau, \tag{3.9}$$

we have the following lemma.

LEMMA 3.1. *Let the process* (X_t) *satisfy* (1.12) *for some* $u \in \mathcal{L}_1$. *Then there exists an* $\mathcal{F}_{\nu_s}$*-Wiener process* (w_s^Z) *such that the process-triple* (Z_s, μ_s, ν_s) *satisfies, for* $s \in [0, \Gamma_T]$,

$$\begin{aligned} dZ_s &= \alpha_s F_{\nu_s}(Z_s)ds + (1-\alpha_s)B_{\nu_s}(Z_s)ds \\ &\quad + (1-\alpha_s)^{1/2} G_{\nu_s}(Z_s)dw_s^Z, \quad Z_0 = X_0, \end{aligned} \tag{3.10a}$$

$$d\mu_s = (1-\alpha_s)\, ds, \quad \mu_0 = 0, \tag{3.10b}$$

$$d\nu_s = \alpha_s ds, \quad \nu_0 = 0, \tag{3.10c}$$

where the functions F, B, *and* G *are as in* (1.12), *the control* α *is given by*

$$\alpha_s = \dot{\nu}_s, \tag{3.11}$$

and it is $\mathcal{F}_s^Z$*-adapted and satisfies* $0 < \alpha_s \le 1$. *Furthermore, the solution of* (3.10) *is unique.*

Proof. By (3.7) we have $0 < \alpha_s \le 1$. By (1.12) the process $Z_s = X_{\nu_s}$ satisfies

$$Z_s = X_{\nu_s} = X_0 + \int_0^{\nu_s} F_t(X_t)dt + \int_0^{\nu_s} B_t(X_t)u_t dt + \int_0^{\nu_s} G_t(X_t)u_t^{1/2} dw_t^X. \tag{3.12}$$

Taking into account the identities

$$\nu_{\Gamma_t} = t, \quad \Gamma_{\nu_s} = s \tag{3.13}$$

valid for $t \in [0,T]$ and $s \in [0, \Gamma_T]$, we derive next a representation for the integrals in the right-hand side of (3.12), namely, performing the change of variables $t = \nu_\tau$,

$$\int_0^{\nu_s} F_t(X_t)dt = \int_0^{\Gamma_{\nu_s}} F_{\nu_\tau}(X_{\nu_\tau})d\nu_\tau = \int_0^s F_{\nu_\tau}(X_{\nu_\tau})\, \alpha_\tau d\tau, \tag{3.14}$$

$$\begin{aligned} \int_0^{\nu_s} B_t(X_t)u_t dt &= \int_0^{\nu_s} B_t(X_t)\frac{u_t}{1+u_t}(1+u_t)\, dt \\ &= \int_0^{\Gamma_{\nu_s}} B_{\nu_\tau}(X_{\nu_\tau}) \left(1 - \frac{1}{1+u_t}\right)_{|t=\nu_\tau} d\Gamma_{\nu_\tau} \\ &= \int_0^s B_{\nu_\tau}(X_{\nu_\tau})\, (1-\alpha_\tau)\, d\tau, \end{aligned} \tag{3.15}$$

$$\int_0^{\nu_s} G_t(X_t)u_t^{1/2} dw_t^X = \int_0^{\Gamma_{\nu_s}} G_{\nu_\tau}(X_{\nu_\tau})\, u_{\nu_\tau}^{1/2} dw_{\nu_\tau}^X. \tag{3.16}$$

The process $w_{\nu_\tau}^X$ is an $\mathcal{F}_{\nu_\tau}^X$-adapted, conditionally Gaussian martingale with continuous trajectories and quadratic variation

$$\langle w^X \rangle_{\nu_\tau} = \nu_\tau = \int_0^\tau \alpha_u du. \tag{3.17}$$

There exists therefore an $\mathcal{F}_{\nu_s}$-Wiener process w_s^Z such that

$$w_{\nu_\tau}^X = \int_0^\tau (\alpha(u))^{1/2} dw_u^Z. \tag{3.18}$$

Substituting (3.18) into (3.16) we then obtain (see also (3.6) and (3.11))

$$\begin{aligned} \int_0^{\nu_s} G_t(X_t) u_t^{1/2} dw_t^X &= \int_0^s G_{\nu_\tau}(X_{\nu_\tau}) \left(\frac{u_{\nu_\tau}}{1+u_{\nu_\tau}}\right)^{1/2} dw_\tau^Z \\ &= \int_0^s G_{\nu_\tau}(X_{\nu_\tau}) (1-\alpha_\tau)^{1/2} dw_\tau^Z. \end{aligned} \tag{3.19}$$

Using (3.14), (3.15), and (3.19) in (3.12) we obtain (3.10a) for Z_s. Analogously, for the process μ_s we obtain

$$\begin{aligned} \mu_s = \int_0^{\nu_s} u_t dt &= \int_0^{\Gamma_{\nu_s}} u_{\nu_\tau} d\nu_\tau \\ &= \int_0^s \frac{u_{\nu_\tau}}{1+u_{\nu_\tau}} d\tau = \int_0^s (1-\alpha_\tau) d\tau. \end{aligned} \tag{3.20}$$

The uniqueness of a strong solution of (3.10) follows analogously to that of system (1.12) (see Theorem 1.2).

Next we shall establish a relationship converse to Lemma 3.1. Therefore consider (3.10) as a system controlled by a process α_s that is $\mathcal{F}_s^Z$-adapted and satisfies $0 < \alpha_s \le 1$ for $0 \le s \le S$, where S is an $\mathcal{F}_s^{Z,\mu}$-stopping time given by

$$S := S_\nu \wedge S_\mu, \tag{3.21}$$

with

$$S_\nu := \inf\{s : \nu_s = T\}, \tag{3.22a}$$

$$S_\mu := \inf\{s : \mu_s = M\}. \tag{3.22b}$$

By the fact that (see (3.10)) $\nu_s + \mu_s = s$, we have $S \le T + M$.

Define $\mathcal{A}$ as the class of $\mathcal{F}_s^Z$-adapted controls α satisfying $\alpha_s \in (0,1]$ for $s \in [0,S]$ and where, if $S = S_\mu$, we extend its definition, letting $\alpha_s = 1$ for $S_\mu < s \le T + M$. We then have the following lemma.

LEMMA 3.2. *Given* (3.10), *let* $\alpha \in \mathcal{A}$. *Then, putting* $\Gamma_t := \inf\{s : \nu_s > t\}$, *there exists an* $\mathcal{F}_{\Gamma_t}^{Z,\mu}$*-adapted control* (u_t) *satisfying* (1.2) (*see also* (3.2)) *and an* $\mathcal{F}_{\Gamma_t}^{Z,\mu}$*-Wiener process* (w_t^X) *such that, for* $t \le T$, *the processes*

$$X_t := Z_{\Gamma_t}, \quad v_t := \mu_{\Gamma_t} \tag{3.23}$$

satisfy (1.12) *with* $v_t = \int_0^t u_\tau d\tau$. *The control* (u_t) *is furthermore given by*

$$u_t = \dot{\Gamma}_t - 1 = \alpha_{\Gamma_t}^{-1} - 1 \tag{3.24}$$

and $\mathcal{F}_{\Gamma_t}^{Z,\mu} = \mathcal{F}_t^{X,v}$.

Proof. From its definition in (3.24), the control u_t is trivially $\mathcal{F}^{Z,\mu}_{\Gamma_t}$-adapted and satisfies (1.2a). Furthermore, under the assumptions of the lemma, we have

$$v_t = \int_0^t u_\tau d\tau = \int_0^t \left(\alpha_{\Gamma_t}^{-1} - 1\right) d\tau = \int_0^t \frac{1-\alpha_{\Gamma_\tau}}{\alpha_{\Gamma_\tau}} d\tau$$
$$= \int_0^t (1-\alpha_{\Gamma_\tau}) d\Gamma_\tau = \int_0^{\Gamma_\tau} (1-\alpha_s)\, ds = \mu_{\Gamma_\tau} \le M$$

so that (1.2b) also is satisfied and $v_t = \mu_{\Gamma_t}$. It remains to show that $X_t = Z_{\Gamma_t}$ satisfies (1.12). For this purpose note that, based on (3.14) and (3.15) as well as (3.24), we may write

$$\int_0^{\Gamma_t} F_{\nu_s}(Z_s)\, \alpha_s\, ds = \int_0^t F_\tau(Z_{\Gamma_\tau})\, d\tau, \tag{3.25}$$

$$\int_0^{\Gamma_t} B_{\nu_s}(Z_s)\, (1-\alpha_s)\, ds = \int_0^t B_\tau(Z_{\Gamma_\tau})\, u_\tau\, d\tau, \tag{3.26}$$

and, finally, based on (3.19)

$$\begin{aligned} \int_0^{\Gamma_t} G_{\nu_s}(Z_s)\, (1-\alpha_s)^{1/2}\, dw_s^Z &= \int_0^t G_\tau(Z_{\Gamma_\tau}) \left(\frac{u_\tau}{1+u_\tau}\right)^{1/2} dw^Z_{\Gamma_\tau} \\ &= \int_0^t G_\tau(Z_{\Gamma_\tau})\, u_\tau^{1/2} dw_\tau^X, \end{aligned} \tag{3.27}$$

where (see also (3.18) and (3.24))

$$w_t^X = \int_0^t \frac{dw^Z_{\Gamma_\tau}}{(1+u_\tau)^{1/2}} \tag{3.28}$$

is a continuous $\mathcal{F}^{Z,\mu}_{\Gamma_t} = \mathcal{F}^{X,v}_t$-martingale with quadratic variation

$$\langle w^X \rangle_t = \int_0^t \frac{d\Gamma_\tau}{1+u_\tau} = t,$$

and therefore an $\mathcal{F}^{X,v}_t$-Wiener process.

The results obtained in the two lemmas above allow us to consider, instead of the original controlled system (1.12) with unbounded controls, the system (3.10) where the controls are bounded. We are now going to define more precisely the control problem corresponding to this latter system, which we shall call the *auxiliary control problem.*

3.2. The auxiliary control problem. This auxiliary problem concerns the controlled system (Z_s, μ_s, ν_s) satisfying (3.10) with controls from an enlarged class $\mathcal{A}_0$ consisting of the controls in $\mathcal{A}$ (defined in Lemma 3.2), where we also allow the value $\alpha_s = 0$, i.e.,

$$\mathcal{A}_0 := \{\alpha \in \mathcal{A} \mid 0 \le \alpha_s \le 1\}. \tag{3.29}$$

This enlargement of the class of controls guarantees, as we shall see, the existence of an optimal solution for the auxiliary problem. On the other hand, through the

correspondence (3.24), this is equivalent to allowing unbounded controls in the original problem.

As a cost functional to be minimized we consider

$$J(\alpha) = E\left\{\int_0^S \left[\alpha_s F^0_{\nu_s}(Z_s) + (1-\alpha_s)F^1_{\nu_s}(Z_s)\right]\, ds + \bar{\Phi}^0_{\nu_S}(Z_S)\right\}, \tag{3.30}$$

where S is the $\mathcal{F}^{Z,\mu}_s$-stopping time defined in (3.21) and (3.22), F^0 and F^1 are as in (3.1) or (2.7), and the terminal cost function is given by

$$\bar{\Phi}^0_\nu(Z) = \begin{cases} \Phi^0(Z) & \text{if } S = S_\nu, \\ \Phi^0(\psi_T(\nu, Z)) + \int_\nu^T F^0_s(\psi_s(\nu, Z))\, ds & \text{if } S = S_\mu, \end{cases} \tag{3.31}$$

with $\psi_s(\nu, Z)$ being the solution on $[\nu, T]$ of the deterministic equation

$$\dot{\psi}_s = F_s(\psi_s), \tag{3.32}$$

having initial condition $\psi_\nu(\nu, Z) = Z$. F is as in (1.12) (see also (3.10a)), and F^0, F^1, and Φ^0 are the same as in $J'(u)$ (see (3.1) or (2.7)).

Remark 3.3. The function $\psi_s(\nu, Z)$ satisfies $\psi_\nu(\nu, Z) = Z$, is continuous in all variables, and has linear growth with respect to Z, since the function $F_s(X)$ is continuous and Lipschitz in X for each s.

3.3. Relationship between the original and auxiliary problems. In this section we will show the correspondence existing between the cost functionals $J'(u)$ in (3.1) and $J(\alpha)$ in (3.30). We have in fact the following proposition.

PROPOSITION 3.4. *If $u \in \mathcal{L}_1$ is given and α is according to* (3.11) *and* (3.6)*, or $\alpha \in \mathcal{A}$ is given, $S = S_\nu$, and u is according to* (3.24)*, then*

$$J'(u) = J(\alpha). \tag{3.33}$$

When $\alpha \in \mathcal{A}$ but $S = S_\mu$ so that $\nu_S < T$, then (3.33) *continues to hold with u according to* (3.24) *if (see the definition of the class $\mathcal{A}$ before Lemma* 3.2*) one puts $\alpha_s = 1$ for $S_\mu < s \le T + M$.*

Proof. For the first part of the statement note the following: given a control $u \in \mathcal{L}_1$ and letting (see (3.11)) $\alpha_s = \dot{\nu}_s$, then since (see (3.8) and (3.9)) $\mu_s = \int_0^{\nu_s} u_\tau\, d\tau$ and $u \in \mathcal{L}_1$ satisfies (3.2), we have $S = S_\nu = \Gamma_T$ with $\Gamma(\cdot)$ as in (3.3) or, equivalently, as in the statement of Lemma 3.2. As a consequence we have

$$\nu_S = T, \quad Z_S = X_T, \tag{3.34}$$

and, by considerations analogous to those leading to (3.14) and (3.15), we then obtain

$$\int_0^T F^0_t(X_t)\, dt = \int_0^S F^0_{\nu_s}(Z_s)\, \alpha_s\, ds, \tag{3.35}$$

$$\int_0^T F^1_t(X_t)\, u_t\, dt = \int_0^S F^1_{\nu_s}(Z_s)\, (1-\alpha_s)\, ds. \tag{3.36}$$

On the other hand, given a control $\alpha \in \mathcal{A}$, if $S = S_\nu$, the same relations (3.34)–(3.36) hold. Combining these considerations with Lemmas 3.1 and 3.2 we obtain the first part of the proposition. The second part follows immediately, taking (3.31) into account.

Remark 3.5. The previous equivalence considerations are valid for $\alpha \in \mathcal{A}$, i.e., such that $\alpha_s > 0$. For the purpose of obtaining existence of an optimal solution for the auxiliary problem, we shall allow also controls in $\mathcal{A}_0$ (see (3.29)) so that α_s might be equal to zero on some subintervals of $[0, S]$. Correspondingly, on these subintervals, ν_s will be constant implying that its inverse $\Gamma_t = \inf\{s : \nu_s > t\}$ jumps. Consequently also $X_t = Z_{\Gamma_t}$ and $v_t = \mu_{\Gamma_t}$ will jump and can therefore not be a solution of (1.12) for any measurable control. In other words, while the auxiliary control problem admits an optimal solution, there may not exist a corresponding optimal solution for the original problem. We shall therefore determine nearly optimal (ε-optimal) solutions for the original problem.

Letting

$$\mathcal{A}_0^L := \{\alpha \in \mathcal{A}_0 : \alpha_t = \alpha_t(Z_t, \mu_t) \text{ a Lipschitz function}\} \tag{3.37}$$

and, analogously, for $\mathcal{A}^L$, we first prove the following.

PROPOSITION 3.6. *For any control $\alpha \in \mathcal{A}_0^L$ there exists a sequence of controls $\alpha^k \in \mathcal{A}^L$ obtained as*

$$\alpha_s^k = \frac{1}{(k+1)} + \frac{k}{(k+1)}\alpha_s, \tag{3.38}$$

where $s \in [0, S]$ if $S = S_\nu$ and $s \in [0, T + M]$ if $S = S_\mu$, such that

$$\lim_{k\to\infty} J(\alpha^k) = J(\alpha). \tag{3.39}$$

Proof. Given $\alpha \in \mathcal{A}_0^L$, let S be the corresponding stopping time defined according to (3.21) and (3.22). Define the sequence $\alpha^k \in \mathcal{A}^L$ as in (3.38). Also let

$$S^k := S_\nu^k \wedge S_\mu^k, \tag{3.40}$$

where S_ν^k and S_μ^k are defined according to (3.22) with $\nu = \nu^k$ and $\mu = \mu^k$ that correspond to α^k via (3.10). The sequence α^k is monotonically decreasing and converges to α.

Let us first show that the sequence S^k converges for all ω to S. In fact, since $\alpha_s^k \geq \alpha_s$ and

$$\alpha_s^{k+1} - \alpha_s^k = \frac{(\alpha_s - 1)}{(k+1)(k+2)} \leq 0$$

for the stopping times $S_\nu^k = \inf\left\{s : \nu_s^k = \int_0^s \alpha_\tau^k d\tau = T\right\}$, we have

$$S_\nu^k \leq S_\nu^{k+1} \leq S_\nu, \tag{3.41}$$

and analogously for $S_\mu^k = \inf\left\{s : \mu_s^k = \int_0^s (1 - \alpha_\tau^k)\, d\tau = M\right\}$ we get

$$S_\mu \leq S_\mu^{k+1} \leq S_\mu^k. \tag{3.42}$$

We can now prove that $\lim_{k\to\infty} S_\nu^k = S_\nu$. From (3.41) the limit of S_ν^k exists and, denoting it by $\bar{S}_\nu$, we have $\bar{S}_\nu \leq S_\nu$. By the uniform convergence of α^k to α we have that ν_s^k converges to ν_s uniformly on compact subintervals of $[0, S_\nu]$, where, we recall, $S_\nu \leq T + M$. In addition

$$\nu_{S_\nu^k}^k = T, \quad \nu_{S_\nu} = T.$$

Let us determine the value of $\nu_{\bar{S}_\nu}$. From

$$\nu^k_{S^k_\nu} - \nu_{\bar{S}_\nu} = \nu^k_{S^k_\nu} - \nu_{S^k_\nu} + \nu_{S^k_\nu} - \nu_{\bar{S}_\nu},$$

the uniform convergence of ν^k_s to ν_s, and the continuity of ν_s it follows that

$$\nu_{\bar{S}_\nu} = \lim_{k\to\infty} \nu^k_{S^k_\nu} = T. \tag{3.43}$$

Now suppose that $\bar{S}_\nu < S_\nu$ for some ω; then (3.43) contradicts the definition of S_ν as the stopping time according to (3.22), and consequently $\lim_{k\to\infty} S^k_\nu = S_\nu$. Analogously we obtain $\lim_{k\to\infty} S^k_\mu = S_\mu$ and finally

$$\lim_{k\to\infty} S^k = \lim_{k\to\infty} \left(S^k_\nu \wedge S^k_\mu\right) = S_\nu \wedge S_\mu = S.$$

Consider next the sequence (Z^k_s), with Z^k_s obtained as solutions of (3.10) corresponding to $\alpha = \alpha^k$. For each given $N > 0$ determine the sequence of stopping times $\theta^{N,k} = \min\{S, S^k, \tau^N, \tau^{N,k}\}, k = 1, 2, \ldots,$ where

$$\tau^N = \inf\{s : ||Z_s|| = N\}, \tag{3.44a}$$

$$\tau^{N,k} = \inf\{s : ||Z^k_s|| = N\}. \tag{3.44b}$$

The properties of the coefficients in the right-hand side of (3.10) guarantee that the trajectories of Z^k_s are a.s. continuous so that, for any $\alpha \in \mathcal{A}^L_0$ and any k,

$$\tau^N,\ \tau^{N,k} \uparrow \infty \qquad \text{a.s. as} \quad N \to \infty. \tag{3.45}$$

Consider next the sequences of processes $Z^k_{s\wedge\theta^{N,k}}$ as well as $Z_{s\wedge\theta^{N,k}}$. By the continuity and the local Lipschitzianity with respect to Z of the functions in the right-hand side of (3.10) as well the uniform convergence of $\nu^k_s \to \nu_s$ and of $\alpha^k_s \to \alpha_s$ on compact subsets of $[0, S_\nu]$, we obtain

$$\begin{aligned} &\sup_{\tau\le s\wedge\theta^{N,k}} E\,||\,Z^k_{\tau\wedge\theta^{N,k}} - Z_{\tau\wedge\theta^{N,k}}\,||^2 \\ &\qquad \le C_1 \int_0^{s\wedge\theta^{N,k}} \sup_{\tau\le u} E\,||\,Z^k_{\tau\wedge\theta^{N,k}} - Z_{\tau\wedge\theta^{N,k}}\,||^2 du + C_2\varepsilon_k, \end{aligned} \tag{3.46}$$

where $\lim_{k\to\infty} \varepsilon_k = 0$. Applying the Gronwall–Bellman inequality to (3.45) we get for all $s \le S$

$$\lim_{k\to\infty} \sup_{\tau\le s\wedge\theta^{N,k}} E\,||\,Z^k_\tau - Z_\tau||^2 = 0. \tag{3.47}$$

Together with k now also let $N \uparrow \infty$; then $\theta^{N,k} \to S$ a.s., implying that for all $s < S$ we have the convergence in $\mathcal{L}^2$ of Z^k_s to Z_s and, by the continuity of Z^k_s in $s = S$, also of $Z^k_{S^k}$ to Z_S.

This convergence in turn implies the convergence in probability of Z^k_s to Z_s for all $s \in [0, S]$ as well as that of $Z^k_{S^k}$ to Z_S. In addition we have the uniform integrability of Z^k_s on $\Omega \times [0, T + M]$ and of $Z^k_{S^k}$ on Ω since, by the linear growth in Z of the functions in the right-hand side of (3.10), we have for $p \ge 1$

$$\int_0^{T+M} E\,||Z^k_s||^p ds \le L < \infty, \quad E\,||Z^k_{S^k}||^p \le L < \infty. \tag{3.48}$$

As a consequence, and due to the polynomial growth in Z of the functions F^0, F^1, and $\bar{\Phi}^0$, we may pass to the limit in (3.30), thus concluding the proof of the proposition.

Combining the result of Proposition 3.6 with that of Proposition 3.4 we immediately obtain the following corollary.

COROLLARY 3.7. *For any control $\alpha \in \mathcal{A}_0^L$ there exists a sequence of Lipschitz Markov (feedback) controls $u_t^k = u_t^k(X_t, v_t)$ satisfying* (1.2) *such that*

$$\lim_{k\to\infty} J'(u^k) = J(\alpha). \tag{3.49}$$

Proof. From (3.38) we immediately have that, if $\alpha_t = \alpha_t(Z_t, \mu_t)$ with $\alpha_t(\cdot)$ Lipschitz, α^k also is Lipschitz. Recall next that the u^k corresponding to α^k is defined by (3.24) and (see Lemma 3.2) satisfies (1.2). Using the relationship (3.23) as well as the fact that $\alpha_s^k \geq \frac{1}{1+k}$, for such a u^k we then have $u_t^k = u_t^k(X_t, v_t)$ with $u_t^k(\cdot)$ Lipschitz. The result then follows by combining the previous considerations with Proposition 3.4.

3.4. Nearly optimal Lipschitz Markov controls. In this section we study first the existence of an optimal solution for the auxiliary control problem in the class $\mathcal{A}_0$ as well as the existence of a nearly optimal Lipschitz Markov control for the original problem. We then return to the relationship between the original and the auxiliary control problems, showing the usefulness of the auxiliary problem to obtain a nearly optimal Lipschitz Markov control in the original problem.

For the first part we have the following theorem.

THEOREM 3.8. *In the class $\mathcal{A}_0$ there exists an optimal control for the auxiliary problem, and it is of the Markov (feedback) type*

$$\alpha_s^0 = \alpha_s^0(Z_s, \mu_s). \tag{3.50}$$

Furthermore, for any $\varepsilon > 0$ there exists a Lipschitz Markov control $\alpha_s^{0,\varepsilon}(Z_s, \mu_s) \in \mathcal{A}_0$ such that

$$J(\alpha^0) = \inf_{\alpha\in\mathcal{A}_0} J(\alpha) \geq J(\alpha^{0,\varepsilon}) - \varepsilon. \tag{3.51}$$

Proof. Concerning the existence of an optimal solution in $\mathcal{A}_0$ note first that, under our assumptions, the set

$$K(\nu, Z) = \left\{ \begin{array}{l} \alpha F + (1-\alpha)B, \quad 1-\alpha, \quad \alpha, \quad (1-\alpha)GG^T \\ \alpha F^0 + (1-\alpha)F^1 \end{array} \right\}_{|\alpha\in[0,1]} \tag{3.52}$$

is, for all (ν, Z), bounded, closed, and convex. This allows us to apply known results on the existence of optimal controls—in particular, Theorem 5.15 in [3]—since the functions F, B, GG^T, F^0, and F^1 satisfy the growth conditions required in that theorem and the admissible control set $\mathcal{A}_0$ is not empty. More precisely, according to Theorem 5.15 in [3] we have that in the auxiliary problem there exists an optimal control in the class $\mathcal{A}_0$ and it is furthermore of the Markov (feedback) type, namely, as in (3.50). The existence of a Lipschitz Markov control can be obtained by using, e.g., results in [7] (see also [6]).

Before coming to the main result of the second part of this section, for later convenience we state the following lemma, whose proof can be obtained via a truncation argument analogous to that used in the proof of Theorem 2.1 concerning the uniqueness of the solution of (1.9).

LEMMA 3.9. *Any Lipschitz Markov control* $u_t = u_t(X_t, v_t)$ *belongs to the class* $\mathcal{L}_1$*; in particular, as a function of* y_0^t *it is Lipschitz in the sense of* (1.5).

THEOREM 3.10. *The following equality holds between the optimal values of the original and the auxiliary control problems*

$$\inf_{u\in\mathcal{L}_0} J(u) = \inf_{u\in\mathcal{L}_1} J'(u) = \inf_{\alpha\in\mathcal{A}_0^L} J(\alpha) \tag{3.53}$$

with $\mathcal{A}_0^L$ *as defined in* (3.37).

Furthermore, given an ε*-optimal control* $\alpha^\varepsilon \in \mathcal{A}_0^L$*, let* k *be so large that for the control* $u^{k,\varepsilon}$*, obtained via* (3.24) *from a Lipschitz control* $\alpha^{k,\varepsilon} \in \mathcal{A}^L$ *that in turn is obtained from* α^ε *via* (3.38)*, we have*

$$J'(u^{k,\varepsilon}) \le J(\alpha^\varepsilon) + \varepsilon. \tag{3.54}$$

Then $u^{k,\varepsilon}$ *belongs to* $\mathcal{L}_1$ *and is a* 4ε*-optimal Lipschitz Markov control for the original problem.*

Proof. The first equality in (3.53) follows from Theorem 2.1. For the second equality note first that, letting u^ε be an ε-optimal control in $\mathcal{L}_1$, α^ε be the corresponding control in $\mathcal{A}_0$ obtained according to (3.11) and (3.6), and α^0 and $\alpha^{0,\varepsilon}$ be the optimal and nearly optimal controls of Theorem 3.8 and using Proposition 3.4 we obtain

$$\begin{aligned}\inf_{u\in\mathcal{L}_1} J'(u) &\ge J'(u^\varepsilon) - \varepsilon = J(\alpha^\varepsilon) - \varepsilon \ge \inf_{\alpha\in\mathcal{A}_0} J(\alpha) - \varepsilon \\ &= J(\alpha^0) - \varepsilon \ge J(\alpha^{0,\varepsilon}) - 2\varepsilon \ge \inf_{\alpha\in\mathcal{A}_0^L} J(\alpha) - 2\varepsilon.\end{aligned} \tag{3.55}$$

On the other hand let $\alpha^\varepsilon \in \mathcal{A}_0^L$ be ε-optimal. Starting from this α^ε, construct the Lipschitz control $\alpha^{k,\varepsilon} \in \mathcal{A}^L$ according to (3.38), and let $u^{k,\varepsilon}$ be the corresponding Lipschitz Markov control obtained according to (3.24). Then, using Corollary 3.7 and the fact that (see Lemma 3.9) if $u_t^{k,\varepsilon} = u_t^{k,\varepsilon}(X_t, v_t)$ is Lipschitz as a function of (X_t, v_t), then it is also in $\mathcal{L}_1$, we have for k sufficiently large

$$\inf_{\alpha\in\mathcal{A}_0^L} J(\alpha) \ge J(\alpha^\varepsilon) - \varepsilon \ge J'(u^{k,\varepsilon}) - 2\varepsilon \ge \inf_{u\in\mathcal{L}_1} J'(u) - 2\varepsilon. \tag{3.56}$$

Combining (3.55) with (3.56) one obtains

$$\inf_{u\in\mathcal{L}_1} J'(u) + 2\varepsilon \ge \inf_{\alpha\in\mathcal{A}_0^L} J(\alpha) \ge \inf_{u\in\mathcal{L}_1} J'(u) - 2\varepsilon, \tag{3.57}$$

from which, due to the arbitrariness of $\varepsilon > 0$, the second equality in (3.53) follows. From (3.55) and (3.56) one also obtains

$$J'(u^{k,\varepsilon}) \le \inf_{u\in\mathcal{L}_1} J'(u) + 4\varepsilon, \tag{3.58}$$

i.e., the 4ε-optimality of $u^{k,\varepsilon}$.

Concluding remarks. From Theorem 3.10 we have that the optimal value $\inf_{u\in\mathcal{L}_0} J(u)$ of the original control problem can be determined by solving the auxiliary problem. However, as mentioned in Remark 3.5, while the auxiliary problem admits an optimal control, there may not exists a control for the original problem for which the optimal value is achieved. There are essentially two reasons for this:

- The filter process X_t corresponding to the optimal solution of the auxiliary problem may jump, so it can be represented as solution of (1.12) only if we allow the control also to have infinite power.
- Even if we allow infinite control power, the representation of the possible jumps of the filter process by means of (1.22) may require anticipative impulse controls.

In fact, due to the linearity in the control α of the Hamilton–Jacobi–Bellman equation of the auxiliary control problem, there will be intervals on which the optimal control α_s^0 for this latter problem will be either zero or one.

If $\alpha_s^0 = 1$, for the corresponding control u_t^0 of the original problem, obtained from α^0 via (3.24), we have $u_t^0 = 0$, and this motivated the extended study of the filter problem in section 1.2.

If, however, $\alpha_s^0 = 0$ on some interval $[s_1, s_2]$, the corresponding ν_s^0 (see (3.10c)) is constant, implying a jump for the inverse function $\Gamma_t^0 = \inf\{s : \nu_s^0 > t\}$. Since (see Lemma 3.2) $X_t = Z_{\Gamma_t}$, $v_t = \mu_{\Gamma_t}$, this then implies that X_t and v_t also jump and can therefore be a solution of (1.12) only if we allow controls u_t with infinite power. This implies an impulse control for the original problem at the moment $t_1 = \nu_{s_1}$, leading (see (1.1b)) to a discrete observation with intensity (see the relation $v_t = \mu_{\Gamma_t}$ in (3.23)) $\Delta v_{t_1} = s_2 - s_1$. Since s_2 is $\mathcal{F}_{s_2}^{Z,\nu} = \mathcal{F}_{t_1^+}^{X,v}$-measurable, this Δv_{t_1} cannot be determined on the basis of the observations up to time t_1; i.e., the control would be anticipative.

Our approach, based on the search of a nearly optimal control, avoids this problem. In fact, the control $u^{k,\varepsilon} \in \mathcal{L}_1$ obtained according to Theorem 3.10 (namely, obtained via (3.24) from a nearly optimal control $\alpha^{k,\varepsilon}$ of the auxiliary problem that belongs to $\mathcal{A}^L$ and thus satisfies $\alpha^{k,\varepsilon} > 0$) has finite power, is $\mathcal{F}_t^{X,v}$-measurable, and allows us to approximate arbitrarily closely the optimal value $\inf_{u\in\mathcal{L}_0} J(u) = \inf_{u\in\mathcal{L}_1} J'(u)$ of the objective function of the original problem.

REFERENCES

[1] F.L. Chernous'ko and V.B. Kolmanovskii, *Optimal Control under Random Perturbations*, Nauka, Moscow, 1978 (in Russian).

[2] F.N. Grigor'ev, N.A. Kuznetsov, and A.P. Serebrovskii, *The Control of Observations in Automatic Systems*, Nauka, Moscow, 1986 (in Russian).

[3] U.G. Haussmann and J.P. Lepeltier, *On the existence of optimal controls*, SIAM J. Control Optim., 28 (1990), pp. 851–902.

[4] H.J. Kushner, *On the optimal timing of observations for linear control systems with unknown initial states*, IEEE Trans. Automat. Control, AC-9 (1964), pp. 144–145.

[5] H.J. Kushner, *Probability Methods for Approximations in Stochastic Control and for Elliptic Equations*, Academic Press, New York, 1977.

[6] H.J. Kushner and W. Runggaldier, *Nearly optimal state feedback controls for stochastic systems with wideband noise disturbances*, SIAM J. Control Optim., 25 (1987), pp. 298–315.

[7] N.V. Krylov, *Controlled Diffusion Processes*, Springer-Verlag, New York, 1980.

[8] R.Sh. Liptser and A.N. Shiryaev, *Statistics of Random Processes* I *and* II, Springer-Verlag, 1977.

[9] N.A. Kuznetsov, R.Sh. Liptser, and A.P. Serebrovskii, *Optimal control and data processing in continuous time (linear system and quadratic functional)*, Automat. Remote Control, 41 (1980), pp. 1369–1374.

[10] B.M. Miller, *Optimization of dynamic systems with a generalized control*, Automat. Remote Control, 50 (1989), pp. 733–742.

[11] B.M. Miller, *Generalized optimization in the problems of observation control*, Automat. Remote Control, 52 (1991), pp. 83–92.

SIAM J. CONTROL OPTIM.
Vol. 35, No. 3, pp. 1053–1091, May 1997

014

STATE MAPS FOR LINEAR SYSTEMS*

PAOLO RAPISARDA† AND J. C. WILLEMS‡

Abstract. Modeling of physical systems consists of writing the equations describing a phenomenon and yields as a result a set of differential-algebraic equations. As such, state-space models are not a natural starting point for modeling, while they have utmost importance in the simulation and control phase. The paper addresses the problem of computing state variables for systems of linear differential-algebraic equations of various forms. The point of view from which the problem is considered is the behavioral one, as put forward in [J. C. Willems, *Automatica J. IFAC*, 22 (1986), pp. 561–580; *Dynamics Reported*, 2 (1989), pp. 171–269; *IEEE Trans. Automat. Control*, 36 (1991), pp. 259–294].

Key words. behavioral system theory, state map, differential-algebraic equations

AMS subject classifications. 93A30, 93B25

PII. S0363012994268412

1. Introduction.

The usual procedure in modeling consists of *tearing* and *zooming*: a system is viewed as an interconnection of subsystems, and modeling consists of describing the subsystems and the interconnection laws. This procedure is often executed hierarchically, with the subsystems in turn viewed as an interconnection. The net result of such a modeling procedure will be a model which involves *manifest* variables (often called *external* variables), which are the variables whose behavior we try to model, and *latent* variables (often called *internal* variables), which are the variables describing the subsystems. The formalization of this modeling procedure is the philosophy underlying the *behavioral* approach to systems theory. These ideas have been explained in detail in [8, 9, 10].

As should be apparent, the resulting model will typically involve many algebraic relations (for example, interconnection constraints, resistors laws, spring and damper characteristics, kinematic constraints), combined with differential equations. These may be first-order (for example, inductors, capacitors, the dynamics of dampers), second-order (for example, the dynamics of masses), or higher-order (for example, subsystems whose dynamic laws have been obtained from an identification procedure).

A state-space model is hence not the natural end result of the modeling phase, while its importance for simulation or for control design is undisputed. This is one of the reasons why the notion of state is one of the most investigated ones in system theory and why its characterization and construction have been the subject of many papers since the beginning of this discipline. The problem that we deal with in this paper is that of computing state variables, from which a state-space model is easily recovered, starting from an arbitrary set of linear differential-algebraic equations.

The paper is organized as follows. In section 2 a set of definitions and results pertaining to the behavioral framework is introduced. In section 3 the consequences of the property of Markovianity, the key to the notion of state, are worked out. In

*Received by the editors May 23, 1994; accepted for publication (in revised form) April 9, 1996.
http://www.siam.org/journals/sicon/35-3/26841.html

†Department of Electronic, Electrical and Computer Engineering, University of Trieste, I-34127 Trieste, Italy (rapisard@univ.trieste.it).

‡Department of Mathematics, University of Groningen, 9700 AV Groningen, The Netherlands (J.C.Willems@math.rug.nl). The research of this author was supported in part by grant SC1*-CT92-0779 on System Identification of the Science Program of the Commission of the European Community.

section 4 the problem at hand is formally stated. In section 5 operators on polynomials are introduced which will be used in sections 6–9, in which state functions for systems of differential-algebraic equations are computed. As we shall see, the systems may be in kernel, in image, or in hybrid form.

The proofs and some of the notation are collected in the appendices.

2. The behavioral framework. In this section we give a brief introduction to behavioral system theory, with emphasis on the definitions and results pertaining to the problem at hand, referring the reader to [8, 9, 10] for a thorough exposition.

In the behavioral framework a *system* is defined as a triple $\Sigma = (T, W, \mathcal{B})$, with T the time set, W the signal space, and $\mathcal{B}$ the behavior of the system, $\mathcal{B} \subseteq W^T$.

Effectively, a system consists of a family of trajectories which take their value in the signal space. In this paper we consider continuous-time ($T = \mathbb{R}$) systems whose variables take values in a finite-dimensional real vector space, $W = \mathbb{R}^q$. A dynamical system will be called *linear* if $\mathcal{B}$ is a linear vector subspace of $(\mathbb{R}^q)^{\mathbb{R}}$, the latter equipped with the usual vector space structure induced by that of $\mathbb{R}^q$, and *time invariant* if the following holds $\forall\ t \in \mathbb{R}$:

$$(w(\cdot)\ \in\ \mathcal{B}) \Longrightarrow (w(\cdot + t)\ \in\ \mathcal{B}). \tag{2.1}$$

In many instances systems are described by differential equations, say,

$$f_1\left(w, \frac{d}{dt}w, \ldots, \frac{d^L}{dt^L}w\right) = f_2\left(w, \frac{d}{dt}w, \ldots, \frac{d^L}{dt^L}w\right). \tag{2.2}$$

A concrete representation of the behavior of a linear, time-invariant, continuous-time differential system $(\mathbb{R}, \mathbb{R}^q, \mathcal{B})$ is then given as the solution set of a system of linear, constant coefficient differential equations:

$$R_0 w + R_1 \frac{d}{dt}w + R_2 \frac{d^2}{dt^2}w + \cdots + R_L \frac{d^L}{dt^L}w = 0 \tag{2.3}$$

with constant matrices $R_i \in \mathbb{R}^{\bullet \times q}$. Equation (2.3) is what we call a *kernel representation* of such a system. A shorthand notation for (2.3) is

$$R\left(\frac{d}{dt}\right) w = 0, \tag{2.4}$$

where $R(\xi) := R_0 + R_1\xi + \cdots + R_L\xi^L \in \mathbb{R}^{\bullet \times q}[\xi]$. Note that (2.4) may involve algebraic equations in addition to ordinary differential equations.

The behavioral framework takes into account the nonuniqueness of the representation of behaviors. This is natural, given the connections between this approach and the actual procedure of modeling physical systems, in which different, although equivalent, sets of equations describing the same phenomenon may be produced.

The formalization of this equivalence concept is given as follows. Two kernel representations $R_1(\frac{d}{dt})w = 0$ and $R_2(\frac{d}{dt})w = 0$ with R_1, $R_2 \in \mathbb{R}^{\bullet \times q}[\xi]$ are *equivalent*—that is, the behaviors associated with them are the same—if and only if there exist polynomial matrices F_1, F_2 with a suitable number of columns such that $R_1 = F_1 R_2$ and $R_2 = F_2 R_1$; in particular, if R_1 and R_2 are of full row rank, this means that there exists a unimodular polynomial matrix F such that $R_1 = F R_2$ (see [10, p. 263]).

As already explained in the introduction, (2.4) is not the most natural result of a modeling process, since normally a number of auxiliary latent variables will have

been introduced. The natural counterpart of (2.4) to cope with this is

$$R\left(\frac{d}{dt}\right)w = M\left(\frac{d}{dt}\right)\ell, \tag{2.5}$$

where $M \in \mathbb{R}^{\bullet\times d}[\xi]$ and where $\ell \in (\mathbb{R}^d)^{\mathbb{R}}$ are the latent variables. The set of equations (2.5) is called a *latent variable* or a *hybrid representation* of the *latent variable system* $(\mathbb{R}, \mathbb{R}^q, \mathbb{R}^d, \mathcal{B}_f)$, where the *full behavior* $\mathcal{B}_f$ is composed of trajectories (w, ℓ) satisfying (2.5) and inducing the *external* or *manifest behavior* $\mathcal{B}_{ext} := \pi_w\mathcal{B}_f$ by projection on the external variables. Actually the external behavior induced by a latent variable system may be described (modulo some closedness problems pointed out in [6] and discussed in detail in the following) in terms of the external variables only by *eliminating the latent variable*, a procedure discussed in the following.

Of course the problem arises what sort of *solution* we want to use for (2.4) and (2.5). Restricting ourselves to $\mathcal{C}^\infty$ (infinitely differentiable signals) would leave out interesting functions such as steps, etc. The space of distributions is a bit too large, leaving us with the problem of defining the value of a solution at a point. The space $\mathcal{L}_1^{loc}$ of locally integrable functions is large enough to accommodate steps, ramps, and so on and still concrete enough to avoid the problems we would have with distributions. Therefore, in (2.4) and (2.5) *w and ℓ are to be intended in $\mathcal{L}_1^{loc}$ and equality in the sense of distributions.*

Let us focus now on the *elimination of the latent variable* from (2.5).

Hybrid representations involve two kind of variables, namely, the manifest and the latent variables; and associated with a hybrid representation are two behaviors, the full behavior $\mathcal{B}_f$, consisting of trajectories with both the latent and the external variables, and the external behavior $\mathcal{B}_{ext}$, composed of trajectories in the manifest variables only.

At the level of trajectories, the relationship between $\mathcal{B}_f$ and $\mathcal{B}_{ext}$ is the following: any trajectory in $\mathcal{B}_{ext}$ is induced by a trajectory in $\mathcal{B}_f$ via the projection operator $\pi_w((w, \ell)) = w$. At the level of representations and of the equations representing the behaviors, things are more complicated. Take a hybrid representation

$$R\left(\frac{d}{dt}\right)w = M\left(\frac{d}{dt}\right)\ell. \tag{2.6}$$

By premultiplication by a unimodular matrix U we can bring $(\,R \;\mid\; -M\,)$ to the form

$$U\,(\,R \;\mid\; -M\,) = \begin{pmatrix} R_1' & 0 \\ R_2' & -M_2' \end{pmatrix} \tag{2.7}$$

with M_2' of full row rank. Unimodularity of U implies that the full behavior represented by (2.6) is not altered by the change of representation and coincides with the behavior represented by

$$R_1'\left(\frac{d}{dt}\right)w = 0, \tag{2.8}$$

$$R_2'\left(\frac{d}{dt}\right)w = M_2'\left(\frac{d}{dt}\right)\ell. \tag{2.9}$$

A natural candidate for representing the external behavior corresponding to (2.6) would be $R_1'(\frac{d}{dt})w = 0$, since if $w \in \mathcal{B}_{ext}$ of (2.6), then $R_1'(\frac{d}{dt})w = 0$. In fact, for

discrete-time systems it has been shown in [10, p. 234] that the analogue of (2.8) in discrete time is indeed a kernel representation of the manifest behavior; this result is referred to in behavioral system theory as the *latent variable elimination theorem*. However, in the continuous-time case there are difficulties. Take, for example, the hybrid representation

$$\begin{aligned} w_1 &= w_2, \\ \frac{d}{dt} w_2 &= \ell. \end{aligned} \tag{2.10}$$

Note that the second equation imposes a smoothness requirement on w_2 not present in the first one: the external behavior does not coincide with the one described by $w_1 = w_2$. When situations like the one exemplified above do not occur, the latent variable ℓ is said to be *properly eliminable* (cf. [6]). Necessary and sufficient conditions for proper eliminability are given in [6].

If the latent variable is not properly eliminable, $\mathcal{B}_{ext}$ is described by $R_1'(\frac{d}{dt})w = 0$ of (2.8) along with some smoothness constraints. These constraints on w cannot be represented by equations involving w alone and the need to circumvent this difficulty arises. The most natural way to do this is to drop them, that is, to consider the closure of $\mathcal{B}_{ext}$ in the topology of $\mathcal{L}_1^{loc}$. This choice has much to recommend it besides its simplicity: it allows to keep $\mathcal{L}_1^{loc}$ as the natural function space in which to operate, and in this way the latent variable can always be eliminated. We summarize this in the following theorem.

THEOREM 2.1. *Let* (2.6) *be a hybrid representation. There exists a unimodular matrix* U *such that*

$$U \left(\begin{array}{c|c} R & -M \end{array} \right) = \begin{pmatrix} R_1' & 0 \\ R_2' & -M_2' \end{pmatrix} \tag{2.11}$$

with M_2' *of full row rank. Then*

$$\begin{aligned} \left\{ w \in \mathcal{L}_1^{loc}(\mathbb{R}; \mathbb{R}^q) \mid R_1' \left(\frac{d}{dt} \right) w = 0 \right\} &= \overline{\pi_w(\mathcal{B}_f)}^{closure} \\ &= \overline{\{w \mid \exists\ \ell\ s.t.\ (2.6)\ holds\}}^{closure} \end{aligned} \tag{2.12}$$

with the closure taken in the topology of $\mathcal{L}_1^{loc}(\mathbb{R}; \mathbb{R}^q)$.

Proof. See the appendix. □

In the following, unless otherwise stated, we will take (2.12) to be the manifest behavior induced by (2.6).

The important notions of controllability and observability emerge in the behavioral context as follows. The time-invariant system $(\mathbb{R}, \mathbb{R}^q, \mathcal{B})$ is said to be *controllable* if for all w_1, w_2 in $\mathcal{B}$, there exists a $T \geq 0$ and a $w \in \mathcal{B}$ such that $w(t) = w_1(t)$ for $t < 0$ and $w(t + T) = w_2(t)$ for $t \geq 0$. The notion of observability deals with latent variable systems and refers to the possibility of deducing the latent variables from the manifest ones. Thus (2.5) defines an *observable* system if there exists a map $F : (\mathbb{R}^q)^{\mathbb{R}} \mapsto (\mathbb{R}^d)^{\mathbb{R}}$ such that $((w, \ell) \in \mathcal{B}_f) \Longrightarrow (\ell = F(w))$. For linear latent variable systems this is equivalent to $((0, \ell) \in \mathcal{B}_f) \Longrightarrow (\ell = 0)$.

The question when a system (2.4) is controllable can be answered effectively in terms of R. Indeed, (2.4) is controllable if and only if $\operatorname{rank}(R(\lambda)) = \operatorname{rank}(R)$ for all $\lambda \in \mathbb{C}$, as shown in [9, p. 238]. (Here one should view $R(\lambda)$ as a matrix over the

field of complex numbers and R as a matrix over the field of real rational functions.) Analogously, (2.5) will be observable if and only if $M(\lambda)$ is right prime (equivalently, if and only if $M(\lambda)$ is of full column rank for all $\lambda \in \mathbb{C}$), as shown in [9, p. 239]. Actually, controllability can also be characterized in terms of (2.5). Take $R = I$ in (2.5), yielding

$$w = M\left(\frac{d}{dt}\right)\ell. \tag{2.13}$$

Let $\mathcal{B}$ be the manifest behavior of (2.13). (More precisely, in view of questions related to closedness, $\mathcal{B} = \overline{M(\frac{d}{dt})\mathcal{C}^\infty(\mathbb{R},\mathbb{R}^d)}^{closure}$, with the closure taken with respect to the topology of $\mathcal{L}_1^{loc}(\mathbb{R},\mathbb{R}^q)$.) This yields the dynamical system $(\mathbb{R},\mathbb{R}^q,\mathcal{B})$, and for obvious reasons we will call (2.13) an *image representation* of $\mathcal{B}$. By the already mentioned latent variable elimination theorem, $\mathcal{B}$ admits a kernel representation as (2.4). However, not every system (2.4) has an image representation. This is the case if and only if the system is controllable! (See [9, p. 238].)

Finally, let us introduce the notion of *input* and *output*. Consider a system (2.4) with $R(\xi)$ of full row rank. Possibly permuting the components of w, assume R partitioned as $R := (\,P \;\mid\; -Q\,)$, with P square, nonsingular, and $P^{-1}Q$ proper. Such a partition of R always exists and can be found as follows. By unimodular premultiplication by a suitable U, bring R in row reduced form (see [3, p. 382]). Let R'_{hc} be the highest row coefficient matrix of $R' := UR$. R'_{hc} has full row rank, and therefore there exists at least one $g \times g$ nonzero minor, corresponding to a choice of the k_1th, k_2th, $\ldots, k_g$th column of R_{hc}. The minor of R corresponding to this column selection is of maximal degree among the $g \times g$ minors of R. This implies that if we take P to be the matrix formed by the k_1th, k_2th, $\ldots, k_g$th column of R, P is nonsingular and $P^{-1}Q$ is proper, $-Q$ being the complementary matrix of P in R.

The partition of R induces a corresponding partition of w in (y,u) so that (2.4) may be rewritten as

$$P\left(\frac{d}{dt}\right)y = Q\left(\frac{d}{dt}\right)u. \tag{2.14}$$

This is an *input-output* representation of the behavior of (2.4), with y the *output* variables and u the *input* variables.

It is important to note that the selection of P and Q of (2.14) is not unique, in general. This implies that for a system whose behavior is described by (2.4), different selections of inputs and outputs can be given. This corresponds to different selections of R'_{hc} to form a nonzero minor in the procedure sketched above. Anyway, it is possible to prove (see [10, p. 243]) that the number of outputs, and consequently the number of inputs, in any representation (2.4) of the behavior of the system is unique and coincides with $\operatorname{rank}(R)$.

3. State, Markovianity, and first-order representations. Let $(\mathbb{R},\mathbb{R}^q,\mathcal{B})$ be a time-invariant dynamical system. We will call it *Markovian* if

$$(w_1, w_2 \in \mathcal{B}) \wedge (w_1, w_2 \text{ continuous at } 0) \wedge (w_1(0) = w_2(0)) \tag{3.1}$$

implies $((w_1 \wedge w_2)) \in \mathcal{B})$; $w_1 \wedge w_2$ stands for concatenation:

$$(f_1 \wedge f_2)(t) := \begin{cases} f_1(t), & t < 0, \\ f_2(t), & t \geq 0. \end{cases} \tag{3.2}$$

Thus in a Markovian system trajectories passing in a continuous way through the same point at $t = 0$ can be concatenated. The very much related notion of state refers to systems with latent variables. Thus let $(\mathbb{R}, \mathbb{R}^q, \mathbb{R}^d, \mathcal{B}_f)$ be a time-invariant latent variable system. Then it is a *state system* if

$$((w_1, \ell_1), (w_2, \ell_2) \in \mathcal{B}_f) \wedge (\ell_1(0) = \ell_2(0)) \wedge (\ell_1, \ell_2 \text{ continuous at } t = 0)$$
$$\implies ((w_1, \ell_1) \wedge (w_2, \ell_2)) \in \mathcal{B}_f). \tag{3.3}$$

We call (3.3) the *axiom of state*. If (3.3) holds, then the latent variable is called the *state*. Thus in a state model trajectories passing in a continuous way through the same state at $t = 0$ can be concatenated. The continuity requirement is inspired by the fact that we are dealing with solutions of (2.4) and (2.5) in $\mathcal{L}_1^{loc}$, in which case the simple requirement $\ell_1(0) = \ell_2(0)$ is of little consequence.

Usually a Markovian or a state variable is denoted by x. We will do so in the following discussion.

It is easy to prove that if the behavior is described by a set of first-order differential equations, as

$$f\left(x, \frac{d}{dt}x\right) = 0, \tag{3.4}$$

then it is Markovian; similarly, if it can be described by a set of differential equations which is first order in the latent variables and zeroth order in the manifest variables, as

$$f\left(w, x, \frac{d}{dt}x\right) = 0, \tag{3.5}$$

then it is a state model (see [9, p. 191]). For linear differential systems this is, in fact, necessary and sufficient, as shown by the following proposition.

PROPOSITION 3.1. *Let Σ_S be a system as in* (2.5). *Then it is a state-space system if and only if there exist matrices E, F, and G such that $\mathcal{B}_f$ has the kernel representation*

$$Gw + Fx + E\frac{d}{dt}x = 0. \tag{3.6}$$

Analogously, $(\mathbb{R}, \mathbb{R}^q, \mathcal{B})$ *as in* (2.4) *is Markovian if and only if there exist matrices E and F such that $\mathcal{B}$ has the kernel representation*

$$Fx + E\frac{d}{dt}x = 0. \tag{3.7}$$

Proof. See the appendix. □

Remark 3.1. The above proposition constitutes an example of application of the following fact. Equation (2.4) determines a representation of the behavior $\mathcal{B}$, but it is not the unique possible representation of $\mathcal{B}$. In fact, if U is a unimodular matrix, then UR determines the same behavior. This allows us to obtain representations which put certain properties in evidence, just like the state property above.

Remark 3.2. A state-space system induces, by projection of $\mathcal{B}_f$ on the external variable w, an external behavior $\mathcal{B} := \pi_w \mathcal{B}_f$. Therefore it will be called a *state-space representation* or a *state-space model* of $\mathcal{B}$.

Besides state-space models of $\mathcal{B}$ whose full behavior is described by equations of the form (3.6), *state-space models with driving variables* and *input/state/output models* can be defined.

A *state-space model with driving variables* is described by

$$\begin{aligned} \frac{d}{dt}x &= Ax + Bv, \\ w &= Cx + Dv, \end{aligned} \tag{3.8}$$

where x is a state variable for $\mathcal{B} = \{w \in \mathcal{L}_1^{loc} | \ \exists\ x \in \mathcal{L}_1^{loc}, v \in \mathcal{L}_1^{loc},$ s.t. (3.8) holds$\}$ and v is composed of free but latent variables which generate, together with the initial conditions, the state trajectory and the external signal. We call v the *driving variable*.

By integrating the state property and the input-output structure in the same representation, an *input/state/output representation* is obtained. It can be computed from a state representation (3.6) by partitioning the w variables in inputs u and outputs y and rearranging the equations (3.6) so that a representation

$$\begin{aligned} \frac{d}{dt}x &= Ax + Bu, \\ y &= Cx + Du \end{aligned} \tag{3.9}$$

is obtained.

Let $\Sigma_S = (\mathbb{R}, \mathbb{R}^q, \mathbb{R}^n, \mathcal{B}_f)$ be a state-space system and $(\mathbb{R}, \mathbb{R}^q, \mathcal{B})$ be its external (i.e., manifest) behavior. We will call Σ_s *minimal* if whenever $\Sigma_{S'} = (\mathbb{R}, \mathbb{R}^q, \mathbb{R}^{n'}, \mathcal{B}'_f)$ is another state-space model with the same external behavior $(\mathbb{R}, \mathbb{R}^q, \mathcal{B})$, then $n \leq n'$. It is possible to prove (see [10, p. 270]) that Σ_S is minimal if and only if it is *observable* (with the state viewed as the latent variable) and *state trim* (meaning that for all $x_0 \in \mathbb{R}^n$ there exists a $(w, x) \in \mathcal{B}_f \bigcap \mathcal{C}^\infty$ such that $x(0) = x_0$). Observability, in particular, implies that there then exists a $F \in \mathbb{R}^{n \times q}[\xi]$ such that $((w, x) \in \mathcal{B}_f) \Longrightarrow (x = F(\frac{d}{dt})w)$. Actually it can further be shown (see [10, p. 271]) that if Σ_S and $\Sigma_{S'} = (\mathbb{R}, \mathbb{R}^q, \mathbb{R}^{n'}, \mathcal{B}'_f)$ are two minimal state space systems with the same external behavior, then there exists a nonsingular matrix $S \in \mathbb{R}^{n \times n}$ such that

$$((w, x) \in \mathcal{B}_S \text{ and } (w, x') \in \mathcal{B}_{S'}) \Longrightarrow (x' = Sx). \tag{3.10}$$

4. Problem statement. The question arises of how to compute a set of state variables when a system is given either in kernel or in hybrid form.

This question may be stated as: *Given a set of equations in either kernel or hybrid form, how do we determine a* state map $X(\frac{d}{dt})$*?* More precisely, given R, determine the integer n and $X \in \mathbb{R}^{n \times q}[\xi]$ such that

$$R\left(\frac{d}{dt}\right) w = 0, \tag{4.1}$$

$$X\left(\frac{d}{dt}\right) w = x \tag{4.2}$$

defines a (minimal) state-space system with external behavior given by (4.1). The problem is to derive X from R. Similarly we want to derive a state map X for the external behavior of a system represented in hybrid form or image form. In this case, in view of the closedness problems discussed in the previous section, we will interpret

the external behavior associated with the hybrid or image representation under study, in the sense of Theorem 2.1.

Remark 4.1. It is of utmost importance at this point to note that the external behavior of the state-space system described by (4.1), (4.2) is assumed to be described by (4.1); that is, the equations (4.2) do not impose any smoothness constraint on the trajectories defined by (4.1). Therefore, when dealing with state-space representations of a given external behavior $\mathcal{B}$, we will consider the (latent) state variable x induced by the state map to be *properly eliminable.*

The next section introduces the tools that we will use to deal with the problem of computing state maps for the various sorts of representations introduced so far.

5. Operators on polynomials. The behavioral framework for linear differential systems is intimately connected to polynomial matrix algebra. These connections are also reflected in the results which will be presented in the following sections, related to the characterization of state maps.

This section is devoted to the introduction of some notational conventions related to polynomials and rational functions.

Any rational function can be written in a unique way as the sum of a polynomial and of a strictly proper rational function. That is, given $q \in \mathbb{R}(\xi)$, there exist unique $p \in \mathbb{R}[\xi]$ and $s \in \mathbb{R}_+(\xi)$, the set of strictly proper rational functions, such that $q = p + s$. Now define

$$(\ )_+ : \mathbb{R}(\xi) \mapsto \mathbb{R}[\xi] \tag{5.1}$$

as

$$(q(\xi))_+ := p(\xi). \tag{5.2}$$

On the set of rational functions in the indeterminate ξ, multiplication by ξ^{-1} defines a map $\xi^{-1} : \mathbb{R}(\xi) \mapsto \mathbb{R}(\xi)$ in the obvious way.

DEFINITION 5.1. *The* shift-and-cut *operator σ_+ is defined as*

$$\sigma_+ : \mathbb{R}(\xi) \mapsto \mathbb{R}[\xi],$$

$$\sigma_+ := (\ )_+ \circ \xi^{-1}. \tag{5.3}$$

The definition of σ_+ is extended to vectors and matrices of rational functions in a componentwise manner.

Iterated application of σ_+ will be considered in the following and denoted as

$$\sigma_+^k := \overbrace{\sigma_+ \circ \sigma_+ \circ \cdots \sigma_+}^{k-times}. \tag{5.4}$$

In the following, special importance will be given to the action of σ_+ on vector polynomials. Therefore, let us examine in detail what the result is of the application of σ_+ to a vector polynomial $p \in \mathbb{R}^{1\times q}[\xi]$. Write

$$p(\xi) := p_\delta \xi^\delta + p_{\delta-1}\xi^{\delta-1} + \cdots + p_1\xi + p_0. \tag{5.5}$$

Then

$$\sigma_+(p(\xi)) = p_\delta \xi^{\delta-1} + p_{\delta-1}\xi^{\delta-2} + \cdots + p_1; \tag{5.6}$$

that is,

$$\sigma_+(p(\xi)) = \xi^{-1}(p(\xi) - p_0). \tag{5.7}$$

Now let $R \in \mathbb{R}^{g\times q}[\xi]$ be given, and assume $R := R_0 + R_1\xi + \cdots + R_L\xi^L$. Define R^k, $k = 0, \ldots, L$, as $R^0 := R$ and $R^k := \sigma_+^k R = \sigma_+ R^{k-1}$, $k = 1, \ldots, L$. Define the Ξ-*matrix* R_Ξ as

$$R_\Xi := \operatorname{col}(R^k)_{k=1,\ldots,L} = \begin{pmatrix} R^1 \\ R^2 \\ \vdots \\ R^L \end{pmatrix}. \tag{5.8}$$

Connected to R_Ξ is the important notion of Ξ-space. Let $r_1, r_2, \ldots, r_g$ denote the rows of R. Then define the $\mathbb{R}$-vector space Ξ_R as

$$\Xi_R := \langle \sigma_+^k(r_i) \rangle, \quad k \in \mathbb{N},\ i = 1, \ldots, g, \tag{5.9}$$

where $\langle\ \rangle$ denotes the span over $\mathbb{R}$. The Ξ-space of R is most easily constructed as the $\mathbb{R}$-vector space generated by the rows of R_Ξ. Note that R_Ξ need not define a basis of the Ξ-space of R.

Introduce now on Ξ_R the equivalence relation $\overset{R}{\sim}$ defined as follows: $p, q \in \Xi_R$ are *equivalent modulo* R, written $p \overset{R}{\sim} q$, if and only if there exists $r(\xi) \in \mathbb{R}^{1\times g}[\xi]$ such that $p(\xi) - q(\xi) = r(\xi)R(\xi)$. It is easily verified that $\overset{R}{\sim}$ is indeed an equivalence relation.

Note that the vector space structure on Ξ_R induces a vector space structure on the set of equivalence classes induced by $\overset{R}{\sim}$ on Ξ_R. We will denote this set of equivalence classes as $\Xi_R \pmod R$. That is,

$$\Xi_R \pmod R = \{[p] \in 2^{\Xi_R} \mid q \in [p] \text{ iff } \exists\ r \in \mathbb{R}^{1\times g}[\xi] \text{ s.t. } p = q + rR\}. \tag{5.10}$$

The following example illustrates the above notions.

Example 5.1. Let

$$R := \begin{pmatrix} \xi^2 + 2\xi - 1 & \xi + 1 \\ \xi - 1 & \xi^2 - 3 \end{pmatrix}, \tag{5.11}$$

and consider its first row, $(\,\xi^2 + 2\xi - 1 \quad \xi + 1\,)$. The shift-and-cut operator acts on this row as

$$\sigma_+(\,\xi^2 + 2\xi - 1 \quad \xi + 1\,) = (\,\xi + 2 \quad 1\,). \tag{5.12}$$

The Ξ_R space is the vector space spanned by

$$(\,\xi + 2 \quad 1\,),\ (\,1 \quad 0\,),\ (\,1 \quad \xi\,),\ (\,0 \quad 1\,), \tag{5.13}$$

which actually form a basis for this space. It is easily verified that the vectors (5.13), interpreted as representing elements of $\Xi_R \pmod R$, are linearly independent as well, and therefore form a basis of $\Xi_R \pmod R$.

Note that selecting from the rows of R_Ξ a maximal set of linearly independent rows and considering these as representatives of elements of $\Xi_R \pmod R$ do not necessarily yield a basis for $\Xi_R \pmod R$, as made explicit by the following example.

Example 5.2. Let

$$R = \begin{pmatrix} 1 & 0 & -1 \\ 0 & 1 & \xi^3 \\ 0 & 0 & \xi \end{pmatrix}. \tag{5.14}$$

A maximal set of linearly independent rows of R_Ξ is

$$(\,0 \quad 0 \quad \xi^2\,),\ (\,0 \quad 0 \quad \xi\,),\ (\,0 \quad 0 \quad 1\,), \tag{5.15}$$

but the first and the second element of this set are equivalent to zero modulo R, since $(\,0 \quad 0 \quad \xi^2\,) = (\,0 \quad 0 \quad \xi\,)\,R$ and $(\,0 \quad 0 \quad \xi\,) = (\,0 \quad 0 \quad 1\,)\,R$.

Equipped with these notions, we are now ready to consider the problem of the determination of state-inducing maps for systems in kernel form.

6. State maps for systems in kernel form. The systems we will consider in this section are those described by equations (2.4). For some systems of this kind the problem of computing a state map is trivial, namely, those corresponding to a behavior coinciding with the zero trajectory only. These systems can be characterized as those corresponding to a right prime polynomial matrix R; this can be readily shown by resorting to the Smith form of R. In the following we assume that R is not right prime.

The main result of this section is a characterization of state-inducing polynomial matrices for systems in kernel form. As a preliminary result, we first consider the conditions under which a trajectory is concatenable with the zero one. These conditions correspond to a system of linear equations involving w and its derivatives and define a polynomial differential operator which, in fact, corresponds to a state map. As we will see, the rows of this polynomial matrix have a nice interpretation in terms of the shift-and-cut operator defined in the previous section.

Before stating Proposition 6.1, we observe the following smoothness result. Consider the polynomial matrices R^1, R^2, ..., and observe that if $w \in \mathcal{L}_1^{loc}$ is a solution of (2.4) in the sense of distributions, then

$$R^k\left(\frac{d}{dt}\right) w \tag{6.1}$$

is continuous for $k = 1, 2, \ldots$. In order to see this, let

$$R(\xi) = R_0 + R_1\xi + \cdots + R_L\xi^L; \tag{6.2}$$

then (2.4) implies

$$\frac{d}{dt}\left(R_1 + \cdots + R_L\frac{d^{L-1}}{dt^{L-1}}\right) w = -R_0 w. \tag{6.3}$$

Since the right-hand side is in $\mathcal{L}_1^{loc}$, this implies that

$$\left(R_1 + \cdots + R_L\frac{d^{L-1}}{dt^{L-1}}\right) w = R^1\left(\frac{d}{dt}\right) w \tag{6.4}$$

is absolutely continuous. Proceeding recursively yields the absolute continuity of (6.1). Note that this implies that $R_\Xi(\frac{d}{dt})w$ is also absolutely continuous.

PROPOSITION 6.1. *Let a kernel representation as in* (2.4) *be given, and let* $\mathcal{B}$ *be its behavior. A trajectory* $w \in \mathcal{B}$ *is concatenable with the zero trajectory; that is,* $0 \wedge w \ \in \mathcal{B}$ *if and only if*

$$\left(R_{\Xi}\left(\frac{d}{dt}\right) w\right)(0) = 0. \tag{6.5}$$

Proof. See the appendix. □

This yields the main result of this section.

THEOREM 6.2. *The polynomial matrix* $X \in \mathbb{R}^{\bullet \times q}[\xi]$ *defines a state-inducing map for* (2.4)*; i.e.,*

$$R\left(\frac{d}{dt}\right) w = 0,$$

$$X\left(\frac{d}{dt}\right) w = x \tag{6.6}$$

defines a state-space system with external behavior $\operatorname{Ker} R(\frac{d}{dt})$ *if and only if there exists a matrix* $A \in \mathbb{R}^{\bullet \times \bullet}$ *and a polynomial matrix* $B(\xi) \in \mathbb{R}^{\bullet \times \bullet}[\xi]$ *such that*

$$R_{\Xi}(\xi) = AX(\xi) + B(\xi)R(\xi) \tag{6.7}$$

and the latent variable x *is properly eliminable from the system with latent variable* (6.6).

Proof. See the appendix. □

Remark 6.1. Proper eliminability of $x = X(\frac{d}{dt})w$ in the system with latent variable (6.6) can be checked as follows (cf. [6, Theorem 2.5]). Without loss of generality, assume $R(\xi)$ to be of full row rank g, and let $X(\xi)$ have n rows. x is properly eliminable if and only if there exists an $(n+g) \times (n+g)$ submatrix of maximal determinantal degree of

$$\begin{pmatrix} R & 0_{g \times n} \\ X & -I_n \end{pmatrix}, \tag{6.8}$$

which includes the last n columns of (6.8).

Remark 6.2. In the discrete-time case, an analogue of Proposition 6.1 has been given in [7, p. 1075]. A necessary condition for a state map, analogous to (6.7), has been given in the continuous-time case (with a solution space other than $\mathcal{L}_1^{loc}$) in [5, p. 77]. In the context of discrete-time *output nulling representations*

$$x(k+1) = Ax(k) + Bw(k),$$

$$0 = Cx(k) + Dw(k), \tag{6.9}$$

a procedure similar to using the shift-and-cut operator to obtain R_{Ξ} from R has been used in [1, p. 3643].

Remark 6.3. If the rows of X of (6.7) are interpreted as representative of elements of $\Xi_R \pmod R$, Theorem 6.2 can be restated as follows: X defines a state-inducing map for (2.4) if and only if the span over $\mathbb{R}$ of its rows contains $\Xi_R \pmod R$ and the latent variable x is properly eliminable from (6.6). This, together with the smoothness result given at the beginning of this section, yields the following corollary of Theorem 6.2.

COROLLARY 6.3. *The polynomial matrix $X \in \mathbb{R}^{\bullet \times q}[\xi]$ defines a minimal state-inducing map for* (2.4) *if and only if its rows, considered as representative of elements of $\Xi_R \pmod R$, form a basis for $\Xi_R \pmod R$.*

Remark 6.4. Note that in the scalar case ($R \in \mathbb{R}[\xi]$, $R \neq 0$) the above theorem corresponds to the usual method of stacking the lower-order derivatives of each component to reduce a system of equations of high order to a system of equations of order one, as made explicit by the following example.

Example 6.1. Let $q = 1$ and a system be described by

$$p\left(\frac{d}{dt}\right) w = 0 \tag{6.10}$$

with

$$p(\xi) := p_0 + p_1\xi + \cdots + p_L\xi^L, \tag{6.11}$$

with $p_L \neq 0$.

The Ξ_p space is generated by

$$p_1 + p_2\xi + p_3\xi^2 + \cdots + p_L\xi^{L-1}, p_2 + p_3\xi + \cdots + p_L\xi^{L-2}, \ldots, p_{L-1} + p_L\xi, p_L, \tag{6.12}$$

which in fact constitutes a basis for the space; another basis for Ξ_p could be chosen as

$$\xi^{L-1}, \xi^{L-2}, \ldots, \xi, 1. \tag{6.13}$$

In fact, it can be verified that both (6.12) and (6.13) are bases of $\Xi_p \pmod p$. Therefore a minimal state is induced by the first $L-1$ derivatives of w, as made apparent by (6.13), or by the differential operators associated with the polynomials (6.12). That is, both

$$x := \begin{pmatrix} \frac{d^{L-1}}{dt^{L-1}} w \\ \frac{d^{L-2}}{dt^{L-2}} w \\ \vdots \\ \frac{d}{dt} w \\ w \end{pmatrix} \tag{6.14}$$

and

$$x := \begin{pmatrix} (p_1 + p_2\frac{d}{dt} + p_3\frac{d^2}{dt^2} + \cdots + p_L\frac{d^{L-1}}{dt^{L-1}})w \\ (p_2 + p_3\frac{d}{dt} + \cdots + p_L\frac{d^{L-2}}{dt^{L-2}})w \\ \vdots \\ (p_{L-1} + p_L\frac{d}{dt})w \\ p_L w \end{pmatrix} \tag{6.15}$$

define minimal state variables.

Remark 6.5. Note that computation of a first-order kernel representation of a state system is easy once the state map is given and amounts to solving a linear system of equations. In fact, once the polynomial matrix $X \in \mathbb{R}^{n \times q}[\xi]$ has been

computed, the equations may be recovered in the following way. Find matrices E, F in $\mathbb{R}^{(n+g)\times n}$, $G \in \mathbb{R}^{(n+g)\times q}$, and $T \in \mathbb{R}^{(n+g)\times g}[\xi]$ which solve the equation

$$(\xi E + F)X(\xi) + G = T(\xi)R(\xi). \tag{6.16}$$

An input/state/output representation is easily computed from the kernel representation of the state system obtained in this way. Note that for simple cases these computations can be done by inspection.

In the context of the nonuniqueness of representation of behaviors pointed out in section 2, a question arises with respect to Theorem 6.2. That is, given two equivalent kernel representations of the same system, which we assume to correspond to full row rank polynomial matrices, what relationships hold between the corresponding Ξ-spaces?

The following result holds.

PROPOSITION 6.4. *Let a kernel representation* (2.4) *be given with R of full row rank, and let an equivalent representation be obtained as UR, U unimodular. Then there exist a constant full column rank matrix A and a polynomial matrix B such that $(UR)_\Xi = AR_\Xi + BR$.*

Proof. See the appendix. ∎

COROLLARY 6.5. *Let a kernel representation* (2.4) *be given with R of full row rank, and let an equivalent representation be obtained as UR, U unimodular. For each polynomial matrix $\bar{\Xi}_{UR}$ whose rows form a basis of Ξ_{UR} and every polynomial matrix $\bar{\Xi}_R$ whose rows form a basis of Ξ_R there exists a polynomial matrix C and a constant nonsingular matrix T such that $\bar{\Xi}_{UR} = T\bar{\Xi}_R + CR$.*

Minimal (in the sense of the minimal possible dimension of the state space) states are induced by the choice of a polynomial matrix X whose rows form a basis of $\Xi_R \pmod R$. A natural question arises as to when minimality of the state space is already guaranteed by directly applying σ_+ to the equations describing the system. The following result holds.

PROPOSITION 6.6. *Let a kernel representation such as* (2.4) *be given. Then the nonzero rows of R_Ξ define a basis for $\Xi_R \pmod R$ if and only if R is in row reduced form. Whence if R is in row reduced form and X is composed of the nonzero rows of R_Ξ,*

$$R\left(\frac{d}{dt}\right)w = 0,$$

$$X\left(\frac{d}{dt}\right)w = x \tag{6.17}$$

defines a minimal state representation.

Proof. See the appendix. ∎

COROLLARY 6.7. *Let* (2.4) *be given with R of full row rank. The minimal dimension of the state space of the system associated to R equals the McMillan degree of R, i.e., the maximal degree of the* $\operatorname{rank}(R) \times \operatorname{rank}(R)$ *minors of R. In the row reduced case, this equals the sum of the row degrees of R.*

Let us now give two examples illustrating the procedure of state construction.

Example 6.2. Consider the system with behavior described by

$$\frac{d^n}{dt^n}w_1 + \cdots + p_1\frac{d}{dt}w_1 + p_0 w_1 = q_n\frac{d^n}{dt^n}w_2 + \cdots + q_1\frac{d}{dt}w_2 + q_0 w_2, \tag{6.18}$$

where w_i, $i = 1, 2$, are scalar functions. Defining

$$p^i := \sigma_+^i(p) = \xi^{n-i} + p_{n-1}\xi^{n-i-1} + \cdots + p_i \tag{6.19}$$

and analogously for q^i, it is easy to see that

$$f_i := \begin{pmatrix} p^i & -q^i \end{pmatrix} \tag{6.20}$$

$i = 1, \ldots, n$, form a basis for $\Xi_{\begin{pmatrix} p & -q \end{pmatrix}} \pmod{\begin{pmatrix} p & -q \end{pmatrix}}$. Stacking the f^i vectors yields a polynomial minimal state-inducing matrix.

Example 6.3. Let a system be described by

$$\frac{d^2}{dt^2}w_1 - \frac{d}{dt}w_2 = 0, \tag{6.21}$$

which corresponds to $R(\xi) = \begin{pmatrix} \xi^2 & -\xi \end{pmatrix}$. The space Ξ_R is generated by $\begin{pmatrix} \xi & -1 \end{pmatrix}$ and $\begin{pmatrix} 1 & 0 \end{pmatrix}$, as is easily seen applying the shift-and-cut operator to $R(\xi)$. Then a state is defined as

$$x := \begin{pmatrix} \frac{d}{dt}w_1 - w_2 \\ w_1 \end{pmatrix}, \tag{6.22}$$

which corresponds to the input/state/output equations

$$\frac{d}{dt}x = \begin{pmatrix} 0 & 0 \\ 1 & 0 \end{pmatrix} x + \begin{pmatrix} 0 \\ 1 \end{pmatrix} w_2.$$

$$w_1 = x_2.$$

We summarize the above results in the following algorithms.

ALGORITHM 1 (Construction of a state map for a kernel representation).

Data: $R \in \mathbb{R}^{g\times q}[\xi]$, of degree L.

Output: $X \in \mathbb{R}^{\bullet\times q}[\xi]$ inducing, through $x = X(\frac{d}{dt})w$, a state for the system described by $R(\frac{d}{dt})w = 0$.

Step 1. Set $R^0 := R$ and compute $R^{k+1} := \sigma_+(R^k)$, for $k = 0, 1, \ldots, L-1$.

Step 2. Find $x_1, \ldots, x_n \in \mathbb{R}^{1\times q}[\xi]$ such that $\langle x_1, \ldots, x_n \rangle$ equals the space spanned by the rows of R_Ξ.

Step 3. $X = \operatorname{col}(x_1, \ldots, x_n)$.

Step 4. *Stop.*

ALGORITHM 2 (Construction of a minimal state map for a kernel representation).

Data: $R \in \mathbb{R}^{g\times q}[\xi]$, of degree L.

Output: $X \in \mathbb{R}^{\bullet\times q}[\xi]$ inducing, through $x = X(\frac{d}{dt})w$, a minimal state for the system described by $R(\frac{d}{dt})w = 0$.

Step 1. As in Algorithm 1.

Step 2. Find $x_1, \ldots, x_n$ forming a basis of $\Xi_R \pmod R$.

Comment: The computation of the vectors x_i can be accomplished by computing R_Ξ and reducing each of its rows modulo R with standard polynomial operations.

Step 3. $X := \operatorname{col}(x_k)_{k=1,\ldots,n}$.

Step 4. *Stop.*

ALGORITHM 3 (Verification of a state map).

Data: $R \in \mathbb{R}^{g\times q}[\xi]$, of degree L and $X \in \mathbb{R}^{n\times q}[\xi]$.

Output: *True* if X is a state map for the system described by R, *False* otherwise.

Step 1. Compute R_Ξ as in Algorithm 1.

Step 2. Find a constant matrix A and a polynomial matrix B such that $R_\Xi(\xi) = AX(\xi) + B(\xi)R(\xi)$.

Comment: The computation of A and B can be accomplished with standard polynomial operations.

Step 3. If A and B exists, then

Step 4. If x is properly eliminable from

$$R\left(\frac{d}{dt}\right) w = 0,$$

$$X\left(\frac{d}{dt}\right) w = x,$$

then Output:=*True* else Output:=*False*.

Comment: Proper eliminability can be checked as described in Remark 6.1.

Step 5. *Stop*.

Remark 6.6. The algorithms described above are of immediate interest for the simulation problem, where by "simulation" we mean a procedure for selecting an arbitrary element of the behavior $\mathcal{B}$, followed by an algorithm for computing it. Before getting to the simulation issue, let us describe how to construct a driving variables representation for a kernel description (2.4). Compute R_Ξ and consider the following set of equations in the unknowns A, B, C, D, $P \in \mathbb{R}^{\bullet\times g}[\xi]$, $P' \in \mathbb{R}^{q\times g}[\xi]$, $U \in \mathbb{R}^{\bullet\times q}[\xi]$:

$$\begin{aligned} \xi R_\Xi &= AR_\Xi + BU + PR, \\ I_q &= CR_\Xi + DU + P'R. \end{aligned} \tag{6.23}$$

Then the equations

$$\begin{aligned} \frac{d}{dt}x &= Ax + Bv, \\ w &= Cx + Dv \end{aligned} \tag{6.24}$$

represent the external behavior of (2.4) with a state-space model with driving variable v, as can be seen applying the latent-variables-elimination theorem. Note that A, B, C, D, U, P, and P' in (6.23) are easily obtained by inspection from R and R_Ξ.

Note that (6.24) leads to the following simulation procedure. Given R, compute (6.24) and choose a vector $x_0 \in \mathbb{R}^n$ and a $v \in \mathcal{L}_1^{loc}(\mathbb{R}, \mathbb{R}^m)$. Generate a trajectory x satisfying the first block of equations of (6.24) with the initial conditions $x(0) = x_0$. Then $w \in \mathcal{B}$ can be computed according to the second block of equations (6.24).

7. State maps for systems in hybrid form. As pointed out in the introduction, hybrid representations are the most natural result of a modeling process. Therefore the characterization of state maps for such representations which we give in this section is especially interesting for applications.

Following Theorem 2.1, given a hybrid representation (2.6), we will consider the problem of computing a state map for the closure of the external behavior described by (2.6). To this purpose, let us make some preliminary comments.

First, note that when considering the simulation or control of a system, the state variables will in general be chosen as function of both the external and the latent variables. In fact, although the former are the quantities we are interested in, in a hybrid representation the two kinds of variables enter the description of the system on an equal footing; this is exemplified by the fact that (2.6) can be considered as a kernel description of the full behavior.

Second, a characterization of w-induced state maps can be given as follows. As discussed in Theorem 2.1, the closure of the external behavior of (2.6) is described by (2.8). This implies that the computation of a w-induced state map for the closure of the external behavior of a hybrid representation (2.6) can be performed as follows. First, the ℓ variable is eliminated by computing a suitable unimodular matrix U such that premultiplying the equations by U yields (2.11). Then a set of generators of the Ξ-space of R_1' of (2.8) is computed, as discussed in section 6.

Note that the computation of w-induced state maps for the closure of the external behavior is obtained by elimination of the latent variables, therefore modifying the original equations. This modification is not a desirable feature of a state construction algorithm: the state variables should reflect as much as possible the physical structure of the system as put in evidence by the original equations.

These considerations motivate us to restrict our attention in the remainder of this section to the characterization of (w, ℓ)-induced state maps.

As a third consideration, note that there are hybrid representations for which the determination of a state variable is trivial, namely, the hybrid representations for which the closure of the external behavior corresponds to $\mathcal{L}_1^{loc}(\mathbb{R}, \mathbb{R}^q)$. These representations are characterized as follows. Note that $\forall\ w \in \mathcal{L}_1^{loc}(\mathbb{R}, \mathbb{R}^q)$ there exists an $\ell \in \mathcal{L}_1^{loc}(\mathbb{R}, \mathbb{R}^d)$ such that (2.6) holds if and only if $M(\xi)$ has full row rank. In the following we will implicitly assume that the systems we are dealing with are not of this kind, and that a nontrivial external behavior corresponds to (2.6).

As a fourth consideration, let us characterize the situations where a state variable for the closure of the external behavior is a state variable for the full behavior as well. This happens if and only if the dimension of the minimal state spaces for the closure of the external and the full behavior are the same. An efficient way of checking this is given in the following proposition.

PROPOSITION 7.1. *Let a system be described in hybrid form as in* (2.6)*, and let ℓ be observable from w. The dimensions of the minimal state spaces for the closure of the external and for the full behavior respectively are the same if and only if there exists an input-output selection in (w, ℓ) such that the variables ℓ are all outputs for the full behavior.*

Proof. See the appendix. □

Remark 7.1. The above proposition implies that if there exists an input-output selection on (w, ℓ) such that the latent variables can all be chosen as outputs, a polynomial differential operator X is a state map for the closure of the external behavior if and only if it is a state map for the full behavior.

Remark 7.2. Existence of an input-output selection on (w, ℓ) such that ℓ is entirely composed of outputs can be checked as follows. Assume that $(\ R \ \ | \ \ -M\ )$ has full row rank g. Then ℓ can be chosen as entirely composed of outputs for the full behavior if and only if one of the $\text{rank}((\ R \ \ | \ \ -M\ )) \times \text{rank}((\ R \ \ | \ \ -M\ ))$ minors of maximal degree among all such minors, contains $-M$ as a submatrix.

This procedure is summarized for future use in the following algorithm.

ALGORITHM ℓ-outputs (Verification if ℓ may be chosen as consisting entirely of outputs for the full system).

Data: $(\,R \quad | \quad -M\,) \in \mathbb{R}^{g\times(q+d)}[\xi]$ of full row rank.

Output: True if there exists an input-output partition such that ℓ is entirely composed of outputs, *False* otherwise.

Step 1. Compute n, the maximal degree of the nonzero $\operatorname{rank}((\,R \quad | \quad -M\,)) \times \operatorname{rank}((\,R \quad | \quad -M\,))$ minors of $(\,R \quad | \quad -M\,)$.

Step 2. For every subset P_i of columns of R such that $(\,P_i \quad | \quad -M\,)$ has $\operatorname{rank}((\,R \quad | \quad -M\,))$ columns, compute the maximal degree of its nonzero $\operatorname{rank}((\,R \quad | \quad -M\,)) \times \operatorname{rank}((\,R \quad | \quad -M\,))$ minors, let it be n_i.

Step 3. If there exists i such that $n_i = n$, then *True* else *False*.

Step 4. *Stop.*

Now assume that ℓ is observable but that in any selection of inputs and outputs for the full system some components of ℓ have to be chosen as inputs. Analogously to what has been done in section 6 for the case of kernel representations, we will first characterize the concatenability of external trajectories. Note that concatenability conditions that involve both w and ℓ are, in general, more restrictive than concatenability conditions involving the external variable only: even if a full trajectory cannot be concatenated with zero at $t = 0$, it could still be possible to concatenate the corresponding external trajectory with the zero one. Therefore, the idea that we pursue in the following is to derive the concatenability conditions for the external trajectories starting from the concatenability conditions for the full trajectories. According to Proposition 6.1, the concatenability conditions of full trajectories can be characterized using the matrix $(\,R \quad -M\,)_\Xi$. As we will see, to derive concatenability conditions for the external trajectory we project $(\,R \quad -M\,)_\Xi$ down with a suitably defined linear map. We call this process the *reduction* of $(\,R \quad -M\,)_\Xi$.

The reduction process involves introducing some new concepts.

Assume that the full row rank matrix $(\,R \quad | \quad -M\,)(\xi) = \sum_{j=0}^{L} (\,R_j \quad -M_j\,)\,\xi^j$ has g rows. Consider

$$\begin{pmatrix} (\,R \quad | \quad -M\,)_\Xi \\ 0_{g\times(d+q)} \end{pmatrix} = \operatorname{col}(\sigma_+^k((\,R \quad | \quad -M\,)))_{k=1,\dots,L+1} \tag{7.1}$$

and the matrix $T := \operatorname{col}(M_i)_{i=0,\dots,L}$.

Define $E := \{r \in \mathbb{R}^{1\times(L+1)g} \mid rT = 0\}$. E is the set of constant left annihilators of T; in fact, E is a vector space.

The following example clarifies the notions just introduced.

Example 7.1. Consider

$$(\,R \quad | \quad -M\,)(\xi) = \begin{pmatrix} \xi & 1 & | & -1 \\ \xi & \xi^2+1 & | & -\xi+1 \end{pmatrix}. \tag{7.2}$$

In this case

$$(\,R \quad | \quad -M\,)_\Xi = \begin{pmatrix} 1 & 0 & 0 \\ 1 & \xi & -1 \\ 0 & 0 & 0 \\ 0 & 1 & 0 \end{pmatrix} \tag{7.3}$$

and

$$T = \begin{pmatrix} 1 \\ -1 \\ 0 \\ 1 \\ 0 \\ 0 \end{pmatrix}. \tag{7.4}$$

The space E is obtained as

$$\begin{aligned} E = \langle &(0 \quad 0 \quad 1 \quad 0 \quad 0 \quad 0), (0 \quad 0 \quad 0 \quad 0 \quad 1 \quad 0), \\ &(0 \quad 0 \quad 0 \quad 0 \quad 0 \quad 1), (1 \quad 1 \quad 0 \quad 0 \quad 0 \quad 0), (1 \quad 0 \quad 0 \quad -1 \quad 0 \quad 0) \rangle. \end{aligned} \tag{7.5}$$

Let us now examine the conditions under which a trajectory $(w, \ell) \in \mathcal{B}_f$ is externally concatenable with zero; that is, $0 \wedge w \in \mathcal{B}_{ext}$. These conditions correspond to a system of linear equations involving w, ℓ, and their derivatives and define a polynomial differential operator which is in fact a state map. The rows of the corresponding polynomial matrix turn out, as stated in the following proposition, to have an interpretation in terms of a set of generators of E, and of the matrix

$$\begin{pmatrix} (R \quad | \quad -M)_{\Xi} \\ 0_{g \times (d+q)} \end{pmatrix}.$$

PROPOSITION 7.2. *Let an hybrid representation* (2.6) *be given with ℓ observable from w. Assume that for every input-output partition of (w, ℓ) there exists at least one component of ℓ chosen as input.*

A trajectory $(w, \ell) \in \mathcal{B}_f$ is externally concatenable with zero, that is, $0 \wedge w \in \mathcal{B}_{ext}$, if and only if given any set $\{v_1, \dots, v_s\}$ of generators of E, there holds

$$\left(v_i \begin{pmatrix} (R \quad | \quad -M)_{\Xi} \\ 0_{g \times (d+q)} \end{pmatrix} \begin{pmatrix} \frac{d}{dt} \end{pmatrix} \begin{pmatrix} w \\ \ell \end{pmatrix} \right)(0) = 0, \tag{7.6}$$

$i = 1, \dots, s$.

Proof. See the appendix. □

We can now state the main result regarding systems in hybrid form with ℓ observable from w.

THEOREM 7.3. *Let a system be described in hybrid form as in* (2.6)*, and let ℓ be observable from w. Assume that for every input-output partition of (w, ℓ) there exists at least one component of ℓ chosen as input. The matrix $X \in \mathbb{R}^{\bullet \times (q+d)}[\xi]$ defines a (w, ℓ)-induced state map for the closure of the external system corresponding to* (2.6)*; that is,*

$$\begin{aligned} (R \quad -M) \begin{pmatrix} \frac{d}{dt} \end{pmatrix} \begin{pmatrix} w \\ \ell \end{pmatrix} &= 0, \\ X \begin{pmatrix} \frac{d}{dt} \end{pmatrix} \begin{pmatrix} w \\ \ell \end{pmatrix} &= x \end{aligned} \tag{7.7}$$

define a state model for the closure of the external behavior corresponding to (2.6) *if and only if for each constant matrix V whose rows generate E there exist a constant matrix A and a polynomial matrix B such that*

$$V \begin{pmatrix} (R \quad | \quad -M)_{\Xi} \\ 0 \end{pmatrix} = AX + B (R \quad | \quad -M)$$

and the variable x does not impose smoothness constraints on the trajectories of the closure of the external behavior of (2.6).

Remark 7.3. To check whether x does not impose smoothness constraints on the trajectories of the closure of the external behavior, we can proceed as follows. Assume without loss of generality that $(\,R \quad -M\,)$ is of full row rank g, and note that by unimodular transformations (7.7) can be brought to the form

$$\begin{aligned} R_1'\left(\frac{d}{dt}\right)w &= 0,\\ R_2'\left(\frac{d}{dt}\right)w &= \ell,\\ X_1\left(\frac{d}{dt}\right)w &= -X_2\left(\frac{d}{dt}\right)\ell + x, \end{aligned} \tag{7.8}$$

where $R_1' \in \mathbb{R}^{g'\times q}$, $R_2' \in \mathbb{R}^{d\times q}$, $X_i \in \mathbb{R}^{n\times \bullet}$, $i = 1, 2$, $X = (\,X_1 \quad X_2\,)$, and $g = g' + d$. Again using unimodular transformations, we can modify this description to

$$R_1'\left(\frac{d}{dt}\right)w = 0, \tag{7.9}$$

$$R_2'\left(\frac{d}{dt}\right)w = \ell, \tag{7.10}$$

$$(X_1 + X_2R_2')\left(\frac{d}{dt}\right)w = x. \tag{7.11}$$

Note that, given (7.9), (7.11), proper eliminability of x could be checked following the procedure illustrated in Remark 6.1. However, this property can be checked on the basis of the original equations, since each $(g' + n) \times (g' + n)$ minor of

$$\begin{pmatrix} R_1' & 0\\ X_1 + X_2R_2' & -I_n \end{pmatrix} \tag{7.12}$$

corresponds uniquely to a $(g' + n + d) \times (g' + n + d)$ minor of

$$\begin{pmatrix} R_1' & 0 & 0\\ R_2' & -I_d & 0\\ X_1 & X_2 & -I_n \end{pmatrix} \tag{7.13}$$

and therefore to a $(g' + n + d) \times (g' + n + d)$ minor of

$$\begin{pmatrix} R & -M & 0\\ X_1 & X_2 & -I_n \end{pmatrix}, \tag{7.14}$$

obtained from a submatrix including the d columns corresponding to ℓ (that is, the columns of (7.14) from the $(q+1)$th up to the $(q+d)$th one). Therefore x is properly eliminable from the equations (7.7) if and only if among all $(g+n)\times(g+n)$ submatrices of (7.14) which include the d columns corresponding to ℓ (that is, the columns of (7.14) from the $(q+1)$th up to the $(q+d)$th one), there exists one of maximal determinantal degree which includes all columns corresponding to x (that is, the columns of (7.14) from the $(q + d + 1)$th up to the $(q + d + n)$th one).

Remark 7.4. Theorem 7.3 may be restated as follows: if no input-output partition of (w, ℓ) exists such that ℓ consists entirely of outputs for the full behavior, X defines a state-inducing map if and only if x does not impose any smoothness constraint on the trajectories of the closure of the external behavior, and the span over $\mathbb{R}$ of the rows of X contains the vector space

$$\left\{ r \begin{pmatrix} \mathrm{col}(\sigma_+^i (R \quad | \quad -M))_{i=1,\dots,L} \\ 0_{g\times(q+d)} \end{pmatrix} \quad | \quad r \in E \right\} \pmod{(R \quad | \quad -M)}, \tag{7.15}$$

defined as the set of equivalence classes determined by the equivalence $\overset{(R|-M)}{\sim}$ on the vector space

$$\left\{ r \begin{pmatrix} \mathrm{col}(\sigma_+^i (R \quad | \quad -M))_{i=1,\dots,L} \\ 0_{g\times(q+d)} \end{pmatrix} \quad | \quad r \in E \right\}. \tag{7.16}$$

This equivalent formulation, together with Proposition 7.1, yields the following characterization of minimality:

COROLLARY 7.4. *X defines a (w, ℓ)-induced minimal state map for the external system corresponding to* (2.6) *if and only if either* (1) *there exists an input-output selection in (w, ℓ) in which ℓ is entirely composed of outputs for the full behavior and the rows of X form a basis for $(R \quad | \quad -M)_\Xi \pmod{(R \quad -M)}$ or* (2) *the rows of X form a basis for the vector space*

$$\left\{ r \begin{pmatrix} \mathrm{col}(\sigma_+^i (R \quad | \quad -M))_{i=1,\dots,L} \\ 0_{g\times(q+d)} \end{pmatrix} \quad | \quad r \in E \right\} \pmod{(R \quad | \quad -M)}. \tag{7.17}$$

Remark 7.5. State-space equations are straightforwardly computed once the state map is given, analogously to the kernel representations case.

The results exposed up to this point suggest the following algorithm for the computation of a state map for a system in hybrid form with ℓ observable from w.

ALGORITHM 4 (Construction of a state map for the external behavior of a system in hybrid form with ℓ observable from w).

Data: $(R \quad | \quad -M) \in \mathbb{R}^{g\times(q+d)}[\xi]$, of degree L and full row rank, M right prime.

Output:] $X \in \mathbb{R}^{\bullet\times(d+q)}[\xi]$ inducing through $x = X(\frac{d}{dt})\binom{w}{\ell}$ a state for the external behavior of the system described in hybrid form by $(R \quad | \quad -M)$.

Step 1. Set $(R \quad | \quad -M)^0 := (R \quad | \quad -M)$ and compute $(R \quad | \quad -M)^{k+1} := \sigma_+^k (R \quad | \quad -M)$, $k = 1, \dots, L+1$.

Step 2. Invoke algorithm ℓ-outputs (cf. Remark 7.2).

Step 3. If *True* then

Step 4. $X := \mathrm{col}((R \quad | \quad -M)^k)_{k=1,\dots,L}$.

Step 5. *Stop.*

Step 6. Else

Step 7. Compute $v_1, \dots, v_s$ in $\mathbb{R}^{1\times g(L+1)}$ such that $\langle v_1, \dots, v_s \rangle$ equals the space $\{r \in \mathbb{R}^{1\times(L+1)g} \mid r\mathrm{col}(M_i)_{i=0\dots L} = 0\}$.

Step 8. $S := \mathrm{col}((R \quad | \quad -M)^k)_{k=1,\dots,L+1}$.

Step 9. $X := VS$.

Comment. Due to the smoothness result given at the beginning of section 6, X defines a properly eliminable latent variable.

Step 10. *Stop.*

The following example illustrates the construction of a state-inducing map for a hybrid representation with observable latent variables.

Example 7.2. Let

$$\begin{pmatrix} 0 & 1 & 1 \\ \frac{d}{dt}-1 & \frac{d}{dt}+1 & 1 \\ \frac{d}{dt} & \frac{d}{dt}-1 & \frac{d}{dt} \end{pmatrix} \begin{pmatrix} w_1 \\ w_2 \\ w_3 \end{pmatrix} = \begin{pmatrix} \frac{d}{dt}-1 & 0 \\ \frac{d}{dt} & 1 \\ 0 & 1 \end{pmatrix} \begin{pmatrix} \ell_1 \\ \ell_2 \end{pmatrix} \tag{7.18}$$

be a hybrid representation of $\mathcal{B}_{ext}$. It is easy to see that ℓ_1 can be chosen as an output for the full system, but in any input-output partition of (w, ℓ), ℓ_2 has to be chosen as an input.

The matrix

$$S = \begin{pmatrix} (R & | & -M)_{\Xi} \\ & 0_{3\times 5} & \end{pmatrix}$$

is

$$S = \begin{pmatrix} 0 & 0 & 0 & -1 & 0 \\ 1 & 1 & 0 & -1 & 0 \\ 1 & 1 & 1 & 0 & 0 \\ 0 & 0 & 0 & 0 & 0 \\ 0 & 0 & 0 & 0 & 0 \\ 0 & 0 & 0 & 0 & 0 \end{pmatrix}, \tag{7.19}$$

and the matrix T is

$$T = \begin{pmatrix} -1 & 0 \\ 0 & 1 \\ 0 & 1 \\ 1 & 0 \\ 1 & 0 \\ 0 & 0 \end{pmatrix}. \tag{7.20}$$

E can be computed as

$$\begin{aligned} E = \langle (0 \quad 1 \quad -1 \quad 0 \quad 0 \quad 0), (1 \quad 0 \quad 0 \quad 1 \quad 0 \quad 0), (0 \quad 0 \quad 0 \quad 1 \quad -1 \quad 0), \\ (0 \quad 0 \quad 0 \quad 0 \quad 0 \quad 1) \rangle, \end{aligned} \tag{7.21}$$

and this yields

$$X = \begin{pmatrix} 0 & 0 & 0 & -1 & 0 \\ 0 & 0 & -1 & -1 & 0 \\ 0 & 0 & 0 & 0 & 0 \\ 0 & 0 & 0 & 0 & 0 \end{pmatrix} \tag{7.22}$$

as a state-inducing map for the external behavior of the system described by (7.18). By choosing X as $X = \left(\begin{smallmatrix} 0 & 0 & 0 & -1 & 0 \\ 0 & 0 & -1 & -1 & 0 \end{smallmatrix} \right)$ a minimal state is obtained.

Remark 7.6. The results exposed up to this point provide us with a technique for computing state maps for systems in hybrid form for which ℓ is not observable from w. For ease of exposition, we will limit the investigation to the case in which M is of full column rank d; the case in which M has column rank $d' < d$ can be dealt with similarly. Note that any non–right prime matrix M can be factored as $M = \bar{M}F$, with $\bar{M}$ a right prime matrix and F a full row rank matrix;. The following result holds.

PROPOSITION 7.5. *Let a hybrid representation of a latent variable system* (2.6) *be given. Let M be factored as $M = \bar{M}F$, with F a full row rank right divisor of M and $\bar{M}$ right prime. Then the closure of the external behavior of* (2.6) *and the closure of that of*

$$R\left(\frac{d}{dt}\right) w = \bar{M}\left(\frac{d}{dt}\right) \ell \tag{7.23}$$

are the same.

Proof. See the appendix. ∎

Let us examine the consequences of Proposition 7.5: given a system with ℓ nonobservable from w, factoring out of M an appropriate right divisor F yields a hybrid representation of a system which has the same external behavior of the original one (modulo the usual closedness issues) and the latent variable observable from the manifest ones. This provides us with a technique to tackle the problem of construction of state maps for nonobservable systems. The underlying idea is the following: given $(\, R \quad | \quad -M \,)$ with M non–right prime, extract a full row rank right factor G from M, getting a representation $(\, R \quad -\bar{M} \,)$ with the same external behavior and $\bar{M}$ right prime. Computation of a polynomial matrix X_{obs} that induces a state for this system can be accomplished according to Algorithm 4. Now partition X_{obs} as $X_{obs} := (\, X_w \quad X_\ell \,)$.

Now note that $\forall\ \bar{\ell} \in \mathcal{L}_1^{loc}(\mathbb{R}, \mathbb{R}^d)$ such that $(w, \bar{\ell})$ belongs to the full behavior associated with $(\, R \quad -\bar{M} \,)$ there exists $\ell \in \mathcal{L}_1^{loc}(\mathbb{R}, \mathbb{R}^d)$ such that $G(\frac{d}{dt})\ell = \bar{\ell}$. Moreover, $(w, \ell)^T \in \mathcal{B}_f((\, R \quad | \quad -M \,))$.

This suggests that

$$x := X_{obs}\left(\frac{d}{dt}\right)\begin{pmatrix} w \\ \bar{\ell} \end{pmatrix} = X_{obs}\left(\frac{d}{dt}\right)\begin{pmatrix} w \\ G(\frac{d}{dt})\ell \end{pmatrix} = (\, X_w \quad X_\ell G \,)\left(\frac{d}{dt}\right)\begin{pmatrix} w \\ \ell \end{pmatrix}$$

is a good candidate for a state for the closure of the external behavior of the system associated with $(\, R \quad | \quad -M \,)$.

PROPOSITION 7.6. *Let $(\, R \quad | \quad -M \,)$ be given, with M a non–right prime matrix of full column rank. Let $M = \bar{M}G$, with G a full row rank matrix and $\bar{M}$ right prime.*

Assume that $X_{obs} := (\, X_w \quad X_l \,)$ is a state-inducing map for the external behavior of the system in hybrid form associated with $(\, R \quad -\bar{M} \,)$. Then

$$X := (\, X_w \quad X_l G \,) \tag{7.24}$$

defines a state-inducing map for the closure of the external behavior of the system described in hybrid form by $(\, R \quad | \quad -M \,)$.

Proof. See the appendix. ∎

Remark 7.7. Minimal state maps are obtained by choosing X_{obs} to be a minimal state-inducing map for the system associated with $(\, R \quad -\bar{M} \,)$.

The above proposition is illustrated in the following example.

Example 7.3. Let the following system in hybrid form be given:

$$\begin{pmatrix} \frac{d}{dt} & 1 \\ \frac{d}{dt} & \frac{d}{dt}^3 + 1 \end{pmatrix} w = \begin{pmatrix} \frac{d}{dt} - 1 \\ \frac{d}{dt}^2 - \frac{d}{dt} \end{pmatrix} \ell. \tag{7.25}$$

The system has ℓ nonobservable from w, since $\xi - 1$ is a nontrivial greatest right factor for $M(\xi)$. The associated system with ℓ observable from w is

$$\begin{pmatrix} \frac{d}{dt} & 1 \\ \frac{d}{dt} & \frac{d}{dt}^3 + 1 \end{pmatrix} w = \begin{pmatrix} 1 \\ \frac{d}{dt} \end{pmatrix} \ell. \tag{7.26}$$

Computing a state-inducing map for the full system described by (7.26) yields

$$(R \quad | \quad -M)_{\Xi} = \begin{pmatrix} 1 & 0 & 0 \\ 1 & \xi^2 & -1 \\ 0 & 0 & 0 \\ 0 & \xi & 0 \\ 0 & 0 & 0 \\ 0 & 1 & 0 \end{pmatrix} \tag{7.27}$$

and the matrix T is

$$T = (1 \quad 0 \quad 0 \quad 1 \quad 0 \quad 0 \quad 0 \quad 0)^T. \tag{7.28}$$

T has left nullspace described by

$$\begin{pmatrix} 1 & 0 & 0 & -1 & 0 & 0 & 0 & 0 \\ 0 & 1 & 0 & 0 & 0 & 0 & 0 & 0 \\ 0 & 0 & 0 & 0 & 0 & 1 & 0 & 0 \\ 0 & 0 & 1 & 0 & 0 & 0 & 0 & 0 \\ 0 & 0 & 0 & 0 & 1 & 0 & 0 & 0 \\ 0 & 0 & 0 & 0 & 0 & 0 & 1 & 0 \\ 0 & 0 & 0 & 0 & 0 & 0 & 0 & 1 \end{pmatrix} \tag{7.29}$$

and therefore a (minimal) state-inducing map for the external observable behavior is obtained multiplying

$$\begin{pmatrix} (R \quad | \quad -M)_{\Xi} \\ 0_{2\times 5} \end{pmatrix}$$

on the left by the first three rows of (7.29), yielding

$$\begin{pmatrix} 1 & -\frac{d}{dt} & 0 \\ 1 & \frac{d^2}{dt^2} & -1 \\ 0 & 1 & 0 \end{pmatrix}. \tag{7.30}$$

The last column of this matrix corresponds to $X_{obs,\ell}$ as in Proposition 7.6. The state for the external behavior of the nonobservable system is therefore induced by

$$\begin{pmatrix} 1 & -\frac{d}{dt} & 0 \\ 1 & \frac{d^2}{dt^2} & -(\frac{d}{dt} - 1) \\ 0 & 1 & 0 \end{pmatrix}. \tag{7.31}$$

The following equations can be written for the state induced by the map in (7.31):

$$\begin{pmatrix} \frac{d}{dt} & 1 & 0 \\ 0 & \frac{d}{dt} & 0 \\ 1 & 0 & \frac{d}{dt} \\ 0 & 0 & 1 \end{pmatrix} x + \begin{pmatrix} -1 & 1 & 0 \\ 0 & 1 & 0 \\ -1 & 0 & 0 \\ 0 & -1 & 0 \end{pmatrix} \begin{pmatrix} w \\ \ell \end{pmatrix} = 0. \tag{7.32}$$

8. State maps for systems in image form. Image representations

$$w = M\left(\frac{d}{dt}\right)\ell, \tag{8.1}$$

where $w \in (\mathbb{R}^q)^{\mathbb{R}}$, $\ell \in (\mathbb{R}^d)^{\mathbb{R}}$, $M \in \mathbb{R}^{q\times d}[\xi]$, of the behavior of a linear time-invariant differential system have been introduced in section 2 in connection with the notion

of controllability: the behavior of a system has an image representation if and only if the system is controllable.

In this section we consider the determination of state maps for the external behavior of systems whose full behavior is described by (8.1). These may be considered to be a special case of systems representable in hybrid form, with $R = I_q$. Therefore, before stating the results pertaining to image representations, let us make some important considerations in the light of the results given in the previous section.

Let us restrict attention to the case in which $M(\xi)$ of (8.1) is right prime; i.e., the latent variable ℓ is observable from w. Note that in a system whose behavior is described by (8.1), the latent variables can be chosen as playing the role of outputs in the full behavior. This is most easily seen by considering that for full column rank M a suitable subset R_1 of the columns of the $q \times q$ identity matrix exists such that

$$(\,R_1 \quad M\,) \tag{8.2}$$

is nonsingular and, arranging the columns of a complementary canonical basis of R_1 in $\mathbb{R}^q$ in a matrix R_2, we have

$$(\,R_1 \quad M\,)^{-1} R_2 \tag{8.3}$$

proper. Then the external variables corresponding to R_2 can be chosen as inputs, while those corresponding to R_1 and the latent variable ℓ can be chosen as outputs for the full behavior.

This result, together with Proposition 7.1, allows us to conclude that for observable image representations the dimensions of the minimal state space for the closure of the external and for the full behavior are equal.

Consider now the problem of determining a state-inducing map for an observable image representation. The following theorem is an immediate consequence of the considerations made so far.

THEOREM 8.1. *Let a system be represented in image form with ℓ observable from w. A polynomial $\bullet \times (q+d)$ matrix X defines a state map for the system* (8.1) *(i.e., $w = M(\frac{d}{dt})\ell$, $x = X(\frac{d}{dt})\binom{w}{\ell}$ defines a state system) if and only if there exist a constant matrix $A \in \mathbb{R}^{\bullet\times\bullet}$ and a polynomial matrix $B \in \mathbb{R}^{\bullet\times q}$ such that $(\,I_q \quad -M\,)_\Xi = AX + B\,(\,I_q \quad -M\,)$ and the variable x is properly eliminable from*

$$(\,I_q \quad -M\,)\begin{pmatrix}\frac{d}{dt}\end{pmatrix}\begin{pmatrix} w \\ \ell \end{pmatrix} = 0, \qquad X\begin{pmatrix}\frac{d}{dt}\end{pmatrix}\begin{pmatrix} w \\ \ell \end{pmatrix} = x. \tag{8.4}$$

Note that $\sigma_+^k\,(\,I_q \quad | \quad -M\,) = (\,0_{q\times q} \quad -\sigma_+^k M\,)$, $k \in \mathbb{N}$, and therefore any state map for the system (8.1), after suitable rearrangement of the equations, may be considered to be ℓ induced. This suggests the following algorithm for the computation of a state map for the external behavior of the system described by (8.1). Note that it is effectively a restatement of Algorithm 4.

ALGORITHM 5 (Construction of a state map for the external behavior of a system in image form with ℓ observable from w).

Data: $M \in \mathbb{R}^{q\times d}[\xi]$, of degree L, M right prime.

Output: $X \in \mathbb{R}^{\bullet\times d}[\xi]$ inducing through $x = X(\frac{d}{dt})\ell$ a state for the external behavior of the system described in image form by (8.1).

Step 1. Set $M^0 := M$ and compute
$M^{k+1} := \sigma_+^k M$, $k = 1, \ldots, L$.

Step 2. $X := \operatorname{col}(M^k)_{k=1,\ldots,L}$.

Step 3. *Stop.*

Remark 8.1. The case in which M of (8.1) is not right prime, i.e., the case in which ℓ is not observable from w, can be dealt with in a manner completely analogous to that described in Remark 7.6.

As noted above, $\sigma_+^k \begin{pmatrix} I_q & | & -M \end{pmatrix} = \begin{pmatrix} 0_{q\times q} & -\sigma_+^k M \end{pmatrix}$, $k \in \mathbb{N}$, and therefore the structure of the space Ξ_M is particularly important for the determination of state maps. In view of the results exposed in the next section, let us pursue further investigation of the structure of the space Ξ_M. Without loss of generality (possibly, permuting the rows) consider $M(\xi)$ partitioned as

$$M = \begin{pmatrix} N \\ D \end{pmatrix} \tag{8.5}$$

with D nonsingular and ND^{-1} proper. Equivalently, choose D as a nonsingular $d \times d$ submatrix of M of maximal determinantal degree.

Let us state the following two propositions, which are of independent interest and yield the main result regarding the structure of the space Ξ_M.

PROPOSITION 8.2. *Let $N \in \mathbb{R}^{p\times d}[\xi]$, $D \in \mathbb{R}^{d\times d}[\xi]$, $\det(D) \neq 0$, be two polynomial matrices such that ND^{-1} is proper. Then $\Xi_N \subseteq \Xi_D$.*

Proof. See the appendix. □

PROPOSITION 8.3. *Let $D \in \mathbb{R}^{d\times d}[\xi]$ be a nonunimodular polynomial matrix with $\det(D) \neq 0$. Then $\Xi_D = \{r \in \mathbb{R}^{1\times d}[\xi] \mid rD^{-1}$ strictly proper$\}$.*

Proof. See the appendix. □

The next proposition states the main result regarding the structure of the space Ξ_M.

PROPOSITION 8.4. *$\Xi_M = \Xi_D = \{r \in \mathbb{R}^{1\times d}[\xi] \mid rD^{-1}$ is strictly proper$\}$.*

Proof. See the appendix. □

The interest in considering state maps for systems represented in image form arises not only from the controllability issue but also from the connections among image representations and the notion of transfer function as given in the behavioral framework. This is the subject of next section.

9. Transfer functions and state maps. The purpose of this section is to make contact with the algebraic approach to the realization problem, put forward in [4] and extensively studied by Fuhrmann [2].

Consider the input-output system

$$P\left(\frac{d}{dt}\right) y = Q\left(\frac{d}{dt}\right) u \tag{9.1}$$

with $P \in \mathbb{R}^{p\times p}[\xi]$, $Q \in \mathbb{R}^{p\times m}[\xi]$, $\det(P) \neq 0$, and $P^{-1}Q$ proper. The function $G := P^{-1}Q \in \mathbb{R}^{p\times m}(\xi)$ is called the *transfer function* of (9.1).

The latent variable system

$$\begin{aligned} u &= D\left(\frac{d}{dt}\right)\ell, \\ y &= N\left(\frac{d}{dt}\right)\ell \end{aligned} \tag{9.2}$$

with $D \in \mathbb{R}^{m\times m}[\xi]$, $N \in \mathbb{R}^{p\times m}[\xi]$, $\det(D) \neq 0$, and ND^{-1} proper defines an input-output system with transfer function $G = ND^{-1}$.

It can be shown that two systems have the same transfer function if and only if they have the same controllable part (see [9, p. 248]). The (unique) controllable

system which has a given transfer function $G \in \mathbb{R}_+^{p\times m}(\xi)$ can be obtained by making a left coprime factorization $G = P^{-1}Q$ of G and considering (9.1) or by making a right factorization $G = ND^{-1}$ and considering (9.2); if the latter factorization is right coprime, then (9.2) will be an observable image representation of the controllable system with transfer function G (see [9, pp. 249, 250]).

Note that the algorithms described in sections 6 and 8 can be directly applied in order to obtain a state-space realization of a system with a given transfer function.

Transfer functions play a prominent role in control theory, since they provide a natural framework in many engineering applications. The concept of realization as put forward in [4], associated with the notion of an input-output map, is intimately connected with the notion of transfer function. Not surprisingly, therefore, many formalizations of the notion of state starting from an input-output or transfer function point of view have been given in the past.

The algorithms proposed in our paper are very akin to those of Fuhrmann [2]. The module structure on which his approach is based, has many connections with left and right factorizations of transfer functions. In particular, the state space corresponding to a right factorization ND^{-1} of a transfer function is defined therein to be isomorphic to the vector space K_D defined as

$$K_D := \{f \in \mathbb{R}^{1\times d}[\xi] \mid fD^{-1} \in \mathbb{R}_+^{1\times d}(\xi)\} \tag{9.3}$$

(cf. [2, Lemma 2-15, p. 11, and Theorem 10-2, p. 41]). The connection with the result of Proposition 8.4 is evident.

10. Conclusions. In modeling physical systems the most natural way of proceeding is to write a set of high-order differential equations possibly with algebraic constraints among the variables. When it comes to simulation of the corresponding system, however, state-space equations are the most natural representation to use. Therefore the need arises to compute the latter from the former. In this paper a characterization of state-inducing maps has been given for systems given in kernel or in hybrid representations. This characterization suggests immediately algorithms to actually perform a computation of the state function from which state-space equations are easily recovered.

Appendix A. Notation.

$\mathbb{N}$ natural numbers (0 is not included).
$\mathbb{Z}_+$ nonnegative integers.
$\mathbb{R}$ real numbers.
2^A set whose members are the subsets of A.
$\mathbb{R}[\xi]$ polynomials with real coefficients.
$\mathbb{R}_+(\xi)$ proper rational functions.
e_i the ith vector of a canonical basis vector in $\mathbb{R}^{1\times\bullet}$.
$\mathbb{R}^{g\times q}$ $g \times q$ real matrices.
$\mathbb{R}^{\bullet\times q}$ real matrices with q columns.
$\operatorname{col}(r_1, \ldots, r_n)$ the matrix

$$\begin{pmatrix} r_1 \\ r_2 \\ \vdots \\ r_n \end{pmatrix}.$$

$\operatorname{diag}(x_k)_{k=1,\ldots,r}$ $r \times r$ diagonal matrix with diagonal elements x_k.

$\mathbb{R}^{g\times q}[\xi]$ $g\times q$ polynomial matrices in the indeterminate ξ.
$\mathbb{R}^{\bullet\times q}$ polynomial matrices in the indeterminate ξ with q columns.
$\mathbb{R}^{g\times q}_{+}(\xi)$ $g\times q$ matrices of strictly proper rational functions.
$(W)^T$ maps from W to T.
$\mathcal{C}^\infty(\mathbb{R},\mathbb{R}^q)$ infinitely differentiable functions from $\mathbb{R}$ to $\mathbb{R}^q$.
$\mathcal{L}_1^{loc}(\mathbb{R};\mathbb{R}^q)$ locally integrable functions from $\mathbb{R}$ to $\mathbb{R}^q$.
$\langle r_1,\dots,r_n\rangle$ space spanned by the vectors r_i.
π_w projection on the w variables: $\pi_w(w,\ell):=w$.
$\circ$ composition of maps.
$[p]$ equivalence class with representative p.

Appendix B. Proofs.

B.1. Proof of Theorem 2.1. That $w\in\overline{\pi_w(\mathcal{B}_f)}^{closure}$ implies $R_1'(\frac{d}{dt})w=0$ is easy to see. To prove the converse, let $\mathcal{B}_1$ be the behavior of $R_1'(\frac{d}{dt})w=0$, and observe that $\mathcal{B}_1\bigcap\mathcal{C}^\infty$ is dense in $\mathcal{B}_1$. Let $M_2'\in\mathbb{R}^{n_1\times n_2}[\xi]$. Obviously $M_2'(\frac{d}{dt})$ maps $\mathcal{C}^\infty(\mathbb{R},\mathbb{R}^{n_2})$ into $\mathcal{C}^\infty(\mathbb{R},\mathbb{R}^{n_1})$. Since M_2' is of full row rank, this map is surjective. (In order to see this, use the Smith form of M_2'.) Hence for all $w\in\mathcal{B}_1\bigcap\mathcal{C}^\infty$ there exists a $(w,\ell)\in\mathcal{B}_f\bigcap\mathcal{C}^\infty$. This shows $\mathcal{B}_1=\overline{\pi_w(\mathcal{B}_f)}^{closure}$. ∎

B.2. Proof of Proposition 3.1. We will prove only the Markovian case, the state-space case being entirely equivalent. The "if" part is trivial. To show the "only if" case, assume that (2.4) satisfies the concatenability condition. Without loss of generality we can assume that R has full row rank. Also, there exists a unimodular $U\in\mathbb{R}^{\bullet\times\bullet}[\xi]$ such that $R':=UR$ is in row reduced form, meaning that the matrix formed by the coefficients of the highest powers in ξ of the rows of $R'(\xi)$ has full row rank. It is easy to see that systems with kernel representations defined by R and R' are the same. We will now show that R' is a first order polynomial matrix. Assume the contrary. Write R' in input-output form:

$$P\left(\frac{d}{dt}\right)w_1=Q\left(\frac{d}{dt}\right)w_2 \tag{B.1}$$

with $\det(P)\neq 0$ and $P^{-1}Q$ proper. The assumption that R' is not first order implies that P is not. From the assumption that $R(\frac{d}{dt})w=0$ is Markovian, it follows that also $P(\frac{d}{dt})w_1=0$ is. Now let w_1', w_1'' be solutions of $P(\frac{d}{dt})w=0$ with $w_1'(0)=w_1''(0)$. Since $\det(P)\neq 0$, w_1' and w_1'' are also $\mathcal{C}^\infty$ and by the state property are concatenable. In order to obtain a contradiction it suffices therefore to prove Proposition 3.1 for autonomous systems. This, however, is an immediate consequence of the following lemma.

LEMMA B.1. *Let the autonomous system $R(\frac{d}{dt})w=0$ with $R\in\mathbb{R}^{q\times q}[\xi]$, $\det R\neq 0$, be Markovian. Then this system admits the kernel representation*

$$Fw+E\frac{d}{dt}w=0 \tag{B.2}$$

with E, $F\in\mathbb{R}^{q\times q}$ and $\det(E\xi+F)\neq 0$

Proof. Let $\mathcal{B}$ be the behavior of $R(\frac{d}{dt})w=0$. Let $P(\frac{d}{dt})w=0$ be the corresponding representation in row reduced form, as in the above proof. Write it as

$$P_0w+P_1\frac{d}{dt}w+\cdots+P_L\frac{d}{dt}w=0. \tag{B.3}$$

We need to prove that it is first order. Assume that this is not the case and that $P_L \neq 0$ and $L \geq 2$.

Denote with L_k, $k = 1, \ldots, q$, the highest order of differentiation of w_k in (B.3). Note that there is at least one $L_k \geq 2$. Introduce the auxiliary variables z_i^k defined as

$$z_i^k := \frac{d^i w_k}{dt^i}, \tag{B.4}$$

$k = 1, \ldots, q$, $i = 0, \ldots, L_k - 1$, and define

$$z := \begin{pmatrix} z_0^1 & z_1^1 & \ldots & z_{L_1-1}^1 & \ldots & z_0^q & \ldots & z_{L_q-1}^q \end{pmatrix}. \tag{B.5}$$

Now consider the system with latent variable z, described by the equations

$$\begin{aligned} \frac{d}{dt} z &= Fz, \\ w_k &= z_0^k, \quad k = 1, \ldots, q, \end{aligned} \tag{B.6}$$

where the entries of the $\sum_{k=1}^q L_k \times \sum_{k=1}^q L_k$ matrix F are determined from (B.3) and the definitions (B.4). Equations (B.6) represent a system in hybrid form with latent variable z; it external behavior coincides with that described by (B.3), as can be checked by applying the latent-variable-elimination theorem.

However, the external behavior of (B.6) does not enjoy the Markovianity property. In fact, (B.6) has exactly one solution (w, z) for each initial condition vector

$$\begin{pmatrix} z_0^1(0) & z_1^1(0) & \ldots & z_{L_1-1}^1(0) & \ldots & z_0^q(0) & \ldots & z_{L_q-1}^q(0) \end{pmatrix}. \tag{B.7}$$

This contradicts Markovianity, since two solutions (w, z), (w', z') of (B.6) with $z_0^k(0) = z_0'^k(0)$, $k = 1, \ldots, q$, cannot be concatenated unless also $z_j^k(0) = z_j'^k(0)$, $j = 1, \ldots, L_k - 1$, $k = 1, \ldots, q$. □

B.3. Proof of Proposition 6.1. The behavior described by $R(\frac{d}{dt})w(t) = 0$ with $w \in \mathcal{L}_1^{loc}$ is the set of all w for which

$$\int_{-\infty}^{+\infty} w^T(t) \left(R\left(-\frac{d}{dt}\right)^T \right) f(t) dt = 0 \tag{B.8}$$

for all testing functions $f(t)$ (that is, f is a $\mathcal{C}^\infty$ vector-valued function with compact support).

(Only if) Assume that $w \in \mathcal{B}$ and $0 \wedge w \in \mathcal{B}$. Define $R^k := \sigma_+^k(R)$. We will show that $(R^k(\frac{d}{dt})w)(0) = 0$ for $k = 1, 2, \ldots$. Consider

$$\int_{-\infty}^{+\infty} (0 \wedge w)^T(t) \left(R\left(-\frac{d}{dt}\right)^T \right) f(t) dt. \tag{B.9}$$

This obviously equals

$$\int_{0}^{+\infty} w^T(t) \left(R\left(-\frac{d}{dt}\right)^T \right) f(t) dt. \tag{B.10}$$

Since $R^k(\frac{d}{dt})w$ is locally integrable $\forall\ k = 1, \ldots, \deg(R)$, (B.10) may be integrated by parts and equals

$$\int_0^{+\infty} \left(R\left(\frac{d}{dt}\right) w(t) \right)^T f(t) dt + \sum_{k=1}^{L} \sum_{j=k}^{L} (-1)^{k-1} (w^{(j-k)}(0))^T R_j^T f^{(k-1)})(0). \tag{B.11}$$

But since $w \in \mathcal{B}$, the integral in (B.11) is zero. Since $0 \wedge w \in \mathcal{B}$, (B.9) and hence (B.11) are also zero, and therefore so is the double sum in (B.11). Hence, due to the arbitrariness of the testing function f,

$$\left(R^k\left(\frac{d}{dt}\right)w\right)(0) = 0 \tag{B.12}$$

$\forall\ k = 1, \ldots, \deg(R)$.

(If) Assume that $w \in \mathcal{B}$ satisfies $(R^k(\frac{d}{dt})w)(0) = 0$ for $k = 1, 2, \ldots$. We want to show that $0 \wedge w \in \mathcal{B}$. To prove this, we have to prove that the integral (B.9) is zero for all testing functions f. Proceeding as above, integrating by parts, (B.11) is obtained. Now the claim is obtained by noting that since $w \in \mathcal{B}$, $R(\frac{d}{dt})w = 0$ holds and therefore the integral is zero, while the double sum in (B.11) is zero by assumption. ∎

B.4. Proof of Theorem 6.2. Let us first prove the following

LEMMA B.2. *Let $\mathcal{B}$ be the behavior of* (2.4). *Let X_1, $X_2 \in \mathbb{R}^{\bullet\times q}[\xi]$. Assume that for all $w \in \mathcal{B} \bigcap \mathcal{C}^\infty(\mathbb{R}, \mathbb{R}^q)$ there holds*

$$\left\{\left(X_1\left(\frac{d}{dt}\right)w\right)(0) = 0\right\} \Longrightarrow \left\{\left(X_2\left(\frac{d}{dt}\right)w\right)(0) = 0\right\}. \tag{B.13}$$

Then there exist $A \in \mathbb{R}^{\bullet\times\bullet}$ and $B \in \mathbb{R}^{\bullet\times\bullet}[\xi]$ such that

$$X_2(\xi) = AX_1(\xi) + B(\xi)R(\xi). \tag{B.14}$$

Proof. We will prove this lemma only in the case that (2.4) defines a controllable system. The general case is left to the reader. Using the Smith form for R it follows that there exist unimodular matrices U and V such that $URV = (\,I \quad 0\,)$. Let $v := V^{-1}(\frac{d}{dt})w$. Then $w \in \mathcal{B}$ if and only if $(\,I \quad 0\,)v = 0$. Define $X_1' := X_1V$ and $X_2' := X_2V$. Partition v, X_1', and X_2' as $\binom{v_1}{v_2}$, $(\,X_{11}' \quad X_{12}'\,)$, $(\,X_{21}' \quad X_{22}'\,)$, with the partition induced by $(\,I \quad 0\,)$. Then for any $v_2 \in \mathcal{C}^\infty$ there holds

$$\left\{\left(X_{12}'\left(\frac{d}{dt}\right)v_2\right)(0) = 0\right\} \Longrightarrow \left\{\left(X_{22}'\left(\frac{d}{dt}\right)v_2\right)(0) = 0\right\}. \tag{B.15}$$

This implies that there exists a matrix $A \in \mathbb{R}^{\bullet\times\bullet}$ such that

$$X_{22}'\left(\frac{d}{dt}\right) = AX_{12}'\left(\frac{d}{dt}\right). \tag{B.16}$$

This yields that X_2' is of the form

$$X_2'(\xi) = AX_1'(\xi) + B(\xi)\,(\,I \quad 0\,). \tag{B.17}$$

Now postmultiply by V^{-1}. ∎

This lemma yields the claim of the theorem as follows.

(Only if) Assume that $X(\frac{d}{dt})$ is a state map. Then $x = X(\frac{d}{dt})w$ is properly eliminable from

$$\begin{aligned} R\left(\frac{d}{dt}\right)w &= 0, \\ X\left(\frac{d}{dt}\right)w &= x. \end{aligned} \tag{B.18}$$

Moreover, $(X(\frac{d}{dt})w)(0) = 0$ implies that w is concatenable with the zero trajectory. Proposition 6.1 states that concatenability with zero is equivalent to $(R_\Xi(\frac{d}{dt})w)(0) = 0$. One has to apply now the above lemma with $X_1 = X$ and $X_2 = R_\Xi$.

(If) Assume that (6.7) holds. Recall (cf. the beginning of section 6) that $R_\Xi(\frac{d}{dt})w$ is absolutely continuous. Now consider $w \in \mathcal{B}$ such that $X(\frac{d}{dt})w$ is continuous at $t = 0$. Then $R_\Xi = AX + BR$ implies that $(BR)(\frac{d}{dt})w$ is continuous at $t = 0$, so that $(R_\Xi(\frac{d}{dt})w)(0) = (AX(\frac{d}{dt})w)(0) + (BR(\frac{d}{dt})w)(0)$ and $(X(\frac{d}{dt})w)(0) = 0$ imply $(R_\Xi(\frac{d}{dt})w)(0) = 0$ since $(BR)(\frac{d}{dt})w = 0$ and $(BR)(\frac{d}{dt})w$ is continuous at $t = 0$. Therefore by Proposition 6.1 one concludes that $0 \wedge w \in \mathcal{B}$. By assumption, $x = X(\frac{d}{dt})w$ is properly eliminable from (B.18), and therefore (B.18) defines a state-space system with external behavior $\operatorname{Ker} R(\frac{d}{dt})$. ∎

B.5. Proof of Proposition 6.4. The claim follows directly by applying the second of the lemmas below. To get to that result, let us first consider the following lemma.

LEMMA B.3. *Let $p \in \mathbb{R}^{1\times q}[\xi]$, $R \in \mathbb{R}^{q\times g}[\xi]$. Then*

$$\sigma_+(pR) = (\sigma_+ p)R + p(0)(\sigma_+ R). \tag{B.19}$$

Proof. Let $p = (\,p_1 \quad \dots \quad p_q\,)$, $p_i = \sum_{j=0}^{n} p_{ji}\xi^j$, and $R = \operatorname{col}(R_i)_{i=1,\dots,q}$, $R_i \in \mathbb{R}^{1\times g}[\xi]$. Note that

$$pR = \sum_{i=1}^{q} \left(\sum_{j=0}^{n} p_{ji}\xi^j R_i \right) \tag{B.20}$$

and therefore that

$$\sigma_+(pR) = \sigma_+ \left(\sum_{i=1}^{q} \left(\sum_{j=0}^{n} p_{ji}\xi^j R_i \right) \right), \tag{B.21}$$

which is equivalent to $\sum_{i=1}^{q} \sigma_+(\sum_{j=0}^{n} p_{ji}\xi^j R_i)$. This is equivalent to

$$\sum_{i=1}^{q} \left(\sum_{j=1}^{n} p_{ji}\xi^{j-1} R_i + p_{0i}\sigma_+(R_i) \right), \tag{B.22}$$

which yields $\sum_{i=1}^{q} (\sigma_+ p_i)R_i + \sum_{i=1}^{q} p_{0i}(\sigma_+ R_i)$ and the claim. ∎

This lemma explains how σ_+ acts on vector multiples of a given matrix. The next one shows how σ_+ acts on unimodular matrix multiples.

LEMMA B.4. *Let R, R' be matrices related as*

$$R = UR' \tag{B.23}$$

for a unimodular U. Then

$$R_\Xi = VR'_\Xi + BR' \tag{B.24}$$

with V a constant full column rank matrix and B a polynomial matrix.

Proof. The proof follows trivially from Lemma B.3 and the fact that a unimodular matrix U has $\det(U(0)) \neq 0$. ∎

B.6. Proof of Proposition 6.6. Let us first prove that the nonzero rows of R_Ξ form a basis of Ξ_R if and only if R is in row reduced form.

Define R'_Ξ to be the submatrix of R_Ξ consisting of the nonzero rows of R_Ξ. Let us prove sufficiency. Assume that R'_Ξ has not full row rank. Then there exists at least one row which is a linear combination of the others. Since $R_\Xi := \mathrm{col}(R^k)_{k=1,\dots,L}$, the highest coefficient vector of this row is the same as that of the corresponding row of R, and is a linear combination of the highest coefficient vectors of the other rows of R. But this contradicts row reducedness. The proof of necessity goes along the same lines.

Note that the result just proven implies necessity of the claim of the proposition. Let us prove sufficiency. Let module(R) be the module of $R^{1\times q}[\xi]$ generated by the rows of R. We have to prove that the intersection of Ξ_R and module(R) consists of the zero vector only.

We prove this as follows. Let ν_i, $i = 1, \dots, g$, be the degree of the ith row R_i of R, and assume that the rows of R have been ordered so that $\nu_1 = \nu_2 = \cdots = \nu_{g'} > \nu_{g'+1} \geq \cdots \geq \nu_g$.

Let now $y \in \Xi_R \bigcap \mathrm{module}(R)$. Note first that, since $y \in \Xi_R$, $\deg(y) \leq \nu_1 - 1$. Then note that $y \in \mathrm{module}(R)$ implies $y = xR$ for some $x = (\, x_1, \dots, x_g\,) \in \mathbb{R}^{1\times g}[\xi]$. From the predictable degree property of R [3, p. 387], we conclude that $\deg(y) = \max_{1\leq i\leq g,\ x_i\neq 0}\{\deg(x_i) + \nu_i\}$ and therefore, since $\deg(y) \leq \nu_1 - 1$, that $x_i = 0$, $1 \leq i \leq g'$.

Assume now that $x_i = 0$ for $g' + 1 \leq i \leq \bar{g} < g$, $x_{\bar{g}+1} \neq 0$. By the predictable degree property of R, $\deg(y) \geq \nu_{\bar{g}+1}$. Since $y \in \Xi_R$, $y = \sum_{j=1}^{g}\sum_{i=1}^{\nu_j} \alpha_{ij}(\sigma_+^i R_j)$, for suitable scalars $\alpha_{ij} \in \mathbb{R}$. Since $\deg(y) \geq \nu_{\bar{g}+1}$, at least one of the α_{ij}'s with $1 \leq j \leq \bar{g}$ must be nonzero, since the only generators of Ξ_R of degree $\geq \nu_{\bar{g}+1}$ are to be found among the vectors $(\sigma_+^i R_j)$, $1 \leq j \leq \bar{g}$. This implies that the highest coefficient of y is a linear combination of the first $\bar{g}$ rows of R_{hc}, the highest row coefficient matrix of R. On the other hand, since $x_i = 0$ for $1 \leq i \leq \bar{g}$, and $x_{\bar{g}+1} \neq 0$, the highest coefficient of $y = xR$ is a linear combination of the last $g - \bar{g}$ rows of R_{hc}. But this implies that the first $\bar{g}$ and the last $g - \bar{g}$ rows of R_{hc} generate the same vector; this, by row reducedness, is possible if and only if this vector is zero. Therefore $x_i = 0$ for all i; that is, $y = 0$ as was to be proven. □

B.7. Proof of Proposition 7.1. The system represented by (2.6) has ℓ observable from w. Therefore it allows a representation of the form

$$N\left(\frac{d}{dt}\right) w = \ell,$$

$$R'_1\left(\frac{d}{dt}\right) w = 0, \tag{B.25}$$

with R'_1 of full row rank. Now partition R'_1 as $(\, P_1 \quad Q_1\,)$ with $P_1^{-1}Q_1$ proper. This induces a partition (y, u) on w in inputs u and outputs y. The proposition follows immediately from examining the minors of

$$\begin{pmatrix} P_1 & Q_1 & 0 \\ N_1 & N_2 & -I \end{pmatrix}. \tag{B.26}$$

In fact, sufficiency can be proven as follows. Since (y, ℓ) consists of outputs for the full system, it follows that

$$\begin{pmatrix} P_1 & 0 \\ N_1 & -I \end{pmatrix} \tag{B.27}$$

has maximal degree among the minors of (B.26). By Corollary 6.7 this implies that the minimal dimension of the state space for the full behavior equals $\deg(\det(\left(\begin{smallmatrix} P_1 & 0 \\ N_1 & -I \end{smallmatrix}\right))) = \deg(\det(P_1))$. Now note that $\deg(\det(P_1))$ equals the minimal dimension of the state space for the external behavior since $\det(P_1)$ has maximal degree among the minors of R_1'. This yields the claim.

As for necessity, note that if the minimal dimensions of the state space for the external and the full behavior are the same, $\deg(\det(P_1))$ equals the maximal degree of the minors of (B.26). Since $\det(\left(\begin{smallmatrix} P_1 & 0 \\ N_1 & -I \end{smallmatrix}\right))$ has degree $\deg(\det(P_1))$, it has maximal degree among the minors of (B.26) and therefore the corresponding partition of the (w, ℓ) variables, that is, (y, ℓ), is a set of outputs for the full system. □

B.8. Proof of Proposition 7.2. Before proving the proposition, let us point out the following general result. Assume that in a hybrid representation ℓ is observable from w; then $w(t) = 0 \ \forall \ t < 0$ implies $\ell(t) = 0 \ \forall \ t < 0$. This is proven as follows. Observability implies that there exists a polynomial differential operator $F(\frac{d}{dt})$ such that $F(\frac{d}{dt})w = \ell$; then for $t < 0$ $(F(\frac{d}{dt})w)(t) = (F(\frac{d}{dt})0)(t) = 0$.

Let us turn to the proof of the proposition.

(Only if) External concatenability with zero is equivalent to

$$\int_{-\infty}^{+\infty} \begin{pmatrix} 0 \wedge w \\ \ell \end{pmatrix}^T (t) \left(\begin{pmatrix} R & -M \end{pmatrix}^T \left(-\frac{d}{dt} \right) \right) f(t) dt = 0 \tag{B.28}$$

for every testing function f.

Observability of ℓ from w and the remark made above imply that integration can be considered in $[0, +\infty)$ only:

$$\int_{0}^{+\infty} \begin{pmatrix} w \\ \ell \end{pmatrix}^T (t) \left(\begin{pmatrix} R & -M \end{pmatrix}^T \left(-\frac{d}{dt} \right) \right) f(t) dt = 0. \tag{B.29}$$

Note that since ℓ is a locally integrable function, there exists an absolutely continuous function $\mathcal{L}$ such that $\frac{d}{dt}\mathcal{L} = \ell$ almost everywhere.

Then $R(\frac{d}{dt})w = M(\frac{d}{dt})\ell$ may be written as

$$R\left(\frac{d}{dt}\right) w = M\left(\frac{d}{dt}\right) \frac{d}{dt}\mathcal{L}, \tag{B.30}$$

and if

$$\begin{pmatrix} R & -M \end{pmatrix} = \begin{pmatrix} R_0 & -M_0 \end{pmatrix} + \begin{pmatrix} R_1 & -M_1 \end{pmatrix} \xi + \cdots + \begin{pmatrix} R_L & -M_L \end{pmatrix} \xi^L, \tag{B.31}$$

(B.30) corresponds to the polynomial matrix

$$\begin{aligned} \begin{pmatrix} R & -M \end{pmatrix}' := \begin{pmatrix} R_0 & 0 \end{pmatrix} + \begin{pmatrix} R_1 & -M_0 \end{pmatrix} \xi + \cdots \\ + \begin{pmatrix} R_L & -M_{L-1} \end{pmatrix} \xi^L + \begin{pmatrix} 0 & -M_L \end{pmatrix} \xi^{L+1}, \end{aligned} \tag{B.32}$$

which is more conveniently written as $\sum_{j=0}^{L+1} \begin{pmatrix} R_j & -M_{j-1} \end{pmatrix} \xi^j$, defining $M_{-1} := 0$, $R_{L+1} := 0$.

Equation (B.29) corresponds then to

$$\int_{0}^{+\infty} \begin{pmatrix} w \\ \mathcal{L} \end{pmatrix}^T (t) \left(\sum_{j=0}^{L+1} (-1)^j \begin{pmatrix} R_j & -M_{j-1} \end{pmatrix}^T \frac{d^j}{dt^j} f(t) \right) dt = 0. \tag{B.33}$$

Analogously to what has been done at the beginning of section 6 it is possible to prove that the Ξ-matrix of (B.32) induces an absolutely continuous function. It is then possible to integrate by parts the left-hand side of (B.33), and this yields

$$\begin{aligned}
&\int_0^{+\infty} \left(\sum_{j=0}^{L+1} \begin{pmatrix} R_j & -M_{j-1} \end{pmatrix} \begin{pmatrix} w \\ \mathcal{L} \end{pmatrix}^{(j)} (t) \right)^T f(t)dt \\
&- \sum_{j=1}^{L+1} \left(\left(\begin{pmatrix} R_j & -M_{j-1} \end{pmatrix} \begin{pmatrix} w \\ \mathcal{L} \end{pmatrix}^{(j-1)} \right)(0) \right)^T f(t) \\
&+ \sum_{j=2}^{L+1} \left(\left(\begin{pmatrix} R_j & -M_{j-1} \end{pmatrix} \begin{pmatrix} w \\ \mathcal{L} \end{pmatrix}^{(j-2)} \right)(0) \right)^T \frac{d}{dt} f(t) + \cdots \\
&+ (-1)^{L+1} \left(\begin{pmatrix} 0 & -M_L \end{pmatrix} \begin{pmatrix} w \\ \mathcal{L} \end{pmatrix}(0) \right)^T f^{(L)}(0), \qquad \text{(B.34)}
\end{aligned}$$

where $g^{(j)}$ denotes the jth derivative of a function g (and the function itself in case $j = 0$).

Equation (B.34) can be rewritten as

$$\begin{aligned}
&\int_0^{+\infty} \left(\sum_{j=0}^{L+1} \begin{pmatrix} R_j & -M_{j-1} \end{pmatrix} \begin{pmatrix} w \\ \mathcal{L} \end{pmatrix}^{(j)} (t) \right)^T f(t)dt \\
&+ \sum_{k=1}^{L+1} \left(\sum_{j=k}^{L+1} (-1)^k \left(\begin{pmatrix} R_j & -M_{j-1} \end{pmatrix} \begin{pmatrix} w \\ \mathcal{L} \end{pmatrix}^{(j-k)} (0) \right)^T \right) f^{(k-1)}(0). \qquad \text{(B.35)}
\end{aligned}$$

Since $(0 \wedge w, \ell) \in \mathcal{B}_f$, (B.35) equals zero for all testing functions f. Now note that the integral in (B.35) is zero, since $(w, \ell) \in \mathcal{B}_f$. Therefore $(0 \wedge w, \ell) \in \mathcal{B}_f$ implies

$$\sum_{k=1}^{L+1} \left(\sum_{j=k}^{L+1} (-1)^k \left(\begin{pmatrix} R_j & -M_{j-1} \end{pmatrix} \begin{pmatrix} w \\ \mathcal{L} \end{pmatrix}^{(j-k)} (0) \right)^T \right) f^{(k-1)}(0) = 0. \qquad \text{(B.36)}$$

Arbitrariness of the testing function f then yields the set of equations

$$\sum_{j=k}^{L+1} (-1)^k \left(\begin{pmatrix} R_j & -M_{j-1} \end{pmatrix} \begin{pmatrix} w \\ \mathcal{L} \end{pmatrix}^{(j-k)} \right)(0) = 0, \qquad \text{(B.37)}$$

$k = 1, \ldots, L+1$, which is more conveniently written as

$$\begin{aligned}
&\sum_{j=k}^{L} \left(\begin{pmatrix} R_j & -M_j \end{pmatrix} \begin{pmatrix} w \\ \ell \end{pmatrix}^{(j-k)} \right)(0) = M_{k-1}\mathcal{L}(0), \quad k = 1, \ldots, L, \\
&0 = M_L \mathcal{L}(0). \qquad \text{(B.38)}
\end{aligned}$$

In matrix form, (B.38) reads as

$$\left(\left(\begin{pmatrix} R & -M \\ & 0 \end{pmatrix}_\Xi \right) \left(\frac{d}{dt} \right) \begin{pmatrix} w \\ \ell \end{pmatrix} \right)(0) = \begin{pmatrix} M_0 \\ M_1 \\ \vdots \\ M_L \end{pmatrix} \mathcal{L}(0). \qquad \text{(B.39)}$$

Now let $V := \mathrm{col}(v_1, \ldots, v_s)$, with $\{v_i\}_{i=1,\ldots,s}$ a set of generators of E. Multiply both sides of (B.39) by V. This yields

$$V\left(\begin{pmatrix} R & -M \end{pmatrix}_{\Xi}\begin{pmatrix} \frac{d}{dt} \end{pmatrix}\begin{pmatrix} w \\ \ell \end{pmatrix}\right)(0) = \left(\left(V\begin{pmatrix} R & -M \end{pmatrix}_{\Xi}\right)\begin{pmatrix} \frac{d}{dt} \end{pmatrix}\begin{pmatrix} w \\ \ell \end{pmatrix}\right)(0) = 0, \tag{B.40}$$

which is the claim of the proposition.

(If) Assume that

$$\left(\left(V\begin{pmatrix} R & -M \end{pmatrix}_{\Xi}\begin{pmatrix} \frac{d}{dt} \end{pmatrix}\right)\begin{pmatrix} w \\ \ell \end{pmatrix}\right)(0) = 0, \tag{B.41}$$

where V is a set of generators of E.

Then in particular

$$\left(V\begin{pmatrix} R & -M \end{pmatrix}_{\Xi}\begin{pmatrix} \frac{d}{dt} \end{pmatrix}\begin{pmatrix} w \\ \ell \end{pmatrix}\right)(0) = V\begin{pmatrix} M_0 \\ M_1 \\ \vdots \\ M_L \end{pmatrix}\mathcal{L}(0), \tag{B.42}$$

with $\mathcal{L}$ such that $\frac{d}{dt}\mathcal{L} = \ell$ almost everywhere. Equation (B.42) is equivalent to

$$V\left(\left(\left(\begin{pmatrix} R & -M \end{pmatrix}_{\Xi}\begin{pmatrix} \frac{d}{dt} \end{pmatrix}\begin{pmatrix} w \\ \ell \end{pmatrix}\right)(0) - \begin{pmatrix} M_0 \\ M_1 \\ \vdots \\ M_L \end{pmatrix}\mathcal{L}(0)\right)\right) = 0 \tag{B.43}$$

and therefore

$$\left(\left(\begin{pmatrix} R & -M \end{pmatrix}_{\Xi}\begin{pmatrix} \frac{d}{dt} \end{pmatrix}\begin{pmatrix} w \\ \ell \end{pmatrix}\right)(0) - \begin{pmatrix} M_0 \\ M_1 \\ \vdots \\ M_L \end{pmatrix}\mathcal{L}(0)\right) \tag{B.44}$$

belongs to the vector space generated by the columns of

$$\begin{pmatrix} M_0 \\ \vdots \\ M_L \end{pmatrix};$$

that is, there exist $\alpha_i \in \mathbb{R}$, $i = 1, \ldots, d$, such that (B.44) equals

$$\begin{pmatrix} M_0 \\ M_1 \\ \vdots \\ M_L \end{pmatrix}\begin{pmatrix} \alpha_1 \\ \vdots \\ \alpha_d \end{pmatrix}. \tag{B.45}$$

Consider now the function $\bar{\mathcal{L}}$ defined as follows: $\bar{\mathcal{L}}(t) := \mathcal{L}(t)\ \forall t \neq 0$ and

$$\bar{\mathcal{L}}(0) := \mathcal{L}(0) + \begin{pmatrix} \alpha_1 \\ \vdots \\ \alpha_d \end{pmatrix}.$$

Then $\frac{d}{dt}\bar{\mathcal{L}} = \ell$ almost everywhere and

$$\left(\begin{pmatrix} (\,R & -M\,)_{\Xi} \\ & 0 \end{pmatrix}\begin{pmatrix} d/dt \end{pmatrix}\begin{pmatrix} w \\ \ell \end{pmatrix}\right)(0) - \begin{pmatrix} M_0 \\ M_1 \\ \vdots \\ M_L \end{pmatrix}\bar{\mathcal{L}}(0) = 0. \tag{B.46}$$

Now consider $(0 \wedge w, 0 \wedge \ell)$. To show that it belongs to the full behavior, we prove that

$$R\left(\frac{d}{dt}\right)(0 \wedge w) = M\left(\frac{d}{dt}\right)\frac{d}{dt}(0 \wedge \bar{\mathcal{L}}). \tag{B.47}$$

Using the notation introduced in (B.32), note that (B.47) holds if and only if

$$\int_{-\infty}^{+\infty}\begin{pmatrix} 0 \wedge w \\ 0 \wedge \bar{\mathcal{L}} \end{pmatrix}^T(t)\left(\sum_{j=0}^{L+1}(-1)^j\,(\,R_j \quad -M_{j-1}\,)^T\frac{d^j f}{dt^j}(t)\right)dt = 0, \tag{B.48}$$

that is, if and only if

$$\int_{0}^{+\infty}\begin{pmatrix} w \\ \bar{\mathcal{L}} \end{pmatrix}^T(t)\left(\sum_{j=0}^{L+1}(-1)^j\,(\,R_j \quad -M_{j-1}\,)^T\frac{d^j f}{dt^j}(t)\right)dt = 0. \tag{B.49}$$

The Ξ-matrix of (B.32) induces an absolutely continuous function. Therefore (B.49) can be integrated by parts and, with manipulations completely analogous to those of the necessity part of the proof, this yields

$$\begin{aligned}&\int_{0}^{+\infty}\left(\sum_{j=0}^{L+1}(\,R_j \quad -M_{j-1}\,)\begin{pmatrix} w \\ \bar{\mathcal{L}} \end{pmatrix}^{(j)}(t)\right)^T f(t)dt \\ &+\sum_{k=1}^{L+1}\left(\sum_{j=k}^{L+1}(-1)^k\left((\,R_j \quad -M_{j-1}\,)\begin{pmatrix} w \\ \bar{\mathcal{L}} \end{pmatrix}^{(j-k)}(0)\right)^T\right)f^{(k-1)}(0).\end{aligned} \tag{B.50}$$

The integral is zero, since $\frac{d}{dt}\bar{\mathcal{L}} = \ell$ and $(w, \ell) \in \mathcal{B}_f$ by assumption. The double sum is zero, since by assumption each addendum of the outermost sum is zero (cf. (B.46)). The claim follows. □

B.9. Proof of Theorem 7.3. Note first that Lemma B.2 holds also for the kernel representation $(\,R \quad -M\,)\left(\frac{d}{dt}\right)\binom{w}{\ell} = 0$ of the full behavior. Let us now prove necessity. If $X(\frac{d}{dt})$ is a (w, ℓ)-induced state map for the external behavior, then x is properly eliminable, and $(X(\frac{d}{dt})\binom{w}{\ell})(0) = 0$ implies external concatenability with zero. Proposition 7.2 states that external concatenability with zero is equivalent to

$$\left(V\begin{pmatrix} (\,R & -M\,)_{\Xi} \\ & 0 \end{pmatrix}\begin{pmatrix} d/dt \end{pmatrix}\begin{pmatrix} w \\ \ell \end{pmatrix}\right)(0) = 0.$$

Now apply Lemma B.2 with $X_1 = X$ and

$$X_2 = V\begin{pmatrix} (\,R & -M\,)_{\Xi} \\ & 0 \end{pmatrix}.$$

Sufficiency is proven as follows.

$$\left(V\left(\begin{pmatrix} R & -M \\ & 0 \end{pmatrix}_{\Xi}\right)\right)\left(\frac{d}{dt}\right)\begin{pmatrix} w \\ \ell \end{pmatrix}$$

is an absolutely continuous function (cf. the remark made at the beginning of section 6). Since

$$\left(V\left(\begin{pmatrix} R & -M \\ & 0 \end{pmatrix}_{\Xi}\right)\right)\left(\frac{d}{dt}\right) = AX\left(\frac{d}{dt}\right) + B\begin{pmatrix} R & -M \end{pmatrix}\left(\frac{d}{dt}\right), \tag{B.51}$$

for each $(w,\ell) \in \mathcal{B}_f$ such that $(X(\frac{d}{dt})\binom{w}{\ell})(0) = 0$ and such that $X(\frac{d}{dt})\binom{w}{\ell}$ is continuous at $t = 0$,

$$\left(V\left(\begin{pmatrix} R & -M \\ & 0 \end{pmatrix}_{\Xi}\right)\right)\left(\frac{d}{dt}\right)\begin{pmatrix} w \\ \ell \end{pmatrix}(0) = 0$$

holds, since

$$B\begin{pmatrix} R & -M \end{pmatrix}\left(\frac{d}{dt}\right)\begin{pmatrix} w \\ \ell \end{pmatrix} = V\left(\begin{pmatrix} R & -M \\ & 0 \end{pmatrix}_{\Xi}\right)\left(\frac{d}{dt}\right)\begin{pmatrix} w \\ \ell \end{pmatrix} - AX\left(\frac{d}{dt}\right)\begin{pmatrix} w \\ \ell \end{pmatrix}$$

is continuous at $t = 0$. Then Proposition 7.2 can be applied, and external concatenability with zero follows. Moreover, since x is properly eliminable, the external behavior of the state-space representation is the same as that of the original hybrid representation. This concludes the proof. □

B.10. Proof of Proposition 7.5. Consider the system described by (2.6). Following Theorem 2.1, computation of a kernel description of (the closure of) its external behavior is done by determining a unimodular matrix U such that $UM = \binom{0}{M_2'}$, with M_2' of full row rank. Partitioning U, R, and M according to the number of rows of M_2' as

$$U := \begin{pmatrix} U_{11} & U_{12} \\ U_{21} & U_{22} \end{pmatrix}, \quad M := \begin{pmatrix} M_1 \\ M_2 \end{pmatrix}, \quad R := \begin{pmatrix} R_1 \\ R_2 \end{pmatrix}, \tag{B.52}$$

the description of the external behavior is given as $R_1'(\frac{d}{dt})w = 0$, with $R_1' = U_{11}R_1 + U_{12}R_2$.

Now assume $M = \bar{M}F$, with F a full row rank right factor of M. Note that

$$0 = U_{11}M_1 + U_{12}M_2 = U_{11}\bar{M}_1F + U_{12}\bar{M}_2F = (U_{11}\bar{M}_1 + U_{12}\bar{M}_2)F \tag{B.53}$$

if and only if $U_{11}\bar{M}_1 + U_{12}\bar{M}_2 = 0$ by the fact that F has full row rank. Therefore U eliminates the latent variable in the description $R(\frac{d}{dt})w = \bar{M}(\frac{d}{dt})\ell$, and $R_1'(\frac{d}{dt})w = 0$ describes the closure of the external behavior of this system as well. □

B.11. Proof of Proposition 7.6. G is a nonsingular $d \times d$ matrix and therefore $\forall\ \bar{\ell} \in \mathcal{L}_1^{loc}(\mathbb{R};\mathbb{R}^d)\ \exists\ \ell \in \mathcal{L}_1^{loc}(\mathbb{R};\mathbb{R}^d)$ such that $\bar{\ell} = G(\frac{d}{dt})\ell$. Therefore

$$\begin{aligned}\left(X_{obs}\left(\frac{d}{dt}\right)\begin{pmatrix} w \\ \bar{\ell} \end{pmatrix}\right)(0) = 0 &\iff \left(X_{obs}\left(\frac{d}{dt}\right)\begin{pmatrix} w \\ G\left(\frac{d}{dt}\right)\ell \end{pmatrix}\right)(0) = 0 \\ &\iff \left(\begin{pmatrix} X_{obs,w} & X_{obs,l}G \end{pmatrix}\left(\frac{d}{dt}\right)\begin{pmatrix} w \\ \ell \end{pmatrix}\right)(0) = 0.\end{aligned}$$

Since X_{obs} induces a state map, $(X_{obs}(\frac{d}{dt})\binom{w}{\bar{\ell}})(0) = 0$ implies that $(w, \bar{\ell})$ is externally concatenable with zero; moreover, external concatenability in zero for $(w, \bar{\ell})$ is equivalent to external concatenability in zero for (w, ℓ), as $\bar{\ell} = G(\frac{d}{dt})\ell$. Let us now prove that $x = X(\frac{d}{dt})\binom{w}{\ell}$ does not impose additional smoothness constraints on the trajectories of the external behavior of $(\,R \quad -M\,)(\frac{d}{dt})\binom{w}{\ell} = 0$. By unimodular transformations, which preserve the proper eliminability of a latent variable, we can bring the equations

$$\begin{aligned} R\left(\frac{d}{dt}\right) w &= M\left(\frac{d}{dt}\right) \ell, \\ X_w\left(\frac{d}{dt}\right) &= -(X_\ell G)\left(\frac{d}{dt}\right) \ell + x \end{aligned} \tag{B.54}$$

to the form

$$\begin{aligned} R_1'\left(\frac{d}{dt}\right) w &= 0, \\ R_2'\left(\frac{d}{dt}\right) w &= G\left(\frac{d}{dt}\right) \ell, \\ X_w\left(\frac{d}{dt}\right) &= -(X_\ell G)\left(\frac{d}{dt}\right) \ell + x \end{aligned} \tag{B.55}$$

and, again by unimodular operations, to

$$\begin{aligned} R_1'\left(\frac{d}{dt}\right) w &= 0, \\ R_2'\left(\frac{d}{dt}\right) w &= G\left(\frac{d}{dt}\right) \ell, \\ (X_w + X_\ell R_2')\left(\frac{d}{dt}\right) w &= x. \end{aligned} \tag{B.56}$$

Observe that $x = (X_w + X_\ell R_2')(\frac{d}{dt})w$ is a state variable for the behavior Ker $R_1'(\frac{d}{dt})$ and therefore that it is properly eliminable from

$$\begin{aligned} R_1'\left(\frac{d}{dt}\right) w &= 0, \\ (X_w + X_\ell R_2')\left(\frac{d}{dt}\right) w &= x; \end{aligned} \tag{B.57}$$

the claim follows. □

B.12. Proof of Proposition 8.2. Let N_i be the ith row of N. Since $N_i D^{-1}$ is proper, there exists a rational vector $n_i := \sum_{k=0}^{\infty} n_{ik}\xi^{-k}$ such that $N_i D^{-1} = n_i$. Write $D = D_0 + D_1\xi + \cdots + D_L\xi^L$ and $N_i = N_{i0} + N_{i1}\xi + \cdots + N_{iL'}$, $L' \le L$. $N_i = n_i D$ yields the following equalities:

$$\begin{aligned} N_{i0} &= n_{i0}D_0 + n_{i1}D_1 + \cdots + n_{iL}D_L, \\ N_{i1} &= n_{i0}D_1 + n_{i1}D_2 + \cdots + n_{iL-1}D_L, \\ &\vdots \\ N_{iL'} &= n_{i0}D_{L'} + n_{i1}D_{L'+1} + \cdots + n_{iL-L'}D_L. \end{aligned} \tag{B.58}$$

These equalities imply $N_i = n_{i0}D + n_{i1}\sigma_+(D) + \cdots + n_{iL}\sigma_+^L(D)$ and therefore $\sigma_+(N_i) = n_{i0}\sigma_+(D) + n_{i1}\sigma_+^2(D) + \cdots + n_{iL-1}\sigma_+^L(D)$. Then $\sigma_+(N_i) \in \Xi_D$, and the same holds for $\sigma_+^2(N_i)$, $\sigma_+^3(N_i), \ldots$. This yields the claim. ∎

B.13. Proof of Proposition 8.3. The inclusion $\Xi_D \subseteq \{r \mid rD^{-1}$ is strictly proper$\}$ can be proven as follows. Let D_i be the ith row of D, $D_i = \sum_{k=0}^{L} D_{ik}\xi^L$. Then $\sigma_+ D_i = \xi^{-1}D_i - \xi^{-1}D_{i0}$, and therefore $\sigma_+ D_i D^{-1} = \xi^{-1}e_i - \xi^{-1}D_{i0}D^{-1} \in \mathbb{R}_+^{1\times d}(\xi)$. Analogously, $\sigma_+^2 D_i = \xi^{-2}D_i - \xi^{-2}D_{i0} - \xi^{-1}D_{i1}$ and therefore $\sigma_+^2 D_i D^{-1} = \xi^{-2}e_i - \xi^{-2}D_{i0}D^{-1} - \xi^{-1}D_{i1}D^{-1} \in \mathbb{R}_+^{1\times d}(\xi)$ and similarly for all iterations of σ_+ and for all rows of D.

The opposite inclusion may be proven as follows. Take $r \in \{r' \mid r'D^{-1}$ is strictly proper$\}$. Then there exists $n \in \mathbb{R}_+^{1\times d}(\xi)$ such that $r = nD$. Write

$$n = \sum_{k=1}^{\infty} n_k \xi^{-k}, \tag{B.59}$$

$n_k \in \mathbb{R}^{1\times d}$, and denote

$$D := D_0 + D_1\xi + \cdots + D_L\xi^L \tag{B.60}$$

and

$$r := r_0 + r_1\xi + \cdots + r_{L'}\xi^{L'}, \tag{B.61}$$

where without loss of generality we can assume $L' \leq L - 1$. $r = nD$ yields, equating powers of ξ,

$$\begin{aligned} r_0 &= n_1 D_1 + n_2 D_2 + \cdots + n_L D_L, \\ r_1 &= n_1 D_2 + n_2 D_3 + \cdots + n_{L-1} D_L, \\ &\vdots \\ r_j &= n_1 D_{j+1} + n_2 D_{j+2} + \cdots + n_{L-j} D_L, \\ &\vdots \end{aligned} \tag{B.62}$$

and this implies

$$r = n_1\sigma_+(D) + n_2\sigma_+^2(D) + \cdots + n_L\sigma_+^L(D) \tag{B.63}$$

and therefore that $r \in \Xi_D$, as we were to prove. ∎

B.14. Proof of Proposition 8.4. The first equality follows from the fact that $\Xi_M = \Xi_N + \Xi_D$ and from Proposition 8.2. The second equality can be proven by applying Proposition 8.3. ∎

Acknowledgment. The first author thanks the Department of Mathematics of the University of Groningen and the Dutch Systems and Control Theory Network for the hospitality and financial support granted for visits to Groningen during which the present results were worked out.

REFERENCES

[1] R. D'Andrea and F. Paganini, *Interconnection of uncertain behavioral systems for robust control*, in Proc. 32nd IEEE Conf. on Decision and Control, San Antonio, TX, 1993, pp. 3642–3647.

[2] P. A. Fuhrmann, *Linear Systems and Operators in Hilbert Space*, McGraw-Hill, New York, 1981.
[3] T. Kailath, *Linear Systems*, Prentice-Hall, Englewood Cliffs, NJ, 1980.
[4] R. Kalman, P. Falb, and M. Arbib, *Topics in Mathematical System Theory*, McGraw-Hill, New York, 1969.
[5] W. Neven, *Polynomial methods in system theory*, Master's thesis, Dept. of Mathematics and Computing Science, Eindhoven University of Technology, Eindhoven, the Netherlands, 1988.
[6] J. W. Polderman, *Proper elimination of latent variables*, in Proc. 12th World Congress IFAC, 1992, pp. 73–76.
[7] C. Praagman, *Inputs, outputs and states in the representation of time series*, in Analysis and Optimization of Systems, A. Bensoussan and J. Lions, eds., Springer-Verlag, Berlin, 1988, pp. 1069–1078.
[8] J. C. Willems, *From time series to linear system. Part* 1*—finite dimensional linear time invariant systems*, Automatica J. IFAC, 22 (1986), pp. 561–580.
[9] J. C. Willems, *Models for dynamics*, Dynamics Reported, 2 (1989), pp. 171–269.
[10] J. C. Willems, *Paradigms and puzzles in the theory of dynamical systems*, IEEE Trans. Automat. Control, 36 (1991), pp. 259–294.

SIAM J. CONTROL OPTIM.
Vol. 35, No. 4, pp. 1093–1115, July 1997

MIN-MAX CHARACTERIZATION OF A SMALL NOISE LIMIT ON RISK-SENSITIVE CONTROL*

A. BENSOUSSAN† AND H. NAGAI‡

Abstract. Stochastic control problems on a finite horizon with exponential cost criteria are considered. By taking a kind of singular limit a Hamilton–Jacobi–Isaacs equation is obtained. Its solution is characterized as the lower value function of a deterministic differential game related to robust control of nonlinear systems.

Key words. risk-sensitive control, small noise limit, H_∞-control, Hamilton–Jacobi–Isaacs equation, differential game

AMS subject classifications. 93E20, 49L20, 35K, 90D

PII. S0363012995279420

Introduction. The large deviation treatment for risk-sensitive control problems has been considered by several authors (cf. [2], [9], [10], [11], [13], [14], [15]) in relation to H_∞-control theory since Whittle's works [22] and [23]. It has been noticed that the large deviation theory of Freidlin–Wentzell in stochastic control problems with the criterion function of the exponential of an additive cost function may link with deterministic control problems called H_∞- or robust-control. In the case of finite-time horizon with full observation, Fleming and McEneaney [9] and, independently, James [13] have given rigorous treatment to the problem. However, those cases are limited by growth conditions which do not cover the linear exponential quadratic Gaussian (LEQG) model. The case has been treated directly. The infinite-horizon case has been considered by Fleming and McEneaney [10] and Fleming and James [11]. By taking a small noise limit, they obtained a deterministic differential game related to H_∞-control theory.

In the present paper we deal with the case of finite-time horizon and relax the growth conditions assumed in [9], [13], [15] so that the LEQG model can be included. We shall comment on this point further in discussion of the assumption (1.14) at the end of subsection 1.2.

Our research is based on [18], where risk-sensitive control problems for nonlinear systems, including the LEQG case, have been treated in a rather general setting, and no breakdown problems and large time asymptotics of the value function have been considered.

More precisely, we consider the controlled stochastic differential equation (SDE) with a small parameter $\epsilon > 0$:

$$\begin{cases} dX_s = \sqrt{\epsilon}\sigma(X_s)dB_s + b(X_s)ds + c(X_s, z_s)ds, \\ X_0 = x \end{cases} \tag{0.1}$$

*Received by the editors January 3, 1995; accepted for publication (in revised form) April 18, 1996.

http://www.siam.org/journals/sicon/35-4/27942.html

†Centre Nationale d'Etudes Spatiales, 2, place Maurice Quentin, 75039 Paris, France (Alain.Bensoussan@cnes.fr).

‡Department of Mathematical Science, Faculty of Engineering Science, Osaka University, Toyonaka, 560, Japan (nagai@sigmath.es.osaka-u.ac.jp).

and take up the following value function of a risk-sensitive control problem:

$$J_\epsilon^*(t,x;T,\theta) = \inf_{z.} E_x\left[\exp\left(\frac{\theta}{\epsilon}\Phi_{T-t}\right)\right], \quad 0 \le t \le T, \tag{0.2}$$

where

$$\Phi_s = \int_0^s \{V(X_\tau) + \phi(X_\tau, z_\tau)\}d\tau. \tag{0.3}$$

The Bellman equation of J^* can be formally written as

$$\begin{cases} \mathcal{L}^\epsilon J_\epsilon + \inf\limits_{z\in Z}\left\{c^i(x,z)D_iJ_\epsilon + \dfrac{\theta}{\epsilon}(V(x)+\phi(x,z))\right\} = 0, \ [0,T)\times R^N, \\ J_\epsilon(T,x) = 1, \end{cases} \tag{0.4}$$

where

$$\mathcal{L}^\epsilon J_\epsilon = \frac{\partial J_\epsilon}{\partial t} + \frac{\epsilon}{2}a^{ij}D_{ij}J_\epsilon + b^iD_iJ_\epsilon.$$

Taking a transformation $J_\epsilon = \exp(\frac{\theta}{\epsilon}w_\epsilon)$, we obtain

$$\begin{cases} \mathcal{L}^\epsilon w_\epsilon + Q_0(x,\nabla w_\epsilon) + V(x) = 0, \\ w_\epsilon(T,x) = 0, \end{cases} \tag{0.5}$$

where

$$\begin{aligned} Q_0(x,p) &\equiv Q_0(x,p;\theta) \\ &= \frac{\theta}{2}a^{ij}p_ip_j + \inf_{z\in Z}\{c^i(x,z)p_i + \phi(x,z)\}. \end{aligned}$$

The existence of a nonnegative solution of (0.5) has been shown under the main assumption (1.14), which indicates the condition on a risk-sensitive parameter for no breakdown. In the present paper we first consider the asymptotic behavior of the solution w_ϵ of (0.5) as $\epsilon \to 0$. To this end, as was done in [9], [13], [15], we employ viscosity methods introduced by Crandall and Lions and developed recently to a large extent (cf. [3], [4], [5], [7], [8], and references therein). Owing to a stability theorem on viscosity solutions, a limit nonlinear equation is obtained by obtaining estimates independent of ϵ on w_ϵ and its first derivatives (cf. section 2). It is to be noted that we cannot expect that w_ϵ is globally Lipshitz, which comes from the fact that we cover the case of polynomial growth $V(x)$, and we utilize some techniques used in [18] to get the estimates.

There are some difficulties in characterizing the limit value, which comes from the fact that the control region is noncompact and the cost functions are unbounded. Those assumptions are necessary to include the LEQG case. We shall take up a more specialized case (cf. section 2.1) than that of section 1 and show that the limit value is characterized as the lower value function of a differential game whose Hamilton–Jacobi–Bellman equation is written as

$$\begin{cases} \dfrac{\partial w}{\partial t} + b^iD_iw + Q_0(x,\nabla w) + V = 0, \\ w(T,x) = 0. \end{cases} \tag{0.6}$$

Following Varaiya [21], Roxin [19], Elliott and Kalton [6], we have formulated a deterministic differential game and shown that the lower value function of the game formulated in this way is the unique nonnegative viscosity solution of (0.6) along the line of Evans and Souganidis [8]. However, because of the noncompactness of the control region and the unboundedness of the cost functions, the finiteness of the lower value function is nontrivial. We need assumption (2.5) (a counterpart of (1.14) for the specialized case) to show it. Moreover, to prove that the lower value function satisfies (0.6) in a viscosity sense, we shall obtain some hitting time estimates on controlled dynamics and develop some technical tricks to show its continuity. We refer to [16], [17], [20], and the references therein, where viscosity solutions of Hamilton–Jacobi–Isaacs equations are directly considered.

1. A small noise limit on risk-sensitive control.

1.1. Setting up. We shall formulate a risk-sensitive control problem depending on a parameter $\epsilon > 0$ and then consider a problem taking a limit as $\epsilon \to 0$. Let $(\Omega, \mathcal{F}, P)$ be a probability space with filtration $\mathcal{F}_t$, $t \geq 0$, B_t a standard N-dimensional $\mathcal{F}_t$ Brownian motion process, and z_t a progressively measurable process taking its value on a Borel subset Z of R^{N_1}. We consider the following SDE with a parameter $\epsilon > 0$:

$$\begin{cases} dX_s^i = \sqrt{\epsilon}\sigma_j^i(X_s)dB_s^j + b^i(X_s)ds + c^i(X_s, z_s)ds, \quad i = 1, ..., N, \\ X_0 = x. \end{cases} \tag{1.1}$$

We then introduce the value function J_ϵ^* with a risk-sensitive parameter $\theta > 0$ as follows:

$$J_\epsilon^*(t, x; T, \theta) = \inf_{\mathcal{A}_{T-t}^0} E_x \left[\exp\left(\frac{\theta}{\epsilon} \Phi_{T-t} \right) \right], \quad 0 \leq t \leq T, \tag{1.2}$$

where

$$\Phi_s = \int_0^s \{V(X_\tau) + \phi(X_\tau, z_\tau)\} d\tau \tag{1.3}$$

and $\mathcal{A}_{T-t}^0$ is the totality of $(\Omega, \mathcal{F}, \mathcal{F}_s, P, B_s, z_s)_{0 \leq s < T-t}$ such that (1.1) has a unique solution for $0 \leq s < T - t$. Note that in this paper we employ the summation convention that if the same indices appear twice in a term, then the symbol of summation is omitted. We assume the following conditions:

(1.4) $\sigma, b, c, V,$ and ϕ are smooth;

(1.5) σ and b satisfy a global Lipshitz condition;

(1.6) $\sigma, b,$ and V and all their derivatives are dominated by $M(1 + |x|^2)^m$ for some $m > 0, M > 0$;

(1.7) $|c(x, z)| \leq c_0(z)$ for some locally bounded function $c_0(z)$;

(1.8) $V(x) \geq 0$ and $\lim_{|x| \to \infty} V(x) = \infty$;

(1.9) $\phi(x, z) \geq 0, \quad \lim_{|z| \to \infty} \phi(x, z) = \infty, \quad \lim_{|z| \to \infty} \dfrac{|c(x, z)|}{\phi(x, z)} = 0,$ uniformly in x;

(1.10) $a^{ij}\xi_i\xi_j \geq \nu|\xi|^2, \quad \xi \in R^N, \quad \exists \nu > 0,$

where $a^{ij} = (\sigma\sigma^*)^{ij}$ and σ^* stands for a transposed matrix of the matrix σ. The above-defined value function J^* is considered to satisfy formally the following Bellman equation:

$$\begin{cases} \mathcal{L}^\epsilon J_\epsilon + \inf_{z\in Z}\left\{c^i(x,z)D_iJ_\epsilon + \dfrac{\theta}{\epsilon}(V(x)+\phi(x,z))\right\} = 0,\ [0,T)\times R^N, \\ J_\epsilon(T,x) = 1, \end{cases} \tag{1.11}$$

where

$$\mathcal{L}^\epsilon J_\epsilon = \frac{\partial J_\epsilon}{\partial t} + \frac{\epsilon}{2}a^{ij}D_{ij}J_\epsilon + b^iD_iJ_\epsilon$$

and $D_i = \frac{\partial}{\partial x^i}$, $D_{ij} = \frac{\partial^2}{\partial x^i \partial x^j}$. Then, by taking a transformation $J_\epsilon = e^{\frac{\theta}{\epsilon}w_\epsilon}$, we obtain the equation

$$\begin{cases} \mathcal{L}^\epsilon w_\epsilon + Q_0(x,\nabla w_\epsilon) + V(x) = 0, \\ w_\epsilon(T,x) = 0, \end{cases} \tag{1.12}$$

where

$$\begin{aligned} Q_0(x,p) &\equiv Q_0(x,p;\theta) \\ &= \frac{\theta}{2}a^{ij}p_ip_j + \inf_{z\in Z}\{c^i(x,z)p_i + \phi(x,z)\}. \end{aligned} \tag{1.13}$$

We furthermore assume that

$$-\frac{k_1}{2}a^{ij}p_ip_j \le Q_0(x,p) \le -\frac{k_2}{2}a^{ij}p_ip_j, \quad p\in R^N, \tag{1.14}$$

for some k_1, $k_2 > 0$ and

$$\left|\frac{\partial Q_0(x,p)}{\partial p}\right| \le M_1|p| + M_2, \quad \left|\frac{\partial Q_0(x,p)}{\partial x}\right| \le M_1|p|^2 + M_2 \tag{1.15}$$

for some locally bounded functions M_1 and M_2. For examples satisfying conditions (1.14) and (1.15) we refer to [18, section 1.5]. Note that according to [18] we have the following theorem.

THEOREM 1.0 (see [18]). *Under the assumptions* (1.4)–(1.10), (1.14), *and* (1.15) *the equation* (1.12) *has a nonnegative solution* $w_\epsilon \in C^{1+\frac{\alpha}{2},2+\alpha}([0,T)\times R^N)\cap C([0,T]\times R^N)$. *Moreover, under those conditions the value function* J_ϵ^* *defined by* (1.2) *has a finite value.*

We shall consider the limit equation of (1.12) as ϵ tends to zero. It can be formally written as

$$\begin{cases} \dfrac{\partial w}{\partial t} + b^iD_iw + Q_0(x,\nabla w) + V = 0, \\ w(T,x) = 0. \end{cases} \tag{1.16}$$

To deduce the equation as the limit of (1.12) we employ viscosity methods (cf. [4], [7], [12]) and then obtain the following theorem.

THEOREM 1.1. *Let* w_ϵ *be a nonnegative solution of* (1.12) *obtained by Theorem* 1.0. *Then there exists a subsequence of* $\{w_\epsilon\}$ *converging uniformly on each compact set to a continuous function* $w(t,x)$, *which is a viscosity solution of equation* (1.16).

Discussion on assumption (1.14). We address here assumption (1.14) and compare it with the literature. Suppose that we are in the LEQG case, where (1.4) to (1.10) and (1.15) are satisfied. Assumption (1.14) remains a restriction which is to be interpreted as a smallness condition on θ. Such a restriction is unavoidable. Indeed, we consider the case where $N = 1$, $Z = R^1$, $\sigma \equiv 1$, $b(x) \equiv 0$, $c(x,z) = z$, $V(x) = \frac{1}{2}x^2$, and $\phi(x,z) = \frac{1}{2}z^2$, namely the case

$$X_s = x + \sqrt{\epsilon}B_s + z_s,$$
$$\Phi_s = \int_0^s \frac{1}{2}(X_s^2 + z_s^2)ds.$$

In this case $Q_0(x,p)$ turns out to be $Q_0(x,p) = -\frac{1-\theta}{2}p^2$, and assumption (1.14) is reduced to the form

$$\theta < 1. \tag{1.14$'$}$$

Moreover, the Bellman equation (1.12) reads

$$\begin{cases} \dfrac{\partial w_\epsilon}{\partial t} + \dfrac{\epsilon}{2}\dfrac{\partial^2 w_\epsilon}{\partial x^2} - \dfrac{1-\theta}{2}\left|\dfrac{\partial w_\epsilon}{\partial x}\right|^2 + \dfrac{1}{2}x^2 = 0, \\ w_\epsilon(T,x) = 0. \end{cases} \tag{1.12$'$}$$

Note that the solution to (1.12$'$) has the following explicit form:

$$w_\epsilon(t,x) = \frac{1}{2}P(t)x^2 + \frac{\epsilon}{2}\int_t^T P(s)ds,$$

provided that the Riccati equation

$$\frac{dP}{dt} - (1-\theta)P^2 + 1 = 0, \quad P(T) = 0, \tag{P}$$

has a solution on $[0,T]$. However, if $\theta > 1$, there may appear a conjugate point (a finite escape time) of T in the equation. In fact, the solution $P(t)$ to (P) is nothing but

$$P(t) = \frac{1}{\sqrt{\theta-1}}\tan(\sqrt{\theta-1}(T-t)),$$

and t_0 satisfying $\sqrt{\theta-1}(T-t_0) = \frac{\pi}{2}$ is the finite escape time. Thus we see that (P) has no solution on $[0,T]$ for T such that $T > \frac{\pi}{2\sqrt{\theta-1}}$ if $\theta > 1$ and the condition (1.14$'$) is to be expected. Such an assumption is unnecessary when V and ϕ are bounded or have linear growth (these are the assumptions in [9], [13], [15]) since the value function (1.2) never diverges in that case and equation (1.12) has a solution regardless of the size of θ. We furthermore consider the case where the controlled process is defined by

$$dX_s = \sqrt{\epsilon}dB_s - \alpha X_s + z_s, \quad X_0 = x,$$

with $\alpha > 0$. Here it remains the case that $N = 1$ and $Z = R^1$. Besides, if $V(x)$ and $\phi(x,z)$ are same as above, then the size of the risk-sensitive parameter ensuring the existence of the solution of the Bellman equation becomes larger than the above; indeed, we can see that it is $\theta < 1 + \alpha^2$. However, our theorem deals with the case of polynomial growth $V(x)$. So, if the growth of $V(x)$ is faster than the quadratic growth, then the effect of such $b(x) = -\alpha x$ disappears and we still need assumption (1.14).

1.2. Proof of Theorem 1.1. We first give an L^∞_{loc} estimate for a solution w_ϵ of (1.12) by using the Feynman–Kac formula. Namely, we shall prove the following.

LEMMA 1.2. *Let w_ϵ be a nonnegative solution of* (1.12) *given by Theorem* 1.0*; then $0 \le w_\epsilon(t,x) \le c_{R,T}$ on $[0,T] \times B_R$ for each $R > 0$, where $c_{R,T}$ is a constant independent of ϵ and $B_R = \{x; |x| < R\}$.*

Proof. Let $\psi_\epsilon(t,x)$ be a solution of the following linear equation:

$$\begin{cases} \mathcal{L}^\epsilon \psi_\epsilon - \dfrac{k_2}{\epsilon} V \psi_\epsilon = 0, \\ \psi_\epsilon(T,x) = 1. \end{cases} \tag{1.17}$$

It is easy to see that $\psi_\epsilon(t,x) \ge 0$, and therefore we may set

$$\tilde{w}_\epsilon(t,x) = -\frac{\epsilon}{k_2} \log \psi_\epsilon(t,x).$$

Then $\tilde{w}_\epsilon(t,x)$ satisfies the following nonlinear equation:

$$\begin{cases} \mathcal{L}^\epsilon \tilde{w}_\epsilon - \dfrac{k_2}{2} a^{ij} D_i \tilde{w}_\epsilon D_j \tilde{w}_\epsilon + V(x) = 0, \\ \tilde{w}_\epsilon(T,x) = 0. \end{cases} \tag{1.18}$$

Because of (1.14), we see that $\tilde{w}_\epsilon(t,x)$ is a supersolution of (1.12) and dominates its solution $w_\epsilon(t,x)$ (cf. the proof of Theorem 1.1 in [18]). Hence we obtain

$$\begin{aligned} w_\epsilon(t,x) &\le \tilde{w}_\epsilon(t,x) \\ &= -\frac{\epsilon}{k_2} \log E_x \left[\exp\left(-\frac{k_2}{\epsilon} \int_0^{T-t} V(Y_s) ds \right) \right] \end{aligned} \tag{1.19}$$

by the Feynman–Kac formula, where Y_s is a solution of the SDE

$$\begin{cases} dY_s = \sqrt{\epsilon}\sigma(Y_s) dB_s + b(Y_s) ds, \\ Y_0 = x. \end{cases} \tag{1.20}$$

Jensen's inequality and (1.19) imply that

$$w_\epsilon(t,x) \le E_x \left[\int_0^{T-t} V(Y_s) ds \right]. \tag{1.21}$$

Standard moment estimates for the solution of the SDE (1.20) give an L^∞_{loc} estimate of the right-hand side of (1.21) because of the assumptions (1.5) and (1.6). Thus we obtain the estimate $0 \le w_\epsilon(t,x) \le c_{R,T}$ on $[0,T] \times B_R$, where $c_{R,T}$ is independent of ϵ such that $0 \le \epsilon \le 1$. □

The following lemma, obtained by using the gradient estimate for a solution w_ϵ of (1.12), plays a key role in the proof of Theorem 1.1.

LEMMA 1.3. *Let w_ϵ be a solution of* (1.12) *given by Theorem* 1.0 *for each $\epsilon > 0$; then $\{w_\epsilon(t,x)\}_{1 \ge \epsilon > 0}$ are equicontinuous on $[0,T] \times B_R$ for each $R > 0$.*

Proof. For the proof of this lemma we note that the following estimates are valid for $u_\epsilon(t,x) = w_\epsilon(T-t,x)$:

$$\frac{\partial u_\epsilon}{\partial t} \ge 0, \tag{1.22}$$

$$t\left(|\nabla u_\epsilon|^2 + \gamma \frac{\partial u_\epsilon}{\partial t} \right) \le t K_{R,\gamma} + L_{R,\gamma} \quad [0,T] \times B_R, \quad \gamma > \frac{2}{k_2 \nu}, \tag{1.23}$$

where $K_{R,\gamma}$ and $L_{R,\gamma}$ are constants independent of ϵ.

Estimate (1.22) is obtained in [18]. As for (1.23), we can follow the proof of Lemma 1.5 in [18]. In fact, we set

$$\Gamma(F) = \frac{\epsilon}{2}a^{ij}D_{ij}F + \frac{\partial Q}{\partial p_i}(x, \nabla u)D_iF - \frac{\partial F}{\partial s} + \frac{F}{s}, \tag{1.24}$$

where $F = t(|\nabla u_\epsilon|^2 + \gamma\frac{\partial u_\epsilon}{\partial t})$ and $Q(x,p) = b^i p_i + Q_0(x,p)$, perform the same procedure as the proof of Lemma 1.5 in [12], and obtain the estimate (1.23). Now we shall complete the proof of Lemma 1.3. We have $0 \le u_\epsilon(t,x) \le E_x[\int_0^t V(Y_s)ds]$ by (1.21), and, consequently, standard moment estimates for the solution of the SDE imply that for each $\eta > 0$ there exists δ_R such that $E_x[\int_0^t V(Y_s)ds] < \eta$ for $t < 2\delta_R$, $x \in B_R$. Take (t_1, x_1) and (t_2, x_2) such that

$$|x_1 - x_2| + |t_1 - t_2| < \eta, \quad |t_1 - t_2| < \delta_R, \quad (t_i, x_i) \in [0,T] \times B_R, \quad i = 1, 2.$$

Then, if $t_1 < \delta_R$ or $t_2 < \delta_R$, we have

$$|u_\epsilon(t_1, x_1) - u_\epsilon(t_2, x_2)| \le 2E_x\left[\int_0^{2\delta_R} V(Y_s)ds\right] \le 2\eta.$$

If $t_1, t_2 > \delta_R$, then from the estimates (1.22) and (1.23) it follows that $|u_\epsilon(t_1, x_1) - u_\epsilon(t_2, x_2)| \le c_R^{'}\eta$ for some constant $c_R^{'}$. □

Proof of Theorem 1.1. Lemmas 1.1 and 1.2 and the Ascoli–Arzelà theorem imply that there exists a subsequence $\{w_{\epsilon_k}\}$ of $\{w_\epsilon\}$ converging uniformly on each compact set to a continuous function $w(t,x)$. Moreover, the stability theorem on viscosity solutions (cf. [4]) asserts that $w(t,x)$ is a viscosity solution of (1.16), since the classical solution $w_\epsilon(t,x)$ of (1.12) is obviously that of viscosity sense. □

2. Interpretation of the limit.

2.1. Differential games. We are going to characterize the solution $w(t,x)$ of the limit equation (1.16) as the lower value function of a differential game. To this end we confine ourselves to the case where $Z = R^{N_1}$:

$$c^i(x,z) = B_k^i(x)z^k, \quad \phi(x,z) = \frac{1}{2}S_{ij}z^i z^j, \tag{2.1}$$

where (S_{ij}) is a symmetric smooth matrix such that

$$S_{ij}(x)\xi^i\xi^j \ge \mu|\xi|^2 \quad \forall \xi \in R^N, \quad \mu > 0, \tag{2.2}$$

and $(B_k^i(x))$ is a bounded and smooth $N \times N_1$ matrix. We assume also (1.4)–(1.6), (1.8), and (1.10). In this case $Q_0(x,p)$ has the following form:

$$\begin{aligned} Q_0(x,p) &= \frac{\theta}{2}a^{ij}p_ip_j + \inf_{z\in R^N}\left\{B_k^i z^k p_i + \frac{1}{2}S_{ij}z^i z^j\right\} \\ &= \frac{\theta}{2}a^{ij}p_ip_j - \frac{1}{2}(BS^{-1}B^*)^{ij}p_ip_j, \end{aligned} \tag{2.3}$$

where B^* stands for the transposed matrix of B. Then equation (1.16) reads

$$\begin{cases} \dfrac{\partial w}{\partial t} + b^i D_i w - \dfrac{1}{2}(BS^{-1}B^* - \theta a)^{ij}D_iwD_jw + V = 0, \\ w(T,x) = 0. \end{cases} \tag{2.4}$$

Moreover, we assume that

$$k_2 a^{ij} p_i p_j \leq (BS^{-1}B^* - \theta a)^{ij} p_i p_j \quad \forall p \in R^N. \tag{2.5}$$

Then conditions (1.14) and (1.15) are satisfied under these assumptions. Thus we see that Theorem 1.1 applies to the present case. Following Varaiya [21], Roxin [19], and Elliot and Kalton [6], we shall formulate a differential game. Let $\mathcal{A}_t$ be the set of all measurable functions $z_1(\cdot) : [0,t] \rightsquigarrow R^N$ such that

$$\int_0^t |z_1(s)|^2 \, ds < \infty, \tag{2.6}$$

and let $\mathcal{B}_t$ be the one of all measurable functions $z_2(\cdot) : [0,t] \rightsquigarrow R^{N_1}$ satisfying (2.6) replaced by $z_2(s)$. Let U be a map $U : \mathcal{A}_t \rightsquigarrow \mathcal{B}_t$ such that whenever for each $0 \leq \sigma \leq t$ and $z_1, \tilde{z}_1 \in \mathcal{A}_t$, $z_1(s) = \tilde{z}_1(s)$ almost everywhere (a.e.) on $0 \leq s \leq \sigma$, then $(Uz_1)(s) = (U\tilde{z}_1)(s)$ a.e. on $0 \leq s \leq \sigma$. The totality of such maps is denoted by Γ_t. We consider the criterion

$$I(t,x;T;z_1,Uz_1) = \int_0^{T-t} \Psi(X(s), z_1(s), (Uz_1)(s)) ds \tag{2.7}$$

for $z_1 \in \mathcal{A}_{T-t}$ and $U \in \Gamma_{T-t}$, where

$$\Psi(x,z_1,z_2) = -\frac{1}{2\theta} z_1^* a^{-1}(x) z_1 + \frac{1}{2} z_2^* S(x) z_2 + V(x) \tag{2.8}$$

and $X(s)$ is the solution of the ordinary differential equation (ODE)

$$\begin{cases} dX(s) = \{b(X(s)) + z_1(s) + B(X(s))(Uz_1)(s)\} ds, \\ X(0) = x. \end{cases} \tag{2.9}$$

We define the lower value function $\underline{w}(t,x)$ by

$$\underline{w}(t,x) = \inf_{U \in \Gamma_{T-t}} \sup_{z_1 \in \mathcal{A}_{T-t}} I(t,x;T;z_1,Uz_1). \tag{2.10}$$

We regard $Q_0(x,p)$ as

$$Q_0(x,p) = \sup_{z_1 \in R^N} \left\{ z_1 \cdot p - \frac{1}{2\theta} z_1^* a^{-1} z_1 \right\} + \inf_{z_2 \in R^{N_1}} \left\{ (Bz_2) \cdot p + \frac{1}{2} z_2^* S z_2 \right\} \tag{2.11}$$

and equation (1.16) as the Hamilton–Jacobi equation associated with the differential game (2.10), where we denote by $z_1 \cdot p$ the inner product of z_1 and p. Then we have the following theorem.

THEOREM 2.1. *We assume* (1.5), (1.6), (1.8), *and* (1.10). *Let* $\underline{w}(t,x)$ *be the lower value function of a differential game defined by* (2.10). *Then* $\underline{w}(t,x)$ *is the unique nonnegative viscosity solution of equation* (2.4).

As a corollary of Theorem 1.1 and 2.1 we have the following.

COROLLARY. *Let* $w_\epsilon(t,x)$ *be a nonnegative solution of* (1.12) *with* $Q_0(x,p)$ *defined by* (2.3) *whose existence is assured by Theorem* 1.0*; then* $w_\epsilon(t,x)$ *converges uniformly on each compact set to the lower value function* $\underline{w}(t,x)$ *of the differential game defined by* (2.10) *as* ϵ *tends to zero.*

2.2. Finiteness of the lower value function. We first prove the following lemma.

LEMMA 2.1. *The lower value function* $\underline{w}(t,x)$ *defined by* (2.10) *has a nonnegative finite value.*

Proof. Note that $V(x) \le M(1+|x|^2)^m$ by assumption (1.6), and set

$$\chi(x) = |x|^{2l}, \tag{2.12}$$

where $l > \max\{\frac{m+1}{2}, 1\}$. Then there exists a constant c such that

$$b \cdot \nabla\chi - \frac{1}{2}(\nabla\chi)^*(BS^{-1}B^* - \theta a)\nabla\chi + V \le c \quad \forall x \in R^N \tag{2.13}$$

since

$$-\frac{1}{2}(\nabla\chi)^*(BS^{-1}B^* - \theta a)\nabla\chi \le -\frac{k_2}{2}(\nabla\chi)^* a \nabla\chi \le -2k_2\nu l^2|x|^{4l-2}$$

because of (2.5) and (1.10). Therefore, for each $z_1 \in R^N$,

$$b \cdot \nabla\chi + z_1 \cdot \nabla\chi - \frac{1}{2\theta} z_1^* a^{-1} z_1 - \frac{1}{2}(\nabla\chi)^* BS^{-1}B^*\nabla\chi + V \le c. \tag{2.14}$$

For each control $z_1(\cdot) \in \mathcal{A}_{T-t}$ we consider the ODE

$$\begin{cases} dX(s) = \{b(X(s)) + z_1(s) - BS^{-1}B^*\nabla\chi(X(s))\}ds, \\ X(0) = x. \end{cases} \tag{2.15}$$

Since b and $BS^{-1}B^*\nabla\chi$ are smooth, (2.15) has a local solution $X(s)$. We shall see that the solution is a global one up to $T-t$ and that $\hat{U}$ defined by

$$\hat{U}z_1(s) = -S^{-1}B^*\nabla\chi(X(s)) \tag{2.16}$$

belongs to Γ_{T-t}. Set

$$\zeta_R = \inf\{s : X(s) \notin B_R\} \tag{2.17}$$

for the local solution $X(s)$ of (2.15). Then

$$\begin{aligned} &\chi(X(s\wedge\zeta_R)) - \chi(x) \\ &= \int_0^{s\wedge\zeta_R} \{b\cdot\nabla\chi + z_1(\tau)\cdot\nabla\chi - (\nabla\chi)^* BS^{-1}B^*\nabla\chi\}(X_\tau)d\tau \\ &\le \int_0^{s\wedge\zeta_R} \left(|b||\nabla\chi| + \frac{1}{2\theta} z_1^* a^{-1} z_1\right) d\tau \\ &\quad - \int_0^{s\wedge\zeta_R} \left\{\frac{1}{2}(\nabla\chi)^*(BS^{-1}B^* - \theta a)\nabla\chi + \frac{1}{2}(\nabla\chi)^* BS^{-1}B^*\nabla\chi\right\}(X(\tau))d\tau \\ &\le \int_0^{s\wedge\zeta_R} \left(c_1(1+|X_\tau|)|\nabla\chi|(X(\tau)) + \frac{\nu}{2\theta}|z_1(\tau)|^2\right) d\tau \\ &\quad - \nu\left(k_2 + \frac{\theta}{2}\right)\int_0^{s\wedge\zeta_R} |\nabla\chi|^2(X(\tau))d\tau \end{aligned} \tag{2.18}$$

for some positive constant c_1. Thus we obtain the following inequality

$$c_2 \int_0^{s\wedge\zeta_R} |X(\tau)|^{4l-2} d\tau \le \chi(x) + \frac{\nu}{2\theta} \int_0^s |z_1(\tau)|^2 d\tau + c_3 s \tag{2.19}$$

for some positive constants c_2 and c_3, from which

$$\int_0^{T-t} |X(\tau)|^{4l-2} d\tau < \infty \tag{2.20}$$

follows, by letting R tend to ∞ and s tend to $T-t$, since $z_1(\cdot) \in \mathcal{A}_{T-t}$. It implies that $\int_0^{T-t} |(\hat{U}z_1)(s)|^2 ds < \infty$ and that (2.15) has a global solution up to $T-t$. Now from (2.14) it follows that

$$\begin{aligned} &\chi(X(T-t)) - \chi(x) \\ &= \int_0^{T-t} \{b \cdot \nabla\chi + z_1(s) \cdot \nabla\chi - (\nabla\chi)^* B S^{-1} B^* \nabla\chi\}(X(s)) ds \\ &\le \int_0^{T-t} \frac{1}{2\theta} z_1^*(s) a^{-1} z_1(s) - \frac{1}{2} (\nabla\chi)^* B S^{-1} B^* \nabla\chi(X(s)) ds \\ &\quad - \int_0^{T-t} V(X(s)) ds + c(T-t), \end{aligned} \tag{2.21}$$

which implies that

$$\int_0^{T-t} \Psi(X(s), z_1(s), \hat{U}z_1(s)) ds \le \chi(x) + c(T-t) \tag{2.22}$$

for each $z_1(\cdot) \in \mathcal{A}_{T-t}$. Hence we obtain our present lemma, since $\underline{w}(t,x)$ is obviously nonnegative. □

2.3. Dynamic programming principle. We shall prove the following lemma.

LEMMA 2.2. *For each $0 \le t_0 < t < T$, $x \in R^N$, and $R > 0$,*

$$\begin{aligned} &\underline{w}(t_0, x_0) \\ &= \inf_{U \in \Gamma_{t-t_0}} \sup_{z_1 \in \mathcal{A}_{t-t_0}} \left\{ \int_0^{t_R - t_0} \Psi(X(s), z_1(s), Uz_1(s)) ds \; + \underline{w}(t_R, X(t_R - t_0)) \right\}, \end{aligned} \tag{2.23}$$

where $\tau_R = \inf\{s; |X(s) - x_0| \ge R\}$ and $t_R = t_0 + (t - t_0) \wedge \tau_R$.

Proof. For each δ there exists $U \in \Gamma_{T-t_0}$ such that

$$\sup_{z_1 \in \mathcal{A}_{T-t_0}} I(t_0, x_0; T; z_1, Uz_1) \le \underline{w}(t_0, x_0) + \delta. \tag{2.24}$$

Note that $\underline{w}(t_0, x_0) < \infty$ for each $(t_0, x_0) \in [0,T] \times R^N$ by Lemma 2.1. Denote the right-hand side of (2.23) by $\tilde{w}(t_0, x_0)$; then by definition for each $U \in \Gamma_{T-t_0}$ satisfying (2.24) we have

$$\tilde{w}(t_0, x_0) \le \sup_{z_1 \in \mathcal{A}_{t-t_0}} \left\{ \int_0^{t_R - t_0} \Psi(X(s), z_1(s), Uz_1(s)) ds \; + \; \underline{w}(t_R, X(t_R - t_0)) \right\}, \tag{2.25}$$

and consequently there exists $\hat{z}_1 \in \mathcal{A}_{t-t_0}$ such that

$$\tilde{w}(t_0,x_0) \le \int_0^{t_R-t_0} \Psi(X(s),\hat{z}_1(s),U\hat{z}_1(s))ds \ + \ \underline{w}(t_R,X(t_R-t_0)) + \delta. \tag{2.26}$$

For each $\tilde{z}_1 \in \mathcal{A}_{T-t_R}$ define $z_1 \in \mathcal{A}_{T-t_0}$ by

$$z_1(s) = \begin{cases} \hat{z}_1(s), & 0 \le s \le (t-t_0)\wedge\tau_R, \\ \tilde{z}_1(s-(t-t_0)\wedge\tau_R), & (t-t_0)\wedge\tau_R \le s \le T-t_0, \end{cases} \tag{2.27}$$

and then define $\tilde{U}$ by

$$\tilde{U}\tilde{z}_1(s) = Uz_1((t-t_0)\wedge\tau_R+s), \quad 0 \le s \le T-t_R. \tag{2.28}$$

Then

$$\underline{w}(t_R,X(t_R-t_0)) \le \sup_{\tilde{z}_1\in\mathcal{A}_{T-t_R}} \int_0^{T-t_R} \Psi(\tilde{X}(s),\tilde{z}_1(s),\tilde{U}\tilde{z}_1(s))ds, \tag{2.29}$$

where $\tilde{X}(s)$ is governed by the ODE

$$\begin{cases} d\tilde{X}(s) = (b(\tilde{X}(s)) + \tilde{z}_1(s) + B\tilde{U}\tilde{z}_1(s))ds, \\ \tilde{X}(0) = X(t_R-t_0) \end{cases} \tag{2.30}$$

and $X(s)$ is the solution of

$$\begin{cases} dX(s) = (b(X(s)) + z_1(s) + BUz_1(s))ds, \quad 0 \le s \le t-t_0, \\ X(0) = x_0. \end{cases} \tag{2.31}$$

The right-hand side of (2.29) is finite because of (2.24) and Lemma 2.1, and there exists $\tilde{z}_1^0 \in \mathcal{A}_{T-t_R}$ such that

$$\underline{w}(t_R,X(t_R-t_0)) \le \int_0^{T-t_R} \Psi(\tilde{X}(s),\tilde{z}_1^0(s),\tilde{U}\tilde{z}_1^0(s))ds + \delta. \tag{2.32}$$

Define $z_1^0 \in \mathcal{A}_{T-t_0}$ by

$$z_1^0(s) = \begin{cases} \hat{z}_1(s), \quad 0 \le s \le (t-t_0)\wedge\tau_R, \\ \tilde{z}_1^0(s-(t-t_0)\wedge\tau_R), \quad (t-t_0)\wedge\tau_R \le s \le T-t_0. \end{cases}$$

Then from (2.26) and (2.32) it follows that

$$\tilde{w}(t_0,x_0) \le \int_0^{T-t_0} \Psi(X^0(s),z_1^0(s),Uz_1^0(s))ds + 2\delta, \tag{2.33}$$

where $X^0(s)$ denotes the trajectory associated with z_1^0 and Uz_1^0, and so (2.24) implies that

$$\tilde{w}(t_0,x_0) \le \ \underline{w}(t_0,x_0) + 3\delta \tag{2.34}$$

for each $\delta > 0$. Hence $\tilde{w}(t_0,x_0) \le \underline{w}(t_0,x_0)$.

Now let us prove the converse inequality. For each $(t_1, x) \in [t_0, T] \times R^N$ and $\delta > 0$ there exists $U^{(t_1,x)} \in \Gamma_{T-t_1}$ such that

$$\sup_{\tilde{z}_1 \in \mathcal{A}_{T-t_1}} I(t_1, x; T; \tilde{z}_1, U^{(t_1,x)}\tilde{z}_1) \leq \underline{w}(t_1, x) + \delta. \tag{2.35}$$

Furthermore, there exists $U^0 \in \Gamma_{t-t_0}$ such that

$$\begin{aligned} \sup_{z_1 \in \mathcal{A}_{t-t_0}} \left\{ \int_0^{t_R - t_0} \Psi(X(s), z_1(s), U^0 z_1(s))ds \ + \ \underline{w}(t_R, X(t_R - t_0)) \right\} \\ \leq \tilde{w}(t_0, x_0) + \delta. \end{aligned} \tag{2.36}$$

For $\hat{z}_1 \in \mathcal{A}_{T-t_0}$ define z_1 and $\tilde{z}_1$ by

$$z_1(s) = \hat{z}_1(s), \quad 0 \leq s \leq t - t_0, \tag{2.37}$$

and

$$\tilde{z}_1(s) = \hat{z}_1((t - t_0) \wedge \tau_R + s), \quad 0 \leq s \leq T - t_R, \tag{2.38}$$

where τ_R is defined by $\tau_R = \inf\{s; |X(s) - x_0| \geq R\}$ and $X(s)$ is a solution of the ODE

$$\begin{cases} dX(s) = (b(X(s)) + \hat{z}_1(s) + BU^0\hat{z}_1(s))ds, \\ X(0) = x_0 \end{cases} \tag{2.39}$$

for $U^0 \in \Gamma_{t-t_0}$. Define $U \in \Gamma_{T-t_0}$ by

$$U\hat{z}_1(s) = \begin{cases} U^0 z_1(s), & 0 \leq s \leq t_R - t_0, \\ U^{(t_R, X(t_R - t_0))}\tilde{z}_1(s - (t_R - t_0)), & t_R - t_0 \leq s \leq T - t_0. \end{cases} \tag{2.40}$$

In (2.35) we set $t_1 = t_R$ and $x = X((t - t_0) \wedge \tau_R)$; then

$$\sup_{\tilde{z}_1 \in \mathcal{A}_{T-t_R}} \int_0^{T-t_R} \Psi(\tilde{X}(s), \tilde{z}_1(s), U^{(t_R,x)}\tilde{z}_1(s))ds \ \leq \underline{w}(t_R, x) + \delta, \tag{2.41}$$

where $x = X(t_R - t_0)$ and $\tilde{X}(s)$ is a solution of

$$\begin{cases} d\tilde{X}(s) = (b(\tilde{X}(s)) + \tilde{z}_1(s) + BU^{(t_R,x)}\tilde{z}_1(s))ds, \\ \tilde{X}(0) = X((t - t_0) \wedge \tau_R). \end{cases} \tag{2.42}$$

Define $\hat{X}(s)$ by

$$\hat{X}(s) = \begin{cases} X(s), & 0 \leq s \leq t_R - t_0, \\ \tilde{X}(s - (t_R - t_0)) & t_R - t_0 \leq s \leq T - t_0. \end{cases} \tag{2.43}$$

Then (2.36) and (2.41) imply that

$$\tilde{w}(t_0, x_0) + \delta \geq \int_0^{T-t_0} \Psi(\hat{X}(s), \hat{z}_1(s), U\hat{z}_1(s))ds - \delta \tag{2.44}$$

for each δ, and therefore we obtain

$$\begin{aligned} \tilde{w}(t_0, x_0) + 2\delta &\geq \sup_{\hat{z}_1 \in \mathcal{A}_{T-t_0}} \int_0^{T-t_0} \Psi(\hat{X}(s), \hat{z}_1(s), U\hat{z}_1(s))ds \\ &\geq \underline{w}(t_0, x_0) \end{aligned} \tag{2.45}$$

for all $\delta > 0$. Hence, the other inequality $\tilde{w}(t_0, x_0) \geq \underline{w}(t_0, x_0)$ holds. □

2.4. Continuity of the lower value function. Let $\mathcal{A}_{T-t}(L)$ be a set defined by

$$\mathcal{A}_{T-t}(L) = \left\{ z_1 \in \mathcal{A}_{T-t} : \int_0^{T-t} |z_1(s)|^2 ds \le L \right\}; \tag{2.46}$$

then we have the following lemma.

LEMMA 2.3. *For each R there exists a constant $L = L(R)$ such that*

$$\underline{w}(t,x) = \inf_{U \in \Gamma_{T-t}} \sup_{z_1 \in \mathcal{A}_{T-t}(L)} \int_0^{T-t} \Psi(X(s), z_1(s), Uz_1(s)) ds \tag{2.47}$$

for $x \in B_R$, $0 \le t \le T$.

Proof. For each control $z_1 \in \mathcal{A}_{T-t}$, we consider the solution of the ODE (2.15) and define $\hat{U}$ by (2.16). Then we have

$$\begin{aligned} &\chi(X(T-t)) - \chi(x) \\ &\le \int_0^{T-t} \left\{ c_1(1+|X(s)|)|\nabla\chi|(X(s)) + \frac{1}{2\theta} z_1^* a^{-1} z_1 - \frac{1}{2}(\hat{U}z_1)^* S \hat{U} z_1 \right\} ds \\ &\quad - \frac{1}{2} \int_0^{T-t} (\nabla\chi)^* (BS^{-1}B^* - \theta a) \nabla\chi(X(s)) ds \\ &\le \int_0^{T-t} \left\{ \frac{1}{2\theta} z_1^* a^{-1} z_1 - \frac{1}{2}(\hat{U}z_1)^* S \hat{U} z_1 \right\} ds \\ &\quad - \frac{k'}{2} \int_0^{T-t} |X(s)|^{4l-2} ds + c'(T-t) \end{aligned} \tag{2.48}$$

for some positive constants k', $c' > 0$. Note that

$$\begin{aligned} \int_0^{T-t} V(X(s)) ds &\le M \int_0^{T-t} (1+|X(s)|^2)^m ds \\ &\le c''(T-t) + \frac{k''}{2} \int_0^{T-t} |X(s)|^{4l-2} ds \end{aligned}$$

hold for some constants $c'' > 0$ and $0 < k'' < k'$ since $l > \frac{m+1}{2}$. Therefore, from (2.48) it follows that

$$\begin{aligned} &\int_0^{T-t} \left\{ -\frac{1}{2\theta} z_1^* a^{-1} z_1 + \frac{1}{2}(\hat{U})^* S \hat{U} z_1 + V(X(s)) \right\} ds \\ &\qquad \le \chi(x) + (c'+c'')(T-t) - \frac{k'-k''}{2} \int_0^{T-t} |X(s)|^{4l-2} ds. \end{aligned}$$

In case

$$\chi(x) + (c'+c'')(T-t) < \frac{k'-k''}{2} \int_0^{T-t} |X(s)|^{4l-2} ds, \tag{2.49}$$

we have

$$\int_0^{T-t} \Psi(X(s), z_1(s), \hat{U}z_1(s)) ds < 0. \tag{2.50}$$

On the other hand, if (2.49) does not hold, then setting $\lambda = \|BS^{-1}B^*\|$

$$\begin{aligned}(2.51)\quad &\int_0^{T-t} \Psi(X(s), z_1(s), \hat{U}z_1(s))ds \\ &\leq \int_0^{T-t} \left\{ -\frac{1}{2\theta\nu}|z_1(s)|^2 + \left(4l^2\lambda + \frac{k''}{2}\right)|X(s)|^{4l-2}ds + c''(T-t) \right\} \\ &\leq -\frac{1}{2\theta}\int_0^{T-t} |z_1(s)|^2 ds + \frac{(8l^2\lambda + k'')}{k' - k''}(\chi(x) + (c' + c'')(T-t)) + c''(T-t).\end{aligned}$$

Therefore, if we take $-\frac{1}{2\theta}\int_0^{T-t}|z_1(s)|^2ds$ so large that the right-hand side of (2.51) becomes negative for all $x \in B_R$, then (2.50) always holds. Thus we obtain our present lemma, since $\underline{w}(t,x) \geq 0$. □

Let $K = K(R)$ be a constant defined by

$$(2.52)\qquad K = \frac{1}{\mu}\sup_{x\in B_R}(\chi(x) + cT),$$

where c is the constant that appeared in (2.13). Define a set $\Gamma_{T-t}(K)$ by

$$(2.53)\qquad \Gamma_{T-t}(K) = \left\{ U \in \Gamma_{T-t} : \sup_{z_1 \in \mathcal{A}_{T-t}(L)} \int_0^{T-t} \left(\frac{1}{2}|Uz_1(s)|^2 - \frac{\nu}{2\mu\theta}|z_1(s)|^2 \right) ds \leq K \right\}.$$

Then we have the following lemma.

LEMMA 2.4. *For $x \in B_R$, $0 \leq t \leq T$,*

$$(2.54)\qquad \underline{w}(t,x) = \inf_{U\in\Gamma_{T-t}(K)} \sup_{z_1\in\mathcal{A}_{T-t}(L)} \int_0^{T-t} \Psi(X(s), z_1(s), Uz_1(s))ds.$$

Proof. If $U \in \Gamma_{T-t} \cap \Gamma_{T-t}(K)^c$, then there exists $z_1 \in \mathcal{A}_{T-t}(L)$ such that

$$\begin{aligned}(2.55)\qquad \int_0^{T-t} \left\{ \frac{1}{2}|Uz_1(s)|^2 - \frac{\nu}{2\mu\theta}|z_1(s)|^2 \right\} ds &> K(R) \\ &\geq \frac{1}{\mu}(\chi(x) + cT).\end{aligned}$$

Therefore

$$\begin{aligned}(2.56)\qquad &\int_0^{T-t} \Psi(X(s), z_1(s), Uz_1(s))ds \\ &\geq -\frac{\nu}{2\theta}\int_0^{T-t} |z_1(s)|^2 ds + \frac{\mu}{2}\int_0^{T-t} |Uz_1(s)|^2 ds \\ &> \chi(x) + cT.\end{aligned}$$

Since $\underline{w}(t,x) \leq \chi(x) + c(T-t)$ by (2.22) we obtain our lemma. □

Now we can see the continuity of the lower value function.

LEMMA 2.5. *The lower value function $\underline{w}(t,x)$ is Hölder continuous on $[0,T] \times R^N$.*

Proof. Let us take $x \in B_R$, $z_1 \in \mathcal{A}_{T-t}(L)$, and $U \in \Gamma_{T-t}(K)$ and consider the following equation:

$$\begin{cases} dX(s) = (b(X(s)) + z_1(s) + BUz_1(s))ds, \\ X(0) = x. \end{cases}$$

Then we have

$$\begin{aligned} |X(s) - x| &\le M \int_0^s |X(s) - x| ds + c_1 s + \int_0^s (|z_1(s)| + |BUz_1(s)|) ds \\ &\le M \int_0^s |X(s) - x| ds + c_2 s^{\frac{1}{2}}, \quad 0 \le s \le T - t, \end{aligned}$$

for some positive constants c_1 and c_2 since

$$\int_0^{T-t} |Uz_1(s)|^2 ds \le 2K + \frac{\nu}{\mu\theta} \int_0^{T-t} |z_1(s)|^2 ds \le 2K + \frac{\nu L}{\mu\theta}.$$

Standard arguments using the Gronwall inequality imply that

$$|X(s) - x| \le c_3 s^{\frac{1}{2}}, \quad 0 \le s \le T - t. \tag{2.57}$$

Let us choose x_1, $x_2 \in B_R$, $0 \le t_1 < t_2 \le T$, such that $|x_1| + c_3|t_2 - t_1|^{\frac{1}{2}} < R$. For each $\epsilon > 0$ there exists $\hat{U} \in \Gamma_{T-t_2}(K)$ such that

$$\underline{w}(t_2, x_2) + \epsilon > \sup_{\tilde{z}_1 \in \mathcal{A}_{T-t_2}} I(t_2, x_2; T; \tilde{z}_1, \hat{U}\tilde{z}_1). \tag{2.58}$$

For each $z_1 \in \mathcal{A}_{T-t_1}(L)$ define $\hat{z}_1 \in \mathcal{A}_{T-t_2}(L)$ by

$$\hat{z}_1(s) = z_1(t_2 - t_1 + s), \quad 0 \le s \le T - t_2,$$

and for some $y_0 \in R^{N_1}$ define $U \in \Gamma_{T-t_1}(K)$ by

$$Uz_1(s) = \begin{cases} y_0, & 0 \le s \le t_2 - t_1, \\ \hat{U}\hat{z}_1(s - (t_2 - t_1)), & t_2 - t_1 \le s \le T - t_1. \end{cases}$$

Let $X_1(s)$ be the solution of

$$\begin{cases} dX_1(s) = (b(X_1(s)) + z_1(s) + BUz_1(s))ds, \quad 0 \le s \le T - t_1, \\ X_1(0) = x_1, \end{cases}$$

and let $X_2(s)$ be the solution of

$$\begin{cases} dX_2(s) = (b(X_2(s)) + \hat{z}_1(s) + B\hat{U}\hat{z}_1(s))ds, \quad 0 \le s \le T - t_2, \\ X_2(0) = x_2. \end{cases}$$

Then we have

$$\begin{aligned} |X_2(s) - X_1(t_2 - t_1 + s)| &\le c_4 |x_2 - X_1(t_2 - t_1)| \\ &\le c_5(|x_2 - x_1| + |t_2 - t_1|^{\frac{1}{2}}), \quad 0 \le s \le T - t_2, \end{aligned} \tag{2.59}$$

because $\hat{z}_1(s) = z_1(t_2 - t_1 + s)$ and $\hat{U}\hat{z}_1(s) = Uz_1(t_2 - t_1 + s)$ for $0 \le s \le T - t_2$. Since for each $\epsilon > 0$ there exists $z_1 \in \mathcal{A}_{T-t}(L)$ such that

$$\underline{w}(t_1, x_1) \le I(t_1, x_1; T, z_1, Uz_1) + \epsilon,$$

we obtain by (2.58)

$$\begin{aligned}\underline{w}(t_1, x_1) - \underline{w}(t_2, x_2) &\le I(t_1, x_1; T; z_1, Uz_1) - I(t_2, x_2; T; \hat{z}_1, \hat{U}\hat{z}_1) + 2\epsilon \\ &= \int_0^{T-t_2} \{\Psi(X_1(t_2 - t_1 + s), \hat{z}_1(s), \hat{U}\hat{z}_1(s)) - \Psi(X_2(s), \hat{z}_1(s), \hat{U}\hat{z}_1(s))\} ds \\ &\quad + \int_0^{t_2-t_1} \Psi(X_1(s), z_1(s), Uz_1(s)) ds + 2\epsilon \\ &\le c_6(|x_2 - x_1| + |t_2 - t_1|^{\frac{1}{2}}) + 2\epsilon.\end{aligned}$$

On the other hand, for each $\epsilon > 0$ there exists $U \in \Gamma_{T-t_1}(K)$ such that

$$\underline{w}(t_1, x_1) + \epsilon > \sup_{z_1 \in \mathcal{A}_{T-t_1}} I(t_1, x_1; T; z_1, Uz_1). \tag{2.60}$$

For each $\hat{z}_1 \in \mathcal{A}_{T-t_2}$ and some $z_0 \in R^N$ define $z_1 \in \mathcal{A}_{T-t_1}$ by

$$z_1(s) = \begin{cases} z_0, & 0 \le s \le t_2 - t_1, \\ \hat{z}_1(s - (t_2 - t_1)), & t_2 - t_1 \le s \le T - t_1. \end{cases}$$

Now define $\hat{U} \in \Gamma_{T-t_2}$ by

$$\hat{U}\hat{z}_1(s) = (Uz_1)(s + t_2 - t_1).$$

Let $X_1(s)$ be the solution of

$$\begin{cases} dX_1(s) = (b(X_1(s)) + z_1(s) + BUz_1(s))ds, & 0 \le s \le T - t_1, \\ X_1(0) = x_1, \end{cases}$$

and let $X_2(s)$ be the solution of

$$\begin{cases} dX_2(s) = (b(X_2(s)) + \hat{z}_1(s) + B\hat{U}\hat{z}_1(s))ds, & 0 \le s \le T - t_2, \\ X_2(0) = x_2. \end{cases}$$

Then we obtain (2.59) for these $X_1(s)$ and $X_2(s)$ in the same way as above. Since for each $\epsilon > 0$ there exists $\hat{z}_1 \in \mathcal{A}_{T-t_2}$ such that

$$\underline{w}(t_2, x_2) \le I(t_2, x_2; T : \hat{z}_1, \hat{U}\hat{z}_1) + \epsilon,$$

we have by (2.60)

$$\begin{aligned}\underline{w}(t_2, x_2) - \underline{w}(t_1, x_1) &\le I(t_2, x_2; T; \hat{z}_1, \hat{U}z_1) - I(t_1, x_1; T; z_1, Uz_1) + 2\epsilon \\ &\le \int_0^{T-t_2} \{\Psi(X_2(s), \hat{z}_1(s), Uz_1(s)) - \Psi(X_1(t_2 - t_1 + s), \hat{z}_1(s), \hat{U}\hat{z}_1(s))\} ds \\ &\quad - \int_0^{t_2-t_1} \Psi(X_1(s), z_1(s), Uz_1(s)) ds + 2\epsilon \\ &\le c_7(|x_2 - x_1| + |t_2 - t_1|^{\frac{1}{2}}) + 2\epsilon.\end{aligned}$$

Thus we see that $\underline{w}(t, x)$ is Hölder continuous. □

2.5. Viscosity solution.

PROPOSITION 2.6. *The lower value function* $\underline{w}(t,x)$ *defined by* (2.10) *is a viscosity solution of* (2.4).

Proof. Let $\phi \in C^\infty((0,T)\times R^N)$ be a function such that $\underline{w}-\phi$ attains its maximum at (t_0,x_0). Then we shall prove that

$$\frac{\partial\phi}{\partial t}(t_0,x_0)+H(x_0,\nabla\phi(t_0,x_0))\geq 0, \tag{2.61}$$

where $H(x,p)=b\cdot p+Q_0(x,p)+V(x)$, $p\in R^N$. Suppose that (2.61) does not hold. Then there exists $\delta>0$ such that

$$\frac{\partial\phi}{\partial t}(t_0,x_0)+H(x_0,\nabla\phi(t_0,x_0))\leq -\delta<0,$$

and so in a neighborhood G of (t_0,x_0),

$$\frac{\partial\phi}{\partial t}(t_0+s,x)+H(x,\nabla\phi(t_0+s,x))\leq -\frac{\delta}{2}<0,\quad (t_0+s,x)\in G. \tag{2.62}$$

Therefore, for each $z_1\in R^N$,

$$\begin{aligned}\frac{\partial\phi}{\partial t}&+b\cdot\nabla\phi+z_1\cdot\nabla\phi-\frac{1}{2\theta}z_1^*a^{-1}z_1-\frac{1}{2}(\nabla\phi)^*BS^{-1}B^*\nabla\phi+V\\&\leq -\frac{\delta}{2}-\frac{1}{2\theta}(z_1-\theta a\nabla\phi)^*a^{-1}(z_1-\theta a\nabla\phi)<0,\quad (t_0+s,x)\in G.\end{aligned} \tag{2.63}$$

Let $X(s)$ be the solution of the ODE

$$\begin{cases} dX(s)=(b(X(s))+z_1(s)-BS^{-1}B^*\nabla\phi(t_0+s,X(s)))ds,\\ X(0)=x_0,\end{cases} \tag{2.64}$$

and set

$$Uz_1(s)=-S^{-1}B^*\nabla\phi(t_0+s,X(s)) \tag{2.65}$$

and

$$f(x,z_1,z_2)=b(x)+z_1+Bz_2,\quad z_1\in R^N,\quad z_2\in R_1^N. \tag{2.66}$$

Then by (2.63) we have

$$\begin{aligned}&\frac{\partial\phi}{\partial t}(t_0+s,X(s))+f(X(s),z_1(s),Uz_1(s))\cdot\nabla\phi(t_0+s,X(s))\\&\quad+\Psi(X(s),z_1(s),Uz_1(s))\\&\leq -\frac{1}{2\theta}(z_1(s)-\theta a\nabla\phi(t_0+s,X(s)))^*a^{-1}(z_1(s)-\theta a\nabla\phi(t_0+s,X(s)))-\frac{\delta}{2},\\&\quad(t_0+s,x)\in G.\end{aligned} \tag{2.67}$$

Take t and R such that $G_0=[t_0,t)\times B_R(x_0)\subset G$, and set

$$\xi(s,x)=\frac{1}{2}|x-x_0|^2+(s-t_0) \tag{2.68}$$

and

$$\tau = \inf\{s : (t_0 + s, X(s)) \notin G_0\} \equiv (t - t_0) \wedge \tau_R. \tag{2.69}$$

Then for each $z_1(s)$

$$\begin{aligned}
\xi(t_0 + \tau, X(\tau)) &= \int_0^\tau \left\{ \frac{\partial \xi}{\partial s}(t_0 + s, X(s)) + f(X, z_1, Uz_1) \cdot \nabla \xi(t_0 + s, X(s)) \right\} ds \\
&= \tau + \int_0^\tau (z_1(s) - \theta a \nabla \phi) \cdot \nabla \xi(t_0 + s, X(s)) ds \\
&\quad + \int_0^\tau (b + \theta a \nabla \phi - BS^{-1}B^* \nabla \phi) \cdot \nabla \xi(t_0 + s, X(s)) ds \\
&\leq c_1 \tau + \frac{1}{2\theta} \int_0^\tau (z_1(s) - \theta a \nabla \phi)^* a^{-1} (z_1(s) - \theta a \nabla \phi)(t_0 + s, X(s)) ds \\
&\quad + \frac{\theta}{2} \int_0^\tau (\nabla \xi)^* a \nabla \xi(t_0 + s, X(s)) ds \\
&\leq c_2 \left(\frac{\delta}{2} \tau + \frac{1}{2\theta} \int_0^\tau (z_1(s) - \theta a \nabla \phi)^* a^{-1} (z_1(s) - \theta a \nabla \phi)(t_0 + s, X(s)) ds \right)
\end{aligned}$$

for some positive constants c_1 and c_2 since b, a, $\nabla \phi$, $BS^{-1}B^*$, and $\nabla \xi$ are bounded in G. Now from $\xi(t_0 + \tau, X(\tau)) \geq (t - t_0) \wedge \frac{R^2}{2}$ we obtain

$$c_2^{-1}(t - t_0) \wedge \frac{R^2}{2} \leq \frac{\delta}{2} \tau + \frac{1}{2\theta} \int_0^\tau (z_1(s) - \theta a \nabla \phi)^* a^{-1} (z_1(s) - \theta a \nabla \phi)(t_0 + s, X(s)) ds \tag{2.70}$$

for each $z_1(s)$, which implies that

$$0 < \inf_{z_1 \in \mathcal{A}_{t-t_0}} \left[\frac{\delta}{2} \tau + \frac{1}{2\theta} \int_0^\tau (z_1(s) - \theta a \nabla \phi)^* a^{-1} (z_1(s) - \theta a \nabla \phi)(t_0 + s, X(s)) ds \right]. \tag{2.71}$$

Thus from (2.67) it follows that

$$\sup_{z_1 \in \mathcal{A}_{t-t_0}} \int_0^\tau \left\{ \frac{\partial \phi}{\partial t} + f(X, z_1, Uz_1) \cdot \nabla \phi(t_0 + s, X(s)) + \Psi(X, z_1, Uz_1) \right\} ds < 0. \tag{2.72}$$

Hence

$$\inf_{U \in \Gamma_{t-t_0}} \sup_{z_1 \in \mathcal{A}_{t-t_0}} \int_0^\tau \left\{ \frac{\partial \phi}{\partial s} + f(X, z_1, Uz_1) \cdot \nabla \phi + \Psi(X, z_1, Uz_1) \right\} ds < 0. \tag{2.73}$$

On the other hand, $\underline{w} - \phi$ attains its local maximum at (t_0, x_0), and we have

$$\underline{w}(t_0, x_0) - \phi(t_0, x_0) \geq \underline{w}(t_0 + \sigma, X(\sigma)) - \phi(t_0 + \sigma, X(\sigma)) \tag{2.74}$$

for sufficiently small σ. Therefore, by (2.23) in section 2.3,

$$\inf_{U \in \Gamma_{t-t_0}} \sup_{z_1 \in \mathcal{A}_{t-t_0}} \left[\int_0^\tau \Psi(X, z_1, Uz_1) ds + \phi(t_0 + \tau, X(\tau)) - \phi(t_0, x_0) \right] \geq 0, \tag{2.75}$$

which means

$$\inf_{U\in\Gamma_{t-t_0}}\sup_{z_1\in\mathcal{A}_{t-t_0}}\int_0^\tau\left\{\frac{\partial\phi}{\partial s}+f(X,z_1,Uz_1)\cdot\nabla\phi+\Psi(X,z_1,Uz_1)\right\}ds\geq 0. \tag{2.76}$$

It contradicts (2.73).

On the other hand, let $\phi\in C^\infty((0,T)\times R^N)$ be a function such that $\underline{w}-\phi$ attains its local minimum at (t_0,x_0). Then we shall prove that

$$\frac{\partial\phi}{\partial t}(t_0,x_0)+H(x_0,\nabla\phi(t_0,x_0))\leq 0. \tag{2.77}$$

Suppose that (2.77) fails to hold. Then there exists $\delta>0$ such that

$$\frac{\partial\phi}{\partial t}(t_0,x_0)+H(x_0,\nabla\phi(t_0,x_0))>\delta>0. \tag{2.78}$$

Then in a neighborhood G of (t_0,x_0),

$$\frac{\partial\phi}{\partial t}(t_0+s,x)+H(x,\nabla\phi(t_0+s,x))>\frac{\delta}{2},\quad (t_0+s,x)\in G. \tag{2.79}$$

Since

$$\sup_{z_1\in R^N}\left\{z_1\cdot\nabla\phi-\frac{1}{2\theta}z_1^*a^{-1}z_1\right\}=\frac{\theta}{2}(\nabla\phi)^*a\nabla\phi, \tag{2.80}$$

$$\inf_{z_2\in R^{N_1}}\left\{(Bz_2)\cdot\nabla\phi+\frac{z_2^*Sz_2}{2}\right\}=-\frac{1}{2}(\nabla\phi)^*BS^{-1}B^*\nabla\phi, \tag{2.81}$$

and $\frac{\theta}{2}(\nabla\phi)^*a\nabla\phi$ is continuous, there exists $\hat{z}_1\in R^N$ and a neighborhood $\hat{G}\subset G$ of (t_0,x_0) such that

$$\begin{aligned}&\frac{\partial\phi}{\partial t}+b\cdot\nabla\phi+\hat{z}_1\cdot\nabla\phi-\frac{1}{2\theta}\hat{z}_1^*a^{-1}\hat{z}_1+(Bz_2)\dot{\nabla}\phi+\frac{1}{2}z_2^*Sz_2+V\\&\qquad\geq\frac{\delta}{4}+\frac{1}{2}(z_2+S^{-1}B^*\nabla\phi)^*S(z_2+S^{-1}B^*\nabla\phi)>0\quad\text{ in }\hat{G}\end{aligned} \tag{2.82}$$

for each $z_2\in R^{N_1}$. Take a control $\hat{z}_1(s)\equiv\hat{z}_1$ and consider the ODE

$$\begin{cases}d\hat{X}(s)=(b(\hat{X}(s))+\hat{z}_1+BU\hat{z}_1)ds,\\ \hat{X}(0)=x_0\end{cases} \tag{2.83}$$

for $U\in\Gamma_{t-t_0}$. Then

$$\begin{aligned}&\frac{\partial\phi}{\partial t}+f(\hat{X},\hat{z}_1,U\hat{z}_1)\cdot\nabla\phi+\Psi(\hat{X},\hat{z}_1,U\hat{z}_1)\\&\geq\frac{\delta}{4}+\frac{1}{2}(U\hat{z}_1+S^{-1}B^*\nabla\phi)^*S(U\hat{z}_1+S^{-1}B^*\nabla\phi),\quad(t_0+s,\hat{X}(s))\in\hat{G}\end{aligned} \tag{2.84}$$

for each $U\in\Gamma_{t-t_0}$. Define $G_0=[t_0,t)\times B_R(x_0)\subset\hat{G}$, and let $\tau=\inf\{s:(t_0+$

$s, \hat{X}(s)) \notin G_0\}$. Take a function ξ defined by (2.68); then

(2.85)
$$\begin{aligned}\xi(t_0+\tau,\hat{X}(\tau)) &= \int_0^\tau \left\{\frac{\partial \xi}{\partial s} + f(\hat{X},\hat{z}_1,U\hat{z}_1)\cdot\nabla\xi(t_0+s,\hat{X}(s))\right\}ds \\ &= \int_0^\tau \{1 + b\cdot\nabla\xi + \hat{z}_1\cdot\nabla\xi + (BU\hat{z}_1)\cdot\nabla\xi\}ds \\ &\le c_1\left(\tau + \int_0^\tau |U\hat{z}_1 + S^{-1}B^*\nabla\phi|ds\right) \\ &\le c_2\left(\frac{\delta}{4}\tau + \frac{1}{2}\int_0^\tau (U\hat{z}_1 + S^{-1}B^*\nabla\phi)^* S(U\hat{z}_1 + S^{-1}B^*\nabla\phi)ds\right)\end{aligned}$$

for some positive constants c_1 and c_2. Thus we see in the same way as above

$$\begin{aligned}&\inf_{U\in\Gamma_{t-t_0}}\left(\frac{\delta}{4}\tau + \frac{1}{2}\int_0^\tau (U\hat{z}_1 + S^{-1}B^*\nabla\phi)^* S(U\hat{z}_1 + S^{-1}B^*\nabla\phi)ds\right) \\ &> c_2^{-1}(t-t_0)\wedge\frac{R^2}{2} > 0.\end{aligned} \tag{2.86}$$

Hence,

$$\inf_{U\in\Gamma_{t-t_0}}\sup_{z_1\in\mathcal{A}_{t-t_0}}\int_0^\tau\left\{\frac{\partial\phi}{\partial s} + f(X,z_1,Uz_1)\cdot\nabla\phi + \Psi(X,z_1,Uz_1)\right\}ds > 0. \tag{2.87}$$

Since $\underline{w} - \phi$ attains its local minimum at (t_0, x_0),

$$\underline{w}(t_0,x_0) - \phi(t_0,x_0) \le \underline{w}(t_0+\sigma, X(\sigma)) - \phi(t_0+\sigma, X(\sigma))$$

for sufficiently small σ. Therefore from (2.23) it follows that

$$\begin{aligned}0 &\ge \inf_{U\in\Gamma_{t-t_0}}\sup_{z_1\in\mathcal{A}_{t-t_0}}\left[\int_0^\tau \Psi(X,z_1,Uz_1)ds + \phi(t_0+\tau,X(\tau)) - \phi(t_0,x_0)\right] \ge 0 \\ &= \inf_{U\in\Gamma_{t-t_0}}\sup_{z_1\in\mathcal{A}_{t-t_0}}\int_0^\tau\left\{\frac{\partial\phi}{\partial s} + f(X,z_1,Uz_1)\cdot\nabla\phi + \Psi(X,z_1,Uz_1)\right\}ds,\end{aligned}$$

which contradicts (2.87). □

2.6. Uniqueness of the positive solution. Let us set

$$\Lambda^{ij} = (BS^{-1}B^* - \theta a)^{ij}; \tag{2.88}$$

then there exist positive constants k_1, $k_2 > 0$, such that

$$k_2|\xi|^2 \le \Lambda^{ij}\xi_i\xi_j \le k_1|\xi|^2 \quad \forall \xi \in R^N \tag{2.89}$$

by our assumptions. Equation (2.4) reads

$$\begin{cases}\dfrac{\partial w}{\partial t} + b\cdot\nabla w - \dfrac{1}{2}(\nabla w)^*\Lambda\nabla w + V(x) = 0, \\ w(T,x) = 0.\end{cases} \tag{2.90}$$

Let $w(t,x)$ be a nonnegative viscosity solution of (2.90), and set

$$w_R(t,x) = \inf_{z\in\mathcal{A}_{T-t}} \left[\int_0^{\tilde{t}_R - t} \left\{ \frac{1}{2} z(s)^* \Lambda^{-1} z(s) + V(X(s)) \right\} ds + w(\tilde{t}_R, X(\tilde{t}_R - t)) \right], \tag{2.91}$$

where $\tilde{t}_R = t + (T-t) \wedge \zeta_R$ and $X(s)$ is the solution of

$$\begin{cases} dX(s) = (b(X(s)) + z(s))ds, \\ X(0) = x \end{cases} \tag{2.92}$$

for $z \in \mathcal{A}_{T-t}$.

LEMMA 2.7. *For each R $w_R(t,x) = w(t,x)$.*

Proof. It is easy to see that w_R is a viscosity solution of the Dirichlet problem

$$\begin{cases} \dfrac{\partial w_R}{\partial t} + b \cdot \nabla w - \dfrac{1}{2} (\nabla w_R)^* \Lambda \nabla w_R + V = 0 \quad \text{in } (0,T) \times B_R, \\ w_R(t,x) = w(t,x), \quad x \in \partial B_R, \quad 0 < t < T, \\ w_R(T,x) = 0, \end{cases} \tag{2.93}$$

and the proof is seen in Fleming and Soner [12, cf. Theorems 7.1 and 16.1, Chapter II]. Since it is known that the viscosity solution of (2.93) is unique (cf. [4]), we see that $w_R(t,x) = w(t,x)$ for each R. □

LEMMA 2.8. *Let $X(s)$ be a solution of* (2.92)*; then*

$$|X(s)|^2 \le c_1 |x|^2 + c_2 |T-t| + c_3 \int_0^s |z(\zeta)|^2 d\zeta, \quad 0 \le s \le T-t, \tag{2.94}$$

where c_1, c_2, and c_3 are positive constants.

Proof.

$$\begin{aligned} |X(s)|^2 - |X(0)|^2 &= 2 \int_0^s \{ b(X(\zeta)) \cdot X(\zeta) + z(\zeta) \cdot X(\zeta) \} d\zeta \\ &\le 2M \int_0^s (1 + |X(\zeta)|) |X(\zeta)| d\zeta + \int_0^s (|z(\zeta)|^2 + |X(\zeta)|^2) d\zeta \\ &\le Ms + \int_0^s |z(\zeta)|^2 d\zeta + (3M+1) \int_0^s |X(\zeta)|^2 d\zeta. \end{aligned} \tag{2.95}$$

Set $m(s) = |X(s)|$; then standard arguments using the Gronwall inequality for $m(s)$ imply our present lemma. □

PROPOSITION 2.9. *A nonnegative viscosity solution of* (2.90) *is unique.*

Proof. By Lemma 2.7 for each R

$$w(t,x) = \inf_{z\in\mathcal{A}_{T-t}} \left[\int_0^{\tilde{t}_R - t} \left\{ \frac{1}{2} z(s)^* \Lambda z(s) + V(X(s)) \right\} ds + w(\tilde{t}_R, X(\tilde{t}_R - t)) \right]. \tag{2.96}$$

Fix (t,x) and take R sufficiently large such that

$$w(t,x) < \frac{1}{2k_1 c_3} (R^2 - c_1 |x|^2 - c_2 (T-t)).$$

Suppose that $\sup_{0\le s\le T-t}|X(s)|\ge R$; then by Lemma 2.8

$$c_3^{-1}(R^2-c_1|x|^2-c_2(T-t))\le\int_0^{\zeta_R}|z(s)|^2ds$$

and consequently

$$\begin{aligned}(2.97)\qquad &\int_0^{\tilde t_R-t}\left\{\frac{1}{2}z(s)^*\Lambda^{-1}z(s)+V(X(s))\right\}ds+w(\tilde t_R,X(\tilde t_R-t))\\ &\ge\int_0^{\tilde t_R-t}\frac{1}{2}z(s)^*\Lambda^{-1}z(s)\ge\frac{1}{2k_1c_3}(R^2-c_1|x|^2-c_2(T-t))\\ &>w(t,x).\end{aligned}$$

Thus we have

$$(2.98)\qquad w(t,x)=\inf_{z\in\mathcal{A}_{T-t}}\int_0^{T-t}\left\{\frac{1}{2}z(s)^*\Lambda^{-1}z(s)+V(X(s))\right\}ds,$$

which implies uniqueness of a viscosity solution of (2.90). □

Theorem 2.1 is a direct consequence of Propositions 2.6 and 2.9.

REFERENCES

[1] T. Basar and P. Bernhard, *H^∞-Optimal Control and Related Minimax Design Problems*, Birkhäuser Boston, Cambridge, MA, 1991.

[2] A. Bensoussan and J. S. Baras, *Output Feedback Risk-Sensitive Control and Differential Games*, INRIA, Rocquencourt, France, 1995, preprint.

[3] M. G. Crandall, L. C. Evans, and P. L. Lions, *Some properties of viscosity solutions of Hamilton-Jacobi equations,* Trans. Amer. Math. Soc., 282 (1984), pp. 487–502.

[4] M. G. Crandall, H. Ishii, and P. L. Lions, *User's guide to viscosity solutions of second order partial differential equations*, Bull. Amer. Math. Soc., 27 (1993), pp. 1–67.

[5] M. G. Crandall and P. L. Lions, *Viscosity solutions of Hamilton-Jacobi equations*, Trans. Amer. Math. Soc., 277 (1983), pp. 1–42.

[6] R. J. Elliott and N. J. Kalton, *The existence of value in differential games*, in Memoirs of the American Mathematical Society 126, AMS, Providence, RI, 1972.

[7] L. C. Evans and H. Ishii, *A pde approach to some asymptotic problems concerning random differential equations with small noise intensities* Ann. Inst. H. Poincaré, Anal. Non Linéaire, 2 (1985), pp. 1–20.

[8] L. C. Evans and P. E. Souganidis, *Differential games and representation formulas for solutions of Hamilton-Jacobi-Isaacs equations*, Indiana Univ. Math. J., 33 (1984), pp. 773–797.

[9] W. H. Fleming and W. M. McEneaney, *Risk Sensitive Optimal Control and Differential Games*, Lecture Notes in Control and Information Science 184, Springer-Verlag, Berlin, New York, 1992, pp. 185–197.

[10] W. H. Fleming and W.M. McEneaney, *Risk-sensitive control on an infinite time horizon*, SIAM J. Control Optim., 33 (1995), pp. 1881–1915.

[11] W. H. Fleming and M. R. James, *The risk-sensitive index and H_2 and H_∞ norms for nonlinear systems*, Math. Control Signals Systems, 8 (1995), pp. 199–221.

[12] W. H. Fleming and M. Soner, *Controlled Markov Processes and Viscosity Solutions*, Springer-Verlag, New York, 1993.

[13] M. R. James, *Asymptotic analysis of nonlinear stochastic risk-sensitive control and differential games*, Math. Control Signals Systems, 5 (1992), pp. 401–417.

[14] M. R. James, J. S. Baras, and R. J. Elliot, *Output feedback risk-sensitive control and differential games for continuous-time nonlinear systems*, in 32nd IEEE CDC, San Antonio, TX, 1993.

[15] W. M. McEneaney, *Connections Between Risk-Sensitive Stochastic Control, Differential Games and H_∞-Control: The Nonlinear Case*, Ph.D. thesis, Brown University, Providence, RI, 1993.

[16] W. M. McEneaney, *Uniqueness for viscosity solutions of nonstationary Hamilton–Jacobi–Bellman equations under some a priori conditions* (*with applications*), SIAM J. Control Optim., 33 (1995), pp. 1560–1576.
[17] W. M. McEneaney, *Robust control and differential games on a finite time horizon*, Math. Control Signals Systems, 8 (1995), pp. 138–166.
[18] H. Nagai, *Bellman equations of risk-sensitive control*, SIAM J. Control Optim., 34 (1996), pp. 74–101.
[19] E. Roxin, *The axiomatic approach in differential games*, J. Optim. Theory Appl., 3 (1969), pp. 153–163.
[20] P. Soravia, H_∞ *Control of nonlinear systems: Differential games and viscosity solutions*, SIAM J. Control Optim., 34 (1996), pp. 1071–1097.
[21] R. P. Varaiya, *The existence of solutions to a differential games*, SIAM J. Control Optim., 5 (1967), pp. 153–162.
[22] P. Whittle, *A risk sensitive maximum principle*, Systems Control Lett., 16 (1990), pp. 183–192.
[23] P. Whittle, *A risk sensitive maximum principle: The case of imperfect state information*, IEEE Trans. Automat. Control, 36 (1991), pp. 793–801.

SIAM J. CONTROL OPTIM.
Vol. 35, No. 4, pp. 1116–1131, July 1997

002

FINITE-DIMENSIONAL FILTERS WITH NONLINEAR DRIFT VII: MITTER CONJECTURE AND STRUCTURE OF η^*

JIE CHEN† AND STEPHEN S.-T. YAU‡

Abstract. The concept of estimation algebra introduced independently by Brockett and Mitter has been playing a fundamental role in the investigation of finite-dimensional nonlinear filters. Mitter conjectured that the observation terms $h_i(x)$ are polynomials of degree one if the corresponding estimation algebra is finite dimensional. Chiou, Leung, and the present authors classify all finite-dimensional estimation algebra of maximal rank with dimension of the state space less than or equal to three. In this paper, we prove the Mitter conjecture for finite-dimensional estimation algebra of maximal rank with arbitrary state space dimension. In the course of our proof, we show that the $\Omega = (\frac{\partial f_j}{\partial x_i} - \frac{\partial f_i}{\partial x_j})$ matrix, where f denotes the drift term, has special linear structure which generalizes our previous result in [J. Chen and S. S.-T. Yau, *Math. Control Signals Systems*, 9 (1996), to appear]. We also give a structure theorem for $\eta = \sum_{i=1}^n \frac{\partial f_i}{\partial x_i} + \sum_{i=1}^n f_i^2 + \sum_{i=1}^m h_i^2$.

Key words. finite-dimensional nonlinear filter, Mitter conjecture, estimation algebras

AMS subject classifications. 17B30, 35J15, 60G35, 93E11

PII. S0363012994272836

1. Introduction.

The idea of using estimation algebras to construct finite-dimensional nonlinear filters was first proposed in Brockett and Clark [Br–Cl], Brockett [Br], and Mitter [Mi]. The concept of estimation algebras has proved to be an invaluable tool in the study of nonlinear filtering problems. In 1983, Brockett proposed classifying all finite-dimensional estimation algebras. As a first step to attack this problem, Mitter conjectured that the observation terms $h_i(x)$ are affine polynomials. In [Ch2–Ya], Chiou and Yau first introduced the concept of estimation algebra of maximal rank. The purpose of this paper is to show that Mitter conjecture is true for all finite-dimensional estimation algebras of maximal rank. In [Wo], the concept of Ω is introduced, defined as the matrix whose (i,j) entry is $\frac{\partial f_j}{\partial x_i} - \frac{\partial f_i}{\partial x_j}$, where f is the drift term of the state evolution equation. Recently, Yau [Ya] has studied a filtering system such that all entries of Ω are constants. He was able to classify all finite-dimensional estimation algebras of maximal rank and proved Mitter conjecture for such a filtering system. If the dimension of the state space is two or three, then Chiou and Yau [Ch2–Ya] and Chen, Leung, and Yau [C–L–Y] have shown, respectively, that all entries of Ω are constants as long as the estimation algebra is of maximal rank and finite dimensional. Thus finite-dimensional estimation algebra of maximal rank is completely classified if the dimension of the state space is at most three.

In [Ch1–Ya], we have shown that Ω is an affine matrix in the sense that every entry in Ω is an affine polynomial if the estimation algebra is of maximal rank and finite-dimensional. This is a fundamental step in classifying finite-dimensional estimation of maximal rank. In fact we proved that Ω has a special affine structure. The purpose of this paper is to give affirmative solution to Mitter conjecture for finite-dimensional estimation algebra of maximal rank. The following is our main theorem.

*Received by the editors August 15, 1994; accepted for publication (in revised form) April 23, 1996. This research was supported by Army Research Office grant DAAH04-93-0006.
http://www.siam.org/journals/sicon/35-4/27283.html
†Department of Management and Information, City of Chicago, 50 W. Washington Boulevard, Cl-54, Chicago, IL 60602 (chen@gauss.math.uic.edu).
‡Control and Information Laboratory, MSCS, M/C 249, University of Illinois at Chicago, 851 South Morgan Street, Chicago, IL 60607-7045 (u32790@uic.edu).

MAIN THEOREM. *Let E be a finite-dimensional estimation algebra of maximal rank. Let k be the quadratic rank of E (cf. Definition 2.2 below). Then*

(1) *the observation terms $h_i(x)$, $1 \le i \le m$, are affine polynomials.*

(2) (a) *ω_{ij}, for $1 \le i \le k$ or $1 \le j \le k$, are constants*

(b) *ω_{ij}, for $k+1 \le i, j \le n$, are degree-one polynomials in $x_{k+1}, \dots, x_n$.*

(3) *$\eta = \sum_{i=1}^n \frac{\partial f_i}{\partial x_i} + \sum_{i=1}^n f_i^2 + \sum_{i=1}^m h_i^2$ is a homogeneous polynomial of degree four. Moreover, η_4 (homogeneous polynomial of degree-four part of η) depends only on $x_{k+1}, \dots, x_n$ variables.*

Notice that in our previous paper, we proved only that ω_{ij} for $1 \le i \le k$ or $1 \le j \le k$ are affine polynomials in $x_1, \dots, x_k$. It is precisely the improvement of the result on ω_{ij} for $1 \le i \le k$ or $1 \le j \le k$ that allows us to solve the Mitter conjecture affirmatively. This paper is, in essence, a continuation of [Ch1–Ya], and we strongly recommend that readers familiarize themselves with the results in [Ch1–Ya]. However, every effort will be made to make this paper as self-contained as possible without too much duplication of the previous paper.

2. Basic concepts. The filtering problem here is based on the following signal observation model:

$$\begin{cases} dx(t) = f\big(x(t)\big)dt + g\big(x(t)\big)dv(t), & x(0) = x_0, \\ dy(t) = h\big(x(t)\big)dt + dw(t), & y(0) = 0, \end{cases} \tag{2.1}$$

in which x, v, y, and w are, respectively, $\mathbb{R}^n$-, $\mathbb{R}^p$-, $\mathbb{R}^m$-, and $\mathbb{R}^m$-valued processes and v and w have components which are independent, standard Brownian process. We further assume that $n = p$, f, h are C^∞ smooth and that g is an orthogonal matrix. We shall refer to $x(t)$ as the state of the system at time t and $y(t)$ as the observation at time t. $\rho(t,x)$, the conditional probability density of the state, $x(t)$, given the observation $\big\{y(s) : 0 \le s \le t\big\}$ is determined by the Duncan–Mortensen–Zakai equation, which in the unnormalized form is given by (see [Da–Ma], for example)

$$\frac{d}{dt}\sigma(t,x) = L_0\sigma(t,x)dt + \sum_{i=1}^m L_i\sigma(t,x)dy_i(t), \quad \sigma(0,x) = \sigma_0, \tag{2.2}$$

where

$$L_0 = \frac{1}{2}\sum_{i=1}^n \frac{\partial^2}{\partial x_i^2} - \sum_{i=1}^n f_i\frac{\partial}{\partial x_i} - \sum_{i=1}^n \frac{\partial f_i}{\partial x_i} - \frac{1}{2}\sum_{i=1}^m h_i^2$$

and for $i = 1, \dots, m$, L_i is the zero-degree differential operator of multiplication by h_i. (If a is a vector, we use the notation a_i to represent the ith component of a.) σ_0 is the probability density of the initial point x_0. When the observation is absent—that is, $h = 0$—then (2.2) is simply the Kolmogorov equation.

It is important to find efficient ways to solve (2.2), which is the subject of many research studies in nonlinear filtering theory. For this purpose, we need to introduce the following definition.

DEFINITION 2.1. *The estimation algebra E of a filtering system* (2.1) *is defined to be the Lie algebra generated by $\{L_0, L_1, \dots, L_m\}$. E is said to be an estimation algebra of maximal rank if for every $1 \le i \le n$ there exists a constant c_i such that $x_i + c_i$ is in E.*

In [Wo], the concept of Ω is introduced, defined as the matrix whose (i,j) element ω_{ij} is $\frac{\partial f_j}{\partial x_i} - \frac{\partial f_i}{\partial x_j}$.

Define

$$D_i = \frac{\partial}{\partial x_i} - f_i$$

and

$$\eta = \sum_{i=1}^{n} \frac{\partial f_i}{\partial x_i} + \sum_{i=1}^{n} f_i^2 + \sum_{i=1}^{m} h_i^2.$$

Then

$$L_0 = \frac{1}{2}\left(\sum_{i=1}^{n} D_i^2 - \eta\right).$$

Recall here that the assumption of maximal rank on E implies that E contains all affine polynomial and the operations $D_1, \ldots, D_n$. This follows from the fact that $D_j = [L_0, x_j + c_j]$ and $1 = [D_j, x_j + c_j]$.

The following theorem proved in [Ya] plays a fundamental roles in Mitter conjecture as well as the classification of finite-dimensional estimation algebra.

THEOREM 2.1. *Let E be a finite-dimensional estimation algebra of* (2.1) *such that $\omega_{ij} = \frac{\partial f_j}{\partial x_i} - \frac{\partial f_i}{\partial x_j}$ are constant functions. Then $h_i(x)$, $1 \le i \le m$, are affine polynomials. If in addition, E is of maximal rank, then E is a real vector space of dimension $2n+2$ with basis given by $1, x_1, x_2, \ldots, x_n, D_1, \ldots, D_n$ and L_0.*

We need the following basic result [Oc] for later discussion.

THEOREM 2.2 (Ocone). *Let E be a finite-dimensional estimation algebra. If φ is a function in E, then φ is a polynomial of degree at most two.*

In our previous paper [Ch1–Ya], we have introduced the following important concepts and notations. Let Q be the space of quadratic forms in n variables, i.e., real vector space spanned by $x_i x_j$, with $1 \le i \le j \le n$. Let $X = (x_1, \ldots, x_n)^T$. For any quadratic form $p \in Q$, there exists a symmetric matrix A such that $p(x) = X^T A X$. The rank of the quadratic form p is denoted by $\mathrm{rk}(p)$ and is defined to be the rank of the matrix A.

DEFINITION 2.2. *A fundamental quadratic form of the estimation algebra E is an element $p_0 \in E \cap Q$ with the greatest positive rank, i.e., $\mathrm{rk}(p_0) \ge \mathrm{rk}(p)$ for any $p \in E \cap Q$. The quadratic rank of the estimation algebra E is defined to be $\mathrm{rk}(p_0)$.*

After an orthogonal transformation on x, p_0 can be written as

$$p_0(x) = c_1 x_1^2 + c_2 x_2^2 + \cdots + c_k x_k^2, \tag{2.3}$$

where $c_i \neq 0$, $1 \le i \le k$, and k is the quadratic rank of E. From $p_0(x)$, we can construct a sequence of quadratic forms in $E \cap Q$ as follows:

$$q_0 = p_0,$$
$$q_j = \big[[L_0, q_{j-1}], q_0\big] = \sum_{i=1}^{k} 4^j c_i^{j+1} x_i^2.$$

In view of the invertibility of the Vandermonde matrix, we can assume that

$$p_0(x) = x_1^2 + x_2^2 + \cdots + x_k^2 \in E. \tag{2.4}$$

The following results were proven in [Ch1–Ya].

LEMMA 2.3. *Let k be the quadratic rank of the estimation algebra E with fundamental quadratic form $p_0(x) = x_1^2 + \cdots + x_k^2$. Then $p(x)$ is independent of $x_{k+1}, x_{k+2}, \ldots, x_n$ for any quadratic form $p(x)$ in E.*

Let $p_1(x)$ be a quadratic form in E with least positive rank, i.e., $\mathrm{rk}(p_1) \le \mathrm{rk}(q)$ for any $q(x) \in E \cap Q$. After an orthogonal transform on X which fixes $x_{k+1}, \ldots, x_n$ (i.e., an orthogonal transform on $x_1, x_2, \ldots, x_k$) and the Vandermonde matrix procedure as before, we can assume

$$p_1(x) = \sum_{i=1}^{k_1} x_i^2 \in E, \quad 1 \le k_1 \le k. \tag{2.5}$$

Note that the orthogonal transform on $x_1, \ldots, x_k$ leaves $p_0(x)$ invariant. By definition, $p_0(x) = \sum_{i=1}^{k} x_i^2$ has the greatest positive rank and $p_1(x) = \sum_{i=1}^{k_1} x_i^2$ has the least positive rank. Define

$$S_1 = \{1, 2, \ldots, k_1\} \subseteq S = \{1, 2, \ldots, k\}, \tag{2.6}$$

$$Q_1 = \text{ real vector space spanned by } \{x_i x_j : k_1 + 1 \le i \le j \le k\} \subseteq Q. \tag{2.7}$$

If $k_1 < k$, then $Q_1 \cap E$ is a nontrivial space since $p_1(x) - p_0(x) \in E \cap Q_1$. In a procedure similar to that above, there exists

$$p_2(x) = \sum_{i=k_1+1}^{k_2} x_i^2 \in E \cap Q_1 \tag{2.8}$$

with the least positive rank in $E \cap Q_1$. By induction, we can construct a series of S_i, Q_i, and $p_i(x)$ such that

$$S_i = \{k_{i-1} + 1, \ldots, k_i\}, \quad k_0 = 0, \quad k_i \le k, \tag{2.9}$$

$$Q_i = \text{ linear span } \{x_i x_j : k_i + 1 \le i \le j \le k\}, \tag{2.10}$$

$$p_i(x) = \sum_{j=k_{i-1}+1}^{k_i} x_j^2 = \sum_{j \in S_i} x_j^2, \quad i > 0, \tag{2.11}$$

and $p_i(x)$ has the least rank in $E \cap Q_{i-1}$.

LEMMA 2.4. *If $p(x) \in E \cap Q$, then*

$$p(0, \ldots, 0, x_{k_{i-1}+1}, \ldots, x_{k_i}, 0, \ldots, 0) = \lambda p_i(x) \quad \textit{for } i > 0,$$
$$p(x_1, \ldots, x_{k_{i-1}}, 0, \ldots, 0, x_{k_i+1}, \ldots, x_k) \in E \quad \textit{for } i > 0.$$

PROPOSITION 2.5. *Suppose that $p(x)$ is a quadratic form in E of the following form, which depends on $\{x_i \ i \in S_{\ell_1} \cup S_{\ell_2}\}$:*

$$p(x) = \left(X_{\ell_1}^T, X_{\ell_2}^T\right) \begin{pmatrix} 0 & A \\ A^T & 0 \end{pmatrix} \begin{pmatrix} X_{\ell_1} \\ X_{\ell_2} \end{pmatrix},$$

where $X_i = \left(x_{k_{i-1}+1}, \ldots, x_{k_i}\right)^T$, i.e., $p(x) = \sum_{i \in S_{\ell_1}} \sum_{j \in S_{\ell_2}} 2a_{ij} x_i x_j$. Suppose that $\ell_1 < \ell_2$. Then $|S_{\ell_1}| = |S_{\ell_2}|$ and $A = bT$, where b is a constant and T is an orthogonal matrix.

THEOREM 2.6. *Suppose that E is a finite-dimensional estimation algebra of maximal rank. Then $\omega_{ij} = \frac{\partial f_j}{\partial x_i} - \frac{\partial f_i}{\partial x_j}$, $1 \le i, j \le n$, are polynomials of degree at most one.*

Notation. From now on we shall write $\omega_{ij} = \beta_{ij} + \gamma_{ij}$, where β_{ij} is the linear part of ω_{ij} while γ_{ij} is the constant part of ω_{ij}.

PROPOSITION 2.7. *Suppose that E is a finite-dimensional estimation algebra of maximal rank. With the same notation as before, if $\ell_1 \neq \ell_2$ and $i \in S_{\ell_1}$, $j \in S_{\ell_2}$, then $\beta_{ij} = 0$; i.e., ω_{ij} is a constant.*

THEOREM 2.8. *Suppose that E is a finite-dimensional estimation algebra of maximal rank. Let k be the quadratic rank of E. With the same notation as before, then*

(i) *for $j \in S_\ell$*

$$\begin{pmatrix} \beta_{jk_{\ell-1}+1} \\ \vdots \\ \beta_{jk_\ell} \end{pmatrix} = A_1^{j,\ell} X_\ell \qquad \text{with} \quad A_1^{j,\ell} = -(A_1^{j,\ell})^T,$$

where $X_\ell = \left(x_{k_{\ell-1}+1}, \dots, x_{k_\ell}\right)$ and $A_1^{j,\ell}$ is a $(k_\ell - k_{\ell-1}) \times (k_\ell - k_{\ell-1})$ matrix;

(ii) *for $j > k$*

$$\begin{pmatrix} \beta_{jk_{\ell-1}+1} \\ \vdots \\ \beta_{jk_\ell} \end{pmatrix} = \lambda_{j,\ell} X_\ell + A_2^{j,\ell} \widetilde{X_\ell},$$

where $\widetilde{X_\ell}$ denote the complementary variable vector of X_ℓ in $(x_1, \dots, x_k)^T$, i.e., $\widetilde{X_\ell} = \left(x_1, \dots, x_{k_{\ell-1}}, x_{k_\ell+1}, \dots, x_k\right)^T$, and $A_2^{j,\ell}$ is a $k_\ell \times (k - k_\ell)$ matrix.

THEOREM 2.9. *Suppose that E is a finite-dimensional estimation algebra of maximal rank. With the same notation as before, then $A_1^{j,\ell} = 0$ in* (i) *of Theorem* 2.8. *This means that $\beta_{ij} = 0$ for $i, j \in S_\ell = \left\{k_{\ell-1} + 1, \dots, k_\ell\right\}$; i.e., $\omega_{ij} =$* constant *for $i, j \in S_\ell$.*

PROPOSITION 2.10. *Suppose that E is a finite-dimensional estimation algebra of maximal rank. Then*

(i) *ω_{ij} is a degree-one polynomial in $x_1, \dots, x_k$ for $1 \le i \le k$ or $1 \le j \le k$;*

(ii) *ω_{ij} is a degree-one polynomial in $x_{k+1}, \dots, x_n$ for $k + 1 \le i, j \le n$.*

It follows from Theorem 2.6, Proposition 2.7, Theorem 2.9, and Proposition 2.10 that we have the following theorem.

THEOREM 2.11. *Let E be a finite-dimensional estimation algebra of maximal rank and k be the quadratic rank of E. Then all the entries $\omega_{ij} = \frac{\partial f_j}{\partial x_i} - \frac{\partial f_i}{\partial x_j}$ of Ω are degree-one polynomials. In fact, for $1 \le i, j \le k$, ω_{ij} are constants; for $1 \le i \le k$ or $1 \le j \le k$, ω_{ij} are degree-one polynomials in $x_1, \dots, x_k$; and for $k + 1 \le i, j \le n$, ω_{ij} are degree-one polynomials in $x_{k+1}, \dots, x_n$.*

For the sake of convenience to the readers, we also provide the following lemma without proof. The proof can be found in [Ya].

LEMMA 2.12. (i) *$[XY, Z] = X[Y, Z] + [X, Z]Y$, where X, Y, and Z are differential operators.*

(ii) *$[gD_i, h] = g\frac{\partial h}{\partial x_i}$, where $D_i = \frac{\partial}{\partial x_i} - f_i$ and g and h are functions defined on $\mathbb{R}^n$.*

(iii) *$[gD_i, hD_j] = -gh\omega_{ij} + g\frac{\partial h}{\partial x_i} D_j - h\frac{\partial g}{\partial x_j} D_i$, where $\omega_{ji} = [D_i, D_j] = \frac{\partial f_i}{\partial x_j} - \frac{\partial f_j}{\partial x_i}$.*

(iv) *$[gD_i^2, h] = 2g\frac{\partial h}{\partial x_i} D_i + g\frac{\partial^2 h}{\partial x_i^2}$.*

(v) $[D_i^2, hD_j] = 2\frac{\partial h}{\partial x_i}D_iD_j - 2h\omega_{ij}D_i + \frac{\partial^2 h}{\partial x_i^2}D_j - h\frac{\partial \omega_{ij}}{\partial x_i}$.

(vi) $[D_i^2, D_j^2] = 4\omega_{ji}D_jD_i + 2\frac{\partial \omega_{ji}}{\partial x_j}D_i + 2\frac{\partial \omega_{ji}}{\partial x_i}D_j + \frac{\partial^2 \omega_{ji}}{\partial x_i \partial x_j} + 2{\omega_{ji}}^2$.

(vii) $[D_k^2, hD_iD_j] = 2\frac{\partial h}{\partial x_k}D_kD_iD_j + 2h\omega_{jk}D_iD_k + 2h\omega_{ik}D_kD_j + \frac{\partial^2 h}{\partial x_k^2}D_iD_j + 2h\frac{\partial \omega_{jk}}{\partial x_i}D_k + h\frac{\partial \omega_{jk}}{\partial x_k}D_i + h\frac{\partial \omega_{ik}}{\partial x_k}D_j + h\frac{\partial^2 \omega_{jk}}{\partial x_i \partial x_k}$.

(viii) $[gD_iD_j, hD_k] = g\frac{\partial h}{\partial x_j}D_iD_k + g\frac{\partial h}{\partial x_i}D_jD_k + gh\omega_{kj}D_i + gh\omega_{ki}D_j + g\frac{\partial^2 h}{\partial x_i \partial x_j}D_k + gh\frac{\partial \omega_{kj}}{\partial x_i} - h\frac{\partial g}{\partial x_k}D_iD_j$.

3. Structure of second-order operators and special linear structure of Ω. Building on our previous results in [Ch1–Ya], in this section, we shall prove that Ω has very special linear structure. It is exactly this result which allows us to prove Mitter conjecture for maximal rank finite-dimensional estimation algebra. To begin with, we need some results on second-order operator in E.

Consider the polynomial algebra $C^\infty(\mathbb{R}^n)[D_1, \dots, D_n]$ in variables $D_1, \dots, D_n$ with coefficients in $C^\infty(\mathbb{R}^n)$ modulo the relations $D_iD_j = D_jD_i + \omega_{ji}$ and $D_ia = aD_i + \frac{\partial a}{\partial x_i}$, where $C^\infty(\mathbb{R}^n)$ is the ring of all C^∞ functions on $\mathbb{R}^n$ and a is a C^∞ function. Every element $A \in C^\infty(\mathbb{R}^n)[D_1, \dots, D_n]$ has a representation in the following:

$$A = \sum a_{i_1 \cdots i_n}(x) D_1^{i_1} \cdots D_n^{i_n}, \tag{3.1}$$

i.e., a polynomial in D_i with coefficients in $C^\infty(\mathbb{R}^n)$. Since $D_iD_j = D_jD_i + \omega_{ji}$, $C^\infty(\mathbb{R}^n)[D_1, \dots, D_n]$ is not a commutative algebra. It is clear that every element of E has a representation of (3.1). For any $A \in E$, let P_A be the principal part of A, i.e., the highest homogeneous part of A in $D_1, \dots, D_n$. For example, for $L_0 \in E$, $P_{L_0} = \frac{1}{2}(D_1^2 + \cdots + D_n^2)$.

LEMMA 3.1. *If $\frac{\partial a}{\partial x_j}(x) \neq 0$, then the principal part of $\left[D_j^\ell, a(x)D_1^{i_1} \cdots D_n^{i_n}\right]$ is given by*

$$P_{\left[D_j^\ell, a(x)D_1^{i_1} \cdots D_n^{i_n}\right]} = \ell \frac{\partial a}{\partial x_j}(x) D_1^{i_1} \cdots D_j^{i_j + \ell - 1} \cdots D_n^{i_n}. \tag{3.2}$$

Proof. If $\ell = 1$, then (3.2) is trivial. Suppose that (3.2) is true for $\ell - 1$; then

$$\begin{aligned} &\left[D_j^\ell, a(x)D_1^{i_1} \cdots D_n^{i_n}\right] \\ &= D_j\left[D_j^{\ell-1}, a(x)D_1^{i_1} \cdots D_n^{i_n}\right] + \left[D_j, a(x)D_1^{i_1} \cdots D_n^{i_n}\right]D_j^{\ell-1}. \end{aligned} \tag{3.3}$$

Hence

$$\begin{aligned} P_{\left[D_j^\ell, a(x)D_1^{i_1} \cdots D_n^{i_n}\right]} &= P_{D_j\left[D_j^{\ell-1}, a(x)D_1^{i_1} \cdots D_n^{i_n}\right]} + P_{\left[D_j, a(x)D_1^{i_1} \cdots D_n^{i_n}\right]D_j^{\ell-1}} \\ &= (\ell - 1)\frac{\partial a}{\partial x_j}(x) D_1^{i_1} \cdots D_j^{i_j + \ell - 1} \cdots D_n^{i_n} \\ &\quad + \frac{\partial a}{\partial x_j}(x) D_1^{i_1} \cdots D_j^{i_j + \ell - 1} \cdots D_n^{i_n} \\ &= \ell \frac{\partial a}{\partial x_j}(x) D_1^{i_1} \cdots D_j^{i_j + \ell - 1} \cdots D_n^{i_n}. \end{aligned}$$

□

PROPOSITION 3.2. *If the principal part of A is of the form*

$$P_A = \sum a_{i_1 \cdots i_n}(x) D_1^{i_1} \cdots D_n^{i_n},$$

where for some $(i_1, \dots, i_n)$ *and* j, $\frac{\partial a_{i_1 \cdots i_n}}{\partial x_j}(x) \neq 0$, *then the principal part of* $[L_0, A]$ *is given by*

$$P_{[L_0,A]} = \sum_j \sum_{i_1 \cdots i_n} \frac{\partial a_{i_1 \cdots i_n}}{\partial x_j}(x) D_1^{i_1} \cdots D_j^{i_j+1} \cdots D_n^{i_n}.$$

Proof. The proposition follows immediately when Lemma 3.1 is applied repeatedly. □

THEOREM 3.3. *Let* E *be a finite-dimensional estimation algebra. Suppose that* A *is a second order in* E *with principal part* $P_A = \sum_{i \leq j} a_{ij}(x) D_i D_j$. *Then* $\frac{\partial a_{ii}}{\partial x_i}(x) = 0$.

Proof. Without loss of generality, we can assume $i = 1$. a_{11} must be a polynomial; otherwise E would be infinite dimensional. If $\frac{\partial a_{11}}{\partial x_1}(x) \neq 0$, then there exists a positive integer ℓ such that

$$\frac{\partial^{\ell+1} a_{11}}{\partial x_1^{\ell+1}}(x) = 0 \quad \text{and} \quad \frac{\partial^{\ell} a_{11}}{\partial x_1^{\ell}}(x) \neq 0.$$

This is true because $a_{11}(x)$ is a polynomial in view of a result of [Wo]. By Proposition 3.2, we have

$$P_{[L_0,A]} = \frac{\partial a_{11}}{\partial x_1}(x) D_1^3 + \text{lower-order term in } D_1.$$

Hence if we let $Ad_{L_0} A = [L_0, A]$ and $Ad_{L_0}^m A = \big[L_0, Ad_{L_0}^{m-1} A\big]$, then

$$P_{Ad_{L_0}^{\ell} A} = \frac{\partial^{\ell} a_{11}}{\partial x_1^{\ell}}(x) D_1^{\ell+2} + \text{lower-order term in } D_1.$$

Let $B = Ad_{L_0}^{\ell-1} A$ (for $\ell = 1$, take $B = A$). Then

$$\begin{aligned} P_{Ad_B^s L_0} =& (-1)^s (\ell+1)(\ell+2)(2\ell+2)(3\ell+2) \cdots ((s-1)\ell+2) \Big(\frac{\partial^{\ell} a_{11}}{\partial x_1^{\ell}}(x) \Big)^s D_1^{s\ell+2} \\ &+ \text{lower-order term in } D_1. \end{aligned}$$

We have produced an infinite sequence of independent elements $\big\{ Ad_B^s L_0 : s = 1, 2, \dots \big\}$ in E. This contradicts our assumption that E is finite dimensional. □

THEOREM 3.4. *Let* E *be a finite-dimensional estimation algebra and* A *be an element in* E *with principal part of* $P_A = \sum_{i \leq j} a_{ij}(x) D_i D_j$. *Suppose that* $\frac{\partial a_{ii}}{\partial x_j}(x) = \frac{\partial a_{jj}}{\partial x_i}(x) = 0$. *Then* $\frac{\partial a_{ij}}{\partial x_i}(x) = \frac{\partial a_{ij}}{\partial x_j}(x) = 0$.

Proof. Without loss of generality, we shall assume $i = 1$ and $j = 2$. Suppose to the contrary that either $\frac{\partial a_{12}}{\partial x_1}(x) \neq 0$ or $\frac{\partial a_{12}}{\partial x_2}(x) \neq 0$. Then

$$P_{[L_0,A]} = \frac{\partial a_{12}}{\partial x_1}(x) D_1^2 D_2 + \frac{\partial a_{12}}{\partial x_2}(x) D_1 D_2^2 + \text{ terms in degree } D_1, D_2 \text{ lower than 3.}$$

If $\frac{\partial a_{12}}{\partial x_1}(x) \neq 0$, then there exists a positive integer ℓ such that

$$\frac{\partial^{\ell} a_{12}}{\partial x_1^{\ell}}(x) \neq 0 \quad \text{and} \quad \frac{\partial^{\ell+1} a_{12}}{\partial x_1^{\ell+1}}(x) = 0.$$

Let $B = Ad_{L_0}^{\ell-1} A \in E$. Then

$$P_B = \frac{\partial^{\ell-1} a_{12}}{\partial x_1^{\ell-1}}(x) D_1^{\ell} D_2 + \text{other terms.}$$

Hence

$$P_{Ad_B^s L_0} = a\Big(\frac{\partial^\ell a_{12}}{\partial x_1^\ell}(x)\Big)^s D_1^{\ell+s} D_2 + \text{other terms},$$

where a is a nonzero constant. So we have produced an infinite sequence $\{Ad_B^s L_0 \in E : s = 1, 2, \dots\}$ of linearly independent elements in E. This contradicts to our hypothesis that $\dim E < \infty$. Therefore we conclude that $\frac{\partial a_{12}}{\partial x_1}(x) = 0$. Similarly, we can prove $\frac{\partial a_{12}}{\partial x_2}(x) = 0$. □

We are now ready to prove the special linear structure of Ω.

THEOREM 3.5. *Suppose that E is a finite-dimensional estimation algebra of maximal rank. Let k be the quadratic rank of E. With the same notation as before let β_{ij} be the linear part of ω_{ij}. Then*

$$\begin{pmatrix} \beta_{jk_{\ell-1}+1} \\ \vdots \\ \beta_{jk_\ell} \end{pmatrix} = A_2^{j,\ell} \widetilde{X_\ell} \qquad \text{for } \ j > k,$$

where $\widetilde{X_\ell} = \big(x_1, \dots, x_{k_{\ell-1}}, x_{k_\ell+1}, \dots, x_k\big)^T$ and $A_2^{j,\ell}$ is a $k_\ell \times (k - k_\ell)$ matrix.

Proof. In view of part (ii) of Theorem 2.8, we have, for $j > k$,

$$\begin{pmatrix} \beta_{jk_{\ell-1}+1} \\ \vdots \\ \beta_{jk_\ell} \end{pmatrix} = \lambda_{j,\ell} X_\ell + A_2^{j,\ell} \widetilde{X_\ell},$$

where $X_\ell = \big(x_{k_{\ell-1}+1}, \dots, x_{k_\ell}\big)^T$. To prove the theorem, we need to prove $\lambda_{j,\ell} = 0$. For this purpose, we need only to prove that $\omega_{mk_{\ell-1}+1}$ does not depend on $x_{k_{\ell-1}+1}$. Since $p_\ell(x) = x_{k_{\ell-1}+1}^2 + \cdots + x_{k_\ell}^2 = \sum_{j \in S_\ell} x_j^2 \in E$, we have

$$\begin{aligned}
\big[L_0, p_\ell\big] \in E &\Longrightarrow \sum_{j \in S_\ell} x_j D_j \in E, \\
\Big[L_0, \sum_{j \in S_\ell} x_j D_j\Big] &= \frac{1}{2} \sum_{i=1}^n \sum_{j \in S_\ell} \big[D_i^2, x_j D_j\big] - \frac{1}{2} E_{k_\ell}(\eta) \\
&= \sum_{i=1}^n \sum_{j \in S_\ell} \left(\delta_{ij} D_i D_j - x_j \omega_{ij} D_i - \frac{1}{2} x_j \frac{\partial \omega_{ij}}{\partial x_i} \right) - \frac{1}{2} E_{k_\ell}(\eta),
\end{aligned}$$

where $E_{k_\ell} = \sum_{j \in S_\ell} x_j \frac{\partial}{\partial x_j}$. Since E is of maximal rank and ω_{ij} is an affine function, we deduce that

$$Z_1 = \sum_{j \in S_\ell} D_j^2 - \sum_{i=1}^n \sum_{j \in S_\ell} x_j \omega_{ij} D_i - \frac{1}{2} E_{k_\ell}(\eta) \in E.$$

It follows from Lemma 2.12 that

$$\begin{aligned}
P_{[L_0, Z_1]} &= P_{\left[\frac{1}{2}\sum_{i=1}^n D_i^2, \sum_{j \in S_\ell} D_j^2\right]} + P_{\left[\frac{1}{2}\sum_{i=1}^n D_i^2, -\sum_{r=1}^n \sum_{j \in S_\ell} x_j \omega_{rj} D_r\right]} \\
&= \sum_{i=1}^n \sum_{j \in S_\ell} 2\omega_{ji} D_j D_i - \sum_{i=1}^n \sum_{r=1}^n \sum_{j \in S_\ell} \frac{\partial (x_j \omega_{rj})}{\partial x_i} D_i D_r.
\end{aligned}$$

Since $\omega_{k_{\ell-1}+1,k_{\ell-1}+1}=0$, the coefficient of $D^2_{k_{\ell-1}+1}$ in $P_{[L_0,Z_1]}$ is

$$-\sum_{j\in S_\ell}\frac{\partial(x_j\omega_{k_{\ell-1}+1,j})}{\partial x_{k_{\ell-1}+1}}=0$$

in view of Theorem 2.9. Similarly, for $m>k=$ quadratic rank of E, we can see that the coefficient of D^2_m is

$$-\sum_{j\in S_\ell}\frac{\partial(x_j\omega_{mj})}{\partial x_m}=0$$

in view of Proposition 2.10. Now the coefficient of $D_{k_{\ell-1}+1}D_m$ is

$$\begin{aligned}
&2\omega_{k_{\ell-1}+1,m}-\sum_{j\in S_\ell}\Big[\frac{\partial(x_j\omega_{k_{\ell-1}+1,j})}{\partial x_m}+\frac{\partial(x_j\omega_{mj})}{\partial x_{k_{\ell-1}+1}}\Big]\\
&=2\omega_{k_{\ell-1}+1,m}-\sum_{j\in S_\ell}\frac{\partial(x_j\omega_{mj})}{\partial x_{k_{\ell-1}+1}}\quad\text{in view of Theorem 2.11}\\
&=3\omega_{k_{\ell-1}+1,m}-\sum_{j\in S_\ell}x_j\frac{\partial\omega_{mj}}{\partial x_{k_{\ell-1}+1}}\\
&=3\omega_{k_{\ell-1}+1,m}-x_{k_{\ell-1}+1}\frac{\partial\omega_{mk_{\ell-1}+1}}{\partial x_{k_{\ell-1}+1}}
\end{aligned}$$

in view of part (ii) of Theorem 2.8. By Theorem 3.4, we know that the coefficient of $D_{k_{\ell-1}+1}D_m$ is independent of $x_{k_{\ell-1}+1}$. This simply means that

$$3\omega_{mk_{\ell-1}+1}+x_{k_{\ell-1}+1}\frac{\partial\omega_{mk_{\ell-1}+1}}{\partial x_{k_{\ell-1}+1}}$$

is independent of $x_{k_{\ell-1}+1}$. Hence $\omega_{mk_{\ell-1}+1}$ does not depend on $x_{k_{\ell-1}+1}$. □

THEOREM 3.6. *With the same hypothesis and notation as in Theorem* 3.5*, then* $A_2^{j,\ell}=0$*, i.e.,* $\beta_{ij}=0$ *for* $i\le k$ *(quadratic rank of* E*)* $<j$.

Proof. Since $[[L_0,p_\ell(x)],D_j]\in E$, we have $\sum_{i\in S_\ell}x_i\beta_{ji}\in E$. Theorem 3.5 says that

$$\begin{pmatrix}\beta_{jk_{\ell-1}+1}\\ \vdots\\ \beta_{jk_\ell}\end{pmatrix}=A_2^{j,\ell}\widetilde{X_\ell},$$

where $\widetilde{X_\ell}=\big(x_1,\dots,x_{k_{\ell-1}},x_{k_\ell+1},\dots,x_k\big)^T$. In view of Lemma 2.3, $X_\ell^TA_2^{j,\ell}\widetilde{X_\ell}=\sum_{i\in S_\ell}x_i\beta_{ji}\in E$ is independent of the $x_{k+1},\dots,x_n$ variable. Hence

$$A_2^{j,\ell}=\big(B_1^{j,\ell},B_2^{j,\ell},\dots\big),$$

where $B_1^{j,\ell},B_2^{j,\ell},\dots$ are constant multiples of some orthogonal matrices by Proposition 2.5. So we have

$$\begin{pmatrix}\beta_{jk_{\ell-1}+1}\\ \vdots\\ \beta_{jk_\ell}\end{pmatrix}=A_2^{j,\ell}\widetilde{X_\ell}=B_1^{j,\ell}X_1+\cdots+B_{\ell-1}^{j,\ell}X_{\ell-1}+B_{\ell+1}^{j,\ell}X_{\ell+1}+\cdots.$$

In order to make clear what our strategy is, in the following we shall list several of these equations explicitly:

$$\begin{pmatrix}\beta_{j1}\\ \vdots\\ \beta_{jk_1}\end{pmatrix} = A_2^{j,1}\widetilde{X_1} = B_2^{j,1}X_2 + B_3^{j,1}X_3 + B_4^{j,1}X_4 + B_5^{j,1}X_5 + \cdots,$$

$$\begin{pmatrix}\beta_{jk_1+1}\\ \vdots\\ \beta_{jk_2}\end{pmatrix} = A_2^{j,2}\widetilde{X_2} = B_1^{j,2}X_1 + B_3^{j,2}X_3 + B_4^{j,2}X_4 + B_5^{j,2}X_5 + \cdots,$$

$$\begin{pmatrix}\beta_{jk_2+1}\\ \vdots\\ \beta_{jk_3}\end{pmatrix} = A_2^{j,3}\widetilde{X_3} = B_1^{j,3}X_1 + B_2^{j,3}X_2 + B_4^{j,3}X_4 + B_5^{j,3}X_5 + \cdots,$$

$$\begin{pmatrix}\beta_{jk_3+1}\\ \vdots\\ \beta_{jk_4}\end{pmatrix} = A_2^{j,4}\widetilde{X_4} = B_1^{j,4}X_1 + B_2^{j,4}X_2 + B_3^{j,4}X_3 + B_5^{j,4}X_5 + \cdots,$$

$$\vdots$$

We first claim that $B_m^{j,\ell} = \big(B_\ell^{j,m}\big)^T$. To see this, we observe that from the cyclic relation $\frac{\partial\omega_{j\ell}}{\partial x_m} + \frac{\partial\omega_{\ell m}}{\partial x_j} + \frac{\partial\omega_{mj}}{\partial x_\ell} = 0$ we deduce that

$$\frac{\partial\beta_{j\ell}}{\partial x_m} + \frac{\partial\beta_{\ell m}}{\partial x_j} + \frac{\partial\beta_{mj}}{\partial x_\ell} = 0.$$

If we take $m, \ell \le k =$ quadratic rank of $E < j$, then $\frac{\partial\beta_{\ell m}}{\partial x_j} = 0$ in view of Theorem 2.11. Therefore the above equation implies

$$\frac{\partial\beta_{j\ell}}{\partial x_m} = \frac{\partial\beta_{jm}}{\partial x_\ell},$$

from which we deduce easily that $B_m^{j,\ell} = \big(B_\ell^{j,m}\big)^T$.

Now we prove $B_1^{j,\ell} = 0$ for $\ell \ge 2$. Since $\sum_{i\in S_\ell} x_i\beta_{ji} = X_\ell^T A_2^{j,\ell}\widetilde{X_\ell} = X_\ell^T B_1^{j,\ell}X_1 + \cdots + X_\ell^T B_{\ell-1}^{j,\ell}X_{\ell-1} + X_\ell B_{\ell+1}^{j,\ell}X_{\ell+1} + \cdots \in E$, we conclude from Lemma 2.4 that $X_\ell^T B_1^{j,\ell}X_1 \in E$. Let $B_1^{j\ell} = \big(b_{ir}^{j\ell}\big), k_{\ell-1}+1 \le i \le k_\ell, 1 \le r \le k_1$. Then

$$\begin{aligned}\big[L_0, X_\ell^T B_1^{j\ell}X_1\big] &= \Big[\frac{1}{2}\sum_{m=1}^n D_m^2, \sum_{i\in S_\ell}\sum_{r\in S_1} x_i b_{ir}^{j\ell}x_r\Big]\\ &= \sum_{m=1}^n\sum_{i\in S_\ell}\sum_{r\in S_1}\big(b_{ir}^{j\ell}x_r\delta_{im}D_m + b_{ir}^{j\ell}x_i\delta_{mr}D_m + b_{ir}^{j\ell}\delta_{im}\delta_{rm}\big)\\ &= \sum_{i\in S_\ell}\sum_{r\in S_1} b_{ir}^{j\ell}x_r D_i + \sum_{i\in S_\ell}\sum_{r\in S_1} b_{ir}^{j\ell}x_i D_r\\ &= \sum_{i\in S_\ell}\sum_{r\in S_1} b_{ir}^{j\ell}\big(x_r D_i + x_i D_r\big) \in E,\end{aligned}$$

$$
\begin{aligned}
W_1 :=& \Big[L_0, \big[L_0, X_\ell^T B_1^{j\ell} X_1\big]\Big] \\
=& \Big[\frac{1}{2}\sum_{m=1}^{n} D_m^2 - \frac{1}{2}\eta, \sum_{i\in S_\ell}\sum_{r\in S_1} b_{ir}^{j\ell}\big(x_r D_i + x_i D_r\big)\Big] \\
=& \frac{1}{2}\sum_{m=1}^{n}\sum_{i\in S_\ell}\sum_{r\in S_1} b_{ir}^{j\ell}\big[D_m^2, x_r D_i + x_i D_r\big] + \text{ function} \\
=& \sum_{m=1}^{n}\sum_{i\in S_\ell}\sum_{r\in S_1} b_{ir}^{j\ell}\big(\delta_{mr} D_m D_i - x_r\omega_{mi} D_m + \delta_{mi} D_m D_r - x_i\omega_{mr} D_m\big) \\
&+ \text{ function} \\
=& 2\sum_{i\in S_\ell}\sum_{r\in S_1} b_{ir}^{j\ell} D_r D_i - \sum_{m=1}^{n}\sum_{i\in S_\ell}\sum_{r\in S_1} b_{ir}^{j\ell}\big(x_r\omega_{mi} + x_i\omega_{mr}\big) D_m \\
&+ \text{ function,} \\
[L_0, W_1] =& \Big[\sum_{m=1}^{n} D_m^2, \sum_{i\in S_\ell}\sum_{r\in S_1} b_{ir}^{j\ell} D_i D_r\Big] \\
&- \Big[\frac{1}{2}\sum_{v=1}^{m} D_v^2, \sum_{m=1}^{n}\sum_{i\in S_\ell}\sum_{r\in S_1} b_{ir}^{j\ell}\big(x_r\omega_{mi} + x_i\omega_{mr}\big) D_m\Big] \\
&+ \text{ first-order term} \\
=& 2\sum_{m=1}^{n}\sum_{i\in S_\ell}\sum_{r\in S_1} b_{ir}^{j\ell}\big(\omega_{rm} D_i D_m + \omega_{im} D_m D_r\big) \\
&- \sum_{v=1}^{n}\sum_{m=1}^{n}\sum_{i\in S_\ell}\sum_{r\in S_1} b_{ir}^{j\ell}\frac{\partial(x_r\omega_{mi} + x_i\omega_{mr})}{\partial x_v} D_v D_m \\
&+ \text{ first-order term.}
\end{aligned}
$$

The coefficient of D_1^2 in $[L_0, W_1]$ is

$$
2\sum_{i\in S_\ell} b_{ir}^{j\ell}\omega_{i1} - \sum_{i\in S_\ell}\sum_{r\in S_1} b_{ir}^{j\ell}\frac{\partial(x_r\omega_{1i} + x_i\omega_{1r})}{\partial x_1} = \sum_{i\in S_\ell} b_{i1}^{j\ell}\omega_{i1},
$$

which is a constant in view of Theorem 2.11.

On the other hand, the coefficient of D_j^2 in $[L_0, W_1]$ for $j > k =$ quadratic rank of E is

$$
-\sum_{i\in S_\ell}\sum_{r\in S_1} b_{ir}^{j\ell}\frac{\partial(x_r\omega_{ji} + x_i\omega_{jr})}{\partial x_j},
$$

which is a zero in view of Theorem 2.11.

We deduce from Theorem 3.4 that the coefficient of $D_1 D_j$ in $[L_0, W_1]$ is independent of x_1 and x_j variables. However, the coefficient of $D_1 D_j$ in $[L_0, W_1]$ is given by

$$
\begin{aligned}
&2\sum_{i\in S_\ell} b_{i1}^{j\ell}\omega_{ij} - \sum_{i\in S_\ell}\sum_{r\in S_1} b_{ir}^{j\ell}\Big(\frac{\partial(x_r\omega_{ji}+x_i\omega_{jr})}{\partial x_1}+\frac{\partial(x_r\omega_{1i}+x_i\omega_{1r})}{\partial x_j}\Big)\\
&= 2\sum_{i\in S_\ell} b_{i1}^{j\ell}\omega_{ij} - \sum_{i\in S_\ell}\sum_{r\in S_1} b_{ir}^{j\ell}\Big(\delta_{1r}\omega_{ji}+x_r\frac{\partial\omega_{ji}}{\partial x_1}+x_i\frac{\partial\omega_{jr}}{\partial x_1}\Big) \text{ in view of Theorem 2.11}\\
&= 2\sum_{i\in S_\ell} b_{i1}^{j\ell}\omega_{ij} - \sum_{i\in S_\ell} b_{i1}^{j\ell}\omega_{ji} - \sum_{i\in S_\ell}\sum_{r\in S_1} b_{ir}^{j\ell}\Big(x_r\frac{\partial\omega_{ji}}{\partial x_1}+x_i\frac{\partial\omega_{jr}}{\partial x_1}\Big)\\
&= 3\sum_{i\in S_\ell} b_{i1}^{j\ell}\omega_{ij} - \sum_{i\in S_\ell}\sum_{r\in S_1} b_{ir}^{j\ell}x_r\frac{\partial\omega_{ji}}{\partial x_1} \quad \text{by Theorem 3.5}\\
&= -3\sum_{i\in S_\ell} b_{i1}^{j\ell}\omega_{ji} - \sum_{i\in S_\ell} b_{i1}^{j\ell}x_1\frac{\partial\omega_{ji}}{\partial x_1} + \text{other terms not involving } x_1\\
&= -3\Big(b_{k_{\ell-1}+1,1}^{j\ell}\omega_{jk_{\ell-1}+1}+b_{k_{\ell-1}+2,1}^{j\ell}\omega_{jk_{\ell-1}+2}+\cdots+b_{k_\ell,1}^{j\ell}\omega_{jk_\ell}\Big)\\
&\quad -\Big(b_{k_{\ell-1}+1,1}^{j\ell}x_1\frac{\partial\omega_{jk_{\ell-1}+1}}{\partial x_1}+b_{k_{\ell-1}+2,1}^{j\ell}x_1\frac{\partial\omega_{jk_{\ell-1}+2}}{\partial x_1}+\cdots+b_{k_\ell,1}^{j\ell}x_1\frac{\partial\omega_{jk_\ell}}{\partial x_1}\Big)\\
&\quad + \text{ other terms not involving } x_1\\
&= -3\Big\{\Big[\Big(b_{k_{\ell-1}+1,1}^{j\ell}\Big)^2x_1+\cdots\Big]+\Big[\Big(b_{k_{\ell-1}+2,1}^{j\ell}\Big)^2x_1+\cdots\Big]+\cdots+\Big[\Big(b_{k_\ell,1}^{j\ell}\Big)^2x_1+\cdots\Big]\Big\}\\
&\quad -\Big[\Big(b_{k_{\ell-1}+1,1}^{j\ell}\Big)^2x_1+\Big(b_{k_{\ell-1}+2,1}^{j\ell}\Big)^2x_1+\cdots+\Big(b_{k_\ell,1}^{j\ell}\Big)^2x_1\Big]\\
&\quad + \text{ other terms not involving } x_1\\
&= -4\sum_{i\in S_\ell}\big(b_{i1}^{j\ell}\big)^2x_1 + \text{other terms not involving } x_1.
\end{aligned}
$$

Therefore we conclude that

$$\sum_{i\in S_\ell}\big(b_{i1}^{j\ell}\big)^2 = 0.$$

This implies that the first column of the matrix $B_1^{j\ell} = \big(b_{ir}^{j\ell}\big)$, $k_{\ell-1}+1 \le i \le k_\ell$, $1 \le r \le k_1$, is a zero vector. Recall that $B_1^{j\ell}$ is either zero or nonsingular. We deduce that $B_1^{j\ell} = 0$ for any $j > k =$ quadratic rank of E and $\ell \ge 2$. By $B_m^{j\ell} = \big(B_\ell^{jm}\big)^T$, we conclude further that

$$
\begin{aligned}
&\begin{pmatrix}\beta_{j1}\\ \vdots\\ \beta_{jk_1}\end{pmatrix} = A_2^{j,1}\widetilde{X_1} = 0,\\
&\begin{pmatrix}\beta_{jk_1+1}\\ \vdots\\ \beta_{jk_2}\end{pmatrix} = A_2^{j,2}\widetilde{X_2} = B_3^{j2}X_3 + B_4^{j2}X_4 + B_5^{j2}X_5 + \cdots,\\
&\begin{pmatrix}\beta_{jk_2+1}\\ \vdots\\ \beta_{jk_3}\end{pmatrix} = A_2^{j,3}\widetilde{X_3} = B_2^{j3}X_2 + B_4^{j3}X_4 + B_5^{j3}X_5 + \cdots,
\end{aligned}
$$

$$\begin{pmatrix} \beta_{jk_3+1} \\ \vdots \\ \beta_{jk_4} \end{pmatrix} = A_2^{j,4}\widetilde{X_4} = B_2^{j4}X_2 + B_3^{j4}X_3 + B_5^{j4}X_5 + \cdots,$$

$$\vdots$$

Similarly, by considering an element in E of the form $\sum_{i\in S_\ell} x_i\beta_{ji} = X_\ell^T A_2^{j,\ell}\widetilde{X_\ell}$ for $\ell \geq 3$, we can prove that $B_2^{j,\ell} = 0$ for $\ell \geq 3$. As before, we deduce $A_2^{j,2} = 0$ in view of $B_m^{j\ell} = \left(B_\ell^{jm}\right)^T$. Hence Theorem 3.6 follows easily by induction. □

As a consequence of Theorem 3.6 and Theorem 2.11, we have proved the following theorem.

THEOREM 3.7. *Let E be a finite-dimensional estimation algebra of maximal rank and k be the maximal rank of quadratic forms in E. Then* (1) *ω_{ij} are constants for $1 \leq i \leq k$ or $1 \leq j \leq k$;* (2) *ω_{ij} are affine polynomials in $x_{k+1}, \ldots, x_n$ for $k+1 \leq i, j \leq n$.*

4. Structure of η. In this section, we shall study the possible structure of η, where

$$\eta = \sum_{i=1}^{n} \frac{\partial f_i}{\partial x_i} + \sum_{i=1}^{n} f_i^2 + \sum_{i=1}^{m} h_i^2. \tag{4.1}$$

LEMMA 4.1. *Let E be a finite-dimensional estimation algebra of maximal rank. Then η is a polynomial of degree at most four.*

Proof. Since E is a finite-dimensional estimation algebra with maximal rank for any $1 \leq i \leq n$, there exists constant c_i such that $x_i + c_i$ is in E. Observe that

$$\begin{aligned} &[L_0, x_j + c_j] = D_j \in E, \\ &[D_j, x_j + c_j] = 1 \in E, \\ &[L_0, D_j] = \sum_{i=1}^{n} \left(\omega_{ji}D_i + \frac{1}{2}\frac{\partial \omega_{ji}}{\partial x_i}\right) + \frac{1}{2}\frac{\partial \eta}{\partial x_j} \in E. \end{aligned}$$

Since ω_{ji}, $1 \leq i, j \leq n$, are polynomial of degree at most one, we deduce that

$$Y_j := \sum_{i=1}^{n} \omega_{ji}D_i + \frac{1}{2}\frac{\partial \eta}{\partial x_j} \in E. \tag{4.2}$$

As

$$[Y_j, D_m] = \sum_{i=1}^{n} \left(\omega_{ji}\omega_{mi} - \frac{\partial \omega_{ji}}{\partial x_m}D_i\right) - \frac{1}{2}\frac{\partial^2 \eta}{\partial x_m \partial x_j} \in E,$$

we deduce that

$$\sum_{\ell=1}^{n} \omega_{j\ell}\omega_{\ell m} + \frac{1}{2}\frac{\partial^2 \eta}{\partial x_m \partial x_j} \in E \tag{4.3}$$

for all $1 \leq j, m \leq n$. Hence $\frac{\partial^2 \eta}{\partial x_m \partial x_j}$ for $1 \leq j, m \leq n$ are polynomials of degree at most two in view of Ocone's theorem. It follows that η is a polynomial of degree at most four. □

Notation. η_4 is denoted the homogeneous part of degree four of η.

PROPOSITION 4.2. *Let E be a finite-dimensional estimation algebra of maximal rank. Then η_4 does not contain $x_{i_1}x_{i_2}x_jx_m$ for $i_1 \le k < i_2$, where k is the quadratic rank of E.*

Proof. Let us assume that this is false. From (4.3), we have

$$\sum_{\ell=1}^{n} \omega_{i_1\ell}\omega_{\ell m} + \frac{1}{2}\frac{\partial^2 \eta}{\partial x_m \partial x_{i_1}} \in E. \tag{4.4}$$

Recall that $\omega_{i_1\ell},\ 1 \le \ell \le n$, are constants, while $\omega_{\ell m},\ 1 \le \ell \le n$, are polynomials of degree one in $x_{k+1}, \dots, x_n$ by Theorem 3.7. Equation (4.4) says that E contains a quadratic form with term of the form $x_{i_2}x_j$. This contradicts Lemma 2.3. □

The following proposition also follows immediately from Theorem 3.7.

PROPOSITION 4.3. *Let E be a finite-dimensional estimation algebra of maximal rank with quadratic rank k. If $i \le k < j$, then $\sum_{\ell=1}^{n} \omega_{i\ell}\omega_{\ell j}$ is a degree-one polynomial in $x_{k+1}, \dots, x_n$.*

THEOREM 4.4. *Let E be a finite-dimensional estimation algebra of maximal rank with quadratic rank k. Then η_4 is an homogeneous polynomial of degree four depending only on $x_{k+1}, \dots, x_n$ variables.*

Proof. Since $p_0(x) = x_1^2 + \cdots + x_k^2 \in E$ by (2.4), we have

$$[L_0, p_0] \in E \Rightarrow \widetilde{E_k} := \sum_{i=1}^{k} x_i D_i \in E.$$

Let $E_k = \sum_{i=1}^{k} x_i \frac{\partial}{\partial x_i}$. Then in view of (4.2), we have

$$\begin{aligned}
[\widetilde{E_k}, Y_j] &= \Big[\sum_{i=1}^{k} x_i D_i, \sum_{i=1}^{k} \omega_{j\ell} D_\ell + \frac{1}{2}\frac{\partial \eta}{\partial x_j}\Big] \\
&= \sum_{i=1}^{k}\sum_{\ell=1}^{n} [x_i D_i, \omega_{j\ell} D_\ell] + \frac{1}{2} E_k\left(\frac{\partial \eta}{\partial x_j}\right) \\
&= \sum_{i=1}^{k}\sum_{\ell=1}^{n} \left(-x_i\omega_{j\ell}\omega_{i\ell} + x_i \frac{\partial \omega_{j\ell}}{\partial x_i} D_\ell - \omega_{j\ell}\delta_{i\ell} D_i\right) + \frac{1}{2} E_k\left(\frac{\partial \eta}{\partial x_j}\right) \\
&= \sum_{\ell=1}^{n} E_k(\omega_{j\ell}) D_\ell - \sum_{i=1}^{k} \omega_{ji} D_i + \frac{1}{2} E_k\left(\frac{\partial \eta}{\partial x_j}\right) - \sum_{\ell=1}^{n}\sum_{i=1}^{k} x_i \omega_{j\ell}\omega_{i\ell} \\
&\in E.
\end{aligned} \tag{4.5}$$

Recall that $\omega_{j\ell},\ 1 \le j, \ell \le n$, are polynomials depending only on $x_{k+1}, \dots, x_n$. So $E_k(\omega_{j\ell}) = 0$ for $1 \le j, \ell \le n$. Since $\omega_{ji},\ 1 \le i \le k$, are constants and $D_i, 1 \le i \le k$, are in E, we deduce from (4.5) that

$$\frac{1}{2} E_k\left(\frac{\partial \eta}{\partial x_j}\right) - \sum_{i=1}^{k} x_i \Big(\sum_{\ell=1}^{n} \omega_{j\ell}\omega_{i\ell}\Big) \in E. \tag{4.6}$$

The second term in (4.6) is a polynomial of degree at most two by Proposition 4.3. This implies that

$$E_k\left(\frac{\partial \eta_4}{\partial x_j}\right) = 0, \qquad 1 \le j \le n, \tag{4.7}$$

because $E_k\left(\frac{\partial \eta_4}{\partial x_j}\right)$ is a polynomial of degree three. By Proposition 4.2, we have

$$\eta_4(x_1, \dots, x_n) = \eta_4^1(x_1, \dots, x_k) + \eta_4^2(x_{k+1}, \dots, x_n),$$

where η_4^1 and η_4^2 are homogeneous polynomials of degree four. Equation (4.7) is equivalent to

$$E_k\left(\frac{\partial \eta_4^1}{\partial x_j}\right) = 0, \qquad 1 \le j \le k. \tag{4.8}$$

Since

$$E_k\left(\frac{\partial \eta_4^1}{\partial x_j}\right) = 3\frac{\partial \eta_4^1}{\partial x_j}, \qquad 1 \le j \le k,$$

we deduce easily that $\eta_4^1(x_1, \dots, x_k) = 0$. So η_4 depends only on $x_{k+1}, \dots, x_n$ variables. □

5. Mitter conjecture. If E is a finite-dimensional estimation algebra, Ocone's theorem says that h_i, $1 \le i \le m$, are polynomials of degree at most two. Mitter conjecture asserts that h_i has to be affine (i.e., degree-one polynomial). In this section, we shall prove the Mitter conjecture for finite-dimensional estimation algebras of maximal rank. For this purpose, let us first recall the following theorem proven in [Ya].

THEOREM 5.1. *Let $F(x_1, \dots, x_n)$ be a polynomial on $\mathbb{R}^n$. Suppose that there exists a polynomial path $c : \mathbb{R} \to \mathbb{R}^n$ such that $\lim_{t\to\infty} \|c(t)\| = \infty$ and $\lim_{t\to\infty} F \circ c(t) = -\infty$. Then there are no c^∞ functions $f_1, \dots, f_n$ on $\mathbb{R}^n$ satisfying the equation*

$$\sum_{i=1}^n \frac{\partial f_i}{\partial x_i} + \sum_{i=1}^n f_i^2 = F.$$

THEOREM 5.2. *If E is a finite-dimensional estimation algebra of maximal rank, then h_i, $1 \le i \le m$, are degree-one polynomials.*

Proof. For $1 \le i \le m$, let $h_i(x) = q_i(x) + \ell_i(x)$, where $q_i(x)$ is a homogeneous degree-two polynomial, while $\ell_i(x)$ is a degree-one polynomial. Since $h_i \in E$ by definition, we deduce that $q_i(x)$ is also in E for all $1 \le i \le m$ by the maximal rank condition of E. In view of Lemma 2.3, we conclude that q_i depends only on $x_1, x_2, \dots, x_k$ variables for all $1 \le i \le m$. Hence

$$h_i(x) = q_i(x_1, \dots, x_k) + \ell_i(x), \qquad 1 \le i \le m.$$

On the other hand, Theorem 4.4 tells us that

$$\sum_{i=1}^n \frac{\partial f_i}{\partial x_i} + \sum_{i=1}^n f_i^2 + \sum_{i=1}^m h_i^2 = \eta_4(x_{k+1}, \dots, x_n) + \text{ polynomial of degree three},$$

which implies

$$\begin{aligned} &\sum_{i=1}^n \frac{\partial f_i}{\partial x_i} + \sum_{i=1}^n f_i^2 \\ &= -\sum_{i=1}^m \big(q_i(x_1, \dots, x_k)\big)^2 + \eta_4(x_{k+1}, \dots, x_n) + \text{ polynomial of degree three}. \end{aligned}$$

The above equation and Theorem 5.1 imply that $q_i(x_i, \dots, x_k) = 0$ for all $1 \le i \le m$; i.e., $h_i(x)$, $1 \le i \le m$, are degree-one polynomials. □

REFERENCES

[Br] R.W. Brockett, *Nonlinear systems and nonlinear estimation theory*, in Stochastic Systems: The Mathematics of Filtering and Identification and Applications, M. Hazewinkel and J.S. Williams, eds., D. Reidel, Dordrecht, The Netherlands, 1981.

[Br–Cl] R.W. Brockett and J.M.C. Clark, *The geometry of the conditional density functions*, in Analysis and Optimization of Stochastic Systems, O.L.R. Jacobs et al., eds., Academic Press, New York, 1980, pp. 299–309.

[C–L–Y] J. Chen, C.W. Leung, and S. S.-T. Yau, *Finite dimensional filters with nonlinear drift* IV: *Classification of finite dimensional estimation algebras of maximal rank with state space dimension* 3, SIAM J. Control Optim., 34 (1996), pp. 179–198.

[Ch1–Ya] J. Chen and S. S.-T. Yau, *Finite dimensional filters with nonlinear drift* VI: *Linear structure of* Ω *matrix*, Math. Control Signals Systems, 6 (1996), to appear.

[Ch2–Ya] W.-L. Chiou and S. S.-T. Yau, *Finite dimensional filters with nonlinear drift* II: *Brockett's problem on classification of finite dimensional estimation algebras*, SIAM J. Control Optim., 32 (1994), pp. 297–310.

[Da–Ma] M.H.A. Davis and S.I. Marcus, *An introduction to nonlinear filtering*, in Stochastic Systems: The Mathematics of Filtering and Identification and Applications, M. Hazewinkel and J.S. Willems, eds., D. Reidel, Dordrecht, The Netherlands, 1981.

[Mi] S.K. Mitter, *On the analogy between mathematical problems of nonlinear filtering and quantum physics*, Ricerche di Automatica, 10 (1979), pp. 163–216.

[Oc] D. Ocone, *Finite dimensional estimation algebras in nonlinear filtering*, in Stochastic Systems: The Mathematics of Filtering and Identification and Application, M. Hazewinkel and J.S. Willems, eds., D. Reidel, Dordrecht, The Netherlands, 1981.

[Wo] W.S. Wong, *Theorems on the structure of finite dimensional estimation algebras*, Systems Control Lett., 9 (1987), pp. 117–124.

[Ya] S. S.-T. Yau, *Finite dimensional filters with nonlinear drift* I: *A class of filters including both Kalman–Bucy filters and Benes filters*, J. Mathematical Systems Estim. Control, 4 (1994), pp. 181–203.

SIAM J. CONTROL OPTIM.
Vol. 35, No. 4, pp. 1132–1141, July 1997

003

FINITE-DIMENSIONAL FILTERS WITH NONLINEAR DRIFT VIII: CLASSIFICATION OF FINITE-DIMENSIONAL ESTIMATION ALGEBRAS OF MAXIMAL RANK WITH STATE-SPACE DIMENSION 4*

JIE CHEN†, STEPHEN S.-T. YAU†, AND CHI-WAH LEUNG‡

Abstract. The idea of using estimation algebra to construct finite-dimensional nonlinear filters was first proposed independently by Brockett and Mitter. Estimation algebra turns out to be a useful concept in the investigation of finite-dimensional nonlinear filters. In his talk at the International Congress of Mathematics in 1983, Brockett proposed classifying all finite-dimensional estimation algebras. Chiou and the present authors classify all finite-dimensional estimation algebras of maximal rank with dimension of the state space less than or equal to three. In this paper, we succeed in classifying all finite-dimensional estimation algebras of maximal rank with state-space dimension equal to four. In fact our method gives classification of all finite-dimensional algebras of maximal rank with state-space dimension equal to or less than four.

Key words. finite-dimensional filter, estimation algebra of maximal rank, nonlinear drift, state-space dimension 4

AMS subject classifications. 17B30, 35J15, 60G35, 93E11

PII. S0363012994273325

1. Introduction.

In the 1960s and early 1970s, the basic approach to nonlinear filtering theory was via the "innovation methods" originally proposed by Kailath and subsequently rigorously developed by Fujisaki, Kallianpur, and Kunita [FKK] in 1972. As pointed out by Mitter [Mi], the difficulty with this approach is that the innovation process is not, in general, explicitly computable (except in the well-known Kalman–Bucy case). In the late 1970s, Brockett and Clark [BrCl], Brockett [Br], and Mitter [Mi] proposed the idea of using estimation algebras to construct a finite-dimensional nonlinear filter. The advantage of this finite-dimensional nonlinear filter is the same as the Kalman–Bucy filter. Moreover it avoids the disadvantages of the Kalman–Bucy filter such as the Gaussian initial condition as well as linearity assumption of the drift term. For more detail, we refer the readers to [TWY], [Ya], and the very interesting Ph.D. thesis by M. Cohen de Lara [La], in which the links between finite-dimensional estimation algebras and finite-dimensional filters were discussed. In [Ya], Yau has studied the general class of nonlinear filtering systems which included both Kalman–Bucy and Benes filtering systems as special cases. He gives necessary and sufficient conditions for an estimation algebra of such filtering systems to be finite dimensional. Using the Wei–Norman approach, he constructed explicitly finite-dimensional recursive filters for such a nonlinear filtering systems.

In his talk at the International Congress of Mathematics in 1983, Brockett proposed classifying all finite-dimensional estimation algebras. Since then, the concept of estimation algebra has been proven to be an invaluable tool in the study of nonlinear filtering problems. If the drift term of the nonlinear filtering system has a

*Received by the editors August 24, 1994; accepted for publication (in revised form) April 23, 1996. This research was supported by U.S. Army grant DAAH0493G006.
http://www.siam.org/journals/sicon/35-4/27332.html
†Control and Information Laboratory, University of Illinois at Chicago, MSCS, M/C 249, 851 South Morgan Street, Chicago, IL 60607-7045 (chen@gauss.math.uic.edu, u32790@uic.edu).
‡Department of Mathematics, National Central University, Chung-Li, Taiwan 32054, Republic of China (leung@math.ncu.edu.tw).

potential function (i.e., drift term is a gradient vector field), then the corresponding estimation algebra is called exact. In [TWY], Tam, Wong, and Yau have classified all finite-dimensional exact estimation algebras of maximal rank with arbitrary state-space dimension. In [ChYa], Chiou and Yau are able to classify all finite-dimensional estimation algebras of maximal rank with state-space dimension less than or equal to two. The novelty of their theorem is that there is no assumption on the drift term of the nonlinear filtering system. In [CLY], Chen, Leung, and Yau classify all finite-dimensional estimation algebras of maximal rank with state-space dimension equal to 3 (without any assumption on the drift term). This paper is a natural continuation of [ChYa] and [CLY]. The following is our main theorem.

MAIN THEOREM. *Suppose that the state space of the filtering system* (2.1) *is of dimension* $n \leq 4$. *If* E *is the finite-dimensional estimation algebra of maximal rank, then the drift term* f *must be a linear vector field (i.e., each component is a polynomial of degree one) plus a gradient vector field and* E *is a real vector space of dimension* $2n + 2$ *with basis given by* $1, x_1, \ldots, x_n, D_1, \ldots, D_n$ *and* L_0. *Moreover* η *is a degree* 2 *polynomial.*

This kind of nonlinear filtering system was studied by Yau [Ya]. Therefore, from Lie algebraic point of view, we have shown that the finite-dimensional filters considered in [Ya] are the most general finite-dimensional filters (cf. [Ch] for a nice review of the Yau filter).

Let $\omega_{ij} = \frac{\partial f_j}{\partial x_i} - \frac{\partial f_i}{\partial x_j}$, which was first introduced by Wong [Wo3]. Our strategy is to prove ω_{ij} constant for all i, j. Then we can apply the result of [Ya] to finish the proof. This involves two steps. The first step is to prove that ω_{ij} is a degree-one polynomial. Let n be the dimension of the state space. In the case $n = 3$ there are three unknowns: ω_{12}, ω_{13}, and ω_{23}. It is easy to see that they are all degree-two polynomials in view of Ocone's theorem. In [YaLe], Leung and Yau showed that the coefficients of the quadratic parts of ω_{12}, ω_{13}, and ω_{23} have to satisfy 90 quadratic equations. It was shown in that paper that the 90 quadratic equations have only a trivial solution. Hence the proof of the first step is completed in this case. Obviously, this approach encounters difficulty when n is greater than 3. Fortunately, Chen and Yau [ChYa1] were able to prove that ω_{ij} is a degree-one polynomial for arbitrary n by means of their new algebraic technique. The second step is to prove that ω_{ij} is actually a constant. This is the hard part of the problem of classification of finite-dimensional estimation algebras of maximal rank. The purpose of this paper is to deal with the hard part of the problem by proving ω_{ij} constant for $n \leq 4$. We introduce a new matrix equation. The key point of this paper is to show that this matrix has no nontrivial solution. The advantage of our new technique is not only that we can obtain entirely new results for $n = 4$ but also that we have simple uniform proof of our previous results for $n \leq 3$.

The paper is in essence a continuation of [Ya], [ChYa], [CLY], [ChYa1], [ChYa2], and we strongly recommend that readers familiarize themselves with the results in [Ya], [ChYa1], [ChYa2]. However, every effort will be made to make this paper as self-contained as possible, with minimal duplication of the previous papers.

2. Basic concepts. In this section, we shall recall some basic concepts and results from [Ya] and [ChYa]. Consider a filtering problem based on the following observation model:

$$\begin{cases} dx(t) = f(x(t))dt + g(x(t))dv(t), & x(0) = x_0, \\ dy(t) = h(x(t))dt + dw(t), & y(0) = 0, \end{cases} \tag{2.1}$$

in which x, v, y, and w are, respectively, $\mathbb{R}^n$-, $\mathbb{R}^p$-, $\mathbb{R}^m$-, and $\mathbb{R}^m$-valued processes and v and w have components which are independent, standard Brownian processes. We further assume that $n = p, f, h$ are C^∞ smooth and that g is an orthogonal matrix. We shall refer to $x(t)$ as the state of the system at time t and to $y(t)$ as the observation at time t.

Let $\rho(t, x)$ denote the conditional density of the state given the observation $\{y(s) : 0 \le s \le t\}$. It is well known (see [DaMa], for example) that $\rho(t, x)$ is given by normalizing a function, $\sigma(t, x)$, which satisfies the Duncan–Mortensen–Zakai equation:

$$d\sigma(t,x) = L_0\sigma(t,x)dt + \sum_{i=1}^{m} L_i\sigma(t,x)dy_i(t), \qquad \sigma(0,x) = \sigma_0, \tag{2.2}$$

where

$$L_0 = \frac{1}{2}\sum_{i=1}^{n}\frac{\partial^2}{\partial x_i^2} - \sum_{i=1}^{n} f_i\frac{\partial}{\partial x_i} - \sum_{i=1}^{n}\frac{\partial f_i}{\partial x_i} - \frac{1}{2}\sum_{i=1}^{m} h_i^2$$

and for $i = 1, \dots, m$, L_i is the zero-degree differential operator of multiplication by h_i, σ_0 is the probability density of the initial point x_0. In this paper, we will assume σ_0 is a C^∞ function.

Equation (2.2) is a stochastic partial differential equation. The stochastic differential is a Stratonovich one and not an Ito one. In real applications, we are interested in constructing state estimators from observed sample paths with some property of robustness. Based on Rozovsky's transformation [Ro],

$$\xi(t,x) = \exp\Big(-\sum_{i=1}^{m} h_i(x)y_i(t)\Big)\sigma(t,x),$$

Davis [Da] proposed studying the following robust Duncan–Mortensen–Zakai equation:

$$\begin{aligned} &\frac{\partial\xi}{\partial t}(t,x) = L_0\xi(t,x) + \sum_{i=1}^{m} y_i(t)[L_0, L_i]\xi(t,x) + \frac{1}{2}\sum_{i,j=1}^{m} y_i(t)y_j(t)[[L_0, L_i], L_j]\xi(t,x), \\ &\xi(0,x) = \sigma_0, \end{aligned} \tag{2.3}$$

which is a time-varying partial differential equation. Here we have used the following notation.

DEFINITION 1. *If X and Y are differential operators, the Lie bracket of X and Y, $[X, Y]$, is defined by $[X, Y]\phi = X(Y\phi) - Y(X\phi)$ for any C^∞ function ϕ.*

DEFINITION 2. *The estimation algebra E of a filtering problem* (2.1) *is defined to be the Lie algebra generated by $\{L_0, L_1, \dots, L_m\}$. E is said to be an estimation algebra of maximal rank if for any $1 \le i \le n$ there exists a constant c_i such that $x_i + c_i$ is in E.*

Most of the known finite-dimensional estimation algebras are maximal. For example, if (2.1) is linear, i.e., $f(x) = Ax$, $g(x) = B$, and $h(x) = Cx$, and if (A, B, C)

also is minimal, then the corresponding estimation algebra is maximal [Ha]. We need the following basic result for later discussion.

THEOREM 2.1 (Ocone). *Let E be a finite-dimensional estimation algebra. If a function ξ is in E, then ξ is a polynomial of degree at most two.*

In [Wo3], the concept of Ω is introduced, defined as the matrix whose (i,j)-element ω_{ij} is $\frac{\partial f_j}{\partial x_i} - \frac{\partial f_i}{\partial x_j}$. Define

$$D_i = \frac{\partial}{\partial x_i} - f_i$$

and

$$\eta = \sum_{i=1}^{n} \frac{\partial f_i}{\partial x_i} + \sum_{i=1}^{n} f_i^2 + \sum_{i=1}^{m} h_i^2.$$

Then

$$L_0 = \frac{1}{2}\left(\sum_{i=1}^{n} D_i^2 - \eta\right).$$

The following theorem proved in [Ya] plays a fundamental role in the classification of finite-dimensional estimation algebras.

THEOREM 2.2 (Yau). *Let E be a finite-dimensional estimation algebra of* (2.1) *such that $\omega_{ij} = \frac{\partial f_j}{\partial x_i} - \frac{\partial f_i}{\partial x_j}$ are constant functions. If E is of maximal rank, then E is a real vector space of dimension $2n+2$ with basis given by $1, x_1, x_2, \dots, x_n, D_1, D_2, \dots, D_n$ and L_0.*

Recently, Chen and Yau [ChYa1] have made important progress in the program of classification of finite-dimensional estimation algebras of maximal rank. Namely, they have shown that Ω matrix is linear in the sense that all ω_{ij} are degree-one polynomials. More recently, in order to prove the Mitter conjecture for finite-dimensional estimation algebra of maximal rank, Chen and Yau [ChYa2] have sharpened the above result. To describe this new result, let us first recall some important concepts and notations introduced in [ChYa1].

Let Q be the space of quadratic forms in n variables, i.e., real vector space spanned by $x_i x_j$, with $1 \le i \le j \le n$. Let $X = (x_1, \dots, x_n)^T$. For any quadratic form $p \in Q$, there exists a symmetric matrix A such that $p(x) = X^T A X$. The rank of the quadratic form p is denoted by $\mathrm{rk}(p)$ and is defined to be the rank of the matrix A.

DEFINITION 3. *A fundamental quadratic form of the estimation algebra E is an element $p_0 \in E \cap Q$ with the greatest positive rank, i.e.,* $\mathrm{rk}(p_0) \ge \mathrm{rk}(p)$ *for any $p \in E \cap Q$. The quadratic rank of the estimation algebra E is defined to be* $\mathrm{rk}(p_0)$. *The following Theorem* 2.3 *and Proposition* 2.4 *are proved in* [ChYa2].

THEOREM 2.3. *Let E be a finite-dimensional estimation algebra of maximal rank. Let k be the quadratic rank of E. Then*

(1) *the observation terms $h_i(x), 1 \le i \le m$, are affine polynomials.*

(2) (a) *ω_{ij}, for $1 \le i \le k$ or $1 \le j \le k$, are constants.*

(b) *ω_{ij}, for $k+1 \le i, j \le n$, are degree-one polynomials in $x_{k+1}, \dots, x_n$.*

(3) *$\eta = \sum_{i=1}^{n} \frac{\partial f_i}{\partial x_i} + \sum_{i=1}^{n} f_i^2 + \sum_{i=1}^{m} h_i^2$ is a homogeneous polynomial of degree four. Moreover, η_4 (homogeneous polynomial of degree-four part of η) depends only on $x_{k+1}, \dots, x_n$ variables.*

PROPOSITION 2.4. *Let E be a finite-dimensional estimation algebra of maximal rank. Let k be the quadratic rank of E. Then η is a polynomial of degree at most four, and any homogeneous polynomial of degree two in E depends only on $x_1, x_2, \dots, x_k$.*

Finally, we need to recall the following theorem proved in [Ya].

THEOREM 2.5. *Let $F(x_1, \dots, x_n)$ be a polynomial on $\mathbb{R}^n$. Suppose that there exists a polynomial path $c : \mathbb{R} \to \mathbb{R}^n$ such that $\lim_{t\to\infty}||c(t)|| = \infty$ and $\lim_{t\to\infty}F(c(t)) = -\infty$. Then there are no C^∞ functions $f_1, \dots, f_n$ on $\mathbb{R}^n$ satisfying the equations*

$$\sum_{i=1}^{n} \frac{\partial f_i}{\partial x_i} + \sum_{i=1}^{n} f_i^2 = F.$$

3. Proof of the main theorem. From Yau's theory [Ya], to classify all finite-dimensional estimation algebras of maximal rank, we need only to prove that the ω_{ij}'s corresponding to these estimation algebras of maximal rank are automatically constants.

We first introduce a new matrix equation which turns out to play an important role in classification of finite-dimensional estimation algebras of maximal rank.

THEOREM 3.1. *Suppose that η_4 is a homogeneous polynomial of degree four in n variables. If $n \le 4$ and $\triangle$ is an antisymmetric matrix with each entry a homogeneous polynomial of degree one such that*

$$\triangle\triangle^T = \frac{1}{2}H(\eta_4), \tag{3.1}$$

where $H(\eta_4) = (\frac{\partial^2\eta_4}{\partial x_i \partial x_j})$ is the Hessian matrix of η_4, then $\triangle = 0$.

Proof. Write

$$\triangle = \sum_{i=1}^{n} A_i x_i, \tag{3.2}$$

$$\frac{1}{2}H(\eta_4) = \sum_{i\le j} H_{ij} x_i x_j, \tag{3.3}$$

where A_i's are real $n \times n$ antisymmetric matrices and H_{ij}'s are real $n \times n$ symmetric matrices, i.e.,

$$A_i = -A_i^T, \tag{3.4}$$

$$H_{ij} = H_{ij}^T. \tag{3.5}$$

Then (3.1) implies

$$A_i^2 = -H_{ii}, \tag{3.6}$$

$$A_iA_j + A_jA_i = -H_{ij}. \tag{3.7}$$

Denote the (i, j) entry of the matrix M by $M(i, j)$. For $i > j$, we let $H_{ij} = H_{ji}$. Note that

$$\frac{\partial^2(x_i^2x_j^2)}{\partial x_i^2} = 2x_j^2, \qquad \frac{\partial^2(x_i^2x_j^2)}{\partial x_j^2} = 2x_i^2, \quad \text{and} \qquad \frac{\partial^2(x_i^2x_j^2)}{\partial x_i\partial x_j} = 4x_ix_j. \tag{3.8}$$

From (3.3) and (3.8), we get

$$2H_{ii}(j, j) = 2H_{jj}(i, i) = H_{ij}(i, j) \quad \text{for } i \ne j. \tag{3.9}$$

Hence, for $i \neq j$, (3.6), (3.7), and (3.9) imply

$$\sum_l A_i(j,l)A_i(l,j) = \sum_l A_j(i,l)A_j(l,i) = \frac{1}{2}\sum_l \big[A_i(i,l)A_j(l,j) + A_j(i,l)A_i(l,j)\big]. \tag{3.10}$$

Recall that each A_i is an antisymmetric matrix. So (3.10) is reduced to the following equation:

$$\sum_l \big[A_i(j,l)\big]^2 = \sum_l \big[A_j(i,l)\big]^2 = \frac{1}{2}\sum_l \big[A_i(i,l)A_j(j,l) + A_j(i,l)A_i(j,l)\big]. \tag{3.11}$$

In view of the Schwarz inequality, we have

$$\begin{aligned} 2\sum_l \big[A_i(j,l)\big]^2 + 2\sum_l \big[A_j(i,l)\big]^2 &= 2\sum_l A_i(i,l)A_j(j,l) + 2\sum_l A_j(i,l)A_i(j,l) \\ &\le \sum_{l\neq i,j} \big[A_i(i,l)\big]^2 + \sum_{l\neq i,j} \big[A_j(j,l)\big]^2 + \sum_{l\neq i,j} \big[A_j(i,l)\big]^2 \\ &\quad + \sum_{l\neq i,j} \big[A_i(j,l)\big]^2. \end{aligned} \tag{3.12}$$

This implies

$$\begin{aligned} &\sum_l \big[A_i(j,l)\big]^2 + \sum_l \big[A_j(i,l)\big]^2 + \sum_{l=i,j} \big[A_j(i,l)\big]^2 + \sum_{l=i,j} \big[A_i(j,l)\big]^2 \\ &\qquad \le \sum_{l\neq i,j} \big[A_i(i,l)\big]^2 + \sum_{l\neq i,j} \big[A_j(j,l)\big]^2. \end{aligned} \tag{3.13}$$

Taking the sum of left-hand side of (3.13) over $i < j$, we get

$$\begin{aligned} &\sum_{i<j}\sum_{l=1}^n \big[A_i(j,l)\big]^2 + \sum_{i<j}\sum_{l=1}^n \big[A_j(i,l)\big]^2 + \sum_{i<j} \big[A_j(i,j)\big]^2 + \sum_{i<j} \big[A_i(j,i)\big]^2 \\ &\quad = \sum_{i<j}\sum_{l=1}^n \big[A_i(j,l)\big]^2 + \sum_{j<i}\sum_{l=1}^n \big[A_j(i,l)\big]^2 + \sum_{i<j} \big[A_j(i,j)\big]^2 + \sum_{j<i} \big[A_j(i,j)\big]^2 \\ &\quad = \sum_{i\neq l, i\neq j} \big[A_i(j,l)\big]^2 + 2\sum_{i\neq l} \big[A_i(i,l)\big]^2. \end{aligned} \tag{3.14}$$

On the other hand, by taking the sum of right-hand side of (3.13) over $i < j$, we get

$$\begin{aligned} &\sum_{i<j}\sum_{l\neq i,j} \big[A_i(i,l)\big]^2 + \sum_{i<j}\sum_{l\neq i,j} \big[A_j(j,l)\big]^2 \\ &\quad = \sum_{i<j}\sum_{l\neq i,j} \big[A_i(i,l)\big]^2 + \sum_{j<i}\sum_{l\neq i,j} \big[A_i(i,l)\big]^2 \\ &\quad = (n-2)\sum_{i\neq l} \big[A_i(i,l)\big]^2. \end{aligned} \tag{3.15}$$

Comparing (3.14) and (3.15), we get

$$\sum_{i\neq l, i\neq j} \big[A_i(j,l)\big]^2 = (n-4)\sum_{i\neq l} \big[A_i(i,l)\big]^2. \tag{3.16}$$

When $n = 4$, we see from (3.16) that

$$\sum_{i\neq l, i\neq j} \big[A_i(j,l)\big]^2 = 0;$$

hence

$$A_i(j,l) = 0 \quad \text{for } i \neq j \quad \text{and } i \neq l. \tag{3.17}$$

Using (3.17) in (3.10) gives

$$\big[A_i(i,j)\big]^2 = \big[A_j(j,i)\big]^2. \tag{3.18}$$

Note that for $i < j$

$$\frac{\partial^2 x_i^3 x_j}{\partial x_i \partial x_j} = 3x_i^2 \quad \text{and} \quad \frac{\partial^2 x_i^3 x_j}{\partial x_i^2} = 6x_i x_j. \tag{3.19}$$

From (3.3) and (3.19), we get

$$2H_{ij}(i,i) = H_{ii}(i,j). \tag{3.20}$$

In view of (3.6), (3.7), and (3.20), we have

$$2\sum_l A_i(i,l)A_j(l,i) = \sum_l A_i(i,l)A_i(l,j). \tag{3.21}$$

Using (3.17) in (3.21) gives

$$A_i(i,j)A_j(j,i) = 0. \tag{3.22}$$

In view of (3.18) and (3.22), we get

$$A_i(i,j) = 0 \quad \text{for all } i,j. \tag{3.23}$$

Therefore $A_1 = A_2 = A_3 = A_4 = 0$ by (3.17) and (3.23). This simply means that $\triangle = 0$. □

PROPOSITION 3.2. *Let E be a finite-dimensional estimation algebra of maximal rank associated with the filtering system* (2.1). *Then E contains the real vector space spanned by $1, x_1, \ldots, x_n, D_1, \ldots, D_n$ and L_0.*

Proof. Since E is a finite-dimensional estimation algebra with maximal rank, there are constants c_i's such that $x_i + c_i$ is in E for $1 \leq i \leq n$:

$$[L_0, x_j + c_j] = \frac{1}{2}\left[\sum_{i=1}^n D_i^2 - \eta, x_j\right] = \frac{1}{2}\sum_{i=1}^n \big[D_i^2, x_j\big] = D_j \in E,$$
$$[D_j, x_j + c_j] = 1 \in E.$$

Hence $x_1, \ldots, x_n \in E$. □

We are now ready to prove the main theorem of this paper stated in section 1. By Theorem 2.3, we know that ω_{ij}'s are constants for $1 \le i \le k$ or $1 \le j \le k$, where k is the quadratic rank of E. We also know that ω_{ij}'s are degree-one polynomials in $x_{k+1}, \dots, x_n$ variables for $k+1 \le i, j \le n$. We are going to prove that ω_{ij}'s are indeed constants for $k+1 \le i, j \le n$ and $n \le 4$. Observe that

$$\begin{aligned}\big[[L_0, D_j], D_l\big] &= \left[\sum_{i=1}^{n}\left(\omega_{ji}D_i + \frac{1}{2}\frac{\partial \omega_{ji}}{\partial x_i}\right) + \frac{1}{2}\frac{\partial \eta}{\partial x_j}, D_l\right] \\ &= \sum_{i=1}^{n}\left(\omega_{ji}\omega_{li} - \frac{\partial \omega_{ji}}{\partial x_l}D_i\right) - \frac{1}{2}\sum_{i=1}^{n}\frac{\partial^2 \omega_{ji}}{\partial x_l \partial x_i} - \frac{1}{2}\frac{\partial^2 \eta}{\partial x_l \partial x_j} \\ &\in E.\end{aligned}$$

In view of Theorem 2.3 and Proposition 3.2, we deduce that

$$\sum_{i=1}^{n}\omega_{ji}\omega_{li} - \frac{1}{2}\frac{\partial^2 \eta}{\partial x_l \partial x_j} \in E. \tag{3.24}$$

Let η_m be the homogeneous polynomial of the degree-m part of η and β_{ij} be the homogeneous polynomial of the degree-one part of ω_{ij}. Then in view of Theorem 2.3 and Proposition 3.2, we have

$$\sum_{i=k+1}^{n}\beta_{ji}\beta_{li} - \frac{1}{2}\frac{\partial^2 \eta_4}{\partial x_l \partial x_j} \in E \tag{3.25}$$

for $k+1 \le l, j \le n$. Observe that $\sum_{i=k+1}^{n}\beta_{ji}\beta_{li} - \frac{1}{2}\frac{\partial^2 \eta}{\partial x_l \partial x_j}$ is a homogeneous polynomial of degree two in E which depends only on $x_1, \dots, x_k$ variables because k is the quadratic rank of E. On the other hand η_4, β_{ji}, and β_{li}, for $k+1 \le i, j, l \le n$, depend only on $x_{k+1}, \dots, x_n$ variables by Theorem 2.3. So the left-hand side of (3.25) depends only on $x_{k+1}, \dots, x_n$. Therefore we deduce that

$$\sum_{i=k+1}^{n}\beta_{ji}\beta_{li} - \frac{1}{2}\frac{\partial^2 \eta}{\partial x_l \partial x_j} = 0 \qquad \text{for } k+1 \le j, l \le n. \tag{3.26}$$

Let $\triangle = (\beta_{ij})$, $k+1 \le i, j \le n$, be an $(n-k) \times (n-k)$ antisymmetric matrix. Then we have

$$\triangle\triangle^T = \frac{1}{2}H(\eta_4), \tag{3.27}$$

where $H(\eta_4) = (\frac{\partial^2 \eta_4}{\partial x_i \partial x_j})$, $k+1 \le i, j \le n$, stands for the Hessian matrix for η_4. In view of Theorem 3.1, we have $\triangle = 0$. So we have shown that ω_{ij}'s are constants for $1 \le i, j \le n$. By Theorem 2.2 of Yau, E is a real vector space of dimension $2n+2$ with basis given by $1, x_1, x_2, \dots, x_n, D_1, D_2, \dots, D_n$ and L_0.

Since $\triangle = 0$, (3.27) implies $\eta_4 = 0$. So

$$\sum_{i=1}^{n}\frac{\partial f_i}{\partial x_i} + \sum_{i=1}^{n}f_i^2 + \sum_{i=1}^{m}h_i^2 = \eta_0 + \eta_1 + \eta_2 + \eta_3,$$

which implies

$$\sum_{i=1}^{n} \frac{\partial f_i}{\partial x_i} + \sum_{i=1}^{n} f_i^2 = \eta_0 + \eta_1 + \eta_2 - \sum_{i=1}^{m} h_i^2 + \eta_3. \tag{3.28}$$

By Theorem 2.3, $\tilde{F} = \eta_0 + \eta_1 + \eta_2 - \sum_{i=1}^{m} h_i^2$ is at most a polynomial of degree two. If η_3 is not identically zero, then we can choose a polynomial path $c : \mathbb{R} \to \mathbb{R}^n$ such that $\lim_{t\to\infty} ||c(t)|| = \infty$ and $\lim_{t\to\infty} F(c(t)) = -\infty$. This is not possible in view of Theorem 2.5. So we conclude that $\eta_3 = 0$; i.e., η is a polynomial of degree two.

REFERENCES

[Be] V. Benes, *Exact finite dimensional filters for certain diffusions with nonlinear drift*, Stochastics, 5 (1981), pp. 65–92.

[BrCl] R.W. Brockett and J.M.C. Clark, *The geometry of the conditional density functions*, in Analysis and Optimization of Stochastic Systems, O.L.R. Jacobs, et al., eds., Academic Press, New York, 1980, pp. 299–309.

[Br] R.W. Brockett, *Nonlinear systems and nonlinear estimation theory*, in Stochastic Systems: The Mathematics of Filtering and Identification and Applications, M. Hazewinkel and J.S. Willems, eds., D. Reidel, Dordrecht, The Netherlands, 1981.

[Ch] J. Chen, *On ubiquity of Yau filters*, in Proc. of the American Control Conference, Baltimore, MD, June 1994, pp. 252–254.

[ChMi] M. Chaleyat-Maurel and D. Michel, *Des resultats de non-existence de filtre de dimension finie*, Stochastics, 13 (1984), pp. 83–102.

[ChYa] W.L. Chiou and S.S.-T. Yau, *Finite dimensional filters with nonlinear drift* II: *Brockett's problem on classification of finite dimensional estimation algebras*, SIAM J. Control Optim., 32 (1994), pp. 297–310.

[CLY] J. Chen, C.-W. Leung, and S.S.-T. Yau, *Finite dimensional filters with nonlinear drift* IV: *Classification of finite dimensional estimation algebras of maximal rank with state space dimension* 3, SIAM J. Control Optim., 34 (1996), pp. 179–198.

[ChYa1] J. Chen and S.S.-T. Yau, *Finite dimensional filters with nonlinear drift* VI: *Linear structure of Ω matrix*, Math. Control Signals Systems, 6 (1996), to appear.

[ChYa2] J. Chen and S.S.-T. Yau, *Finite dimensional filters with nonlinear drift* VII: *Mitter conjecture and structure of η*, SIAM J. Control Optim., 35 (1997), pp. 1116–1131.

[Co] P.C. Collingwood, *Some remarks on estimation algebras*, Systems Control Lett., 7 (1986), pp. 217–224.

[Da] M.H.A. Davis, *On a multiplicative functional transformation arising in nonlinear filtering theory*, Z. Wahrsch. Verw. Gebiete, 54 (1980), pp. 125–139.

[DaMa] M.H.A. Davis and S.I. Marcus, *An introduction to nonlinear filtering*, in Stochastic Systems: The Mathematics of Filtering and Identification and Applications, M. Hazewinkel and J.S. Willems, eds., D. Reidel, Dordrecht, The Netherlands, 1981.

[DTWY] R.T. Dong, L.F. Tam, W.S. Wong, and S.S.-T. Yau, *Structure and classification theorems of finite dimensional exact estimation algebras*, SIAM J. Control Optim., 29 (1991), pp. 866–877.

[Fr] A. Friedman, *Stochastic Differential Equations and Applications, Vol.* 1, Academic Press, New York, 1975.

[FKK] M. Fujisaki, G. Kallianpur, and H. Kunita, *Stochastic differential equations for the nonlinear filtering problem*, Osaka J. Math., 1 (1972), pp. 19–40.

[Ha] M. Hazewinkel, *Lecture on linear and nonlinear filtering*, in Analysis and Estimation of Stochastic Mechanical Systems, CISM Courses and Lectures 303, W. Shiehlen and W. Wedig, eds., Springer-Verlag, Vienna, 1998.

[La] M. Cohen de Lara, *Contribution des methodes geometriques au filtrage de dimension finie*, Ph.D. thesis, Ecole des Mines de Paris, 1991.

[Mi] S.K. Mitter, *On the analogy between mathematical problems of nonlinear filtering and quantum physics*, Ricerche di Automatica, 10 (1979), pp. 163–216.

[Oc] D.L. Ocone, *Finite dimensional estimation algebras in nonlinear filtering*, in Stochastic Systems: The Mathematics of Filtering and Identification and Applications, M. Hazewinkel and J.S. Willems, eds., D. Reidel, Dordrecht, The Netherlands, 1981.

[Ro] B.L. Rozovsky, *Stochastic partial differential equations arising in nonlinear filtering problem*, Uspekhi Mat. Nauk, 27 (1972), pp. 213–214 (in Russian).

[St] S. Steinberg, *Applications of the Lie algebraic formulas of Baker, Campbell, Hausdorff and Zassenhaus to the calculation of explicit solutions of partial differential equations*, J. Differential Equations, 26 (1979), pp. 404–434.

[TWY] L.F. Tam, W.S. Wong, and S.S.-T. Yau, *On a necessary and sufficient condition for finite dimensionality of estimation algebras*, SIAM J. Control Optim., 28 (1990), pp. 173–185.

[W] E. Wong, *Stochastic Processes in Information and Dynamic Systems*, McGraw-Hill, New York, 1971.

[WeNo] J. Wei and E. Norman, *On global representations of the solutions of linear differential equations as a product of exponentials*, Proc. Amer. Math. Soc., 15 (1964), pp. 327–334.

[Wi] D.V. Widder, *The Heat Equation*, Mathematics 67, Academic Press, New York, 1975.

[Wo1] W.S. Wong, *New classes of finite dimensional nonlinear filters*, Systems Control Lett., 3 (1983), pp. 155–164.

[Wo2] W.S. Wong, *On a new class of finite dimensional estimation algebras*, Systems Control Lett., 9 (1987), pp. 79–83.

[Wo3] W.S. Wong, *Theorems on the structure of finite dimensional estimation algebras*, Systems Control Lett., 9 (1987), pp. 117–124.

[Ya] S.S.-T. Yau, *Finite dimensional filters with nonlinear drift* I: *A class of filters including both Kalman-Bucy filters and Benes filters*, J. Math. Systems Estim. Control, 4 (1994), pp. 181–203.

[YaCh] S.S.-T. Yau and W.L. Chiou, *Recent results on classification of finite dimensional estimation algebras: Dimension of state space* ≤ 2, in Proc. 30th IEEE Conf. on Decision and Control, Brighton, England, Dec. 11–13, 1991.

[YaLe] S.S.-T. Yau and C.-W. Leung, *Recent results on classification of finite dimensional maximal rank estimation algebras with state space dimension* 3, in Proc. 31st Conf. on Decision and Control, Tucson, AZ, Dec. 1992, pp. 2247–2250.

SIAM J. CONTROL OPTIM.
Vol. 35, No. 4, pp. 1142–1168, July 1997

004

PROXIMAL MINIMIZATION METHODS WITH GENERALIZED BREGMAN FUNCTIONS*

KRZYSZTOF C. KIWIEL†

Abstract. We consider methods for minimizing a convex function f that generate a sequence $\{x^k\}$ by taking x^{k+1} to be an approximate minimizer of $f(x)+D_h(x,x^k)/c_k$, where $c_k>0$ and D_h is the D-function of a Bregman function h. Extensions are made to B-functions that generalize Bregman functions and cover more applications. Convergence is established under criteria amenable to implementation. Applications are made to nonquadratic multiplier methods for nonlinear programs.

Key words. convex programming, nondifferentiable optimization, proximal methods, Bregman functions, B-functions

AMS subject classifications. 65K05, 90C25

PII. S0363012995281742

1. Introduction.

We consider the convex minimization problem

$$f_* = \inf\{\, f(x) : x \in X \,\}, \tag{1.1}$$

where $f : \mathbb{R}^n \to (-\infty,\infty]$ is a closed proper convex function and X is a nonempty closed convex set in $\mathbb{R}^n$. One method for solving (1.1) is the proximal point algorithm (PPA) [Mar70, Roc76b], which generates a sequence

$$x^{k+1} = \arg\min\{\, f(x) + |x-x^k|^2/2c_k : x \in X \,\} \quad \text{for } k = 1,2,\dots, \tag{1.2}$$

starting from any point $x^1 \in \mathbb{R}^n$, where $|\cdot|$ is the Euclidean norm and $\{c_k\}$ is a sequence of positive numbers. The convergence and applications of the PPA are discussed in, e.g., [Aus86, CoL93, EcB92, GoT89, Gül91, Lem89, Roc76a, Roc76b].

Several proposals have been made for replacing the quadratic term in (1.2) with other distancelike functions [BeT94, CeZ92, ChT93, Eck93, Egg90, Ius95, IuT93, Teb92, TsB93]. In [CeZ92], (1.2) is replaced by

$$x^{k+1} = \arg\min\{\, f(x) + D_h(x,x^k)/c_k : x \in X \,\}, \tag{1.3}$$

where $D_h(x,y) = h(x) - h(y) - \langle \nabla h(y), x-y\rangle$ is the D-function of a Bregman function h [Bre67, CeL81], which is continuous, strictly convex, and differentiable in the interior of its domain; here $\langle\cdot,\cdot\rangle$ is the usual inner product and ∇h is the gradient of h. Accordingly, this is called *Bregman proximal minimization* (BPM). The convergence of the BPM method is discussed in [CeZ92, ChT93, Eck93, Ius95, TsB93], a generalization for finding zeros of monotone operators is given in [Eck93], and applications to convex programming are presented in [Cha94, Eck93, Ius95, NiZ92, NiZ93a, NiZ93b, Teb92, TsB93].

This paper discusses convergence of the BPM method using the *B-functions* of [Kiw97] that generalize Bregman functions, being possibly nondifferentiable and infinite on the boundary of their domains (cf. section 2; another recent generalization of

*Received by the editors February 17, 1995; accepted for publication (in revised form) April 24, 1996. This research was supported by Polish State Committee for Scientific Research grant 8T11A02610.

http://www.siam.org/journals/sicon/35-4/28174.html

†Systems Research Institute, Newelska 6, 01–447 Warsaw, Poland (kiwiel@ibspan.waw.pl).

Bregman functions is given in [BaB97] for convex feasibility problems). Then (1.3) involves $D_h^k(x, x^k) = h(x) - h(x^k) - \langle \gamma^k, x - x^k \rangle$, where γ^k is a subgradient of h at x^k. We establish for the first time convergence of versions of the BPM method that relax the requirement for exact minimization in (1.3). (The alternative approach of [Flå94], being restricted to Bregman functions with Lipschitz continuous gradients, cannot handle the applications of sections 7–9.) We note that in several important applications, strictly convex problems of the form (1.3) may be solved by dual ascent methods; cf. references in [Kiw97, Tse90].

The application of the BPM method to the dual functional of a convex program yields nonquadratic multiplier methods [Eck93, Teb92]. By allowing h to have singularities, we extend this class of methods to include, e.g., *shifted* Frish and Carroll barrier function methods [FiM68]. We show that our criteria for inexact minimization can be implemented similarly as in the nonquadratic multiplier methods of [Ber82, Chap. 5]. Our convergence results extend those in [Eck93, TsB93] to quite general *shifted penalty functions*, including twice continuously differentiable ones.

We add that the continuing interest in nonquadratic modified Lagrangians stems from the fact that, in contrast with the quadratic one, they are twice continuously differentiable, and this facilitates their minimization [Ber82, BTYZ92, BrS93, BrS94, CGT92, CGT94, GoT89, IST94, JeP94, Kiw96, NPS94, Pol92, PoT97, Teb92, TsB93]. By the way, our convergence results seem stronger than ones in [IST94, PoT97] for *modified barrier functions*, resulting from a dual application of (1.3) with $D_h^k(x, x^k)$ replaced by an entropylike ϕ-divergence.

The paper is organized as follows. In section 2 we recall several elementary properties of B-functions. In section 3 we present an inexact BPM method. Its global convergence under various conditions is established in sections 4 and 5. In section 6 we show that the exact BPM method converges finitely when (1.1) enjoys a sharp minimum property. Applications to multiplier methods are given in section 7. Convergence of general multiplier methods is studied in section 8, while section 9 focuses on two classes of shifted penalty methods. Additional aspects of multiplier methods are discussed in section 10. The appendix contains proofs of certain technical results.

Our notation is fairly standard. $\mathbb{R}_+$ and $\mathbb{R}_>$ are the nonnegative and positive reals, respectively. For any set C in $\mathbb{R}^n$, $\operatorname{cl} C$, $\overset{\circ}{C}$, $\operatorname{ri} C$, and ∂C denote the closure, interior, relative interior, and boundary of C, respectively. The *indicator* function of C is denoted by $\imath_C$ ($\imath_C(x) = 0$ if $x \in C$, ∞ otherwise), and its *support* function by $\imath_C^*(\cdot) = \sup_{x \in C} \langle \cdot, x \rangle$.

2. *B*-functions. We first recall some useful concepts from convex analysis (see, e.g., [Roc70]).

For any proper convex function h on $\mathbb{R}^n$, $\mathcal{D}_h = \{x : f(x) < \infty\}$ denotes its *effective domain*, $\partial_\epsilon h(\cdot) = \{p : h(y) \geq h(\cdot) + \langle p, y - \cdot \rangle - \epsilon \ \forall y\}$ its *ϵ-subdifferential* for each $\epsilon \geq 0$, $\partial h = \partial_0 h$ its *subdifferential*, and $h'(x; d) = \lim_{t \downarrow 0} [h(x + td) - h(x)]/t$ its *derivative* in any direction d at a point x where h is finite. By [Roc70, Thm. 23.2],

$$h'(x; d) \geq \imath^*_{\partial h(x)}(d) = \sup\{\langle g, d \rangle : g \in \partial h(x)\}. \tag{2.1}$$

The *domain* and *range* of ∂h are denoted by $\mathcal{D}_{\partial h}$ and $\operatorname{im} \partial h$, respectively. If h is proper, then $\operatorname{ri} \mathcal{D}_h \subset \mathcal{D}_{\partial h} \subset \mathcal{D}_h$ [Roc70, Thm. 23.4]. h is called *cofinite* when its *conjugate* $h^*(\cdot) = \sup_x \langle \cdot, x \rangle - h(x)$ is real valued. If h is closed proper convex, its *recession* function $h0^+(\cdot) = \lim_{t \to \infty} [h(x + t\cdot) - h(x)]/t$ ($\forall x \in \mathcal{D}_h$) is positively homogeneous [Roc70, Thm. 8.5]. h is called *essentially strictly convex* if h is strictly convex on

every convex subset of $\mathcal{D}_{\partial h}$. A proper convex function h is called *essentially smooth* if $\mathcal{D}_{\nabla h} = \mathring{\mathcal{D}}_h \neq \emptyset$ and $|\nabla h(x^k)| \to \infty$ whenever $\{x^k\} \subset \mathring{\mathcal{D}}_h$, $x^k \to x \in \partial\mathcal{D}_h$.

FACT 2.1 (see [Roc70, Thm. 23.5]). *If h is closed proper convex, then the following are equivalent*: $g \in \partial h(x)$, $x \in \partial h^*(g)$, $x \in \operatorname{Arg\,min}\{h(\cdot) - \langle g, \cdot\rangle\}$, $h(x) + h^*(g) = \langle g, x\rangle$.

FACT 2.2 (see [Roc70, Thms. 25.1 and 25.5]). *If h is proper convex, then ∇h is continuous on $\mathcal{D}_{\nabla h} \subset \mathring{\mathcal{D}}_h$, and $x \in \mathcal{D}_{\nabla h}$, i.e., h is differentiable at x, iff $\partial h(x) = \{\nabla h(x)\}$.*

FACT 2.3 (see [Roc70, Thms. 26.1 and 26.3]). *Suppose that h is closed proper convex. Then*

(i) *h is essentially smooth iff $\partial h(x) = \{\nabla h(x)\}$, $\forall x \in \mathring{\mathcal{D}}_h$, and $\partial h(x) = \emptyset$, $\forall x \in \partial\mathcal{D}_h$.*

(ii) *h is essentially strictly convex iff h^* is essentially smooth.*

For any proper convex function h on $\mathbb{R}^n$, we define its *difference functions*:

$$\begin{aligned} D_h^\flat(x,y) &= h(x) - h(y) - \imath^*_{\partial h(y)}(x-y) \quad \forall x, y \in \mathcal{D}_h, \\ D_h^\sharp(x,y) &= h(x) - h(y) + \imath^*_{\partial h(y)}(y-x) \quad \forall x, y \in \mathcal{D}_h. \end{aligned} \tag{2.2}$$

By convexity (cf. (2.1)), $h(x) \geq h(y) + \imath^*_{\partial h(y)}(x-y)$ and

$$0 \leq D_h^\flat(x,y) \leq h(x) - h(y) - \langle \gamma, x-y\rangle \leq D_h^\sharp(x,y) \quad \forall x, y \in \mathcal{D}_h, \gamma \in \partial h(y). \tag{2.3}$$

$D_h^\flat$ and $D_h^\sharp$ generalize the usual *D-function* of h [Bre67, CeL81], defined by

$$D_h(x,y) = h(x) - h(y) - \langle \nabla h(y), x-y\rangle \quad \forall x \in \mathcal{D}_h,\ y \in \mathcal{D}_{\nabla h}, \tag{2.4}$$

since

$$D_h(x,y) = D_h^\flat(x,y) = D_h^\sharp(x,y) \quad \forall x \in \mathcal{D}_h,\ y \in \mathcal{D}_{\nabla h}. \tag{2.5}$$

DEFINITION 2.4 (see [Kiw97]). *A closed proper (possibly nondifferentiable) convex function h is called a B-*function (generalized Bregman function) *if*

(i) *h is strictly convex on $\mathcal{D}_h$.*

(ii) *h is continuous on $\mathcal{D}_h$.*

(iii) *For every $\alpha \in \mathbb{R}$ and $x \in \mathcal{D}_h$, the set $\mathcal{L}_h^\flat(x,\alpha) = \{y \in \mathcal{D}_{\partial h} : D_h^\flat(x,y) \leq \alpha\}$ is bounded.*

(iv) *For every $\alpha \in \mathbb{R}$ and $x \in \mathcal{D}_h$, if $\{y^k\} \subset \mathcal{L}_h^\flat(x,\alpha)$ is a convergent sequence with limit $y^* \in \mathcal{D}_h \setminus \{x\}$, then $D_h^\sharp(y^*, y^k) \to 0$.*

Remarks 2.5.

(i) $D_f^\flat$ and $D_f^\sharp$ are used like distances, because for $x \in \mathcal{D}_f$ and $y \in \mathcal{D}_{\partial f}$, $0 \leq D_f^\flat(x,y) \leq D_f^\sharp(x,y)$, and $D_f^\flat(x,y) = 0 \iff D_f^\sharp(x,y) = 0 \iff x = y$ by strict convexity.

(ii) Our generalization of Bregman functions [CeL81] has two components. First, we allow singularities of h at $\partial\mathcal{D}_h$ to cover more examples. Second, handling nondifferentiable functions only requires slightly more complex notation. In many cases h will be essentially smooth, so $D_h = D_h^\flat = D_h^\sharp$ on $\mathcal{D}_h \times \mathring{\mathcal{D}}_h$ (Fact 2.3(i) and (2.5)).

(iii) Several results of [Kiw97] help in checking if a given function is a B-function. We only recall the following results and examples from [Kiw97].

LEMMA 2.6 (see [Kiw97, Lem. 2.8]). *Let $h = \sum_{i=1}^k h_i$, where $h_1, \ldots, h_k$ are closed proper convex functions such that (s.t.) $h_{j+1}, \ldots, h_k$ ($j \geq 0$) are polyhedral and*

$\cap_{i=1}^{j} \operatorname{ri}(\mathcal{D}_{h_i}) \cap_{i=j+1}^{k} \mathcal{D}_{h_i} \neq \emptyset$. *If* h_1 *is a B-function,* $h_2, \ldots, h_j$ *are continuous on* $\mathcal{D}_h = \cap_{i=1}^{k} \mathcal{D}_{h_i}$ *and satisfy condition* (iv) *of Definition* 2.4, *then* h *is a B-function. In particular,* h *is a B-function if* $h_1, \ldots, h_j$ *are.*

LEMMA 2.7 (see [Kiw97, Lems. 2.10 and 2.11]).

(i) *Let* h *be a proper convex function on* $\mathbb{R}$. *Then* h *is a B-function iff* h *is closed and essentially strictly convex and* $\mathcal{D}_{h^*} = \mathring{\mathcal{D}}_{h^*}$.

(ii) *Let* $h_1, \ldots, h_n$ *be B-functions on* $\mathbb{R}$. *Then* $h(x) = \sum_{i=1}^{n} h_i(x_i)$ *is a B-function.*

Examples 2.8 (see [Kiw97, Exs. 2.12]). If $h(x) = \sum_{i=1}^{n} h_i(x_i)$ with $h_i : \mathbb{R} \to (-\infty, \infty]$, then $h^*(y) = \sum_i h_i^*(y_i)$ and $D_h(x, y) = \sum_i D_{h_i}(x_i, y_i)$, so we consider only $n = 1$. In each example, it can be verified that h is an essentially smooth B-function.

(1 [Eck93]). $h(x) = |x|^\alpha/\alpha$ on $\mathcal{D}_h = \mathbb{R}$ for $\alpha > 1$. Then $h^*(\cdot) = |\cdot|^\beta/\beta$ with $\alpha + \beta = \alpha\beta$ (cf. [Roc70, sect. 12]). For $\alpha = 2$, $h(x) = |x|^2/2$ and $D_h(x, y) = |x - y|^2/2$.

(2) $h(x) = -x^\alpha/\alpha$ on $\mathcal{D}_h = \mathbb{R}_+$ with $\alpha \in (0, 1)$. Then $h^*(y) = -(-y)^\beta/\beta$ on $\mathcal{D}_{h^*} = (-\infty, 0)$ with $\alpha + \beta = \alpha\beta$. For $h(x) = (\alpha x - x^\alpha)/(1 - \alpha)$ [Teb92], $h^*(y) = (1 + y/\beta)^\beta$ on $\mathcal{D}_h^* = (-\infty, -\beta)$. For $\alpha = \frac{1}{2}$, $D_h(x, y) = (x^{1/2} - y^{1/2})^2/y^{1/2}$.

(3 ("$x \log x$"-entropy) [Bre67]) $h(x) = x \ln x$ on $\mathcal{D}_h = \mathbb{R}_+$ ($0 \ln 0 = 0$). Then $h^*(y) = \exp(y - 1)$ on $\mathcal{D}_{h^*} = \mathbb{R}$ and $D_h(x, y) = x \ln(x/y) + y - x$ (the Kullback–Leibler entropy). For $h(x) = x \ln x - x$ (Boltzmann–Shannon), $h^*(y) = \exp y$ and D_h is the same.

(4 (Burg's entropy) [CDPI91]) $h(x) = -\ln x$ on $\mathcal{D}_h = \mathbb{R}_>$. Then $h^*(y) = -\ln(-y) - 1$ on $\mathcal{D}_{h^*} = (-\infty, 0)$, and $D_h(x, y) = -\ln(x/y) + x/y - 1$.

(5 (Hellinger)) $h(x) = -\sqrt{1 - x^2}$ on $\mathcal{D}_h = [-1, 1]$. Then $h^*(y) = \sqrt{1 + y^2}$ on $\mathcal{D}_{h^*} = \mathbb{R}$ and $D_h(x, y) = (1 - xy)/\sqrt{1 - y^2} - \sqrt{1 - x^2}$ on $[-1, 1] \times (-1, 1)$. For $h(x) = -\sqrt{x(1 - x)}$ on $\mathcal{D}_h = [0, 1]$, $h^*(y) = \frac{1}{2}(y + \sqrt{1 + y^2})$.

(6 (Fermi–Dirac)) $h(x) = x \ln x + (1 - x) \ln(1 - x)$ on $\mathcal{D}_h = [0, 1]$. Then $h^*(y) = \ln(1 + \exp y)$ on $\mathcal{D}_{h^*} = \mathbb{R}$.

(7) $h(x) = -\ln x - \ln(1 - x)$ on $\mathcal{D}_h = [0, 1]$. Then $h^*(y) = \frac{1}{2}[(y^2 + 4)^{1/2} + y - 2] + \ln[(y^2 + 4)^{1/2} - 2] - \ln y^2$ for $y \neq 0$, $h^*(0) = -\ln 4$, $\mathcal{D}_{h^*} = \mathbb{R}$.

We now provide a dual complement of Lemma 2.7(i).

LEMMA 2.9.

(i) *If* ψ *is a B-function on* $\mathbb{R}$, *then* ψ^* *is essentially smooth and* $\mathcal{D}_{\psi^*} = \mathring{\mathcal{D}}_{\psi^*}$.

(ii) *If* $\phi : \mathbb{R} \to (-\infty, \infty]$ *is closed proper convex essentially smooth and* $\mathcal{D}_\phi = \mathring{\mathcal{D}}_\phi$, *then* ϕ^* *is a B-function with* $\operatorname{ri} \mathcal{D}_{\phi^*} \subset \operatorname{im} \nabla\phi \subset \mathcal{D}_{\phi^*}$.

Proof. (i) This follows from Definition 2.4, Fact 2.3(ii), and Lemma 2.7(i). (ii) By Facts 2.1 and 2.3, $\operatorname{ri} \mathcal{D}_{\phi^*} \subset \mathcal{D}_{\partial\phi^*} = \operatorname{im} \partial\phi = \operatorname{im} \nabla\phi \subset \mathcal{D}_{\phi^*}$ and ϕ^* is closed proper essentially strictly convex with $\phi^{**} = \phi$ [Roc70, Thm. 12.2]. The conclusion follows from Lemma 2.7(i). ∎

We need the following consequence of [Kiw97, Lems. 2.15 and 2.16] formulated in terms of

$$0 \leq D'_h(x, y) := h(x) - h(y) - h'(y; x - y) \leq D_h^\flat(x, y) \quad \forall x, y \in \mathcal{D}_h. \tag{2.6}$$

LEMMA 2.10. *Let* h *be a B-function on* $\mathbb{R}^n$. *Then*

(i) *If* $\{x^k\}$ *is a sequence in* $\mathcal{L}_h^\flat(x, \alpha)$ *for some* $x \in \mathcal{D}_h$, $\alpha \in \mathbb{R}$, *then* $\{x^k\}$ *is bounded and every limit point of* $\{x^k\}$ *is in* $\mathcal{D}_h$.

(ii) *If* $\{x^k\} \subset \mathcal{D}_h$ *is bounded and* $D'_h(x, x^k) \to 0$ *for some* x, *then* $x^k \to x$.

3. The BPM method. We make the following standing assumptions about problem (1.1) and the algorithm.

Assumption 3.1.
(i) f is a closed proper convex function.
(ii) X is a nonempty closed convex set.
(iii) h is a (possibly nonsmooth) B-function.
(iv) $\mathcal{D}_{f_X} \cap \mathcal{D}_h \neq \emptyset$, where $f_X = f + \imath_X$ is the *essential objective* of (1.1).
(v) $\{c_k\}$ is a sequence of positive numbers satisfying $\sum_{k=1}^{\infty} c_k = \infty$.
(vi) $\{\epsilon_k\}$ is a sequence of nonnegative numbers satisfying

$$\lim_{l\to\infty} \sum_{k=1}^{l} c_k\epsilon_k / \sum_{k=1}^{l} c_k = 0.$$

Consider the following *inexact BPM method*. At iteration $k \geq 1$, having

$$x^k \in \mathcal{D}_{f_X} \cap \mathcal{D}_{\partial h}, \tag{3.1}$$

$$\gamma^k \in \partial h(x^k), \tag{3.2}$$

$$D_h^k(x, x^k) = h(x) - h(x^k) - \langle \gamma^k, x - x^k \rangle \quad \forall x, \tag{3.3}$$

find x^{k+1}, γ^{k+1}, and p^{k+1} satisfying

$$\gamma^{k+1} \in \partial h(x^{k+1}), \tag{3.4}$$

$$c_k p^{k+1} + \gamma^{k+1} - \gamma^k = 0, \tag{3.5}$$

$$p^{k+1} \in \partial_{\epsilon_k} f_X(x^{k+1}), \tag{3.6}$$

$$f_X(x^{k+1}) \leq f_X(x^k). \tag{3.7}$$

We note that $x^{k+1} \approx \arg\min\{f_X + D_h^k(\cdot, x^k)/c_k\}$ with

$$0 \leq D_h^\flat(\cdot, x^k) \leq D_h^k(\cdot, x^k) \leq D_h^\sharp(\cdot, x^k) \tag{3.8}$$

by (2.2), (2.3), (3.2), and (3.3); in fact x^{k+1} is an ϵ_k-minimizer of

$$\phi_k(x) = f_X(x) + D_h^k(x, x^k)/c_k, \tag{3.9}$$

as shown after the following (well-known) technical result (cf. [Roc70, Thm. 27.1]).

LEMMA 3.2. *A closed proper and strictly convex function ϕ on $\mathbb{R}^n$ has a unique minimizer iff ϕ is inf-compact, i.e., the α-level set $\mathcal{L}_\phi(\alpha) = \{x : \phi(x) \leq \alpha\}$ is bounded for any $\alpha \in \mathbb{R}$, and this holds iff $\mathcal{L}_\phi(\alpha)$ is nonempty and bounded for one $\alpha \in \mathbb{R}$.*

Proof. If $x \in \operatorname{Arg\,min} \phi$, then by strict convexity of ϕ, $\mathcal{L}_\phi(\phi(x)) = \{x\}$ is bounded, so ϕ is inf-compact (cf. [Roc70, Cor. 8.7.1]). If for some $\alpha \in \mathbb{R}$, $\mathcal{L}_\phi(\alpha) \neq \emptyset$ is bounded, then it is closed (cf. [Roc70, Thm. 7.1]) and contains $\operatorname{Arg\,min} \phi \neq \emptyset$ because ϕ is closed. ∎

LEMMA 3.3. *Under the above assumptions, we have the following:*
(i) *ϕ_k is closed proper and strictly convex.*
(ii) *$\phi_k(x^{k+1}) \leq \inf \phi_k + \epsilon_k$ (i.e., $0 \in \partial_{\epsilon_k}\phi_k(x^{k+1})$).*
(iii) *If $f_* = \inf_X f > -\infty$, then ϕ_k is inf-compact.*
(iv) *ϕ_k is inf-compact if $(\gamma^k - c_k \operatorname{im} \partial f_X) \cap \operatorname{im} \partial h \neq \emptyset$, where $\operatorname{im} \partial h = \mathring{\mathcal{D}}_{h^*}$, so that $\operatorname{im} \partial h = \mathbb{R}^n$ iff h is cofinite. In particular, ϕ_k is inf-compact if $(\gamma^k - c_k \operatorname{ri} \mathcal{D}_{f_X^*}) \cap \operatorname{ri} \mathcal{D}_{h^*} \neq \emptyset$.*

(v) *If ϕ_k is inf-compact and either* $\operatorname{ri}\mathcal{D}_{f_X}\cap\operatorname{ri}\mathcal{D}_h\neq\emptyset$, *or* $\mathcal{D}_{f_X}\cap\operatorname{ri}\mathcal{D}_h\neq\emptyset$ *and f_X is polyhedral, then there exist* $\hat{x}^{k+1}=\arg\min\phi_k$, $\hat{p}^{k+1}\in\partial f_X(\hat{x}^{k+1})$, *and* $\hat{\gamma}^{k+1}\in\partial h(\hat{x}^{k+1})$ *s.t.* $f_X(\hat{x}^{k+1})+D_h^k(\hat{x}^{k+1},x^k)/c_k\le f_X(x^k)$ *and* $c_k\hat{p}^{k+1}+\hat{\gamma}^{k+1}-\gamma^k=0$; *also* $\hat{x}^{k+1}\in\mathring{\mathcal{D}}_h$ *if* $\mathcal{D}_{\partial f_X}\subset\mathring{\mathcal{D}}_h$ *or* $\mathcal{D}_{\partial h}=\mathring{\mathcal{D}}_h$; *e.g., h is essentially smooth.*

(vi) *The assumptions of* (v) *hold if either* $\operatorname{ri}\mathcal{D}_{f_X}\subset\mathring{\mathcal{D}}_h$ *and* $\inf_X f>-\infty$, *or* $\mathcal{D}_{\partial f_X}\subset\mathring{\mathcal{D}}_h$ *and* $\operatorname{im}\partial h=\mathbb{R}^n$.

Proof. (i) Since f, $\imath_X$, and h are closed proper convex, so are $f_X=f+\imath_X$, $D_h^k(\cdot,x^k)$ and $\phi_k=f_X+D_h^k(\cdot,x^k)/c_k$ (cf. [Roc70, Thm. 9.3]), having nonempty domains $\mathcal{D}_f\cap X$, $\mathcal{D}_h$, and $\mathcal{D}_{f_X}\cap\mathcal{D}_h$ respectively (cf. Assumption 3.1(iv)). $D_h^k(\cdot,x^k)$ and ϕ_k are strictly convex, since so is h (cf. Definition 2.4(i)).

(ii) For any x, add the inequality (cf. (3.3), (3.4)) $D_h^k(x,x^k)\ge D_h^k(x^{k+1},x^k)+\langle\gamma^{k+1}-\gamma^k,x-x^{k+1}\rangle$ divided by c_k to $f_X(x)\ge f_X(x^{k+1})+\langle p^{k+1},x-x^{k+1}\rangle-\epsilon_k$ (cf. (3.6)) and use (3.5) to get $\phi_k(x)\ge\phi_k(x^{k+1})-\epsilon_k$.

(iii) By part (i), $\psi=D_h^k(\cdot,x^k)$ is closed proper strictly convex, and $\mathcal{L}_\psi(0)=\{x^k\}$ by strict convexity of h (cf. Definition 2.4(i), (2.3), and (2.6)), so ψ is inf-compact (cf. Lemma 3.2). Let $\beta=\inf\phi_k$. Since $\psi\ge0$ (cf. (3.8)), $\beta\ge f_*$ and $\emptyset\neq\mathcal{L}_{\phi_k}(\beta+1)\subset\mathcal{L}_\psi(c_k(\beta-f_*+1))$ (cf. (3.9)). The last set is bounded, since ψ is inf-compact, so ϕ_k is inf-compact by part (i) and Lemma 3.2.

(iv) Let $\hat{y}\in\mathcal{D}_{\partial f_X}$, $\hat{\gamma}\in\partial f_X(\hat{y})$, $\tilde{x}\in\mathcal{D}_{\partial h}$, and $\tilde{\gamma}\in\partial h(\tilde{x})$ satisfy $\gamma^k-c_k\hat{\gamma}=\tilde{\gamma}$. Then $\tilde{\psi}(\cdot)=f_X(\hat{y})+\langle\hat{\gamma},\cdot-\hat{y}\rangle+D_h^k(\cdot,x^k)/c_k$ is closed proper and strictly convex (as is $D_h^k(\cdot,x^k)$; cf. part (i)), and $\tilde{x}=\arg\min\tilde{\psi}$ because $0\in\partial\tilde{\psi}(\tilde{x})=\hat{\gamma}+(\partial h(\tilde{x})-\gamma^k)/c_k$ (cf. [Roc70, Thm. 23.8]). Hence $\tilde{\psi}$ is inf-compact (cf. Lemma 3.2), and so is ϕ_k, since $\phi_k\ge\tilde{\psi}$ from $f_X(\cdot)\ge f_X(\hat{y})+\langle\hat{\gamma},\cdot-\hat{y}\rangle$. By strict convexity of h (cf. Definition 2.4(i)) and Facts 2.1 and 2.3, $\operatorname{im}\partial h=\mathcal{D}_{\partial h^*}=\mathring{\mathcal{D}}_{h^*}$. Of course, $\mathring{\mathcal{D}}_{h^*}=\mathbb{R}^n$ iff $\mathcal{D}_{h^*}=\mathbb{R}^n$, i.e., iff h is cofinite. The second assertion follows from $\operatorname{ri}\mathcal{D}_{f_X^*}\subset\mathcal{D}_{\partial f_X^*}=\operatorname{im}\partial f_X$.

(v) By part (i) and Lemma 3.2, $\hat{x}^{k+1}=\arg\min\phi_k$ is well defined. The rest follows from $D_h^k(\cdot,x^k)\ge0$ (cf. (3.8)), the fact $0\in\partial\phi_k(\hat{x}^{k+1})=\partial f(\hat{x}^{k+1})+c_k[\partial h(\hat{x}^{k+1})-\gamma^k]$ due to our assumptions on $\mathcal{D}_{f_X}$ and $\operatorname{ri}\mathcal{D}_h$ (cf. [Roc70, Thm. 23.8]), and Fact 2.3(i).

(vi) If $\inf_X f>-\infty$ or $\operatorname{im}\partial h=\mathbb{R}^n$, then ϕ_k is inf-compact by parts (iii) and (iv). If $\operatorname{ri}\mathcal{D}_{f_X}\subset\mathring{\mathcal{D}}_h$ then $\operatorname{ri}\mathcal{D}_{f_X}\cap\operatorname{ri}\mathcal{D}_h=\operatorname{ri}\mathcal{D}_{f_X}\neq\emptyset$, since $\mathcal{D}_{f_X}\neq\emptyset$ (cf. Assumption 3.1(iv)). □

Remark 3.4. Lemma 3.3(v) and (vi) state conditions under which the *exact* BPM method (with $x^{k+1}=\hat{x}^{k+1}=\arg\min\phi_k$ and $\epsilon_k=0$ in (3.6)) is well defined. Our conditions are slightly weaker than those in [Eck93, Thm. 5], which correspond to $\operatorname{ri}\mathcal{D}_{f_X}\subset\mathring{\mathcal{D}}_h$, and either $\operatorname{cl}\mathcal{D}_{f_X}\subset\mathring{\mathcal{D}}_h$ and $\operatorname{im}\partial h=\mathbb{R}^n$ or, f being finite, continuous and bounded below on X.

Example 3.5. Let $X=\{x\ge0:Ax=b\}$, $f=\langle\hat{c},\cdot\rangle+\imath_X$, and $h(x)=-\sum_{i=1}^n\ln x_i$, where $A\in\mathbb{R}^{m\times n}$, $b\in\mathbb{R}^m$, and $\hat{c}\in\mathbb{R}^n$. Suppose $f_*>-\infty$ and $Ax=b$ for some $x>0$. Since $\mathring{\mathcal{D}}_h=\{x:x>0\}$, Lemma 3.3(iii) and (v) implies that $\hat{x}^{k+1}$ is well defined.

Example 3.6. Let $n=1$, $X=\mathbb{R}$, $f(x)=-x$, and $h(x)=e^{-x}+x$. Then $f^*=\imath_{\{-1\}}$, $\operatorname{ri}\mathcal{D}_{f^*}=\operatorname{im}\partial f=\{-1\}$, $\mathring{\mathcal{D}}_{h^*}=\operatorname{im}\partial h=(-\infty,1)$, and $\operatorname{ri}\mathcal{D}_{f^*}\cap\mathring{\mathcal{D}}_{h^*}\neq\emptyset$. Clearly, $\phi_k(x)=e^{-x}+x(e^{-x^k}-1)+\text{const}$ for $c_k=1$, so $\arg\min\phi_k\neq\emptyset$ iff $x^k<0$. Although h is not a Bregman function, this is a counterexample to [Teb92, Thm. 3.1].

4. Convergence of the BPM method. We first derive a global convergence rate estimate for the BPM method. We follow the analysis of [ChT93], which generalized that in [Gül91]. Let $s_k=\sum_{j=1}^k c_j$ for all k.

LEMMA 4.1. *For all $x \in \mathcal{D}_h$ and $k \le l$, we have*

$$D_h^{k+1}(x,x^{k+1}) + D_h^k(x^{k+1},x^k) - D_h^k(x,x^k) = \langle \gamma^k - \gamma^{k+1}, x - x^{k+1}\rangle \le c_k[f_X(x) - f_X(x^{k+1})] + c_k\epsilon_k, \tag{4.1}$$

$$D_h^{k+1}(x,x^{k+1}) \le D_h^k(x,x^k) - D_h^k(x^{k+1},x^k) + c_k\epsilon_k \quad \text{if} \quad f_X(x) \le f_X(x^{k+1}), \tag{4.2}$$

$$s_l[f_X(x^{l+1}) - f_X(x)] \le D_h^1(x,x^1) - D_h^{l+1}(x,x^{l+1}) - \sum_{k=1}^{l} D_h^k(x^{k+1},x^k) + \sum_{k=1}^{l} c_k\epsilon_k, \tag{4.3}$$

$$f_X(x^{l+1}) - f_X(x) \le D_h^1(x,x^1)/s_l + \sum_{k=1}^{l} c_k\epsilon_k/s_l. \tag{4.4}$$

Proof. The equality in (4.1) follows from (3.3), and the inequality from $\gamma^k - \gamma^{k+1} = c_k p^{k+1}$ (cf. (3.5)) and $p^{k+1} \in \partial_{\epsilon_k} f_X(x^{k+1})$ (cf. (3.6)), i.e., $\langle p^{k+1}, x - x^{k+1}\rangle \le f_X(x) - f_X(x^{k+1}) + \epsilon_k$, since $c_k > 0$. (4.2) is a consequence of (4.1). Summing (4.1) over $k = 1{:}\,l$ we obtain

$$D_h^{l+1}(x,x^{l+1}) - D_h^1(x,x^1) + \sum_{k=1}^{l} D_h^k(x^{k+1},x^k) \le s_l f_X(x) - \sum_{k=1}^{l} c_k f_X(x^{k+1}) + \sum_{k=1}^{l} c_k\epsilon_k. \tag{4.5}$$

Use $f_X(x^{l+1}) \le f_X(x^k)$ for $k = 1{:}\,l$ (cf. (3.7)) in (4.5) to get (4.3). Equation (4.4) follows from (4.3) and the fact $D_h^k(\cdot,x^k) \ge 0$ for all k (cf. (3.8)). ∎

By (2.6) and (3.8), for all k

$$0 \le D_h'(\cdot,x^k) \le D_h^\flat(\cdot,x^k) \le D_h^k(\cdot,x^k) \le D_h^\sharp(\cdot,x^k). \tag{4.6}$$

LEMMA 4.2. *If $\sum_{k=1}^{\infty} c_k\epsilon_k < \infty$ and $x \in \mathcal{D}_h$ is s.t. $f_X(x^k) \ge f_X(x)$ for all k, then*

(i) *$\{x^k\}$ is bounded and $\{x^k\} \subset \mathcal{L}_h^\flat(x,\alpha)$, where $\alpha = D_h^1(x,x^1) + \sum_{k=1}^{\infty} c_k\epsilon_k$.*

(ii) *every limit point of $\{x^k\}$ is in $\mathcal{D}_h$.*

(iii) *$\{x^k\}$ converges to some $x^\infty \in \mathcal{D}_{f_X} \cap \mathcal{D}_h$ s.t. $f_X(x^k) \ge f_X(x^\infty)$ for all k.*

Proof. (i) We have $D_h^l(x,x^l) \le D_h^1(x,x^1) + \sum_{k=1}^{l-1} c_k\epsilon_k \le \alpha$ for all l (cf. (4.2), (4.6)) and $\{x^k\} \subset \mathcal{D}_{\partial h}$ (cf. (3.1)), so $\{x^k\} \subset \mathcal{L}_h^\flat(x,\alpha)$, a bounded set (cf. Definition 2.4(iii)).

(ii) Since $\{x^k\} \subset \mathcal{L}_h^\flat(x,\alpha)$ by part (i), this follows from Lemma 2.10(i).

(iii) By parts (i) and (ii), a subsequence $\{x^{l_j}\}$ converges to some $x^\infty \in \mathcal{D}_h$. Suppose $x^\infty \ne x$. Since $\{x^k\} \subset \mathcal{L}_h^\flat(x,\alpha)$, $D_h^\sharp(x^\infty,x^{l_j}) \to 0$ (cf. Definition 2.4(iv)) and $D_h^{l_j}(x^\infty,x^{l_j}) \to 0$ (cf. (4.6)). But $f_X(x^k) \ge f_X(x^\infty)$ for all k, since $x^{l_j} \to x^\infty$, $f_X(x^{k+1}) \le f_X(x^k)$ (cf. (3.7)) and f_X is closed (cf. Assumption 3.1(i), (ii)). Hence for $l > l_j$, $D_h^l(x^\infty,x^l) \le D_h^{l_j}(x^\infty,x^{l_j}) + \sum_{k=l_j}^{l-1} c_k\epsilon_k$ (cf. (4.2)) with $\sum_{k=l_j}^{\infty} c_k\epsilon_k \to 0$ as $j \to \infty$ yield $D_h^l(x^\infty,x^l) \to 0$ as $l \to \infty$. Thus $D_h'(x^\infty,x^k) \to 0$ (cf. (4.6)) and $x^k \to x^\infty$ by Lemma 2.10(ii). Finally, if $x^\infty = x$ but $\{x^k\}$ does not converge, it has a limit point $x' \in \mathcal{D}_h \setminus \{x\}$ (cf. parts (i) and (ii)), and replacing x^∞ by x' in the preceding argument yields a contradiction. ∎

We may now prove our main result for the inexact BPM descent method (3.1)–(3.7).

THEOREM 4.3. *Suppose that Assumption* 3.1(i), (ii), (iv), (v) *holds with h closed proper convex.*

(i) *If* $\lim_{l\to\infty} \sum_{k=1}^{l} c_k\epsilon_k / \sum_{k=1}^{l} c_k = 0$, *then*

$$f_X(x^k) \downarrow \inf_{\mathcal{D}_h} f_X = \inf_{\mathrm{cl}(\mathcal{D}_h \cap \mathcal{D}_{f_X})} f.$$

Hence $f_X(x^k) \downarrow \inf_X f$ *if* $\mathcal{D}_{f_X} \subset \mathcal{D}_h$. *If* $\mathrm{ri}\,\mathcal{D}_h \cap \mathrm{ri}\,\mathcal{D}_{f_X} \neq \emptyset$ (*e.g.*, $\mathring{\mathcal{D}}_h \cap \mathcal{D}_{f_X} \neq \emptyset$), *then* $\inf_{\mathcal{D}_h} f_X = \inf_{(\mathrm{cl}\,\mathcal{D}_h)\cap(\mathrm{cl}\,\mathcal{D}_{f_X})} f = \inf_{\mathrm{cl}\,\mathcal{D}_h} f_X$. *If* $\mathrm{ri}\,\mathcal{D}_{f_X} \subset \mathrm{cl}\,\mathcal{D}_h$ (*e.g.*, $\mathcal{D}_{\partial f_X} \subset \mathrm{cl}\,\mathcal{D}_h$), *then* $\mathrm{cl}\,\mathcal{D}_h \supset \mathrm{cl}\,\mathcal{D}_{f_X}$ *and* $\mathrm{Arg\,min}_X f \subset \mathrm{cl}\,\mathcal{D}_h$.

(ii) *If h is a B-function,* $f_X(x^k) \to \inf_{\mathcal{D}_h} f_X$, $\sum_{k=1}^{\infty} c_k\epsilon_k < \infty$, *and* $X_* = \mathrm{Arg\,min}_{\mathcal{D}_h} f_X$ *is nonempty, then* $\{x^k\}$ *converges to some* $x^\infty \in X_*$, *and*

$$x^\infty \in \mathrm{Arg}\min_X f \qquad \textit{if} \qquad \mathcal{D}_{f_X} \subset \mathcal{D}_h.$$

(iii) *If* $f_X(x^k) \to \inf_{\mathcal{D}_h} f_X$, $\mathcal{D}_{f_X} \subset \mathcal{D}_h$, *and* $X_* = \emptyset$, *then* $|x^k| \to \infty$.

Proof. (i) For any $x \in \mathcal{D}_h$, taking the limit in (4.4) yields $\lim_{l\to\infty} f_X(x^l) \leq f_X(x)$, using $f_X(x^{l+1}) \leq f_X(x^l)$ (cf. (3.7)), $s_l \to \infty$ (Assumption 3.1(v)), and $\sum_{k=1}^{l} c_k\epsilon_k/s_l \to 0$. Hence $f_X(x^k) \to \inf_{\mathcal{D}_h} f_X = \inf_{\mathcal{D}_h\cap\mathcal{D}_{f_X}} f = \inf_{\mathrm{cl}(\mathcal{D}_h\cap\mathcal{D}_{f_X})} f$ (cf. [Roc70, Cor. 7.3.2]). If $\mathrm{ri}\,\mathcal{D}_h \cap \mathrm{ri}\,\mathcal{D}_{f_X} \neq \emptyset$ (e.g., $\mathring{\mathcal{D}}_h \cap \mathcal{D}_{f_X} \neq \emptyset$; cf. [Roc70, Cor. 6.3.2]), then $\mathrm{cl}(\mathcal{D}_h \cap \mathcal{D}_{f_X}) = \mathrm{cl}(\mathcal{D}_h) \cap \mathrm{cl}(\mathcal{D}_{f_X})$ (cf. [Roc70, Thm. 6.5]) and $\inf_{\mathcal{D}_h} f_X = \inf_{(\mathrm{cl}\,\mathcal{D}_h)\cap(\mathrm{cl}\,\mathcal{D}_{f_X})} f \leq \inf_{\mathcal{D}_{f_X}\cap\mathrm{cl}\,\mathcal{D}_h} f = \inf_{\mathrm{cl}\,\mathcal{D}_h} f_X$, so $\inf_{\mathcal{D}_h} f_X = \inf_{\mathrm{cl}\,\mathcal{D}_h} f_X$. If $\mathrm{ri}\,\mathcal{D}_{f_X} \subset \mathrm{cl}\,\mathcal{D}_h$, then $\mathrm{cl}\,\mathcal{D}_{f_X} \subset \mathrm{cl}\,\mathcal{D}_h$ (cf. [Roc70, Thm. 6.5]).

(ii) If $x \in X_*$, then $f_X(x^k) \to f_X(x)$. But $f_X(x^k) \geq f_X(x)$ for all k (cf. (3.1)), so $x^k \to x^\infty \in \mathcal{D}_{f_X} \cap \mathcal{D}_h$ and $\lim_{k\to\infty} f_X(x^k) \geq f_X(x^\infty)$ by Lemma 4.2, and thus $x^\infty \in X_*$.

(iii) If $|x^k| \not\to \infty$, $\{x^k\}$ has a limit point x with $f_X(x) \leq \inf_{\mathcal{D}_h} f_X \leftarrow f_X(x^k)$ (f_X is closed; cf. Assumption 3.1(i), (ii)), so $\mathcal{D}_{f_X} \subset \mathcal{D}_h$ yields $x \in \mathcal{D}_h \cap X_*$, i.e., $X_* \neq \emptyset$. ∎

Remark 4.4. For the exact BPM method (with $\epsilon_k \equiv 0$), Theorem 4.3(i), (ii) subsume [ChT93, Thm. 3.4], which assumes $\mathrm{ri}\,\mathcal{D}_{f_X} \subset \mathring{\mathcal{D}}_h$ and $\mathcal{D}_h = \mathrm{cl}\,\mathcal{D}_h$. Theorem 4.3(ii), (iii) strengthen [Eck93, Thm. 5], which shows only that $\{x^k\}$ is unbounded if $\mathrm{cl}\,\mathcal{D}_{f_X} \subset \mathring{\mathcal{D}}_h$ and $X_* = \emptyset$. Theorem 4.3(i), (ii) and Lemma 3.3 subsume [Ius95, Thm. 4.1], which assumes that h is essentially smooth, f is continuous on $\mathcal{D}_f$, $\mathcal{D}_f \cap \mathring{\mathcal{D}}_h \neq \emptyset$, $X = \mathrm{cl}\,\mathcal{D}_h$, $\mathrm{Arg\,min}_X f \neq \emptyset$, and $\inf_k c_k > 0$.

For choosing $\{\epsilon_k\}$ (cf. Assumption 3.1(vi)), one may use the following simple result.

LEMMA 4.5.

(i) *If* $\epsilon_k \to 0$, *then* $\sum_{k=1}^{l} c_k\epsilon_k/s_l \to 0$ *as* $l \to \infty$.

(ii) *If* $\sum_{k=1}^{\infty} \epsilon_k < \infty$ *and* $\{c_k\} \subset (0, c_{\max}]$ *for some* $c_{\max} < \infty$, *then* $\sum_{k=1}^{\infty} c_k\epsilon_k < \infty$.

Proof. (i) For any $\epsilon > 0$, pick $\bar{k}$ and $\bar{l} > \bar{k}$ s.t. $\epsilon_k \leq \epsilon$ for all $k \geq \bar{k}$ and $\sum_{k=1}^{\bar{k}} c_k\epsilon_k/s_l \leq \epsilon$ for all $l \geq \bar{l}$; then

$$\sum_{k=1}^{l} c_k\epsilon_k/s_l \leq \sum_{k=1}^{\bar{k}} c_k\epsilon_k/s_l + \epsilon \sum_{k=\bar{k}+1}^{l} c_k / \sum_{k=1}^{l} c_k \leq 2\epsilon$$

for all $l \geq \bar{l}$.

(ii) We have $\sum_{k=1}^{\infty} c_k\epsilon_k \leq c_{\max} \sum_{k=1}^{\infty} \epsilon_k < \infty$. ∎

5. Convergence of a nondescent BPM method. In certain applications (cf. section 7) it may be difficult to satisfy the descent requirement (3.7). Hence we now consider a *nondescent BPM method*, in which (3.7) is replaced by

$$f_X(x^{k+1}) + D_h^k(x^{k+1}, x^k)/c_k \le f_X(x^k) + \epsilon_k. \tag{5.1}$$

By Lemma 3.3(ii), (5.1) holds *automatically*, since it means $\phi_k(x^{k+1}) \le \phi_k(x^k) + \epsilon_k$.

LEMMA 5.1. *For all $x \in \mathcal{D}_h$ and $k \le l$, we have*

$$f_X(x^{k+1}) \le f_X(x^k) + \epsilon_k, \tag{5.2}$$

$$s_l[f_X(x^{l+1}) - f_X(x)] \le D_h^1(x, x^1) - D_h^{l+1}(x, x^{l+1}) - \sum_{k=1}^{l}(s_k/c_k)D_h^k(x^{k+1}, x^k) + \sum_{k=1}^{l} s_k\epsilon_k, \tag{5.3}$$

$$f_X(x^{l+1}) - f_X(x) \le D_h^1(x, x^1)/s_l + \sum_{k=1}^{l} s_k\epsilon_k/s_l. \tag{5.4}$$

Proof. Equations (4.1) and (4.2) still hold. Equation (5.2) follows from $D_h^k(x^{k+1}, x^k) \ge 0$ (cf. (3.8)) and (cf. (5.1)) $f_X(x^k) - f_X(x^{k+1}) \ge D_h^k(x^{k+1}, x^k)/c_k - \epsilon_k$. Multiplying this inequality by $s_{k-1} = s_k - c_k$ and summing over $k = 1{:}\, l$ yield

$$-s_l f_X(x^{l+1}) + \sum_{k=1}^{l} c_k f_X(x^{k+1}) \ge \sum_{k=1}^{l}(s_{k-1}/c_k)D_h^k(x^{k+1}, x^k) - \sum_{k=1}^{l} s_{k-1}\epsilon_k. \tag{5.5}$$

Subtract (5.5) from (4.5) and rearrange, using $s_k = s_{k-1} + c_k$, to get (5.3). Equation (5.4) follows from (5.3) and the fact $D_h^k(\cdot, x^k) \ge 0$ for all k (cf. (3.8)). ∎

THEOREM 5.2. *Suppose that Assumption* 3.1(i)–(ii), (iv)–(v) *holds with h closed proper convex.*

(i) *If $\sum_{k=1}^{l} s_k\epsilon_k/s_l \to 0$ (see Lemma* 5.3 *for sufficient conditions), then*

$$f_X(x^k) \to \inf_{\mathcal{D}_h} f_X.$$

Hence the assertions of Theorem 4.3(i) *hold.*

(ii) *If h is a B-function, $f_X(x^k) \to \inf_{\mathcal{D}_h} f_X$, $\sum_{k=1}^{\infty} c_k\epsilon_k < \infty$, and $X_* = \operatorname{Arg\,min}_{\mathcal{D}_h} f_X$ is nonempty then $\{x^k\}$ converges to some $x^\infty \in X_*$, and*

$$x^\infty \in \operatorname*{Arg\,min}_X f \qquad \text{if} \qquad \mathcal{D}_{f_X} \subset \mathcal{D}_h.$$

(iii) *If $f_X(x^k) \to \inf_{\mathcal{D}_h} f_X$, $\mathcal{D}_{f_X} \subset \mathcal{D}_h$, and $X_* = \emptyset$, then $|x^k| \to \infty$.*

Proof. (i) The upper limit in (5.4) for any $x \in \mathcal{D}_h$ yields $\limsup_{l\to\infty} f_X(x^l) \le \inf_{\mathcal{D}_h} f_X$, using $\sum_{k=1}^{l} s_k\epsilon_k/s_l \to 0$. But $\{x^k\} \subset \mathcal{D}_h$ (cf. (3.1)), so $\liminf_{l\to\infty} f_X(x^l) \ge \inf_{\mathcal{D}_h} f_X$.

(ii) If $x \in X_*$, then $f_X(x^k) \to f_X(x)$ and $f_X(x^k) \ge f_X(x)$ for all k (cf. (3.1)). Assertions (i)–(iii) of Lemma 4.2 still hold, since the proofs of (i)–(ii) remain valid, whereas in the proof of (iii) we have $x^\infty \in \mathcal{D}_h$ and $f_X(x^\infty) \le \lim_{j\to\infty} f_X(x^{l_j}) = f_X(x)$ (f_X is closed), so $x^\infty \in X_*$ and $f_X(x^k) \ge f_X(x^\infty)$ for all k as before yield $x^k \to x^\infty$.

(iii) Use the proof of Theorem 4.3(iii). ∎

LEMMA 5.3.

(i) *Let $\{\alpha_k\}$, $\{\beta_k\}$, and $\{\varepsilon_k\}$ be sequences in $\mathbb{R}$ s.t. $0 \le \alpha_{k+1} \le (1-\beta_k)\alpha_k + \varepsilon_k$, $\alpha_1 \ge 0$, $0 < \beta_k \le 1$, $\varepsilon_k \ge 0$ for $k = 1, 2, \ldots$, $\sum_{k=1}^{\infty} \beta_k = \infty$, and $\lim_{k\to\infty} \varepsilon_k/\beta_k = 0$. Then $\lim_{k\to\infty} \alpha_k = 0$.*

(ii) *If* $\sum_{l=1}^{\infty} c_l/s_l = \infty$ *and* $\lim_{k\to\infty} \epsilon_k s_k/c_k = 0$, *then* $\lim_{l\to\infty} \sum_{k=1}^{l} s_k\epsilon_k/s_l = 0$.

(iii) *If* $\{c_k\} \subset [c_{\min}, c_{\max}]$ *for some* $0 < c_{\min} \le c_{\max}$ *and* $k\epsilon_k \to 0$, *then* $\sum_{k=1}^{l} s_k\epsilon_k/s_l \to 0$.

Proof. (i) See, e.g., [Pol83, Lem. 2.2.3].

(ii) Use part (i) with $\alpha_l = \sum_{k=1}^{l} s_k\epsilon_k/s_l$, $s_l = \sum_{k=1}^{l} c_k$, and $\alpha_{l+1} = (1 - c_{l+1}/s_{l+1})\alpha_l + \epsilon_{l+1}$.

(iii) Use part (ii) with $c_l/s_l \in [c_{\min}/lc_{\max}, c_{\max}/lc_{\min}]$ for all l. □

6. Finite termination for sharp minima. We now extend to the exact BPM method the finite convergence property of the PPA in the case of sharp minima (cf. [Fer91, Roc76b, BuF93]).

THEOREM 6.1. *Let* f *have a sharp minimum on* X, *i.e.,* $X_* = \operatorname{Arg\,min}_X f \ne \emptyset$, *and there exists* $\alpha > 0$ *s.t.* $f_X(x) \ge \min_X f + \alpha \min_{y\in X_*} |x - y|$ *for all* x. *Consider the exact BPM method applied to* (1.1) *with a B-function* h *s.t.* $\mathcal{D}_{f_X} \subset \mathcal{D}_{\nabla h}$, $\epsilon_k \equiv 0$, *and* $\inf_k c_k > 0$. *Then there exists* k *s.t.* $p^k = 0$ *and* $x^k \in X_*$.

Proof. By Theorem 4.3, $x^k \to x^\infty \in X_*$, so $x^\infty \in \mathcal{D}_{\nabla h}$, $\gamma^k = \nabla h(x^k) \to \nabla h(x^\infty)$ (cf. (3.2) and Fact 2.2) and $\partial f_X(x^k) \ni p^k = (\gamma^{k-1} - \gamma^k)/c_{k-1} \to 0$ (cf. (3.5), (3.6)). But if $x \notin X_*$ and $p \in \partial f_X(x)$, then $|p| \ge \alpha$ (cf. [Ber82, sect. 5.4]) (since for $y = \arg\min_{y\in X_*} |x - y|$, $\min_X f = f_X(y) \ge f_X(x) + \langle p, y - x\rangle$ yields $|p||x - y| \ge \langle p, x - y\rangle \ge \alpha|x - y|$). Hence for some k, $|p^k| < \alpha$ implies $p^k = 0$ and $x^k \in X_*$. □

We note that piecewise linear programs have sharp minima, if any (cf. [Ber82, section 5.4]).

7. Inexact multiplier methods. Following [Eck93, Teb92], this section considers the application of the BPM method to dual formulations of convex programs of the form presented in [Roc70, sect. 28]:

$$\text{minimize} \quad f(x) \quad \text{subject to} \quad g_i(x) \le 0,\ i = 1\colon m, \tag{7.1}$$

under the following

Assumption 7.1. f, $g_1, \dots, g_m$ are closed proper convex functions on $\mathbb{R}^n$ with $\mathcal{D}_f \subset \bigcap_{i=1}^m \mathcal{D}_{g_i}$ and $\operatorname{ri}\mathcal{D}_f \subset \bigcap_{i=1}^m \operatorname{ri}\mathcal{D}_{g_i}$.

Letting $g(\cdot) = (g_1(\cdot), \dots, g_m(\cdot))$, we define the *Lagrangian* of (7.1):

$$L(x,\pi) = \begin{cases} f(x) + \langle \pi, g(x)\rangle & \text{if } x \in \mathcal{D}_f \text{ and } \pi \in \mathbb{R}^m_+, \\ -\infty & \text{if } x \in \mathcal{D}_f \text{ and } \pi \notin \mathbb{R}^m_+, \\ \infty & \text{if } x \notin \mathcal{D}_f, \end{cases}$$

and the *dual functional* $d(\pi) = \inf_x L(x,\pi)$. Then $d(\pi) = -\infty$ if $\pi \notin \mathbb{R}^m_+$. Assume that $d(\pi) > -\infty$ for some π. The *dual problem* to (7.1) is to maximize d, or equivalently to minimize $q(\pi)$ over $\pi \ge 0$, where $q = -d$ is a closed proper convex function. We will apply the BPM method to this problem, using some B-function h on $\mathbb{R}^m$.

We assume that $\mathbb{R}^m_> \subset \mathcal{D}_h$, so that $h_+ = h + \iota_{\mathbb{R}^m_+}$ is a B-function (cf. Lemma 2.6). The *monotone conjugate* of h (cf. [Roc70, p. 111]) defined by $h^+(\cdot) = \sup_{\pi\ge 0}\{\langle \pi, \cdot\rangle - h(\pi)\}$ is *nondecreasing* (i.e., $h^+(u) \le h^+(u')$ if $u \le u'$, since $\langle \pi, u\rangle \le \langle \pi, u'\rangle$ $\forall \pi \ge 0$) and coincides with the convex conjugate h^*_+ of h_+, since $h^+(\cdot) = \sup_\pi\{\langle \pi, \cdot\rangle - h_+(\pi)\} = h^*_+(\cdot)$. We need the following variation on [Eck93, Lem. A3]. Its proof is given in the appendix.

LEMMA 7.2. *If* h *is a closed proper essentially strictly convex function on* $\mathbb{R}^m$ *with* $\mathbb{R}^m_+ \cap \operatorname{ri}\mathcal{D}_h \ne \emptyset$, *then* h^+ *is closed proper convex and essentially smooth,* $\partial h^+(u) = \{\nabla h^+(u)\}$ *for all* $u \in \mathcal{D}_{\partial h^+}$, $\partial h^+ = (\partial h_+)^{-1}$, *and* ∇h^+ *is continuous on* $\mathcal{D}_{\partial h^+} =$

$\mathring{\mathcal{D}}_{h^+} = \operatorname{im}\partial h_+$. *Further,* $\mathcal{D}_{h^+} = \mathcal{D}_{h^+} - \mathbb{R}^m_+$, $\mathring{\mathcal{D}}_{h^+} = \mathring{\mathcal{D}}_{h^+} - \mathbb{R}^m_+$, $\partial h_+ = \partial h + N_{\mathbb{R}^m_+}$, *and* $\nabla h^+ = \nabla h^+ \circ (I + N_{\mathbb{R}^m_+} \circ \nabla h^+)$, *where* I *is the identity operator and* $N_{\mathbb{R}^m_+} = \partial \imath_{\mathbb{R}^m_+}$ *is the normal cone operator of* $\mathbb{R}^m_+$, *i.e.,* $N_{\mathbb{R}^m_+}(\pi) = \{\gamma \le 0 : \langle \gamma, \pi\rangle = 0\}$ *if* $\pi \ge 0$, $N_{\mathbb{R}^m_+}(\pi) = \emptyset$ *if* $\pi \not\ge 0$. *If additionally* $\operatorname{im}\partial h \supset \mathbb{R}^m_>$, *then* h_+ *is cofinite,* $\mathcal{D}_{h^+} = \mathbb{R}^m$, *and* h^+ *is continuously differentiable.*

Since $\mathbb{R}^m_> \subset \mathcal{D}_{h_+} \subset \mathbb{R}^m_+$, to find $\inf_{\pi\ge 0} q(\pi)$ via the BPM method we replace in (3.1)–(3.6) f, X, h, and x^k by q, $\mathbb{R}^m$, h_+, and π^k, respectively. Given $\pi^k \in \mathcal{D}_q \cap \mathcal{D}_{\partial h_+}$ and $\gamma^k \in \partial h_+(\pi^k)$, our *inexact multiplier method* requires finding π^{k+1} and x^{k+1} s.t.

$$L(x^{k+1}, \pi^{k+1}) \le \inf_x L(x, \pi^{k+1}) + \epsilon_k = d(\pi^{k+1}) + \epsilon_k, \tag{7.2}$$

$$\pi^{k+1} = \nabla h^+(\gamma^k + c_k g(x^{k+1})) \tag{7.3}$$

with

$$p^{k+1} \in \partial_{\epsilon_k} q(\pi^{k+1}), \tag{7.4}$$

$$\gamma^{k+1} = \gamma^k - c_k p^{k+1} \in \partial h_+(\pi^{k+1}) \tag{7.5}$$

for some p^{k+1} and γ^{k+1}. Note that (7.2) implies

$$-g(x^{k+1}) \in \partial_{\epsilon_k} q(\pi^{k+1}) = \partial_{\epsilon_k} q(\pi^{k+1}) + \partial \imath_{\mathbb{R}^m_+}(\pi^{k+1}), \tag{7.6}$$

since $-d = q \ge \tilde{q} := -f(x^{k+1}) - \langle \cdot, g(x^{k+1})\rangle = \tilde{q}(\pi^{k+1}) + \langle -g(x^{k+1}), \cdot - \pi^{k+1}\rangle$ and $\mathcal{D}_q \subset \mathbb{R}^m_+$ from $q = \sup_x -L(x,\cdot)$, and $\tilde{q}(\pi^{k+1}) \ge q(\pi^{k+1}) - \epsilon_k$ (cf. (7.2)). Next, (7.3) gives $\pi^{k+1} \in \mathcal{D}_{\partial h_+} \subset \mathcal{D}_{h_+} \subset \mathbb{R}^m_+$, whereas $q(\pi^{k+1}) \le q(\pi^k) + \epsilon_k$ (cf. (5.1)) yields $\pi^{k+1} \in \mathcal{D}_q$. By (7.6), (7.4) and (7.5) hold if we take $p^{k+1} = (\gamma^k - \gamma^{k+1})/c_k$ and

$$\gamma^{k+1} = \gamma^k + c_k g(x^{k+1}) - \tilde{\gamma}^{k+1} \in \partial h_+(\pi^{k+1}) \quad \text{with} \quad \tilde{\gamma}^{k+1} \in N_{\mathbb{R}^m_+}(\pi^{k+1}), \tag{7.7}$$

since then

$$p^{k+1} = -g(x^{k+1}) + \tilde{\gamma}^{k+1}/c_k \in \partial_{\epsilon_k} q(\pi^{k+1}). \tag{7.8}$$

Using (7.3) and $(\partial h_+)^{-1} = \nabla h^+$ (Lemma 7.2), we have

$$\gamma^k + c_k g(x^{k+1}) \in \partial h_+(\pi^{k+1}) = \partial h(\pi^{k+1}) + N_{\mathbb{R}_+}(\pi^{k+1}), \tag{7.9}$$

so we may take $\tilde{\gamma}^{k+1} = 0$; other choices will be discussed later.

Further insight may be gained as follows. Rewrite (7.3) as

$$\pi^{k+1} = \nabla P_k(g(x^{k+1})), \tag{7.10}$$

where

$$P_k(u) = h^+(\gamma^k + c_k u)/c_k \quad \forall u \in \mathbb{R}^m. \tag{7.11}$$

Let

$$L_k(x) = f(x) + \frac{1}{c_k}[h^+(\gamma^k + c_k g(x)) - h^+(\gamma^k)] \tag{7.12}$$

if $x \in \mathcal{D}_f$ ($\subset \mathcal{D}_g = \bigcap_{i=1}^m \mathcal{D}_{g_i}$; cf. Assumption 7.1), $L_k(x) = \infty$ otherwise.

LEMMA 7.3. *Suppose* $\inf_{\mathcal{D}_f} \max_{i=1}^m g_i \le 0$, *e.g., the* feasible set $C_0 = \{x \in \mathcal{D}_f : g(x) \le 0\}$ *of* (7.1) *is nonempty. Then* L_k *is a proper convex function and*

$$\partial L_k(x) = \partial f(x) + \sum_{i=1}^m [\nabla P_k(g(x))]_i \partial g_i(x) \quad \forall x \in \mathcal{D}_{L_k} \supset \mathcal{D}_{\partial L_k}. \tag{7.13}$$

If $\partial L_k(x) \neq \emptyset$, *then* $\pi = \nabla P_k(g(x))$ *is well defined,* $\pi \geq 0$, *and* $\partial L_k(x) = \partial_x L(x,\pi)$, *where*

$$\partial_x L(x,\pi) = \partial f(x) + \sum_{i=1}^{m} \pi_i \partial g_i(x) \quad \forall x \in \mathbb{R}^n, \forall \pi \in \mathbb{R}^m_+. \tag{7.14}$$

If $\hat{x} \in \operatorname{Arg\,min} L_k$ *then* $\hat{x} \in \operatorname{Arg\,min}_x L(x,\hat{\pi})$ *for* $\hat{\pi} = \nabla P_k(g(\hat{x}))$. *The preceding assertions hold when* $\inf_{\mathcal{D}_f} \max_{i=1}^m g_i > 0$ *but* $\mathcal{D}_{h^+} = \mathbb{R}^m$, *e.g., if* $\operatorname{im} \partial h \supset \mathbb{R}^m_>$ (*cf. Lemma* 7.2).

Proof. Using $\gamma^k \in \partial h_+(\pi^k) \subset \mathring{\mathcal{D}}_{h^+}$ (cf. Lemma 7.2) and $\mathring{\mathcal{D}}_{P_k} = (\mathring{\mathcal{D}}_{h^+} - \gamma^k)/c_k$, pick $\tilde{u} \in \mathcal{D}_{P_k} \cap \mathbb{R}^m_>$ and $\tilde{x} \in \mathcal{D}_f$ s.t. $g(\tilde{x}) < \tilde{u}$. Then, since P_k is nondecreasing (as is h^+) and $\operatorname{ri} \mathcal{D}_f \subset \bigcap_i \operatorname{ri} \mathcal{D}_{g_i}$ (cf. Assumption 7.1), Lemma A.1 in the appendix yields $\operatorname{im} \partial P_k \subset \mathbb{R}^m_+$ and (7.13), using $\partial P_k = \{\nabla P_k\}$ (cf. Lemma 7.2). Hence if $\partial L_k(x) \neq \emptyset$, then $\pi = \nabla P_k(g(x)) \geq 0$, so $\operatorname{ri} \mathcal{D}_f \subset \bigcap_i \operatorname{ri} \mathcal{D}_{g_i}$ implies (cf. [Roc70, Thm. 23.8]) $\partial_x L(x,\pi) = \partial f(x) + \sum_i \pi_i \partial g_i(x) = \partial L_k(x)$. If $\hat{x} \in \operatorname{Arg\,min} L_k$, then $0 \in \partial L_k(\hat{x}) = \partial_x L(\hat{x},\hat{\pi})$ for $\hat{\pi} = \nabla P_k(g(\hat{x}))$ yields $\hat{x} \in \operatorname{Arg\,min}_x L(x,\hat{\pi})$. Finally, when $\mathcal{D}_{h^+} = \mathbb{R}^m$, then for any $\tilde{x} \in \mathcal{D}_f$ we may pick $\tilde{u} \in \mathcal{D}_{P_k}$ with $g(\tilde{x}) < \tilde{u}$, since $\mathcal{D}_f \subset \bigcap_i \mathcal{D}_{g_i}$ (Assumption 7.1) and $\mathcal{D}_{P_k} = \mathbb{R}^m$. □

The *exact* multiplier method of [Eck93, Thm. 7] takes $x^{k+1} \in \operatorname{Arg\,min} L_k$ and $\pi^{k+1} = \nabla P_k(g(x^{k+1}))$, assuming h is smooth, $\mathring{\mathcal{D}}_h \supset \mathbb{R}^m_>$ and $\operatorname{im} \nabla h \supset \mathbb{R}^m_>$. Then (7.2) holds with $\epsilon_k = 0$ (cf. Lemma 7.3). Our inexact method requires only that $x^{k+1} \tilde{\in} \operatorname{Arg\,min} L_k$ in the sense that (7.2) holds for a given $\epsilon_k \geq 0$. Thus we have derived the following algorithm.

ALGORITHM 7.4. At iteration $k \geq 1$, having $\pi^k \in \mathcal{D}_q$ and $\gamma^k \in \partial h_+(\pi^k)$, find

$$\begin{aligned} x^{k+1} &\tilde{\in} \operatorname{Arg\,min}_{x \in \mathcal{D}_f} \left\{ f(x) + \frac{1}{c_k} h^+(\gamma^k + c_k g(x)) \right\}, \\ \pi^{k+1} &= \nabla h^+(\gamma^k + c_k g(x^{k+1})) \end{aligned}$$

s.t. (7.2) holds, choose γ^{k+1} satisfying (7.7), and set $p^{k+1} = (\gamma^k - \gamma^{k+1})/c_k$.

Remark 7.5. To find x^{k+1} as in [Ber82, sect. 5.3], suppose that f is *strongly convex*, i.e.,

$$\exists \breve{\alpha} > 0 \quad \text{s.t.} \quad f(x) \geq f(\bar{x}) + \langle \gamma, x - \bar{x} \rangle + \breve{\alpha}|x - \bar{x}|^2/2 \quad \forall x, \bar{x}, \forall \gamma \in \partial f(\bar{x}). \tag{7.15}$$

Adding subgradient inequalities of g_i, $i = 1\colon m$, and using (7.14) yield for all x

$$L(x,\pi^{k+1}) \geq L(x^{k+1},\pi^{k+1}) + \langle \gamma, x - x^{k+1} \rangle + \breve{\alpha}|x - x^{k+1}|^2/2 \quad \forall \gamma \in \partial_x L(x^{k+1},\pi^{k+1}). \tag{7.16}$$

Assuming $\partial L_k(x^{k+1}) \neq \emptyset$ and $\partial_x L(x^{k+1},\pi^{k+1}) = \partial L_k(x^{k+1})$ (e.g., $C_0 \neq \emptyset$ or $\mathcal{D}_{h^+} = \mathbb{R}^m$; cf. Lemma 7.3), let $\Delta_x L_k(x^{k+1}) = \arg\min_{\gamma \in \partial L_k(x^{k+1})} |\gamma|$. Minimization in (7.16) yields

$$d(\pi^{k+1}) \geq L(x^{k+1},\pi^{k+1}) - |\Delta_x L_k(x^{k+1})|^2/2\breve{\alpha}, \tag{7.17}$$

so (7.2) holds if

$$|\Delta_x L_k(x^{k+1})|^2/2\breve{\alpha} \leq \epsilon_k. \tag{7.18}$$

Thus, as in the multiplier methods of [Ber82, sect. 5.3], one may use any algorithm for minimizing L_k that generates a sequence $\{z^j\}$ such that $\liminf_{j\to\infty} |\Delta_x L_k(z^j)| = 0$, setting $x^{k+1} = z^j$ when (7.18) occurs. (If $\breve{\alpha}$ is unknown, it may be replaced in (7.18) by any fixed $\bar{\alpha} > 0$; this only scales $\{\epsilon_k\}$.) Of course, the strong convexity assumption

is not necessary if one can employ the direct criterion (7.2), i.e., $L(z^j,\pi)\le d(\pi)+\epsilon_k$ with $\pi=\nabla P_k(g(z^j))$ (cf. (7.10)), where $d(\pi)$ may be computed with an *error* that can be absorbed in ϵ_k.

Some examples are now in order. We start with three quite abstract examples: the first one treats general separable penalties, whereas the next two employ different assumptions on the behavior of ∇h near $\partial\mathbb{R}^m_>$ (this aspect will be studied in sections 8 and 9). They are followed by four examples of more concrete methods.

Example 7.6. Suppose that $h(\pi)=\sum_{i=1}^m h_i(\pi_i)$, where h_i are B-functions on $\mathbb{R}$ with $\mathcal{D}_{h_i}\supset\mathbb{R}_>$, $i=1\!:\!m$ (cf. Lemma 2.7(ii)). For each i, let $\bar u_i=h_i'(0;1)$ if $0\in\mathcal{D}_{h_i}$ and $\bar u_i=-\infty$ if $0\notin\mathcal{D}_{h_i}$ so that (cf. [Eck93, Ex. 6]) $h_i^+(u_i)=h_i^*(\max\{u_i,\bar u_i\})$ and $\nabla h_i^+(u_i)=\max\{0,\nabla h_i^*(u_i)\}$. Using (7.9) and "maximal" γ^{k+1} in (7.7), Algorithm 7.4 may be written as

$$x^{k+1}\ \tilde\in\ \operatorname{Arg\,min}_x\left\{f(x)+\frac{1}{c_k}\sum_{i=1}^m h_i^*(\max\{\bar u_i,\gamma_i^k+c_kg_i(x)\})\right\},\tag{7.19a}$$

$$\pi_i^{k+1}=\max\{0,\nabla h_i^*(\gamma_i^k+c_kg_i(x^{k+1})\},\quad i=1\!:\!m,\tag{7.19b}$$

$$\gamma_i^{k+1}=\max\{\bar u_i,\gamma_i^k+c_kg_i(x^{k+1})\},\quad i=1\!:\!m.\tag{7.19c}$$

Remark 7.7. To justify (7.19c), note that if we had $\gamma^k<\bar u\in\mathbb{R}^m$, then (7.19a) would not penalize constraint violations $g_i(x)\in(0,(\bar u_i-\gamma_i^k)/c_k]$. An *ordinary* penalty method (cf. [Ber82, p. 354]) would use (7.19a) and (7.19b) with $\gamma^k\equiv\bar u$ and $c_k\uparrow\infty$. Thus (7.19) is a *shifted* penalty method in which the *shifts* γ^k should ensure convergence even for $\sup_k c_k<\infty$, thus avoiding the ill-conditioning of ordinary penalty methods.

Example 7.8. Suppose that $\mathcal{D}_{\partial h}\cap\mathbb{R}^m_+=\mathcal{D}_{\nabla h}\cap\mathbb{R}^m_+$ so that $\partial h_+=\nabla h+\partial\imath_{\mathbb{R}^m_+}$ from $\mathbb{R}^m_>\subset\mathcal{D}_h$ (cf. [Roc70, Thm. 23.8] and Fact 2.2). Then we may use $\gamma^k=\nabla h(\pi^k)$ for all k, since the maximal shift $\gamma^{k+1}=\nabla h(\pi^{k+1})$ satisfies (7.7) due to (7.9). Thus Algorithm 7.4 becomes

$$\begin{aligned}&x^{k+1}\ \tilde\in\ \operatorname{Arg\,min}_x\left\{f(x)+\frac{1}{c_k}h^+(\nabla h(\pi^k)+c_kg(x))\right\},\\&\pi^{k+1}=\nabla h^+(\nabla h(\pi^k)+c_kg(x^{k+1})).\end{aligned}$$

In the separable case of Example 7.6, the formulas specialize to

$$\begin{aligned}&x^{k+1}\ \tilde\in\ \operatorname{Arg\,min}_x\left\{f(x)+\frac{1}{c_k}\sum_{i=1}^m h_i^*(\max\{\bar u_i,\nabla h_i(\pi_i^k)+c_kg_i(x)\})\right\},\\&\pi_i^{k+1}=\max\{0,\nabla h_i^*(\nabla h_i(\pi_i^k)+c_kg_i(x^{k+1}))\},\quad i=1\!:\!m,\end{aligned}$$

where $\bar u_i=\nabla h_i(0)$ if $0\in\mathcal{D}_{\partial h_i}$, $\bar u_i=-\infty$ if $0\notin\mathcal{D}_{\partial h_i}$, $i=1\!:\!m$.

Example 7.9. Let $h(\pi)=\sum_{i=1}^m\psi(\pi_i)$, where ψ is a B-function on $\mathbb{R}$ with $\mathcal{D}_{\nabla\psi}\supset\mathbb{R}_>$. Let $\bar v=\psi'(0;1)$ if $0\in\mathcal{D}_\psi$ and $\bar v=-\infty$ if $0\notin\mathcal{D}_\psi$. Then $\partial\psi_+(t)=\{\psi'(t;1)\}$ for $t>0$, $\partial\psi_+(0)=(-\infty,\bar v]$ if $\bar v>-\infty$, $\partial\psi_+(0)=\emptyset$ if $\bar v=-\infty$. Using (7.7) and (7.9) as in Example 7.6, we may let $\gamma_i^{k+1}=\psi'(\pi_i^{k+1};1)$, $i=1\!:\!m$. Thus Algorithm 7.4 becomes

$$\begin{aligned}&x^{k+1}\ \tilde\in\ \operatorname{Arg\,min}_x\left\{f(x)+\frac{1}{c_k}\sum_{i=1}^m\psi^+(\psi'(\pi_i^k;1)+c_kg_i(x))\right\},\\&\pi_i^{k+1}=\nabla\psi^+(\psi'(\pi_i^k;1)+c_kg_i(x^{k+1})),\quad i=1\!:\!m,\end{aligned}\tag{7.20}$$

i.e.,

$$\begin{aligned} x^{k+1} &\tilde{\in} \operatorname{Arg\,min}_x \left\{ f(x) + \frac{1}{c_k} \sum_{i=1}^{m} \psi^*(\max\{\bar{v}, \psi'(\pi_i^k; 1) + c_k g_i(x)\}) \right\}, \\ \pi_i^{k+1} &= \max\{0, \nabla\psi^*(\psi'(\pi_i^k; 1) + c_k g_i(x^{k+1}))\}, \quad i = 1{:}\, m. \end{aligned} \tag{7.21}$$

Example 7.10. For $\psi(t) = |t|^\alpha/\alpha$ with $\alpha > 1$ and $\beta = \alpha/(\alpha - 1)$ (cf. Example 2.8(1)), (7.21) becomes

$$x^{k+1} \tilde{\in} \operatorname{Arg\,min}_x \left\{ f(x) + \frac{1}{\beta c_k} \sum_{i=1}^{m} \max\{0, (\pi_i^k)^{1/(\beta-1)} + c_k g_i(x)\}^\beta \right\}, \tag{7.22a}$$

$$\pi_i^{k+1} = \max\{0, (\pi_i^k)^{1/(\beta-1)} + c_k g_i(x^{k+1})\}^{\beta-1}, \quad i = 1{:}\, m. \tag{7.22b}$$

Even if f and all g_i are smooth, for $\beta = 2$ the objective of (7.22a) is, in general, only once continuously differentiable. This is a well-known drawback of quadratic augmented Lagrangians (cf. [Ber82, TsB93]). However, for $\beta = 3$ we obtain a *cubic multiplier method* [Kiw97] with a *twice* continuously differentiable objective.

Example 7.11 (see [Eck93, Ex. 7]). For $\psi(t) = t \ln t - t$ (cf. Example 2.8(3)), (7.21) reduces to

$$\begin{aligned} x^{k+1} &\tilde{\in} \operatorname{Arg\,min}_x \left\{ f(x) + \frac{1}{c_k} \sum_{i=1}^{m} \pi_i^k \exp[c_k g_i(x)] \right\}, \\ \pi_i^{k+1} &= \pi_i^k \exp[c_k g_i(x^{k+1})], \quad i = 1{:}\, m, \end{aligned} \tag{7.23}$$

i.e., to an inexact *exponential multiplier method* (cf. [Ber82, sect. 5.1.2; TsB93]).

Example 7.12. For $\psi(t) = -\ln t$ (cf. Example 2.8(4)), (7.21) reduces to

$$\begin{aligned} x^{k+1} &\tilde{\in} \operatorname{Arg\,min}_x \left\{ f(x) - \frac{1}{c_k} \sum_{i=1}^{m} \ln[1/\pi_i^k - c_k g_i(x)] \right\}, \\ \pi_i^{k+1} &= \pi_i^k / [1 - c_k \pi_i^k g_i(x^{k+1})], \quad i = 1{:}\, m, \end{aligned}$$

i.e., to an inexact *shifted logarithm barrier method* (which was also derived heuristically in [Cha94, Ex. 4.2]). This method is related, but not identical, to ones in [CGT92, GMSW88]; cf. [CGT94].

Example 7.13. If $\psi(t) = -t^\alpha/\alpha$, $\alpha \in (0, 1)$ (cf. Example 2.8(2)), (7.21) reduces to

$$\begin{aligned} x^{k+1} &\tilde{\in} \operatorname{Arg\,min}_x \left\{ f(x) - \frac{1}{\beta c_k} \sum_{i=1}^{m} [(\pi_i^k)^{1/(\beta-1)} - c_k g_i(x)]^\beta \right\}, \\ \pi_i^{k+1} &= [(\pi_i^k)^{1/(\beta-1)} - c_k g_i(x^{k+1})]^{\beta-1}, \quad i = 1{:}\, m, \end{aligned}$$

where $\beta = \alpha/(\alpha - 1)$; $\beta = -1$ corresponds to a *shifted Carroll barrier method*.

8. Convergence of multiplier methods. Ideally, the sequences $\{x^k\}$ and $\{\pi^k\}$ generated by Algorithm 7.4 should solve the primal and dual problems asymptotically. Convergence of $\{\pi^k\}$ may be studied in the framework of section 5, but that of $\{x^k\}$ will require more work. For the framework of section 5, in addition to Assumption 7.1, we make the following standing assumptions.

Assumption 8.1.

(i) h_+ is a B-function s.t. $\mathcal{D}_{h_+} \supset \mathbb{R}^m_>$ (e.g., so is h; cf. Lemma 2.6).

(ii) Either $\mathcal{D}_q \cap \mathbb{R}^m_> \neq \emptyset$ or $\emptyset \neq \mathcal{D}_q \subset \mathcal{D}_{h_+}$, where $-q = d = \inf_x L(x,\cdot)$.

(iii) $\{c_k\}$ is a sequence of positive numbers s.t. $s_k = \sum_{j=1}^k c_j \to \infty$.

Remark 8.2. Under Assumption 8.1, q is closed proper convex, $\mathring{\mathcal{D}}_{h_+} = \mathbb{R}^m_> \subset \mathcal{D}_{h_+} \subset \mathbb{R}^m_+$, $\operatorname{cl}\mathcal{D}_{h_+} = \mathbb{R}^m_+ \supset \mathcal{D}_q$, $\mathcal{D}_q \cap \mathring{\mathcal{D}}_{h_+} \neq \emptyset$ if $\mathcal{D}_q \cap \mathbb{R}^m_> \neq \emptyset$, and $\inf_{\mathcal{D}_{h_+}} q = \inf q = \inf_{\operatorname{cl}\mathcal{D}_{h_+}} q$. Hence for the BPM method applied to the dual problem $\sup d = -\inf q$ with a B-function h_+ we may invoke the results of sections 3–6 (replacing f, X, and h with q, $\mathbb{R}^m$, and h_+, respectively).

We first translate Theorem 5.2 for the dual variables $\{\pi^k\}$ of Algorithm 7.4.

THEOREM 8.3. *If* $\sum_{j=1}^k s_j\epsilon_j/s_k \to 0$ (*cf. Lemma* 5.3), *then* $d(\pi^k) \to \sup d$. *If* $d(\pi^k) \to \sup d$, $\mathcal{D}_{h_+} \cap \operatorname{Arg\,max} d \neq \emptyset$, *and* $\sum_{k=1}^\infty c_k\epsilon_k < \infty$, *then* $\pi^k \to \pi^\infty \in \operatorname{Arg\,max} d$. *If* $d(\pi^k) \to \sup d$, $\mathcal{D}_q \subset \mathcal{D}_{h_+}$ *and* $\operatorname{Arg\,max}_{\mathcal{D}_{h_+}} d = \emptyset$ (*e.g.*, $\mathcal{D}_{h_+} = \mathbb{R}^m_+$ *and* $\operatorname{Arg\,max} d = \emptyset$), *then* $|\pi^k| \to \infty$.

Proof. This follows from Remark 8.2 and Theorem 5.2, since $\mathcal{D}_{h_+} \cap \operatorname{Arg\,max} d \subset \operatorname{Arg\,max}_{\mathcal{D}_{h_+}} d \subset \operatorname{Arg\,max} d$ if $\mathcal{D}_{h_+} \cap \operatorname{Arg\,max} d \neq \emptyset$. ∎

Thanks to Bregman proximal regularization, $\{\pi^k\}$ in Theorem 8.3 may converge to a particular dual solution even when $\operatorname{Arg\,max} d$ is not a singleton. Since the primal subproblems of Algorithm 7.4 are not regularized, convergence of $\{x^k\}$ will require additional assumptions (cf. Remark 8.5). We start with the "standard" case of $\mathcal{D}_{\nabla h} \supset \mathbb{R}^m_+$ (as in Example 7.10).

THEOREM 8.4. *Let* $\mathcal{D}_{\nabla h} \supset \mathbb{R}^m_+$, $\gamma^k = \nabla h(\pi^k)$ *for all* k (*cf. Example* 7.8) *and* $\sum_{j=1}^k s_j\epsilon_j/s_k \to 0$. *Then* $d(\pi^k) \to \sup d$. *If* $\operatorname{Arg\,max} d \neq \emptyset$ *and* $\sum_{k=1}^\infty c_k\epsilon_k < \infty$, *then* $\pi^k \to \pi^\infty \in \operatorname{Arg\,max} d$, *and if* $\inf_k c_k > 0$, *then*

$$\limsup_{k\to\infty} f(x^k) \le \sup_\pi d(\pi) \quad \textit{and} \quad \limsup_{k\to\infty} g_i(x^k) \le 0, \quad i = 1\colon m, \tag{8.1}$$

and every limit point of $\{x^k\}$ *solves* (7.1). *If* $\operatorname{Arg\,max} d = \emptyset$, *then* $|\pi^k| \to \infty$.

Proof. Since $\mathcal{D}_h \supset \mathcal{D}_{\nabla h} \supset \mathbb{R}^m_+$, the assertions about $\{\pi^k\}$ follow from Theorem 8.3. Suppose $\pi^k \to \pi^\infty \in \operatorname{Arg\,max} d$, $\inf_k c_k > 0$. Since $p^k = (\gamma^{k-1} - \gamma^k)/c_{k-1}$ with $p^k + g(x^k) \in N_{\mathbb{R}^m_+}(\pi^k)$ (cf. Example 7.8), we have (cf. Lemma 7.2) $\langle \pi^k, g(x^k)\rangle = -\langle \pi^k, p^k\rangle$ and $g(x^k) \le -p^k$ $\forall k > 1$, with $p^k \to 0$, since $\pi^k \to \pi^\infty$, ∇h is continuous on $\mathbb{R}^m_+$, and $c_k \ge c_{\min}$ $\forall k$. Hence $\langle \pi^k, g(x^k)\rangle \to 0$ and $\limsup_{k\to\infty} g_i(x^k) \le 0$ $\forall i$. Since $L(x^k,\pi^k) \le \inf_x L(x,\pi^k) + \epsilon_{k-1}$ (cf. (7.2)) means $f(x^k) + \langle \pi^k, g(x^k)\rangle \le f(x) + \langle \pi^k, g(x)\rangle + \epsilon_{k-1}$ for any x, in the limit $\limsup_k f(x^k) \le L(x,\pi^\infty)$ ($\epsilon_k \to 0$), so $\limsup_k f(x^k) \le d(\pi^\infty)$. Suppose $x^k \xrightarrow{K} x^\infty$ for some x^∞ and $K \subset \{1,2,\ldots\}$. By (8.1), $f(x^\infty) \le \sup d$ and $g(x^\infty) \le 0$ (f and g are closed), so by weak duality, $f(x^\infty) \ge \sup d$, $f(x^\infty) = \max d$, and x^∞ solves (7.1). ∎

Remark 8.5. Let C_* denote the optimal solution set for (7.1). If (7.1) is consistent (i.e., $C_0 \neq \emptyset$), then C_* is nonempty and compact iff f and g_i, $i = 1\colon m$, have no common direction of recession [Ber82, sect. 5.3], in which case (8.1) implies that $\{x^k\}$ is bounded and hence has limit points. In particular, if $C_* = \{x^*\}$, then $x^k \to x^*$ in Theorem 8.4.

Remark 8.6. Theorems 8.3 and 8.4 subsume [Eck93, Thm. 7], which additionally requires that $\epsilon_k \equiv 0$, $\operatorname{im}\nabla h \supset \mathbb{R}^m_>$ and each g_i is continuous on $\mathcal{D}_f$.

We now establish finite convergence of an exact version of Algorithm 7.4 (with $\epsilon_k \equiv 0$) in the case of a dual sharp minimum (cf. [Ber82, sect. 5.4] and [BuF93]).

THEOREM 8.7. *Let* (7.1) *be s.t.* $q = -d$ *has a sharp minimum. Let* $\mathcal{D}_{\nabla h} \supset \mathbb{R}^m_+$, $\inf_k c_k > 0$, $\epsilon_k = 0$, *and* $\gamma^k = \nabla h(\pi^k)$ (*cf. Example* 7.8) *for all* k. *Then there exists* k *s.t.* $p^k = 0$, $\pi^k \in \operatorname{Arg\,max} d$, *and* x^k *solves* (7.1).

Proof. Using the proof of Theorem 6.1 with $\pi^k \to \pi^\infty \in \operatorname{Arg\,max} d \subset \mathcal{D}_{\nabla h}$ and $\gamma^k = \nabla h(\pi^k) \to \nabla h(\pi^\infty)$, we get k s.t. $\pi^k \in \operatorname{Arg\,max} d$ and $p^k = 0$; the conclusion follows from the proof of Theorem 8.4. ∎

Remark 8.8. Results on finite convergence of other multiplier methods are restricted to only once continuously differentiable augmented Lagrangians [Ber82, sect. 5.4], whereas Theorem 8.7 covers Example 7.10 also with $\beta > 2$. Applications include polyhedral programs.

We shall need the following result, similar to ones in [Ber82, sect. 5.3] and [TsB93].

LEMMA 8.9. *With $u^{k+1} := g(x^{k+1})$, for each k, we have*

$$L(x^{k+1}, \pi^{k+1}) = L_k(x^{k+1}) + D_{h^+}(\gamma^k, \gamma^k + c_k u^{k+1})/c_k \geq L_k(x^{k+1}), \tag{8.2}$$

$$L_k(x^{k+1}) = L(x^{k+1}, \pi^k) + D_{h^+}(\gamma^k + c_k u^{k+1}, \gamma^k)/c_k \geq L(x^{k+1}, \pi^k), \tag{8.3}$$

$$\begin{aligned} L(x^{k+1}, \pi^{k+1}) - L(x^{k+1}, \pi^k) &= \left\langle \pi^{k+1} - \pi^k, u^{k+1} \right\rangle \\ &= \left\langle \nabla h^+(\gamma^k + c_k u^{k+1}) - \nabla h^+(\gamma^k), u^{k+1} \right\rangle \geq 0, \end{aligned} \tag{8.4}$$

$$d(\pi^k) \leq L(x^{k+1}, \pi^k) \leq L_k(x^{k+1}) \leq L(x^{k+1}, \pi^{k+1}) \leq d(\pi^{k+1}) + \epsilon_k. \tag{8.5}$$

Proof. As for (8.2), use (7.12), (7.3), (2.4), and convexity of h^+ to develop

$$\begin{aligned} L(x^{k+1}, \pi^{k+1}) - L_k(x^{k+1}) &= \left\langle \pi^{k+1}, u^{k+1} \right\rangle - [h^+(\gamma^k + c_k u^{k+1}) - h^+(\gamma^k)]/c_k \\ &= [h^+(\gamma^k) - h^+(\gamma^k + c_k u^{k+1}) \\ &\qquad - \left\langle \nabla h^+(\gamma^k + c_k u^{k+1}), -c_k u^{k+1} \right\rangle]/c_k \\ &= D_{h^+}(\gamma^k, \gamma^k + c_k u^{k+1})/c_k \geq 0. \end{aligned}$$

Since $\nabla h^+ = (\partial h_+)^{-1}$ (cf. Lemma 7.2) and $\gamma^k \in \partial h_+(\pi^k)$ (cf. (7.5)), $\pi^k = \nabla h^+(\gamma^k)$, so

$$\begin{aligned} L_k(x^{k+1}) - L(x^{k+1}, \pi^k) &= [h^+(\gamma^k + c_k u^{k+1}) - h^+(\gamma^k)]/c_k - \left\langle \pi^k, u^{k+1} \right\rangle \\ &= [h^+(\gamma^k + c_k u^{k+1}) - h^+(\gamma^k) - \left\langle \nabla h^+(\gamma^k), c_k u^{k+1} \right\rangle]/c_k \\ &= D_{h^+}(\gamma^k + c_k u^{k+1}, \gamma^k)/c_k \geq 0 \end{aligned}$$

yields (8.3), and (8.4) holds with $\left\langle \nabla h^+(\gamma^k + c_k u^{k+1}) - \nabla h^+(\gamma^k), c_k u^{k+1} \right\rangle / c_k \geq 0$ by the convexity of h^+. Equation (8.5) follows from (8.2)–(8.4) and (7.2). ∎

Theorem 8.4 covers only methods with $\mathcal{D}_{\nabla h} \supset \mathbb{R}^m_+$, such as Example 7.10. To handle other examples in section 9, we shall use the following abstraction of the ergodic framework of [TsB93]. For each k, define the aggregate primal solution

$$\breve{x}^{k+1} = \sum_{j=1}^{k} c_j x^{j+1}/s_k, \quad \text{where} \quad s_k = \sum_{j=1}^{k} c_j. \tag{8.6}$$

Since g is convex and $c_j g(x^{j+1}) \leq -c_j p^{j+1} = \gamma^{j+1} - \gamma^j$ for $j = 1\colon k$ by (7.7) and (7.8),

$$g(\breve{x}^{k+1}) \leq \sum_{j=1}^{k} c_j g(x^{j+1})/s_k \leq (\gamma^{k+1} - \gamma^1)/s_k. \tag{8.7}$$

LEMMA 8.10. *Suppose that $\sup_{i,k} \gamma_i^k < \infty$, $\epsilon_k \to 0$, $\left\langle \pi^k, u^k \right\rangle \to 0$, and $d(\pi^k) \to d^\infty < \infty$. Then*

$$\limsup_{k\to\infty} f(\breve{x}^k) \leq d^\infty \quad \textit{and} \quad \limsup_{k\to\infty} g_i(\breve{x}^k) \leq 0, \quad i = 1\colon m. \tag{8.8}$$

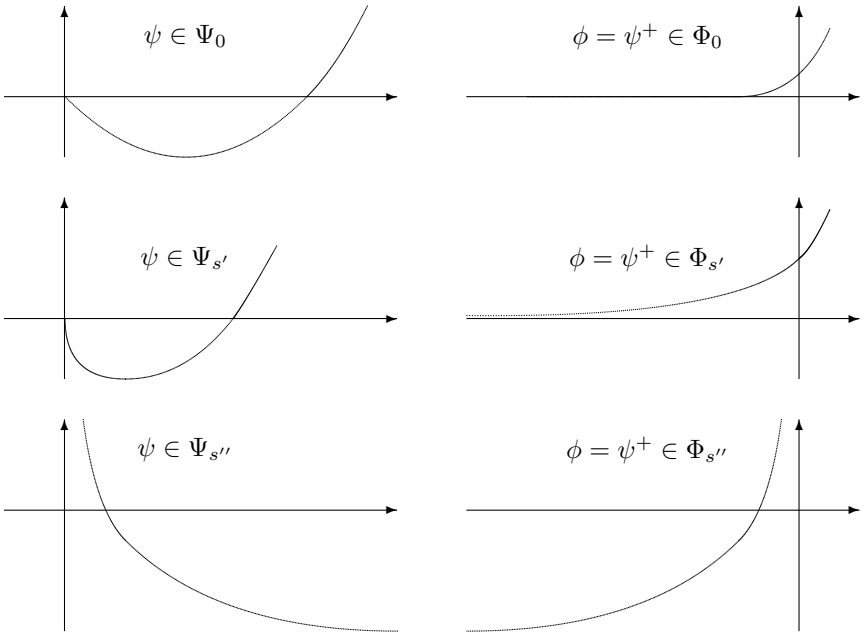


FIG. 9.1. *Examples of B-functions and corresponding penalty functions.*

If $\{\breve{x}^k\}$ *has a limit point* x^∞ (*e.g.,* $C_* \neq \emptyset$ *is bounded; cf. Remark* 8.11), *then* x^∞ *solves* (7.1), $f(x^\infty) = d^\infty = \max d$, *and each limit point of* $\{\pi^k\}$ *maximizes* d.

Proof. By (8.7), $\limsup_k g_i(\breve{x}^k) \leq 0\ \forall i$, since $s_k \to \infty$. By (8.6) and convexity of f, $f(\breve{x}^{k+1}) \leq \sum_{j=1}^k c_j f(x^{j+1})/s_k$, while $f(x^k) = L(x^k, \pi^k) - \langle \pi^k, u^k \rangle \to d^\infty$ from (8.5), so $\limsup_k f(\breve{x}^k) \leq d^\infty$. Suppose $\breve{x}^k \xrightarrow{K} x^\infty$. By (8.8), $f(x^\infty) \leq d^\infty$ and $g(x^\infty) \leq 0$ (f and g are closed). Hence by weak duality, $f(x^\infty) \geq d^\infty \leftarrow d(\pi^k)$, $f(x^\infty) = d^\infty = \max d$, and x^∞ solves (7.1). Since $d(\pi^k) \to d^\infty$ and d is closed, each cluster of $\{\pi^k\}$ maximizes d. ∎

Remark 8.11. If $C_* \neq \emptyset$ is bounded, then (8.8) implies that $\{\breve{x}^k\}$ is bounded (cf. Remark 8.5). In particular, if $C_* = \{x^*\}$, then $\breve{x}^k \to x^*$ in Lemma 8.10.

9. Classes of penalty functions. Examples 7.11–7.13 stem from B-functions of the form $h(\pi) = \sum_{i=1}^m \psi(\pi_i)$, where ψ is a B-function on $\mathbb{R}$ s.t. $\psi_+ = \psi$. Since $\psi_+ = (\psi^+)^*$, such examples may also be derived by choosing suitable penalty functions ϕ on $\mathbb{R}$ and letting $\psi = \phi^*$ (cf. Lemma 2.9). Let us first define several useful classes of B-functions and penalty functions.

DEFINITION 9.1. *We say that* $\psi \in \Psi$ *iff* ψ *is a B-function on* $\mathbb{R}$ *with* $\mathcal{D}_\psi \supset \mathbb{R}_>$. *Let* $\tilde{\Psi} = \{\psi \in \Psi : \mathcal{D}_{\nabla\psi} \supset \mathbb{R}_>\}$, $\Psi_0 = \{\psi \in \tilde{\Psi} : \mathcal{D}_{\partial\psi} \supset \mathbb{R}_+\}$, $\Psi_s = \{\psi \in \tilde{\Psi} : \mathcal{D}_{\partial\psi} = \mathbb{R}_>\}$, $\Psi_{s'} = \{\psi \in \Psi_s : 0 \in \mathcal{D}_\psi\}$, $\Psi_{s''} = \{\psi \in \Psi_s : 0 \notin \mathcal{D}_\psi\}$ (*cf. Fig.* 9.1).

Remark 9.2. We have $\tilde{\Psi} = \Psi_0 \cup \Psi_s$, $\Psi_0 = \{\psi \in \tilde{\Psi} : 0 \in \mathcal{D}_{\partial\psi}\}$, $\Psi_s = \{\psi \in \tilde{\Psi} : 0 \notin \mathcal{D}_{\partial\psi}\}$ (since $\mathcal{D}_{\partial\psi} \supset \mathcal{D}_{\nabla\psi} \supset \mathbb{R}_>$ if $\psi \in \tilde{\Psi}$) and $\psi_+ = \psi$ if $\psi \in \Psi_s$ (from $\mathcal{D}_\psi \subset \operatorname{cl} \mathcal{D}_{\partial\psi} = \mathbb{R}_+$).

DEFINITION 9.3. *We say that* $\phi \in \Phi$ *iff* $\phi : \mathbb{R} \to (-\infty, \infty]$ *is closed proper convex essentially smooth,* $\mathring{\mathcal{D}}_\phi = \mathcal{D}_\phi$, *and* $\mathbb{R}_> \subset \operatorname{im} \nabla\phi \subset \mathbb{R}_+$. *Let* $t_\phi = \sup_{t \in \mathcal{D}_\phi} t$, $t_\phi^0 = \sup_{\nabla\phi(t)=0} t$, $\Phi_s = \{\phi \in \Phi : \phi$ *is strictly convex*$\}$, $\Phi_0 = \{\phi \in \Phi : \phi$ *is strictly convex on* $(t_\phi^0, t_\phi), t_\phi^0 > -\infty\}$, $\Phi_{s'} = \{\phi \in \Phi_s : \inf \phi > -\infty\}$, $\Phi_{s''} = \{\phi \in \Phi_s : \inf \phi = -\infty\}$, $\tilde{\Phi} = \Phi_0 \cup \Phi_s$.

Remark 9.4. If $\phi \in \Phi$, then ϕ is nondecreasing ($\operatorname{im}\nabla\phi \subset \mathbb{R}_+$), $\mathcal{D}_\phi = (-\infty, t_\phi)$, $t_\phi^0 = -\infty$ iff $\operatorname{im}\nabla\phi = \mathbb{R}_>$, $\phi \in \Phi_s$ iff $\nabla\phi$ is increasing, $\phi \in \Phi_0$ iff $\nabla\phi$ is increasing on (t_ϕ^0, t_ϕ), and $t_\phi^0 > -\infty$ (cf. [Roc70, p. 254]). Also $\phi \in \Phi$ iff ϕ is closed proper convex, $\mathcal{D}_{\nabla\phi} = \mathring{\mathcal{D}}_\phi = \mathcal{D}_\phi$ (cf. Fact 2.3(i)), and $\mathbb{R}_> \subset \operatorname{im}\nabla\phi \subset \mathbb{R}_+$.

We now discuss duality relations between Definitions 9.1 and 9.3.

LEMMA 9.5. *If $\phi \in \Phi$, then ϕ^* is a B-function with $\mathbb{R}_> \subset \mathcal{D}_{\phi^*} \subset \mathbb{R}_+$, $(\phi^*)^+ = \phi^{**} = \phi$, $\lim_{t\downarrow-\infty}\nabla\phi(t) = 0$, $\lim_{t\uparrow t_\phi}\nabla\phi(t) = \lim_{t\uparrow t_\phi}\phi(t) = \infty$, and $\phi 0^+ = \imath^*_{\mathbb{R}_+}$. If $\phi \in \Phi_s$, then ϕ^* is essentially smooth, $\nabla\phi^* = (\nabla\phi)^{-1}$, and $\mathcal{D}_{\partial\phi^*} = \mathcal{D}_{\nabla\phi^*} = \mathbb{R}_>$. If $\phi \in \Phi_0$, then $\nabla\phi^* = (\nabla\phi)^{-1}$, $\mathcal{D}_{\nabla\phi^*} = \mathbb{R}_>$ and $\partial\phi^*(0) = (-\infty, t_\phi^0]$.*

Proof. By Definition 9.3 and Lemma 2.9, $\mathbb{R}_> \subset \operatorname{im}\nabla\phi \subset \mathbb{R}_+$ and ϕ^* is a B-function with $\operatorname{ri}\mathcal{D}_{\phi^*} \subset \operatorname{im}\nabla\phi \subset \mathcal{D}_{\phi^*}$, so $\mathbb{R}_> \subset \mathcal{D}_{\phi^*} \subset \mathbb{R}_+$. $\mathcal{D}_{\phi^*} \subset \mathbb{R}_+$ yields $(\phi^*)^+ = \phi^{**} = \phi$. Since $\mathbb{R}_> \subset \operatorname{im}\nabla\phi \subset \mathbb{R}_+$ and $\nabla\phi$ is nondecreasing, $\lim_{t\downarrow-\infty}\nabla\phi(t) = 0$ and $\lim_{t\uparrow t_\phi}\nabla\phi(t) = \infty$. Since ϕ is closed and proper, $\phi 0^+ = \imath^*_{\mathcal{D}_{\phi^*}}$ [Roc70, Thm. 13.3] with $\imath^*_{\mathcal{D}_{\phi^*}} = \imath^*_{\operatorname{cl}\mathcal{D}_{\phi^*}}$ and $\operatorname{cl}\mathcal{D}_{\phi^*} = \mathbb{R}_+$ from $\mathbb{R}_> \subset \mathcal{D}_{\phi^*} \subset \mathbb{R}_+$. If $t_\phi < \infty$, then $\lim_{t\uparrow t_\phi}\phi(t) = \infty$ from $t_\phi \notin \mathcal{D}_\phi$ and closedness of ϕ; otherwise $\lim_{t\uparrow t_\phi}\phi(t) = \infty$ from $\infty = \phi 0^+(1) = \lim_{t\uparrow\infty}[\phi(t) - \phi(0)]/t$ [Roc70, Thm. 8.5]. If $\phi \in \Phi_s$, then ϕ^* is essentially smooth, $\nabla\phi^* = (\nabla\phi)^{-1}$, and $\mathcal{D}_{\nabla\phi^*} = \mathcal{D}_{\partial\phi^*} = \mathring{\mathcal{D}}_{\phi^*} = \mathbb{R}_>$ (Facts 2.1 and 2.3). If $\phi \in \Phi_0$, then $\partial\phi^* = (\partial\phi)^{-1} = \{(\nabla\phi)^{-1}\}$ (Facts 2.1 and 2.3(i)) yields $\partial\phi^*(0) = \{t : \nabla\phi(t) = 0\} = (-\infty, t_\phi^0]$ ($0 \le \nabla\phi(t) \le \nabla\phi(t_\phi^0)\ \forall t \le t_\phi^0$), whereas $\nabla\phi$ is increasing on (t_ϕ^0, t_ϕ) (Remark 9.4), so $\partial\phi^* = \{(\nabla\phi)^{-1}\}$ is single valued on $\mathbb{R}_> = (\nabla\phi(t_\phi^0), \infty) \subset \operatorname{im}\nabla\phi$, and hence $\partial\phi^* = \{\nabla\phi^*\}$ on $\mathbb{R}_>$ (Fact 2.2). ∎

LEMMA 9.6. *Let ψ be a B-function on $\mathbb{R}$ s.t. $\mathcal{D}_\psi \supset \mathbb{R}_>$. Then $\psi^+ \in \Phi$. Suppose that $\mathcal{D}_{\nabla\psi} \supset \mathbb{R}_>$. If $\partial\psi(0) = \emptyset$ (i.e., $0 \notin \mathcal{D}_\psi$ or $\psi'(0;1) = -\infty$), then ψ_+ is essentially smooth and $\psi^+ \in \Phi_s$. If $\partial\psi(0) \ne \emptyset$ (i.e., $\psi'(0;1) > -\infty$), then $\psi^+ \in \Phi_0$ with $t^0_{\psi^+} = \psi'(0;1)$, and there exists a B-function $\breve{\psi}$ s.t. $\psi_+ = \breve{\psi}_+$, $\psi^+ = \breve{\psi}^+$, $\mathcal{D}_{\nabla\breve{\psi}} \supset \mathbb{R}_+$, and $\nabla\breve{\psi}(0) = t^0_{\psi^+}$.*

Proof. $\psi_+ = \psi + \imath_{\mathbb{R}_+}$ is a B-function (Lemma 2.6) and $\psi^+ = \psi^*_+$, so $\mathcal{D}_{\psi^+} = \mathring{\mathcal{D}}_{\psi^+}$ (Lemma 2.9(i)). Also ψ^+ is nondecreasing and essentially smooth (Lemma 7.2), so $\operatorname{im}\nabla\psi^+ \subset \mathbb{R}_+$, whereas $\mathbb{R}_> \subset \mathcal{D}_{\psi_+}$ yields $\mathbb{R}_> \subset \mathring{\mathcal{D}}_{\psi_+} \subset \mathcal{D}_{\partial\psi_+} = \operatorname{im}\partial\psi^+ = \operatorname{im}\nabla\psi^+$. Suppose $\mathcal{D}_{\nabla\psi} \supset \mathbb{R}_>$. By strict convexity of ψ (cf. Definition 2.4(i)), $\nabla\psi_+ = \nabla\psi$ is increasing on $\mathbb{R}_>$, so $\nabla\psi^+ = (\nabla\psi_+)^{-1}$ is increasing on $(t^0, \infty) \cap \mathcal{D}_{\psi^+}$ with $t^0 = \lim_{t\downarrow 0}\nabla\psi(t)$, and hence ψ^+ is strictly convex on (t^0, ∞) (cf. [Roc70, p. 254]). If $\partial\psi(0) = \emptyset$, then $t^0 = -\infty$, $\psi^+ \in \Phi_s$, and ψ_+ is essentially smooth (Fact 2.3(ii)). Otherwise, $t^0 = \psi'(0;1) = t^0_{\psi^+}$. Let $\breve{\psi}(t) = \psi(t)\ \forall t \ge 0$, and let $\breve{\psi}(t)$ for $t \le 0$ be a strictly convex quadratic function s.t. $\breve{\psi}(0) = \psi(0)$ and $\breve{\psi}'(0;-1) = -\psi'(0;1)$. Then $\breve{\psi}_+ = \psi_+$ and $\nabla\breve{\psi}(0) = t^0_{\psi^+}$. ∎

Remark 9.7. Lemmas 9.5 and 9.6 imply the following duality relations: $\Phi^* = \Psi_+$, $\tilde{\Phi}^* = \tilde{\Psi}_+$, $\Phi_s^* = \Psi_s$, $\Phi_{s'}^* = \Psi_{s'}$, $\Phi_{s''}^* = \Psi_{s''}$, $\Phi_0^* = \Psi_{0+}$, using $\Psi_+ = \{\psi_+ : \psi \in \Psi\} = \{\psi \in \Psi : \psi_+ = \psi\}$ and $\Psi_{s+} = \Psi_s$ (cf. Remark 9.2). (If $\psi \in \Psi_+$, then $\psi^* = \psi_+^* = \psi^+ \in \Phi$ (Lemma 9.6) and $\psi = \psi^{**} \in \Phi^*$; use $\inf\psi^* = -\psi(0)$ for the remaining "$\supset$" inclusions.) Similarly, $\Psi^+ = \Phi$, $\tilde{\Psi}^+ = \tilde{\Phi}$, $\Psi_s^+ = \Phi_s$, $\Psi_{s'}^+ = \Phi_{s'}$, $\Psi_{s''}^+ = \Phi_{s''}$, $\Psi_0^+ = \Phi_0$. (If $\phi \in \Phi$, then $\phi^* \in \Psi$ and $\phi = (\phi^*)^+ \in \Psi^+$ by Lemma 9.5; use $\inf\phi = -\phi^*(0)$ for the remaining "$\supset$" inclusions.)

We now extend Theorems 8.4 and 8.7 to Example 7.9 with $\psi \in \Psi_0$ (allowing $\mathcal{D}_{\nabla\psi} \not\supset \mathbb{R}_+$).

COROLLARY 9.8. *If $\psi \in \Psi_0$ (e.g., $\psi = \phi^*$ with $\phi \in \Phi_0$; cf. Lemma 9.5), then the method of Example 7.9 coincides with that of Example 7.8 with $h(\pi) = \sum_{i=1}^m \breve{\psi}(\pi_i)$,*

where $\breve{\psi}$ is the smooth extension of ψ described in Lemma 9.6*, so that $\mathcal{D}_{\nabla h} \supset \mathbb{R}^m_+$ and Theorems* 8.4 *and* 8.7 *apply.*

Proof. Use $\psi^+ = \breve{\psi}^+$ and $\psi'(t,1) = \nabla\breve{\psi}(t)\ \forall t \geq 0$ (cf. Lemma 9.6) in Example 7.9. ∎

Remarks 9.9.

(i) Example 7.9 with $\psi \in \Psi_0$ has $\mathcal{D}_{\partial\psi} \supset \mathbb{R}_+$ and $\bar{v} = \psi'(0;1) > -\infty$. It is not changed if we replace ψ with $\tilde{\psi} = \psi_+$ (since $\tilde{\psi}^+ = \psi^+$); e.g., use ϕ^* with $\phi = \psi^+ \in \Phi_0$ (Lemma 9.6), since $\phi^* = (\psi^+)^* = (\psi^*_+)^* = \psi_+$.

(ii) In terms of $\phi \in \Phi_0$, the method of Example 7.9 with $\psi = \phi^*$ becomes

$$x^{k+1} \in \operatorname{Arg\,min}_x \left\{ f(x) + \frac{1}{c_k}\sum_{i=1}^m \phi\left(\phi^{*'}(\pi_i^k;1) + c_k g_i(x)\right) \right\},$$
$$\pi_i^{k+1} = \nabla\phi\left(\phi^{*'}(\pi_i^k;1) + c_k g_i(x^{k+1})\right), \quad i = 1\!:\!m,$$

where $\phi^{*'}(\pi_i^k;1) = (\nabla\phi)^{-1}(\pi_i^k)$ if $\pi_i^k > 0$, $\phi^{*'}(\pi_i^k;1) = t^0_\phi$ if $\pi_i^k = 0$ (cf. Lemma 9.5).

In view of Corollary 9.8, we restrict attention to Example 7.9 with $\psi \in \Psi_s$.

Remark 9.10. Choosing $\psi \in \Psi_s$ corresponds to using $\psi = \phi^*$ with $\phi \in \Phi_s$ (Remark 9.7).

Example 9.11. Choosing $\phi \in \Phi_s$ and $\psi = \phi^*$ in Example 7.9 yields the method

$$x^{k+1} \in \operatorname{Arg\,min}_x \left\{ f(x) + \frac{1}{c_k}\sum_{i=1}^m \phi\left((\nabla\phi)^{-1}(\pi_i^k) + c_k g_i(x)\right) \right\},$$
$$\pi_i^{k+1} = \nabla\phi\left((\nabla\phi)^{-1}(\pi_i^k) + c_k g_i(x^{k+1})\right), \quad i = 1\!:\!m,$$

with $\gamma_i^k = (\nabla\phi)^{-1}(\pi_i^k)$, $\pi_i^k = \nabla\phi(\gamma_i^k)$, $i = 1\!:\!m$, for all k. (Indeed, $\psi^+ = \phi$, $\mathcal{D}_{\nabla\psi} = \mathbb{R}_>$ and $\nabla\psi = (\nabla\phi)^{-1}$ (Lemma 9.5), while $\pi^k = \nabla h^+(\gamma^k)$ yields $\pi_i^k = \nabla\psi^+(\gamma_i^k) = \nabla\phi(\gamma_i^k) > 0$ (Remark 9.2), $i = 1\!:\!m$.) Note that $\phi(t) = e^t$ for Example 7.11, $\phi(t) = -1 - \ln(-t)$ $(t < 0)$ for Example 7.12, $\phi(t) = -(-t)^\beta/\beta$ $(t < 0,\ \beta < 0)$ for Example 7.13.

We now explore a forcing property of $\phi \in \Phi_s$ needed by Lemma 9.16 below to ensure that $\langle \pi^k, u^k \rangle \to 0$ in Example 9.11, as required in Lemma 8.10.

DEFINITION 9.12. *We say that $\phi \in \Phi$ is* forcing *on $[t'_\phi, t''_\phi]$ if $[\phi'(t'_k) - \phi'(t''_k)](t'_k - t''_k) \to 0$ implies $\phi'(t''_k)(t'_k - t''_k) \to 0$ for any sequences $\{t'_k\}, \{t''_k\} \subset [t'_\phi, t''_\phi] \cap \mathcal{D}_\phi$, where $\phi' = \nabla\phi$.*

LEMMA 9.13. *If $\phi \in \Phi_s$, $\inf\phi > -\infty$, and $t''_\phi \in \mathcal{D}_\phi$, then ϕ is forcing on $[-\infty, t''_\phi]$.*

Proof. Replace ϕ with $\phi - \inf\phi$ so that $\inf\phi = 0$. Since $\phi' = \nabla\phi$ is positive and increasing (cf. Remark 9.4), so is ϕ. Let $[\phi'(t'_k) - \phi'(t_k)]\tau_k \to 0$, $\tau_k > 0$, $t'_k = t_k + \tau_k \leq t''_\phi$. If $\phi'(t_k)\tau_k \not\to 0$, there are $\epsilon > 0$ and $K \subset \{1, 2, \ldots\}$ s.t. $\phi'(t_k)\tau_k \geq \epsilon\ \forall k \in K$, so $\frac{\phi'(t'_k)}{\phi'(t_k)} \xrightarrow{K} 1$. Since $\phi'(t_k) < \phi'(t''_\phi)$ and $\phi(t'_k) \geq \phi(t_k) + \phi'(t_k)\tau_k \geq \epsilon$, $\tau_k \geq \epsilon/\phi'(t''_\phi)$ and $t'_k \geq \phi^{-1}(\epsilon)\ \forall k \in K$. Pick t_∞ and $K' \subset K$ s.t. $t'_k \xrightarrow{K'} t_\infty$. Then $t_k + \epsilon/2\phi'(t''_\phi) \leq t_\infty$ and $\phi'(t_k) \leq \phi'(t_\infty - \epsilon/2\phi'(t''_\phi)) < \phi'(t_\infty) = \lim_{k\in K'} \phi'(t'_k)$ for large $k \in K'$ contradict $\frac{\phi'(t'_k)}{\phi'(t_k)} \xrightarrow{K} 1$. Therefore, $\phi'(t_k)\tau_k \to 0$; i.e., ϕ is forcing. ∎

Example 9.14. The following functions are forcing on $[-\infty, t''_\phi]$: $\phi_1(t) = e^t$ with $t''_\phi \in \mathbb{R}$, $\phi_2(t) = -1 - \ln(-t)$ $(t < 0)$ with $t''_\phi \leq 0$, $\phi_3(t) = -(-t)^\beta/\beta$ $(t < 0,\ \beta < 0)$ with $t''_\phi < 0$. Indeed, let $\phi = \phi_2$. Suppose $\frac{\phi'(t_k+\tau_k) - \phi'(t_k)}{\phi'(t_k)}\phi'(t_k)\tau_k \to 0$. Since

$\phi'(t_k)\tau_k = -\tau_k/t_k$ and $\frac{\phi'(t_k+\tau_k)-\phi'(t_k)}{\phi'(t_k)} = \frac{-1}{1+t_k/\tau_k}$, $\phi'(t_k)\tau_k \to 0$; i.e., ϕ is forcing. Invoke Lemma 9.13 for ϕ_1 and ϕ_3.

Example 9.15. Let $\phi \in \Phi_s$ be s.t. $\phi(t) = -\frac{(-t)^\beta - 1}{\beta}$ for $t < -\frac{1}{2}$, $\beta \in (0,1)$. Let $t_k = -k$, $\tau_k = 1/\phi'(t'_k)$. Then $[\phi'(t_k+\tau_k) - \phi'(t_k)]\tau_k = (1-k^{-\beta})^{\beta-1} - 1 \to 0$, but $\phi'(t_k)\tau_k \to 1$; i.e., ϕ is not forcing on $[-\infty,-1]$, although $\lim_{\beta\downarrow 0} -\frac{(-t)^\beta-1}{\beta} = -\ln(-t)$ is (cf. Example 9.14).

LEMMA 9.16. *Suppose that in Example* 9.11 $\phi \in \Phi_s$ *is forcing on* $(-\infty, t_\gamma]$ *with* $t_\gamma = \sup_{i,k}\gamma_i^k$, $c_k \geq c_{\min} > 0$ *for all* k, *and* $\langle \pi^{k+1} - \pi^k, u^{k+1}\rangle \to 0$. *Then* $\langle \pi^k, u^k\rangle \to 0$.

Proof. Since $\nabla\phi$ is nondecreasing and $h^+(u) = \sum_i \phi(u_i)$, we deduce from (8.4) that

$$
\begin{aligned}
0 \leftarrow \langle \pi^{k+1} - \pi^k, u^{k+1}\rangle &= \sum_{i=1}^m [\phi'(\gamma_i^k + c_k u_i^{k+1}) - \phi'(\gamma_i^k)]u_i^{k+1} \\
&\geq \sum_{i=1}^m [\phi'(\gamma_i^k + c_{\min} u_i^{k+1}) - \phi'(\gamma_i^k)]u_i^{k+1} \geq 0
\end{aligned} \tag{9.1}
$$

and $[\phi'(\gamma_i^k + c_{\min}u_i^{k+1}) - \phi'(\gamma_i^k)]c_{\min}u_i^{k+1} \to 0$, $i = 1\colon m$. But $\gamma^{k+1} = \gamma^k + c_k u^{k+1}$ for all k (cf. Example 9.11) yields $\sup_{i,k}\{\gamma_i^k + c_{\min}u_i^{k+1}\} \leq t_\gamma$, so the preceding relation and the forcing property of ϕ give $\pi_i^k u_i^{k+1} = \phi'(\gamma_i^k)u_i^k \to 0\ \forall i$; hence $\langle \pi^{k+1}, u^{k+1}\rangle \to 0$ by (9.1). ∎

The following result relates the quantity t_γ of Lemma 9.16 with boundedness of $\{\pi^k\}$.

LEMMA 9.17. *Consider Example* 9.11 *with* $\phi \in \Phi_s$, $t_\phi = \sup_{t\in\mathcal{D}_\phi} t$, *and* $t_\gamma = \sup_{i,k}\gamma_i^k$. *Then* $t_\gamma \leq t_\phi$ (*so that* $t_\gamma < \infty$ *if* $t_\phi < \infty$). *In general,* $t_\gamma < t_\phi$ *iff* $\{\pi^k\}$ *is bounded.*

Proof. This follows from the facts $\pi_i^k = \nabla\phi(\gamma_i^k) \geq 0$, $\gamma_i^k \in \mathcal{D}_\phi = (-\infty, t_\phi)$, $\lim_{t\uparrow t_\phi}\nabla\phi(t) = \infty$ and monotonicity of $\nabla\phi$; cf. Remark 9.4, Lemma 9.5, and Example 9.11. ∎

We may now establish ergodic convergence of Example 9.11 with $\phi \in \Phi_{s'}$ (cf. Definition 9.3).

THEOREM 9.18. *Consider Example* 9.11 *with* $\phi \in \Phi_s$ *s.t.* $\inf\phi > -\infty$. *Suppose* $\operatorname{Arg\,max} d \neq \emptyset$, $\sum_{j=1}^k s_j\epsilon_j/s_k \to 0$, $\sum_{k=1}^\infty c_k\epsilon_k < \infty$, *and* $\inf_k c_k > 0$. *Then* $\pi^k \to \pi^\infty \in \operatorname{Arg\,max} d$, $d(\pi^k) \to d^\infty = d(\pi^\infty)$ *and* (8.8) *holds. If* $\{\breve{x}^k\}$ *has a limit point* x^∞ (*e.g.,* $C_* \neq \emptyset$ *is bounded; cf. Remark* 8.11), *then* x^∞ *solves* (7.1) *and* $f(x^\infty) = d^\infty$.

Proof. Let $\psi = \phi^*$. We have $\psi(0) = -\inf\phi < \infty$, $\mathcal{D}_\psi = \mathbb{R}_+$ (cf. Lemma 9.5), $\mathcal{D}_{\psi_+} = \mathbb{R}_+$, and $\mathcal{D}_{h_+^*} = \mathbb{R}_+^m$, so the assertions about $\{\pi^k\}$ follow from Theorem 8.3. Then $t_\gamma = \sup_{i,k}\gamma_i^k < t_\phi$ by Lemma 9.17 ($\{\pi^k\}$ is bounded), so ϕ is forcing on $[-\infty, t_\gamma]$ (Lemma 9.13). Since $d(\pi^k) \to d^\infty < \infty$ and $0 \leq \epsilon_k \leq \sum_{j=1}^k s_j\epsilon_j/s_k \to 0$, (8.4) and (8.5) yield $\langle \pi^{k+1} - \pi^k, u^{k+1}\rangle \to 0$. Then $\langle \pi^k, u^k\rangle \to 0$ by Lemma 9.16. The conclusion follows from Lemma 8.10. ∎

Remarks 9.19.

(i) For the exponential multiplier method (Example 7.11 with $\phi(t) = e^t$), Theorems 8.3 and 9.18 subsume [TsB93, Prop. 3.1] (in which $\operatorname{Arg\,max} d \neq \emptyset$, $C_* \neq \emptyset$ is bounded, $\epsilon_k \equiv 0$) and [IST94, Thm. 7.3] (in which $x^k \to x^\infty$ implies $\breve{x}^k \to x^\infty$).

(ii) Theorem 9.18 holds for Example 7.9 with $\psi \in \Psi_{s'}$, since $\Psi_{s'} = \Phi_{s'}^*$ (Remark 9.7).

For $\psi \in \Psi_0 \cup \Psi_{s'}$, Corollary 9.8 and Theorem 9.18 hinge on $\pi^k \to \pi^\infty \in \operatorname{Arg\,max} d \subset \mathcal{D}_{h^+} = \mathbb{R}^m_+$. Since $\mathcal{D}_h = \mathbb{R}^m_>$ for $\psi \in \Psi_{s''}$, we now present another approach.

THEOREM 9.20. *Consider Example* 9.11 *with* $\phi \in \Phi_s$ *forcing on* $(-\infty, t_\phi) \neq \mathbb{R}$ *(e.g.,* $\phi(t) = -1 - \ln(-t)$*; cf. Example* 9.14*). Suppose* $\epsilon_k \to 0$, $\inf_k c_k > 0$, *and* $d(\pi^k) \to d^\infty < \infty$. *Then* (8.8) *holds. If* $\{\breve{x}^k\}$ *has a limit point* x^∞ *(e.g.,* $C_* \neq \emptyset$ *is bounded; cf. Remark* 8.11*), then* x^∞ *solves* (7.1), $f(x^\infty) = d^\infty = \max d$, *and each limit point of* $\{\pi^k\}$ *maximizes* d.

Proof. By Lemma 9.17, $t_\gamma = \sup_{i,k} \gamma^k_i \leq t_\phi$, so ϕ is forcing on $(-\infty, t_\gamma]$. Since $d(\pi^k) \to d^\infty < \infty$ and $\epsilon_k \to 0$, (8.4) and (8.5) yield $\langle \pi^{k+1} - \pi^k, u^{k+1} \rangle \to 0$. Then $\langle \pi^k, u^k \rangle \to 0$ by Lemma 9.16. Since $t_\gamma \leq t_\phi < \infty$, the conclusion follows from Lemma 8.10. ∎

Remark 9.21. Suppose $\sum_{k=1}^\infty \epsilon_k < \infty$. Then $d^{k+1} \geq d^k - \epsilon_k \; \forall k$ (cf. (8.4)) yields $d(\pi^k) \to d^\infty \in (\infty, \infty]$ (cf. [Pol83, Lem. 2.2.3]). If $d^\infty = \infty$, then $C_0 = \emptyset$ by weak duality. If $d^\infty < \infty$, then $\{\pi^k\}$ is bounded if $\operatorname{Arg\,max} d \neq \emptyset$ is (cf. [Roc70, Cor. 8.7.1]), whereas if $C_0 \neq \emptyset$, then $\operatorname{Arg\,max} d \neq \emptyset$ is bounded iff Slater's condition holds, i.e., $g(x) < 0$ for some $x \in \mathcal{D}_f$ [GoT89, Thm. 1.3.4]. This observation may be used in Lemma 8.10 and Theorem 9.20.

The following two results use Slater's condition to complement Theorems 9.18 and 9.20.

THEOREM 9.22. *Consider Example* 9.11 *with* $\phi \in \Phi_s$ *s.t.* $\inf \phi > -\infty$. *Suppose* $g(x) < 0$ *for some* $x \in \mathcal{D}_f$, $\sum_{k=1}^\infty \epsilon_k < \infty$, *and* $\inf_k c_k > 0$. *Then* $d(\pi^k) \to d^\infty < \infty$ *and* (8.8) *holds. If* $\{\breve{x}^k\}$ *has a limit point* x^∞ *(e.g.,* $C_* \neq \emptyset$ *is bounded; cf. Remark* 8.11*), then* x^∞ *solves* (7.1), $f(x^\infty) = d^\infty = \max d$, *and each limit point of* $\{\pi^k\}$ *maximizes* d. *If* $d^\infty = \sup d$ *and* $\sum_{k=1}^\infty c_k \epsilon_k < \infty$, *then* $\pi^k \to \pi^\infty \in \operatorname{Arg\,max} d$.

Proof. Since $\epsilon_k \to 0$, $d(\pi^k) \to d^\infty < \infty$, $\{\pi^k\}$ and $\operatorname{Arg\,max} d \neq \emptyset$ are bounded (Remark 9.21), we get, as in the proof of Theorem 9.18, $\mathcal{D}_{h_+} = \mathbb{R}^m_+$, $t_\gamma < t_\phi$, and $\langle \pi^k, u^k \rangle \to 0$. Hence the first two assertions follow from Lemma 8.10, and the third one from Theorem 8.3. ∎

THEOREM 9.23. *Consider Example* 9.11 *with* $\phi \in \Phi_s$ *forcing on* $(-\infty, t''_\phi] \; \forall t''_\phi \in \mathbb{R}$. *Suppose that* $g(x) < 0$ *for some* $x \in \mathcal{D}_f$, $\sum_{k=1}^\infty \epsilon_k < \infty$, *and* $\inf_k c_k > 0$. *Then* $d(\pi^k) \to d^\infty < \infty$ *and* (8.8) *holds. If* $\{\breve{x}^k\}$ *has a limit point* x^∞ *(e.g.,* $C_* \neq \emptyset$ *is bounded; cf. Remark* 8.11*), then* x^∞ *solves* (7.1), $f(x^\infty) = d^\infty = \max d$, *and each limit point of* $\{\pi^k\}$ *maximizes* d. *If* $d^\infty = \sup d$, $\operatorname{Arg\,max} d \cap \mathcal{D}_{h_+} \neq \emptyset$, *and* $\sum_{k=1}^\infty c_k \epsilon_k < \infty$, *then* $\pi^k \to \pi^\infty \in \operatorname{Arg\,max} d$.

Proof. Use the proof of Theorem 9.22, without asserting that $\mathcal{D}_{h_+} = \mathbb{R}^m_+$. ∎

Remark 9.24. It is easy to see that we may replace $\phi \in \Phi_s$ with $\phi \in \Phi_0$ and Example 9.11 with Example 7.9 with $\psi = \phi^*$ in Lemmas 9.13, 9.16, and 9.17 and Theorems 9.18, 9.20, 9.22, and 9.23. (In the proof of Lemma 9.13, $t_\infty \geq \psi^{-1}(\epsilon) > t^0_\phi$, since ϕ' and ϕ are positive and increasing on (t^0_ϕ, t_ϕ); in the proof of Lemma 9.16, use $\gamma^{k+1} \geq \gamma^k + c_k u^{k+1}$ (cf. (7.7)); in proving Lemma 9.17, recall Remark 9.9(ii).) Such results complement Theorems 8.4 and 8.7; cf. Corollary 9.8.

10. Additional aspects of multiplier methods. Modified barrier functions can be extrapolated quadratically to facilitate their minimization; cf. [BTYZ92, BrS93, BrS94, NPS94, PoT97]. We now extend such techniques to our penalty functions, starting with a technical result.

LEMMA 10.1. *Let* $\phi_1, \phi_2 \in \Phi$ *be s.t. for some* $t_s \in (t^0_{\phi_1}, t_{\phi_1})$, $\phi_1(t_s) = \phi_2(t_s)$, $\phi'_1(t_s) = \phi'_2(t_s)$, ϕ_1 *is forcing on* $(-\infty, t_s]$, *and* ϕ_2 *is forcing on* $[t_s, t''_{\phi_2}]$ *with* $t''_{\phi_2} \in$

$[t_s, t_{\phi_2}]$. *Let* $\phi(t) = \phi_1(t)$ *if* $t \le t_s$, $\phi(t) = \phi_2(t)$ *if* $t > t_s$. *Then* ϕ *is forcing on* $(-\infty, t''_{\phi_2}]$. *If* $\phi_2 \in \Phi_s \cup \Phi_0$, *then* $\phi \in \Phi_s$ *iff* $\phi_1 \in \Phi_s$ *and* $\phi \in \Phi_0$ *iff* $\phi_1 \in \Phi_0$.

Proof. Suppose $[\phi'(t''_k) - \phi'(t'_k)](t''_k - t'_k) \to 0$ with $t'_k \le t_s \le t''_k \le t''_{\phi_2}$ (other cases being trivial). Since ϕ'_1 and ϕ'_2 are nondecreasing, so is ϕ'; therefore, all terms in

$$\begin{aligned}[\phi'(t''_k) - \phi'(t'_k)](t''_k - t'_k) &\ge [\phi'(t''_k) - \phi'(t_s)](t''_k - t_s) + [\phi'(t_s) - \phi'(t'_k)](t_s - t'_k)\\ &= [\phi'_2(t''_k) - \phi'_2(t_s)](t''_k - t_s) + [\phi'_1(t_s) - \phi'_1(t'_k)](t_s - t'_k)\end{aligned}$$

are nonnegative and tend to zero. Thus $\phi'_2(t_s)(t''_k - t_s) \to 0$ and $\phi'_1(t_s)(t_s - t'_k) \to 0$ (Definition 9.12). Hence $t'_k, t''_k \to t_s$ ($\phi'_2(t_s) = \phi'_1(t_s) > 0$), $\phi'(t''_k)(t''_k - t'_k) \to \phi'(t_s)0$, and $\phi'(t'_k)(t''_k - t'_k) \to 0$ yield the first assertion. For the second one, use Definition 9.3 and Remark 9.4. ∎

Examples 10.2. Using the notation of Lemma 10.1, we add the condition $\phi''_1(t_s) = \phi''_2(t_s)$ to make ϕ twice continuously differentiable. In each example, $\phi \in \Phi_s \cup \Phi_0$ is forcing on $(-\infty, t''_\phi]$ $\forall t''_\phi \in \mathbb{R}$; cf. Remark 9.4, Lemma 9.13, Example 9.14, and Remark 9.24.

(1 (cubic-quadratic)) $\phi(t) = \frac{\max\{0, t+t_s\}^3}{12 t_s} - \frac{t_s^2}{6}$ if $t \le t_s$, $\phi(t) = \frac{\max\{0,t\}^2}{2} = \phi_2(t)$ if $t > t_s$, $t_s > 0$. This ϕ grows only as fast as ϕ_2 in Example 7.10 with $\beta = 2$ but is smoother.

(2 (exponential-quadratic)) $\phi(t) = e^t$ if $t \le t_s > 0$, $\phi(t) = e^{t_s}(\frac{t^2}{2} + (1 - t_s)t + 1 - t_s - \frac{t_s^2}{2})$ if $t > t_s$, $\phi_2(\cdot) = a \max\{0, \cdot - t^0_{\phi_2}\}^2 + b$. This ϕ does not grow as fast as e^t in Example 7.11.

(3 (log-quadratic)) $\phi(t) = -\ln(-t) - 1 = \phi_1(t)$ if $t \le t_s < 0$, $\phi(t) = \frac{t^2}{2t_s^2} - \frac{2t}{t_s} + \frac{1}{2} - \ln(-t_s)$ if $t > t_s$. This ϕ allows arbitrarily large infeasibilities, in contrast to ϕ_1 in Example 7.12.

(4 (hyperbolic-quadratic)) $\phi(t) = -\frac{1}{t} = \phi_1(t)$ if $t \le t_s < 0$, $\phi(t) = \frac{t^2}{|t_s|^3} + \frac{3t}{t_s^2} - \frac{3}{t_s}$ if $t > t_s$. Again, this ϕ has $\mathcal{D}_\phi = \mathbb{R}$, in contrast to ϕ_1 in Example 7.13.

(5 (hyperbolic-log-quadratic)) $\phi(t) = \frac{-4t'_s}{-t'_s - t} - 2 - \ln(-t'_s)$ if $t \le t'_s < 0$, $\phi(t) = -\ln(-t)$ if $t'_s \le t \le t_s < 0$, $\phi(t) = \frac{t^2}{2t_s^2} - \frac{2t}{t_s} + \frac{3}{2} - \ln(-t_s)$ if $t > t_s$.

Remark 10.3. Other smooth penalty functions (e.g., cubic-log-quadratic) are easy to derive. Such functions are covered by the various results of section 9. Their properties, e.g., $\inf \phi > -\infty$, may also have practical significance; this should be verified experimentally.

The following result (inspired by [Ber82, Prop. 5.7]) shows that minimizing L_k (cf. (7.12)) in Algorithm 7.4 is well posed under mild conditions (see the appendix for its proof).

LEMMA 10.4. *Let* $h(\pi) = \sum_{i=1}^m \psi(\pi_i)$, *where* ψ *is a B-function with* $\mathcal{D}_\psi \supset \mathbb{R}_>$. *Suppose that* $L_k \not\equiv \infty$ (*e.g.,* $\inf_{\mathcal{D}_f} \max_{i=1}^m g_i \le 0$). *Then* $\operatorname{Arg\,min} L_k$ *is nonempty and compact iff* f *and* $g_1, \dots, g_m$ *have no common direction of recession, and if* $C_0 \ne \emptyset$, *then this is equivalent to* (7.1) *having a nonempty and compact set of solutions.*

We now consider a variant of condition (7.18), inspired by one in [Ber82, p. 328].

LEMMA 10.5. *Under the strong convexity assumption* (7.15), *consider* (7.17) *with*

$$|\Delta_x L_k(x^{k+1})|^2 \le \eta_k [L(x^{k+1}, \pi^{k+1}) - L_k(x^{k+1})] \tag{10.1}$$

and $\epsilon_k = |\Delta_x L_k(x^{k+1})|^2 / 2\breve{\alpha}$ *replacing* (7.18), *where* $\eta_k \ge 0$. *Then*

$$L(x^{k+1}, \pi^{k+1}) - d(\pi^{k+1}) \le \epsilon_k \le \frac{\eta_k}{2\breve{\alpha}} [L(x^{k+1}, \pi^{k+1}) - L_k(x^{k+1})], \tag{10.2}$$

$$d(\pi^k) \le L(x^{k+1}, \pi^k) \le L_k(x^{k+1}) \le d(\pi^{k+1}) \quad \text{if} \quad \eta_k \le 2\breve{\alpha}, \tag{10.3}$$

$$\epsilon_k \le \frac{\eta_k}{\breve{\alpha}}[d(\pi^{k+1}) - d(\pi^k)] \le d(\pi^{k+1}) - d(\pi^k) \quad \text{if} \quad \eta_k \le \breve{\alpha}. \tag{10.4}$$

Next, suppose $\eta_k \to 0$ *in* (10.1). *Then* $d(\pi^k) \to d^\infty \in (-\infty, \infty]$. *If* $d^\infty < \infty$, *then* $\sum_{k=1}^\infty \epsilon_k < \infty$, $\epsilon_k \to 0$, $\sum_{j=1}^k c_j\epsilon_j/s_k \to 0$; *further,* $\sum_{k=1}^\infty c_k\epsilon_k < \infty$ *if* $\{c_k\eta_k\}$ *is bounded.*

Proof. By (7.17) and (10.1), (10.2) holds with $L(x^{k+1}, \pi^{k+1}) \ge L_k(x^{k+1})$ by (8.2). Thus $\eta_k \le 2\breve{\alpha}$ yields $L_k(x^{k+1}) \le d(\pi^{k+1})$ and (10.3) follows from (8.5). Similarly, $L(x^{k+1}, \pi^{k+1}) - d(\pi^{k+1}) \le \frac{1}{2}[L(x^{k+1}, \pi^{k+1}) - L_k(x^{k+1})]$ for $\eta_k \le \breve{\alpha}$ yields $L(x^{k+1}, \pi^{k+1}) - L_k(x^{k+1}) \le 2[d(\pi^{k+1}) - L_k(x^{k+1})]$, so (10.4) follows from (10.2) and $d(\pi^k) \le L_k(x^{k+1})$ (cf. (10.3)). Next, let $\eta_k \to 0$. Pick $\bar{k}$ s.t. $\eta_k \le \breve{\alpha}$ $\forall k \ge \bar{k}$. Equations (10.3)–(10.4) yield $d(\pi^k) \to d^\infty$, $\sum_{k=\bar{k}}^\infty \epsilon_k \le [d^\infty - d(\pi^{\bar{k}})]$, $\sum_{k=\bar{k}}^\infty c_k\epsilon_k \le \sup_k \frac{c_k\eta_k}{\breve{\alpha}}[d^\infty - d(\pi^{\bar{k}})]$. If $d^\infty < \infty$, then $\epsilon_k \to 0$ gives $\sum_{j=1}^k c_j\epsilon_j/s_k \to 0$ (Lemma 4.5(i)). ∎

Remark 10.6. In view of Lemma 10.5, suppose in the strongly convex case of (7.15), (10.1) is used with $\eta_k \to 0$. Since $q(\pi^{k+1}) \le q(\pi^k)$ for all large k (cf. (10.3)), the results of sections 8 and 9 may invoke, instead of Theorem 5.2 with $\sum_{j=1}^k s_j\epsilon_j/s_k \to 0$, Theorem 4.3 with $\sum_{j=1}^k c_j\epsilon_j/s_k \to 0$. The latter condition holds automatically if $\lim_{k\to\infty} d(\pi^k) < \infty$, e.g., $\sup d < \infty$. Thus we may drop the conditions

$$\sum_{j=1}^k s_j\epsilon_j/s_k \to 0$$

from Theorems 8.3, 8.4, and 9.18, $\epsilon_k \to 0$ from Lemma 8.10 and Theorem 9.20, and $\sum_{k=1}^\infty \epsilon_k < \infty$ from Theorems 9.22 and 9.23. Instead of $\sum_{k=1}^\infty c_k\epsilon_k < \infty$, we may assume that $\{c_k\eta_k\}$ is bounded in Theorems 8.3, 8.4, 9.18, 9.22, and 9.23.

Condition (10.1) can be implemented as in [Ber82, Prop. 5.7(b)].

LEMMA 10.7. *Suppose that* f *is strongly convex,* $\inf_{\mathcal{D}_f} \max_{i=1}^m g_i \le 0$, *and* g *is continuous on* $\mathcal{D}_f$. *Consider iteration* k *of Example* 7.6 *with* $h(\pi) = \sum_{i=1}^m \psi(\pi_i)$, *where* ψ *is a B-function s.t.* $\mathcal{D}_{\nabla\psi} \supset \mathbb{R}_>$. *If* $\eta_k > 0$, π^k *is not a Lagrange multiplier of* (7.1), $\{z^j\}$ *is a sequence converging to* $\hat{x} = \arg\min L_k$, *and* $\Delta_x L_k(z^j) \to 0$, *then there exists* $x^{k+1} \in \{z^1, z^2, \ldots\}$ *satisfying the stopping criterion* (10.1).

Proof. By Lemmas 9.5 and 9.6, Example 7.6 has $\bar{u}_i = t_\phi^0$, $\pi_i^k = \nabla\phi(\gamma_i^k)$, $\gamma_i^k \ge t_\phi^0$, $i = 1{:}\,m$, $h^+(u) = \sum_{i=1}^m \phi(u_i)$, where $\phi = \psi^+ \in \Phi_s \cup \Phi_0$. Let $\hat{u} = g(\hat{x})$ and $\hat{\pi} = \nabla h^+(\gamma^k + c_k\hat{u})$. Then, as in (8.2),

$$L(\hat{x}, \hat{\pi}) - L_k(\hat{x}) = D_{h^+}(\gamma^k, \gamma^k + c_k\hat{u})/c_k \ge 0. \tag{10.5}$$

Suppose that $L(\hat{x}, \hat{\pi}) = L_k(\hat{x})$. By (10.5), (2.4), and convexity of h^+, $\phi(\gamma_i^k) - \phi(\gamma_i^k + c_k\hat{u}_i) - \nabla\phi(\gamma_i^k + c_k\hat{u}_i)(-c_k\hat{u}_i) = 0$, $i = 1{:}\,m$. Therefore, since ϕ is strictly convex on $[t_\phi^0, t_\phi)$ with $\nabla\phi(t) = 0$ iff $t \le t_\phi^0$ (Definition 9.3), and $\gamma_i^k \ge t_\phi^0$, for each i, either $\gamma_i^k + c_k\hat{u}_i = \gamma_i^k > t_\phi^0$ yields $\hat{u}_i = 0$ and $\hat{\pi}_i = \pi_i^k = \nabla\phi(\gamma_i^k)$ or $\gamma_i^k + c_k\hat{u}_i \le t_\phi^0 = \gamma_i^k$ yields $\hat{u}_i \le 0$ and $\hat{\pi}_i = \pi_i^k = \nabla\phi(\gamma_i^k) = 0$. Hence $\hat{\pi} = \pi^k$, $\hat{u} \le 0$ and $\langle\hat{\pi}, \hat{u}\rangle = 0$. Combining this with $0 \in \partial L_k(\hat{x}) = \partial_x L(\hat{x}, \hat{\pi})$ (Lemma 7.3), we see (cf. [Roc70, Thm. 28.3]) that π^k is a Lagrange multiplier, a contradiction. Therefore, we must have strict inequality in (10.5). Since $g(z^j) \to \hat{u}$ and $D_{h^+}(\gamma^k, \gamma^k + c_k g(z^j)) \to D_{h^+}(\gamma^k, \gamma^k + c_k\hat{u}) > 0$ by continuity, whereas $\eta_k > 0$ and $\Delta_x L_k(z^j) \to 0$, the stopping criterion will be satisfied for sufficiently large j. ∎

Appendix A. We now give proofs of certain technical results.

Proof of Lemma 7.2. $\mathbb{R}^m_+ \cap \operatorname{ri} \mathcal{D}_h \neq \emptyset$ implies $\partial h_+ = \partial h + \partial \iota_{\mathbb{R}^m_+}$ [Roc70, Thm. 23.8], so $\mathcal{D}_{\partial h_+} = \mathcal{D}_{\partial h} \cap \mathbb{R}^m_+$ and h_+ is essentially strictly convex. Hence $h^+ = h^*_+$ is closed proper essentially smooth (Fact 2.3(ii)), $\partial h^+(u) = \{\nabla h^+(u)\}\ \forall u \in \mathring{\mathcal{D}}_{h^+} = \mathcal{D}_{\partial h^+}$ (Fact 2.3(i)), ∇h^+ is continuous on $\mathring{\mathcal{D}}_{h^+}$ (Fact 2.2), $\partial h^*_+ = (\partial h_+)^{-1}$, and $\operatorname{im} \partial h_+ = \mathcal{D}_{\partial h^+}$ (Fact 2.1). Since h^+ is nondecreasing, $\mathcal{D}_{h^+} = \mathcal{D}_{h^+} - \mathbb{R}^m_+$, so $\mathring{\mathcal{D}}_{h^+} = \mathring{\mathcal{D}}_{h^+} - \mathbb{R}^m_+$ as the union of open sets. That $N_{\mathbb{R}^m_+}(\pi) = \{\gamma \leq 0 : \langle \gamma, \pi \rangle = 0\}$ for $\pi \geq 0$ is elementary (cf. [Roc70, p. 226]). If $\pi = \nabla h^+(\gamma)$ and $\tilde{\gamma} \in N_{\mathbb{R}^m_+}(\pi)$, then $\gamma \in \partial h_+(\pi)$ and $\gamma + \tilde{\gamma} \in \partial h_+(\pi)$, so $\pi = \nabla h^+(\gamma + \tilde{\gamma})$. If $\operatorname{im} \partial h \supset \mathbb{R}^m_>$ and $u \in \mathbb{R}^m$, then $-h^+(u) = \inf \phi$, where $\phi = h_+ - \langle u, \cdot \rangle$ is inf-compact. Indeed, pick $\tilde{\pi}$ and $\tilde{u} \in \partial h(\tilde{\pi})$ s.t. $\tilde{u} > u$. Then $\tilde{\phi}(\pi) = h(\tilde{\pi}) + \langle \tilde{u}, \pi - \tilde{\pi} \rangle - \langle u, \pi \rangle \leq \phi(\pi)$ for all $\pi \geq 0$, and if $\{\pi^k\} \subset \mathbb{R}^m_+$, $|\pi^k| \to \infty$, then $\tilde{\phi}(\pi^k) \to \infty$ since $\tilde{u} - u > 0$. Hence ϕ is inf-compact and $u \in \mathcal{D}_{h^+}$, so $\mathcal{D}_{h^+} = \mathbb{R}^m$. ∎

We need the following slightly sharpened version of [GoT89, Thm. 1.5.4].

LEMMA A.1 (subdifferential chain rule). *Let $f_1, \ldots, f_m$ be proper convex functions on $\mathbb{R}^n$ with $\bigcap_{i=1}^m \operatorname{ri} \mathcal{D}_{f_i} \neq \emptyset$. Let $f(\cdot) = (f_1(\cdot), \ldots, f_m(\cdot))$ and $\mathcal{D}_f = \bigcap_{i=1}^m \mathcal{D}_{f_i}$. Let ϕ be a proper convex nondecreasing function on $\mathbb{R}^m$ s.t. $f(\tilde{x}) < \tilde{y}$ for some $\tilde{x} \in \mathcal{D}_f$ and $\tilde{y} \in \mathcal{D}_\phi$. Let $\psi(x) = \phi(f(x))$ if $x \in \mathcal{D}_f$, $\psi(x) = \infty$ if $x \notin \mathcal{D}_f$. Then ψ is proper convex, $\operatorname{im} \partial \phi \subset \mathbb{R}^m_+$, and for each $\bar{x} \in \mathcal{D}_f$ and $\bar{y} = f(\bar{x})$*

$$\partial \psi(\bar{x}) = \bigcup \{ \sum_{i=1}^m \gamma_i \partial f_i(\bar{x}) : \gamma \in \partial \phi(\bar{y}) \}. \tag{A.1}$$

Proof. For any $x^1, x^2 \in \mathcal{D}_f$ and $\lambda \in [0,1]$, $f(\lambda x^1 + (1-\lambda)x^2) \leq \lambda f(x^1) + (1-\lambda) f(x^2)$ and hence $\psi(\lambda x^1 + (1-\lambda)x^2) \leq \phi(\lambda f(x^1) + (1-\lambda) f(x^2)) \leq \lambda \psi(x^1) + (1-\lambda)\psi(x^2)$, so ψ is convex. Since $\psi(x) > -\infty$ for all x, ψ is proper. Let $Q = \bigcup_{\gamma \in \partial \phi(\bar{y})} \sum_{i=1}^m \gamma_i \partial f_i(\bar{x})$. Let $\gamma \in \partial \phi(\bar{y})$, $\gamma^i \in \partial f_i(\bar{x})$, $i = 1\colon m$, $\Gamma = [\gamma^1, \ldots, \gamma^m]^T$. For any x, $f(x) \geq f(\bar{x}) + \Gamma(x - \bar{x})$ yields $\psi(x) \geq \phi(f(\bar{x}) + \Gamma(x - \bar{x})) \geq \psi(\bar{x}) + \gamma^T \Gamma(x - \bar{x})$, i.e., $\Gamma^T \gamma \in \partial \psi(\bar{x})$, so $Q \subset \partial \psi(\bar{x})$. To prove the opposite inclusion, let $\bar{\gamma} \in \partial \psi(\bar{x})$. Consider the convex program

$$\text{minimize} \quad \phi(y) - \langle \bar{\gamma}, x \rangle \ \text{ s.t. } f(x) - y \leq 0,\ x \in \mathcal{D}_f,\ y \in \mathcal{D}_\phi. \tag{A.2}$$

By the monotonicity of ϕ and the definition of subdifferential, $(\bar{x}, \bar{y})$ solves (A.2), which satisfies Slater's condition (cf. $f(\tilde{x}) < \tilde{y}$), so (cf. [Roc70, Cor. 28.2.1]) it has a Kuhn–Tucker point $\bar{\pi} \in \mathbb{R}^m_+$ s.t. (cf. [Roc70, Thm. 28.3])

$$\phi(y) - \langle \bar{\gamma}, x \rangle + \langle \bar{\pi}, f(x) - y \rangle \geq \phi(\bar{y}) - \langle \bar{\gamma}, \bar{x} \rangle + \langle \bar{\pi}, f(\bar{x}) - \bar{y} \rangle \quad \forall x \in \mathcal{D}_f, y \in \mathcal{D}_\phi.$$

Then $\phi(y) \geq \phi(\bar{y}) + \langle \bar{\pi}, y - \bar{y} \rangle\ \forall y$ yields $\bar{\pi} \in \partial \phi(\bar{y})$, whereas $\langle \bar{\pi}, f(x) \rangle \geq \langle \bar{\pi}, f(\bar{x}) \rangle + \langle \bar{\gamma}, x - \bar{x} \rangle\ \forall x$ yields $\bar{\gamma} \in \partial(\sum_{i=1}^m \bar{\pi}_i f_i)(\bar{x}) = \sum_{i=1}^m \bar{\pi}_i \partial f_i(\bar{x})$ from $\bigcap_{i=1}^m \operatorname{ri} \mathcal{D}_{f_i} \neq \emptyset$ (cf. [Roc70, Thm. 23.8]). Thus $\partial \psi(\bar{x}) \subset Q$, i.e., $\partial \psi(\bar{x}) = Q$. To see that $\operatorname{im} \partial \phi \subset \mathbb{R}^m_+$, note that if $\gamma \in \partial \phi(y^1)$, then $\phi(y^1) \geq \phi(y^2) \geq \phi(y^1) + \langle \gamma, y^2 - y^1 \rangle$ for all $y^2 \leq y^1$ implies $\gamma \geq 0$. ∎

Proof of Lemma 10.4. Let $\phi_i(x) = \psi^+(\gamma^k_i + c_k g_i(x))$ if $x \in \mathcal{D}_{g_i}$, $\phi_i(x) = \infty$ if $x \notin \mathcal{D}_{g_i}$, $i = 1\colon m$. Each ϕ_i is closed: for any $\alpha \in \mathbb{R}$, $\{t : \psi^+(t) \leq \alpha\} = (-\infty, \beta]$ for some $\beta < \infty$ (ψ^+ is closed nondecreasing and $\lim_{t \uparrow t_{\psi^+}} \psi^+(t) = \infty$ by Lemmas 9.5 and 9.6) and $\{x : \phi_i(x) \leq \alpha\} = \{x : g_i(x) \leq (\beta - \gamma^k_i)/c_k\}$ is closed (as is g_i). We have $L_k = f + \frac{1}{c_k} \sum_{i=1}^m [\phi_i - \psi^+(\gamma^k_i)]$ with f and ϕ_i closed proper and $L_k \not\equiv \infty$, so L_k is closed and $L_k 0^+ = f 0^+ + \frac{1}{c_k} \sum_{i=1}^m \phi_i 0^+$ [Roc70, Thm. 9.3]. Suppose that $g_i 0^+(y) \leq 0$. Since $L_k \not\equiv \infty$, $\mathcal{D}_{\psi^+} = (-\infty, t_{\psi^+})$ (cf. Lemma 9.6 and Definition 9.3),

and g_i is closed, there is $x \in \operatorname{ri} \mathcal{D}_{g_i}$ s.t. $\gamma_i^k + c_k g_i(x) \in \mathcal{D}_{\psi^+}$. Let $\gamma \in \partial g_i(x)$. Then $g_i(x)+t\,\langle \gamma, y\rangle \le g_i(x+ty) \le g_i(x)\ \forall t \ge 0$, so $\langle \gamma, y\rangle \le 0$ and, since ψ^+ is nondecreasing, $\psi^+(\gamma_i^k + c_k[g_i(x)+t\,\langle \gamma, y\rangle]) \le \psi^+(\gamma_k^k + c_k g_i(x+ty)) \le \psi^+(\gamma_i^k + c_k g_i(x))\ \forall t \ge 0$. Hence $\psi^+0^+(c_k\,\langle \gamma, y\rangle) \le \phi_i 0^+(y) \le 0$, so $\langle \gamma, y\rangle \le 0$ and $\psi^+0^+ = \imath^*_{\mathbb{R}_+}$ (cf. Lemmas 9.5 and 9.6) yield $\phi_i 0^+(y) = 0$. Now suppose $g_i 0^+(y) > 0$. Pick $\bar t > 0$ and $\bar\alpha > 0$ s.t. $[g_i(x+ty) - g_i(x)]/t \ge \bar\alpha\ \forall t \ge \bar t$. Then

$$\begin{aligned}
\phi_i 0^+(y) &= \lim_{t\uparrow\infty}[\psi^+(\gamma_i^k + c_k g_i(x+ty)) - \psi^+(\gamma_i^k + c_k g_i(x))]/t \\
&\ge \lim_{t\uparrow\infty}[\psi^+(\gamma_i^k + c_k (g_i(x) + t\bar\alpha)) - \psi^+(\gamma_i^k + c_k g_i(x))]/t \\
&= \psi^+0^+(c_k\bar\alpha) = \infty
\end{aligned}$$

from $\psi^+0^+ = \imath^*_{\mathbb{R}_+}$. Thus $\phi_i 0^+(y) = 0$ if $g_i 0^+(y) \le 0$, $\phi_i 0^+(y) = \infty$ if $g_i 0^+(y) > 0$. Therefore, $L_k 0^+(y) = f 0^+(y)$ if $g_i 0^+(y) \le 0$ for $i = 1{:}\,m$, $L_k 0^+(y) = \infty$ otherwise. The proof may be finished as in [Ber82, sect. 5.3]. □

Acknowledgments. I would like to thank the associate editor and the anonymous referee for their valuable comments.

REFERENCES

[Aus86] A. Auslender, *Numerical methods for nondifferentiable convex optimization*, Math. Programming Stud., 30 (1986), pp. 102–126.

[BaB97] H. H. Bauschke and J. M. Borwein, *Legendre functions and the method of random Bregman projections*, J. Convex Anal., 4 (1997), to appear.

[Ber82] D. P. Bertsekas, *Constrained Optimization and Lagrange Multiplier Methods*, Academic Press, New York, 1982.

[BeT94] D. P. Bertsekas and P. Tseng, *Partial proximal minimization algorithms for convex programming*, SIAM J. Optim., 4 (1994), pp. 551–572.

[Bre67] L. M. Bregman, *The relaxation method of finding the common point of convex sets and its application to the solution of problems in convex programming*, Zh. Vychisl. Mat. i Mat. Fiz., 7 (1967), pp. 620–631 (in Russian). English transl. in U.S.S.R. Comput. Math. and Math. Phys., 7 (1967), pp. 200–217.

[BrS93] M. G. Breitfeld and D. F. Shanno, *Computational Experience with Modified Log-Barrier Methods for Nonlinear Programming*, Research Report RRR 17-93, RUTCOR, Rutgers Univ., New Brunswick, NJ, 1993. Revised March 1994.

[BrS94] M. G. Breitfeld and D. F. Shanno, *A Globally Convergent Penalty-Barrier Algorithm for Nonlinear Programming and Its Computational Performance*, Research Report RRR 12-94, RUTCOR, Rutgers Univ., New Brunswick, NJ, 1994.

[BTYZ92] A. Ben-Tal, I. Yuzefovich, and M. Zibulevsky, *Penalty/Barrier Multiplier Methods for Minimax and Constrained Smooth Convex Programs*, Research Report 9/92, Optimization Laboratory, Technion, Haifa, Israel, 1992.

[BuF93] J. V. Burke and M. C. Ferris, *Weak sharp minima in mathematical programming*, SIAM J. Control Optim., 31 (1993), pp. 1340–1359.

[CDPI91] Y. Censor, A. R. De Pierro, and A. N. Iusem, *Optimization of Burg's entropy over linear constraints*, Appl. Numer. Math., 7 (1991), pp. 151–165.

[CeL81] Y. Censor and A. Lent, *An iterative row action method for interval convex programming*, J. Optim. Theory Appl., 34 (1981), pp. 321–353.

[CeZ92] Y. Censor and S. A. Zenios, *Proximal minimization algorithm with D-functions*, J. Optim. Theory Appl., 73 (1992), pp. 451–464.

[CGT92] A. R. Conn, N. I. M. Gould, and Ph. L. Toint, *A Globally Convergent Lagrangian Barrier Algorithm for Optimization with General Inequality Constraints and Simple Bounds*, Report 92/07, Département de Mathématique, Facultés Universitaires de Namur, Namur, 1992.

[CGT94] A. R. Conn, N. I. M. Gould, and Ph. L. Toint, *Large-scale nonlinear constrained optimization: A current survey*, in Algorithms for Continuous Optimization: The State of the Art, E. Spedicato, ed., Kluwer, Dordrecht, the Netherlands, 1994, pp. 287–327.

[Cha94] I. Chabini, *Nouvelles Méthodes Séquentielles et Parralléles pour l'Optimisation de Réseaux à Coôts Linéaires et Convexes*, Ph.D. thesis, Département d'informatique et de recherche opérationelle, Université de Montréal, Montréal, Québec, Canada, 1994.

[ChT93] G. Chen and M. Teboulle, *Convergence analysis of a proximal-like minimization algorithm using Bregman functions*, SIAM J. Optim., 3 (1993), pp. 538–543.

[CoL93] R. Correa and C. Lemaréchal, *Convergence of some algorithms for convex minimization*, Math. Programming, 62 (1993), pp. 261–275.

[EcB92] J. Eckstein and D. P. Bertsekas, *On the Douglas-Rachford splitting method and the proximal point algorithm for maximal monotone operators*, Math. Programming, 55 (1992), pp. 293–318.

[Eck93] J. Eckstein, *Nonlinear proximal point algorithms using Bregman functions, with applications to convex programming*, Math. Oper. Res., 18 (1993), pp. 202–226.

[Egg90] P. P. B. Eggermont, *Multiplicative iterative algorithms for convex programming*, Linear Algebra Appl., 130 (1990), pp. 25–42.

[Fer91] M. C. Ferris, *Finite termination of the proximal point algorithm*, Math. Programming, 50 (1991), pp. 359–366.

[FiM68] A. V. Fiacco and G. P. McCormick, *Nonlinear Programming: Sequential Unconstrained Minimization Techniques*, Wiley, New York, 1968.

[Flå94] S. D. Flåm, *Equilibrium Programming Using Proximal-like Algorithms*, Working Paper, Dept. of Economics, Univ. of Bergen, Bergen, Norway, 1994.

[GMSW88] P. E. Gill, W. Murray, M. A. Saunders, and M. H. Wright, *Shifted Barrier Methods for Linear Programming*, Report SOL 88-9, Department of Operations Research, Stanford Univ., Stanford, CA, 1988.

[GoT89] E. G. Golshtein and N. V. Tretyakov, *Modified Lagrange Functions; Theory and Optimization Methods*, Nauka, Moscow, 1989 (in Russian).

[Gül91] O. Güler, *On the convergence of the proximal point algorithm for convex minimization*, SIAM J. Control Optim., 29 (1991), pp. 403–419.

[IST94] A. N. Iusem, B. Svaiter, and M. Teboulle, *Entropy-like proximal methods in convex programming*, Math. Oper. Res., 19 (1994), pp. 790–814.

[Ius95] A. N. Iusem, *On some properties of generalized proximal point methods for quadratic and linear programming*, J. Optim. Theory Appl., (1995), to appear.

[IuT93] A. N. Iusem and M. Teboulle, *On the convergence rate of entropic proximal optimization algorithms*, Math. Apl. Comput., 12 (1993), pp. 153–168.

[JeP94] D. L. Jensen and R. A. Polyak, *The convergence of a modified barrier method for convex programming*, IBM J. Res. Develop., 38 (1994), pp. 307–321.

[Kiw96] K. C. Kiwiel, *On the twice differentiable cubic augmented Lagrangian*, J. Optim. Theory Appl., 88 (1996), to appear.

[Kiw97] K. C. Kiwiel, *Free-steering relaxation methods for problems with strictly convex costs and linear constraints*, Math. Oper. Res., (1996), to appear.

[Lem89] B. Lemaire, *The proximal algorithm*, in New Methods in Optimization and Their Industrial Uses, Internat. Ser. Numer. Math. 87, J. P. Penot, ed., Birkhäuser, Basel, 1989, pp. 73–87.

[Mar70] B. Martinet, *Régularisation d'inéquations variationelles par approximations successives*, RAIRO Rech. Opér., 4(R3) (1970), pp. 154–158.

[NiZ92] S. S. Nielsen and S. A. Zenios, *Massively parallel algorithms for singly constrained convex programs*, ORSA J. Comput., 4 (1992), pp. 166–181.

[NiZ93a] S. S. Nielsen and S. A. Zenios, *A massively parallel algorithm for nonlinear stochastic network problems*, Oper. Res., 41 (1993), pp. 319–337.

[NiZ93b] S. S. Nielsen and S. A. Zenios, *Proximal minimizations with D-functions and the massively parallel solution of linear network programs*, Comput. Optim. Appl., 1 (1993), pp. 375–398.

[NPS94] S. G. Nash, R. Polyak, and A. Sofer, *A numerical comparison of barrier and modified barrier methods for large-scale bound-constrained optimization*, in Large Scale Optimization: State of the Art, W. W. Hager, D. W. Hearn, and P. M. Pardalos, eds., Kluwer, Dordrecht, the Netherlands, 1994, pp. 319–338.

[Pol83] B. T. Polyak, *Introduction to Optimization*, Nauka, Moscow, 1983. English transl., Optimization Software Inc., New York, 1987.

[Pol92] R. Polyak, *Modified barrier functions (theory and methods)*, Math. Programming, 54 (1992), pp. 177–222.

[PoT97] R. Polyak and M. Teboulle, *Nonlinear rescaling and proximal-like methods in convex optimization*, Math. Programming, 76 (1997), pp. 265–284.

[Roc70] R. T. Rockafellar, *Convex Analysis*, Princeton University Press, Princeton, NJ, 1970.

[Roc76a] R. T. Rockafellar, *Augmented Lagrangians and applications of the proximal point algorithm in convex programming*, Math. Oper. Res., 1 (1976), pp. 97–116.

[Roc76b] R. T. Rockafellar, *Monotone operators and the proximal point algorithm*, SIAM J. Control Optim., 14 (1976), pp. 877–898.

[Teb92] M. Teboulle, *Entropic proximal mappings with applications to nonlinear programming*, Math. Oper. Res., 17 (1992), pp. 670–690.

[TsB93] P. Tseng and D. P. Bertsekas, *On the convergence of the exponential multiplier method for convex programming*, Math. Programming, 60 (1993), pp. 1–19.

[Tse90] P. Tseng, *Dual ascent methods for problems with strictly convex costs and linear constraints: A unified approach*, SIAM J. Control Optim., 28 (1990), pp. 214–242.

SIAM J. CONTROL OPTIM.
Vol. 35, No. 4, pp. 1169–1203, July 1997

005

TURNPIKE PROPERTY OF OPTIMAL SOLUTIONS OF INFINITE-HORIZON VARIATIONAL PROBLEMS*

A. J. ZASLAVSKI†

Dedicated to Valery L. Makarov on his 60th birthday.

Abstract. Given $x_0 \in R^n$ we study the infinite-horizon problem of minimizing the expression $\int_0^T f(x(t), x'(t))dt$ as T grows to infinity, where a function $x\colon [0,\infty) \to R^n$ is absolutely continuous and satisfies the initial condition $x(0) = x_0$ and $f = f(x,u)$ is an integrand. We study the structure of (f)-weakly optimal solutions and establish the turnpike property for a generic integrand f.

Key words. infinite horizon, weakly optimal function, good function, turnpike property

AMS subject classification. 49J99

PII. S036301299528935X

1. Introduction.

In this paper we consider a special class of extremals, the so-called *weakly optimal solutions* of infinite-horizon autonomous variational problems with vector-valued functions.

Given $x_0 \in R^n$ we study the infinite-horizon problem of minimizing the expression $\int_0^T f(x(t), x'(t))dt$ as T grows to infinity, where a function $x\colon [0,\infty) \to R^n$ is absolutely continuous (a.c.) and satisfies the initial condition $x(0) = x_0$ and $f = f(x,u)$ is an integrand.

The study of variational and optimal control problems defined on infinite intervals has recently been a rapidly growing area of research. These problems arise in engineering (see Anderson and Moore [1], Artstein and Leizarowitz [2], Leizarowitz [14]), in models of economic growth (see Rockafellar [18], Brock and Haurie [3], Leizarowitz [12], Haurie [10], Carlson [5]), and in theory of thermodynamical equilibrium for materials (see Leizarowitz and Mizel [16], Coleman, Marcus, and Mizel [8], Zaslavski [23, 24]).

The following notion, known as *the overtaking optimality criterion*, was introduced in the economics literature by Gale [9] and von Weizsacker [20] and has been used in control theory by Artstein and Leizarowitz [2]; Brock and Haurie [3]; Carlson [5]; Carlson, Haurie, and Jabrane [6]; and Carlson, Haurie, and Leizarowitz [7].

An a.c. function $x\colon [0,\infty) \to R^n$ is called (f)-*overtaking optimal* if, for any a.c. function $y\colon [0,\infty) \to R^n$ satisfying $y(0) = x(0)$,

$$\limsup_{T\to\infty} \int_0^T [f(x(t), x'(t)) - f(y(t), y'(t))]\, dt \le 0.$$

Various existence results of overtaking optimal functions are nicely collected in Carlson, Haurie, and Leizarowitz [7]. The most typical infinite-horizon optimization problem for which the existence of overtaking optimal function has been established is an autonomous variational problem with a convex integrand $\int_0^T f(x(t),\ x'(t))dt$, studied by Rockafellar [18], Brock and Haurie [3], and Leizarowitz [12].

For convex integrands the existence of overtaking optimal solutions may follow from the fact that all good trajectories converge to a unique steady state (Brock and

*Received by the editors July 24, 1995; accepted for publication April 24, 1996.
http://www.siam.org/journals/sicon/35-4/28935.html.
†Department of Mathematics, Technion-Israel Institute of Technology, Haifa 32000, Israel (mar9319@technion.technion.ac.il).

Haurie [3], Leizarowitz [12]). For nonconvex integrands the existence of overtaking optimal solutions is not guaranteed, and in this situation we look for *weakly optimal solutions*.

In this paper we employ the following weakened version of the overtaking optimality criterion.

An a.c. function $x\colon [0,\infty) \to R^n$ is called (f)-*weakly optimal* if, for any a.c. function $y\colon [0,\infty) \to R^n$ satisfying $y(0) = x(0)$,

$$\liminf_{T\to\infty} \int_0^T [f(x(t), x'(t)) - f(y(t), y'(t))]\, dt \le 0.$$

The first existence result of weakly optimal solutions without convexity assumptions was obtained by Carlson [4] for autonomous optimal control problems with vector-valued functions. Under the assumptions posed in Carlson [4] for every (f)-good trajectory $x(\cdot)$ defined on $[0,\infty)$ the following holds:

$$\tau^{-1} \int_0^\tau x(t)\, dt \to \bar{x} \quad \text{as } \tau \to \infty,$$

where $\bar{x}$ is a unique steady state. Using this fact Carlson established the existence of weakly optimal solutions.

In a general situation, when we do not have any kind of a convergence property of all good trajectories to a unique steady state, our consideration can be based on the following observation, which was described in Leizarowitz [15]: *For every initial state there exists a weakly optimal solution if all good trajectories have the same limit point set.*

Recently this was used by Zaslavski [25] to establish, for a class of variational problems described below, the existence of weakly optimal solutions for a generic integrand and any initial state.

Denote by $|\cdot|$ the Euclidean norm in R^n, and denote by $\mathfrak{A}$ the set of continuous functions $f\colon R^n \times R^n \to R^1$ which satisfy the following assumptions.

Assumption A.

(i) For each $x \in R^n$ the function $f(x,\cdot)\colon R^n \to R^1$ is convex.

(ii) $f(x,u) \ge \sup\{\psi(|x|), \psi(|u|)|u|\} - a$ for each $(x,u) \in R^n \times R^n$, where $a > 0$ is a constant and $\psi\colon [0,\infty) \to [0,\infty)$ is an increasing function such that $\psi(t) \to +\infty$ as $t \to \infty$ (here a and ψ are independent on f).

(iii) For each $M, \varepsilon > 0$ there exist $\Gamma, \delta > 0$ such that

$$|f(x_1,u_1) - f(x_2,u_2)| \le \varepsilon \sup\{f(x_1,u_1), f(x_2,u_2)\}$$

for each $u_1, u_2, x_1, x_2 \in R^n$ which satisfy

$$|x_i| \le M, \quad |u_i| \ge \Gamma \quad (i = 1,2), \quad \sup\{|x_1 - x_2|, |u_1 - u_2|\} \le \delta.$$

It is an elementary exercise to show that an integrand $f = f(x,u) \in C^1(R^{2n})$ belongs to $\mathfrak{A}$ if f satisfies assumptions A(i) and A(ii) with a constant $a > 0$ and a function $\psi\colon [0,\infty) \to [0,\infty)$ and there exists an increasing function $\psi_0\colon [0,\infty) \to [0,\infty)$ such that for each $x, u \in R^n$

$$\sup\{|\partial f/\partial x(x,u)|, \ \ |\partial f/\partial u(x,u)|\} \le \psi_0(|x|)(1 + \psi(|u|)|u|).$$

For the set $\mathfrak{A}$ we consider the uniformity which is determined by the following base:

$$E(N,\varepsilon,\lambda)=\{(f,g)\in\mathfrak{A}\times\mathfrak{A}:\ \ |f(x,u)-g(x,u)|\le\varepsilon\ \ (u,x\in R^n,\ \ |x|,|u|\le N),$$
$$(|f(x,u)|+1)(|g(x,u)|+1)^{-1}\in[\lambda^{-1},\lambda]\ (x,u\in R^n,\ |x|\le N)\},$$

where $N>0$, $\varepsilon>0$, $\lambda>1$ (see Kelley [11]).

It was shown in Zaslavski [25] that the uniform space $\mathfrak{A}$ is metrizable and complete. We consider functionals of the form

$$I^f(T_1,T_2,x)=\int_{T_1}^{T_2}f(x(t),\ x'(t))\,dt, \tag{1.1}$$

where $f\in\mathfrak{A}$, $0\le T_1<T_2<+\infty$ and $x\colon [T_1,T_2]\to R^n$ is an a.c. function.

For $f\in\mathfrak{A}$, y, $z\in R^n$, and numbers T_1,T_2 satisfying $0\le T_1<T_2$ we set

$$U^f(T_1,T_2,y,z)=\inf\{I^f(T_1,T_2,x)\colon x\colon[T_1,T_2]\to R^n \text{ is an a.c. function}$$
$$\text{satisfying } x(T_1)=y,\ x(T_2)=z\}. \tag{1.2}$$

It is easy to see that $-\infty<U^f(T_1,T_2,y,z)<+\infty$ for each $f\in\mathfrak{A}$, each $y,z\in R^n$, and each numbers T_1,T_2 satisfying $0\le T_1<T_2$.

Let $f\in\mathfrak{A}$. For any a.c. function $x\colon[0,\infty)\to R^n$ we set

$$J(x)=\liminf_{T\to\infty}T^{-1}I^f(0,T,x). \tag{1.3}$$

Of special interest is the *minimal long-run average cost growth rate*

$$\mu(f)=\inf\{J(x)\colon x\colon[0,\infty)\to R^n \text{ is an a.c. function}\}. \tag{1.4}$$

Clearly $-\infty<\mu(f)<+\infty$. Here we follow Leizarowitz [13] in defining "good functions" for the infinite-horizon variational problem with the integrand f.

An a.c. function $x\colon[0,\infty)\to R^n$ is called an *(f)-good function* if the function $\Phi_x^f\colon T\to I^f(0,T,x)-\mu(f)T$, $T\in(0,\infty)$ is bounded.

Propositions 1.1 and 3.1 in Zaslavski [25] imply the following result.

PROPOSITION 1.1. *For any a.c. function $x\colon[0,\infty)\to R^n$ either*

$$I^f(0,T,x)-T\mu(f)\to+\infty \text{ as } T\to\infty$$

or

$$\sup\{|I^f(0,T,x)-T\mu(f)|\colon T\in(0,\infty)\}<\infty.$$

Moreover any (f)-good function $x\colon[0,\infty)\to R^n$ is bounded.

We denote $d(x,B)=\inf\{|x-y|\colon\ y\in B\}$ for $x\in R^n$, $B\subset R^n$. Denote by $\operatorname{dist}(A,B)$ the Hausdorff metric for two sets $A\subset R^n$, $B\subset R^n$. For every bounded a.c. function $x\colon[0,\infty)\to R^n$ define

$$\Omega(x)=\{y\in R^n\colon\text{there exists a sequence } (t_i)_{i=0}^{\infty}\subset(0,\infty) \text{ for which}$$
$$t_i\to\infty,\ x(t_i)\to y \text{ as } i\to\infty\}. \tag{1.5}$$

We say that an integrand $f\in\mathfrak{A}$ has Property B if $\Omega(v_2)=\Omega(v_1)$ for each (f)-good functions $v_i\colon[0,\infty)\to R^n$, $i=1,2$.

In Zaslavski [25, Theorem 2.1] we established the following result, which describes the limit behavior of (f)-good functions for a generic $f \in \mathfrak{A}$.

THEOREM 1.1. *There exists a set $\mathcal{F} \subset \mathfrak{A}$ which is a countable intersection of open everywhere dense subsets of $\mathfrak{A}$ and such that each $f \in \mathcal{F}$ has Property* B.

Moreover, it was shown in [25, Theorem 2.2] that for each integrand $f \in \mathfrak{A}$ which has Property B and each initial condition $x_0 \in R^n$ there exists an (f)-weakly optimal function $x\colon [0,\infty) \to R^n$ satisfying $x(0) = x_0$.

Since there exists an example of a function $f \in \mathfrak{A}$ which does not have Property B (see [25, sect. 14]) it is very important for applications to know that Property B, which provides the existence of a weakly optimal solution with any initial value, holds for "almost all" integrands $f \in \mathfrak{A}$.

Consider any $f \in \mathfrak{A}$. In analyzing the infinite-horizon variational problem with the integrand f in [25] we studied the function $U^f(T_1,\ T_2,\ y,\ z)$ $(T_2 > T_1 \geq 0,\ y,\ z \in R^n)$ defined by (1.2). By a simple modification of the proof of Proposition 4.4 in Leizarowitz and Mizel [16] (see [25, Theorems 8.1 and 8.2]) we established the representation formula

$$U^f(0,T,x,y) = T\mu(f) + \pi^f(x) - \pi^f(y) + \bar{\theta}^f_T(x,y), \qquad x,y \in R^n, T \in (0,\infty), \tag{1.6}$$

where $\pi^f\colon\ R^n \to R^1$ is a continuous function and $(T,x,y) \to \bar{\theta}^f_T(x,y) \in R^1$ is a continuous nonnegative function defined for $T > 0$, $x,\ y \in R^n$,

$$\begin{aligned}\pi^f(x) = \inf\{\liminf_{T\to+\infty}[I^f(0,T,v) - \mu(f)T]\colon v\colon\ [0,\infty) \to R^n \text{ is an a.c. function}\\ \text{satisfying } v(0) = x\}, \qquad x \in R^n,\end{aligned} \tag{1.7}$$

and for every $T > 0$, every $x \in R^n$ there is $y \in R^n$ satisfying $\bar{\theta}^f_T(x,y) = 0$.

Assume that there exists a compact set $H(f) \subset R^n$ such that $\Omega(v) = H(f)$ for each (f)-good function $v\colon\ [0,\infty) \to R^n$. (By Theorem 1.1 this assumption holds for a generic $f \in \mathfrak{A}$).

By Theorems 8.3 and 8.4 in [25] for every $x \in R^n$ there exists an (f)-good function $v\colon\ [0,\infty) \to R^n$ such that $v(0) = x$ and the relation

$$I^f(T_1,T_2,v) = (T_2 - T_1)\mu(f) + \pi^f(v(T_1)) - \pi^f(v(T_2)) \tag{1.8}$$

holds for each $T_1 \in [0,\infty)$, $T_2 \in (T_1,\infty)$, and moreover, each a.c. function $v\colon\ [0,\infty) \to R^n$ such that (1.8) holds for each $T_1 \in [0,\infty)$, $T_2 \in (T_1,\infty)$ is an (f)-weakly optimal function.

Denote by $\mathcal{A}(f)$ the set of all a.c. functions $v\colon\ [0,\infty) \to R^n$ which satisfy (1.8) for each $T_1 \in [0,\infty)$, $T_2 \in (T_1,\infty)$.

For a given $x_0 \in R^n$ we can construct a function $x \in \mathcal{A}(f)$, which is (f)-weakly optimal and satisfies initial condition $x(0) = x_0$, using the following scheme.

Let $x_0 \in R^n$. We fix a positive number T and obtain $\mu(f)$ by (1.4), $U^f(0,T,y,z)$ $(y,z \in R^n)$ by (1.2), $\pi^f(y)(y \in R^n)$ by (1.7), and $\bar{\theta}^f_T(y,z)(y,z \in R^n)$ by (1.6). By the properties of the functions $\bar{\theta}^f_T$ we define a sequence $\{x_i\}_{i=0}^{\infty} \subset R^n$ such that $\bar{\theta}^f_T(x_i,x_{i+1}) = 0$, $i = 0,1,\ldots$. Then we construct an a.c. function $x\colon\ [0,\infty) \to R^n$ for which

$$x(iT) = x_i, \quad I^f(iT,(i+1)T,x) = U^f(0,T,x_i,x_{i+1}), \quad i = 0,1,\ldots.$$

It follows from this construction that $x \in \mathcal{A}(f)$ and it is an (f)-weakly optimal function. Since the above construction uses the set $\mathcal{A}(f)$, the function π^f, and the

minimal long-run average cost growth rate $\mu(f)$, and since we are not aware of any other way of constructing weakly optimal solutions, it follows that these concepts are of great importance for us.

In this paper we will establish the following propositions for any integrand f which has Property B:

(a) f is a continuity point of the mapping $g \to (\mu(g), \pi^g) \in R^1 \times C(R^n)$, $g \in \mathfrak{A}$, where $C(R^n)$ is the space of all continuous functions $\phi\colon R^n \to R^1$ with the topology of the uniform convergence on bounded subsets.

(b) Given any $\varepsilon > 0$ there exist numbers ℓ, $\delta > 0$ such that for each $v \in \mathcal{A}(f)$ satisfying $d(v(0), H(f)) \le \delta$ and each $T \ge 0$

$$\operatorname{dist}(H(f), \quad \{v(t)\colon t \in [T, T+\ell]\}) \le \varepsilon. \tag{1.9}$$

(c) For each ε, $K > 0$ there exist numbers ℓ, $Q > 0$ such that for each $v \in \mathcal{A}(f)$ satisfying $|v(0)| \le K$ relation (1.9) holds for each $T \ge Q$.

Assume that we have some algorithm to calculate $\mu(f)$ and π^f. Indeed using this algorithm, instead of f, we deal with a function g, which is an approximation of f. Proposition (a) shows that if g is close enough to f, then the results we obtain are a good approximation of $\mu(f)$, π^f. Propositions (b) and (c) establish the turnpike property, which is well known in mathematical economics (see [7], [17], and survey [19]) for (f)-weakly optimal solutions $v \in \mathcal{A}(f)$. In [25, Theorem 2.4] a weak version of the turnpike property was established. For an optimal solution $v \in \mathcal{A}(f)$ Theorem 2.4 in [25] implies that relation (1.9), with ℓ which depends on ε and $|v(0)|$, holds for all $t \in [0, \infty) \backslash E$, where $E \subset [0, \infty)$ is a measurable subset and the Lebesgue measure of E does not exceed a constant which depends on ε and $|v(0)|$. Clearly Propositions (b) and (c) are an essential improvement of Theorem 2.4 in [25] for optimal solutions $v \in \mathcal{A}(f)$.

Propositions (b) and (c) are also important for applications. Suppose that we know (f)-weakly optimal solutions $x_i \in \mathcal{A}(f), i = 1, \dots, q$, with initial conditions $x_i(0) \in \{z \in R^n\colon |z| \le K\}$, where K is a positive constant. Therefore we know the turnpike $H(f)$, or at least its approximation, and the constant Q (see Proposition (c)) which is an estimate for the time period required to reach the turnpike. This information can be useful if we need to find an (f)-weakly optimal solution with a new initial value from the set $\{z \in R^n\colon |z| \le K\}$.

Propositions (a), (b), and (c) are extensions of the main results in Zaslavski [22] which were established for discrete-time control systems. In the approach used in [21, 22] the following property played a crucial role.

Property C. In the space of integrands (or cost functions) there exists an everywhere dense subset E such that for each $f \in E$ there exists an (f)-good periodic trajectory.

This approach was also used in Zaslavski [24] for a class of one-dimensional variational problems arising in continuum mechanics which was discussed in Leizarowitz and Mizel [16] and Coleman, Marcus, and Mizel [8]. It is not clear whether Property C holds in general. In the present paper we develop a more general approach based on the following idea which arose in our discussions with Moshe Marcus.

The validity of Property B for an integrand $f \in \mathfrak{A}$ (namely, if all (f)-good trajectories have the same limit point set) implies that f has the additional interesting properties: for any initial condition there exists an (f)-weakly optimal solution; the turnpike property holds for optimal functions $v \in \mathcal{A}(f)$; f is a continuity point of the mapping $g \to (\mu(g), \pi^g)$, $g \in \mathfrak{A}$.

We hope that this approach is universal and can be applied to various classes of variational and optimal control problems, in particular when Property C does not hold.

2. Main results. For each function $f \in \mathfrak{A}$ denote by $\mathcal{A}(f)$ the set of all a.c. functions $v\colon [0,\infty) \to R^n$ which satisfy (1.8) for each $T_1 \in [0,\infty)$, $T_2 \in (T_1,\infty)$. It follows from Theorem 8.3 in [25] that for every $f \in \mathfrak{A}$ and every $x \in R^n$ there exists an (f)-good function $v \in \mathcal{A}(f)$ satisfying $v(0) = x$.

If $f \in \mathfrak{A}$ and for each (f)-good function $v_i\colon [0,\infty) \to R^n$, $i = 1, 2$,

$$\Omega(v_2) = \Omega(v_1),$$

then by Theorem 8.4 in [25] any function $v \in \mathcal{A}(f)$ is an (f)-weakly optimal function.

We will establish the following results.

THEOREM 2.1. *Assume that $f \in \mathfrak{A}$ and there exists a compact set $H(f) \subset R^n$ such that $\Omega(v) = H(f)$ for each (f)-good function $v\colon [0,\infty) \to R^n$. Then f is a continuity point of the mapping $g \to (\mu(g), \pi^g) \in R^1 \times C(R^n)$, $g \in \mathfrak{A}$, where $C(R^n)$ is the space of all continuous functions $\phi\colon R^n \to R^1$ with the topology of the uniform convergence on bounded subsets.*

Theorems 2.2 and 2.3 establish the turnpike property for a generic $f \in \mathfrak{A}$.

THEOREM 2.2. *Assume that $f \in \mathfrak{A}$ and there exists a compact set $H(f) \subset R^n$ such that $\Omega(v) = H(f)$ for each (f)-good function $v\colon [0,\infty) \to R^n$. Let ε be a positive number. Then there exist numbers $L, \delta > 0$ and a neighborhood $\mathcal{U}$ of f in $\mathfrak{A}$ such that for each $g \in \mathcal{U}$, each $v \in \mathcal{A}(g)$ satisfying $d(v(0), H(f)) \le \delta$, and each $T \in [0,\infty)$*

$$\operatorname{dist}(H(f),\ \{v(t)\colon t \in [T, T+L]\}) \le \varepsilon. \tag{2.1}$$

THEOREM 2.3. *Assume that $f \in \mathfrak{A}$ and there exists a compact set $H(f) \subset R^n$ such that $\Omega(v) = H(f)$ for each (f)-good function $v\colon [0,\infty) \to R^n$. Let ε and K be positive numbers. Then there exist numbers $L, Q > 0$ and a neighborhood $\mathcal{U}$ of f in $\mathfrak{A}$ such that for each $g \in \mathcal{U}$, each $v \in \mathcal{A}(g)$ satisfying $|v(0)| \le K$ relation (2.1) holds for all $T \in [Q,\infty)$.*

For each $f \in \mathfrak{A}$ denote by $\mathcal{B}(f)$ the set of all a.c. functions $v\colon R^1 \to R^n$ such that (1.8) holds for each $T_1 \in R^1$, $T_2 \in (T_1,\infty)$, and $\liminf_{t\to-\infty} |v(t)| < \infty$.

THEOREM 2.4. *Assume that $f \in \mathfrak{A}$ and there exists a compact set $H(f) \subset R^n$ such that $\Omega(v) = H(f)$ for each (f)-good function $v\colon [0,\infty) \to R^n$. Then the following properties hold:*

(1) *for each $h \in H(f)$ there exists $v \in \mathcal{B}(f)$ satisfying $v(0) = h$;*

(2) *for each $v \in \mathcal{B}(f)$ the relation $v(t) \in H(f)$ holds for all $t \in R^1$;*

(3) *for each $\varepsilon > 0$ there exists $L > 0$ such that (2.1) holds for each $v \in \mathcal{B}(f)$ and each $T \in R^1$.*

3. Auxiliary results. In [25] we established the following results.

PROPOSITION 3.1 ([25, Proposition 3.1]). *For each $f \in \mathfrak{A}$ there exists a neighborhood $\mathcal{U}$ of f in $\mathfrak{A}$ and a number $M > 0$ such that, for each $g \in \mathcal{U}$ and each (g)-good function $x\colon [0,\infty) \to R^n$,*

$$\limsup_{t\to\infty} |x(t)| < M.$$

PROPOSITION 3.2 ([25, Proposition 3.2]). *Let $f \in \mathfrak{A}$ and $M_1,\ M_2,\ c > 0$. Then there exist a neighborhood $\mathcal{U}$ of f in $\mathfrak{A}$ and $S > 0$ such that for each $g \in \mathcal{U}$, each $T_1 \in [0,\infty)$, and each $T_2 \in [T_1 + c, \infty)$ the following property holds: for each*

x, $y \in R^n$ satisfying $|x|, |y| \le M_1$ and each a.c. function $v\colon [T_1, T_2] \to R^n$ satisfying $v(T_1) = x, v(T_2) = y, I^g(T_1, T_2, v) \le U^g(T_1, T_2, x, y) + M_2$, the relation $|v(t)| \le S$ $(t \in [T_1, T_2])$ holds.

PROPOSITION 3.3 ([25, Proposition 3.3]). *Let $f \in \mathfrak{A}$, $0 < c_1 < c_2 < \infty, M, \varepsilon > 0$. Then there exists $\delta > 0$ such that for each $T_1, T_2 \ge 0$ satisfying $T_2 - T_1 \in [c_1, c_2]$ and each $y_1, y_2, z_1, z_2 \in R^n$ satisfying $|y_i|, |z_i| \le M (i = 1, 2)$, $\sup\{|y_1 - y_2|, |z_1 - z_2|\} \le \delta$ the relation $|U^f(T_1, T_2, y_1, z_1) - U^f(T_1, T_2, y_2, z_2)| \le \varepsilon$ holds.*

PROPOSITION 3.4 ([25, Proposition 3.4]). *Assume that $f \in \mathfrak{A}, M_1 > 0$, $0 \le T_1 < T_2, x_i\colon [T_1, T_2] \to R^n, i = 1, 2, \ldots,$ is a sequence of a.c. functions such that $I^f(T_1, T_2, x_i) \le M_1, i = 1, 2, \ldots$. Then there exist a subsequence $\{x_{i_k}\}_{k=1}^{\infty}$ and an a.c. function $x\colon [T_1, T_2] \to R^n$ such that $I^f(T_1, T_2, x) \le M_1, x_{i_k}(t) \to x(t)$ as $k \to \infty$ uniformly in $[T_1, T_2]$ and $x'_{i_k} \to x'$ as $k \to \infty$ weakly in $L^1(R^n; (T_1, T_2))$.*

PROPOSITION 3.5 ([25, Proposition 3.5]). *For each $f \in \mathfrak{A}$, each number T_1, T_2 satisfying $0 \le T_1 < T_2$, and each z_1, $z_2 \in R^n$ there is an a.c. function $x\colon [T_1, T_2] \to R^n$ such that $x(T_i) = z_i$, $i = 1, 2$, $I^f(T_1, T_2, x) = U^f(T_1, T_2, z_1, z_2)$.*

PROPOSITION 3.6 ([25, Proposition 3.7]). *Let $f \in \mathfrak{A}, 0 < c_1 < c_2 < \infty, c_3 > 0$. Then there exists a neighborhood $\mathcal{U}$ of f in $\mathfrak{A}$ such that the set $\{U^g(T_1, T_2, z_1, z_2)\colon g \in \mathcal{U}, T_1 \in [0, \infty), T_2 \in [T_1 + c_1, T_1 + c_2], z_1, z_2 \in R^n, |z_1| \le c_3 (i = 1, 2)\}$ is bounded.*

PROPOSITION 3.7 ([25, Proposition 3.8]). *Let $f \in \mathfrak{A}, 0 < c_1 < c_2 < \infty, D, \varepsilon > 0$. Then there is a neighborhood V of f in $\mathfrak{A}$ such that for each $g \in V$, each T_1, $T_2 \ge 0$ satisfying $T_2 - T_1 \in [c_1, c_2]$, and each a.c. function $x\colon [T_1, T_2] \to R^n$ satisfying $\inf\{I^f(T_1, T_2, x), I^g(T_1, T_2, x)\} \le D$ the relation $|I^f(T_1, T_2, x) - I^g(T_1, T_2, x)| \le \varepsilon$ holds.*

PROPOSITION 3.8 ([25, Proposition 3.9]). *Let $f \in \mathfrak{A}, 0 < c_1 < c_2 < \infty, c_3, \varepsilon > 0$. Then there exists a neighborhood V of f in $\mathfrak{A}$ such that for each $g \in V$, each T_1, $T_2 \ge 0$ satisfying $T_2 - T_1 \in [c_1, c_2]$, and each z, $y \in R^n$ satisfying $|y|, |z| \le c_3$ the relation $|U^f(T_1, T_2, y, z) - U^g(T_1, T_2, y, z)| \le \varepsilon$ holds.*

PROPOSITION 3.9 ([25, Theorem 5.1]). *Assume that $f \in \mathfrak{A}$ and there exists a compact set $H(f) \subset R^n$ such that $\Omega(v) = H(f)$ for each (f)-good function $v\colon [0, \infty) \to R^n$. Let ε be a positive number. Then there exists an integer $L \ge 1$ such that for each (f)-good function $v\colon [0, \infty) \to R^n$*

$$\operatorname{dist}(H(f), \quad \{v(t)\colon\ t \in [T, T + L]\}) \le \varepsilon \quad \text{for all large } T.$$

PROPOSITION 3.10 ([25, Theorem 6.1]). *Assume that $f \in \mathfrak{A}$. Then the mapping $(T_1, T_2, x, y) \to U^f(T_1, T_2, x, y)$ is continuous for $T_1 \in [0, \infty), T_2 \in (T_1, \infty), x, y \in R^n$.*

PROPOSITION 3.11 ([25, Theorem 8.3]). *For each $f \in \mathfrak{A}$ and each $x \in R^n$ there exists an (f)-good function $v \in \mathcal{A}(f)$ satisfying $v(0) = x$.*

PROPOSITION 3.12 ([25, Theorem 8.4]). *Assume that $f \in \mathfrak{A}$ and there exists a compact set $H(f) \in R^n$ such that $\Omega(x) = H(f)$ for each (f)-good function $x\colon [0, \infty) \to R^n$. Then each $v \in \mathcal{A}(f)$ is an (f)-weakly optimal function.*

PROPOSITION 3.13 ([25, Theorem 2.3]). *Assume that $f \in \mathfrak{A}$ and there exists a compact set $H(f) \subset R^n$ such that $\Omega(v) = H(f)$ for each (f)-good function $v\colon [0, \infty) \to R^n$. Let ε be a positive number. Then there exist an integer $L \ge 1$ and a neighborhood $\mathcal{U}$ of f in $\mathfrak{A}$ such that for each $g \in \mathcal{U}$ and each (g)-good function $v\colon [0, \infty) \to R^n$*

$$\operatorname{dist}(H(f), \quad \{v(t)\colon\ t \in [T, T + L]\}) \le \varepsilon \quad \text{for all large } T.$$

Theorem 8.1, equation (8.2), and Proposition 7.3 in [25] imply the following result.

PROPOSITION 3.14. *Let $f \in \mathfrak{A}$. Then $\pi^f(x) \to +\infty$ as $|x| \to \infty$.*

4. Proof of Theorems 2.1–2.3. Assume that $f \in \mathfrak{A}$ and $H(f) \subset R^n$ is a compact set such that $\Omega(v) = H(f)$ for each (f)-good function $v\colon [0,\infty) \to R^n$.

In [25] we established the following results.

LEMMA 4.1 ([25, Lemma 10.2]). *Let $\varepsilon_0 \in (0,1)$, $K_0, M_0 > 0$, and let ℓ be a positive integer such that, for each (f)-good function $x\colon [0,\infty) \to R^n$,*

$$\operatorname{dist}(H(f), \{x(t)\colon t \in [T, T+\ell]\}) \le 8^{-1}\varepsilon_0 \tag{4.1}$$

for all large T (the existence of ℓ follows from Proposition 3.9). Then there exist an integer $N \ge 10$ and a neighborhood $\mathcal{U}$ of f in $\mathfrak{A}$ such that for each $g \in \mathcal{U}$, each $S \in [0,\infty)$, and each a.c. function $x\colon [S, S+N\ell] \to R^n$ satisfying

$$|x(S)|, \quad |x(S+N\ell)| \le K_0, \quad I^g(S, S+N\ell, x) \le U^g(S, S+N\ell, x(S), x(S+N\ell)) + M_0$$

there exists an integer $i_0 \in [0, N-8]$ such that for all $T \in [S+i_0\ell, S+(i_0+7)\ell]$

$$\operatorname{dist}(H(f), \{x(t)\colon t \in [T, T+\ell]\}) \le \varepsilon_0.$$

LEMMA 4.2 ([25, Lemma 10.3]). *Let $\varepsilon > 0$. Then there exists $\delta > 0$ such that, for each $x_1, x_2 \in R^n$ which satisfy $d(x_i, H(f)) \le \delta$, $i = 1, 2$, there exists an a.c. function $v\colon [0,T] \to R^n$ for which*

$$T \ge 1, \quad v(0) = x_1, \quad v(T) = x_2, \quad I^f(0,T,v) - \pi^f(x_1) + \pi^f(x_2) - T\mu(f) \le \varepsilon.$$

LEMMA 4.3 ([25, Lemma 10.4]). *Let $\varepsilon \in (0,1)$ and let L be a positive integer such that for each (f)-good function $v\colon [0,\infty) \to R^n$*

$$\operatorname{dist}(H(f), \{v(t)\colon t \in [S, S+L]\}) \le \varepsilon \tag{4.2}$$

for all large S (the existence of L follows from Proposition 3.9).

Then there exists $\delta > 0$ such that for each $T \in [L, \infty)$ and each a.c. function $v\colon [0,T] \to R^n$ which satisfies

$$d(v(0), H(f)) \le \delta, \quad d(v(T), H(f)) \le \delta, \quad I^f(0,T,v) - T\mu(f) - \pi^f(v(0)) + \pi^f(v(T)) \le \delta$$

relation (4.2) holds for every $S \in [0, T-L]$.

Lemma 4.1 and Proposition 3.2 imply the following.

LEMMA 4.4. *Let $\varepsilon_0 \in (0,1)$, $K_0, M_0 > 0$, and let ℓ be a positive integer such that each (f)-good function $x\colon [0,\infty) \to R^n$ satisfies (4.1) for all large T. Then there exist an integer $N \ge 10$, a neighborhood $\mathcal{U}$ of f in $\mathfrak{A}$, and a number $M_1 > 0$ such that for each $g \in \mathcal{U}$, each $T_1 \ge 0, T_2 \ge T_1 + N\ell$, and each a.c. function $x\colon [T_1, T_2] \to R^n$ which satisfies*

$$|x(T_i)| \le K_0, \quad i = 1, 2, \quad I^g(T_1, T_2, x) \le U^g(T_1, T_2, x(T_1), x(T_2)) + M_0$$

the following properties hold: $|x(t)| \le M_1$ for all $t \in [T_1, T_2]$; for each $S \in [T_1, T_2 - N\ell]$ there exists an integer $i_0 \in [0, N-8]$ such that

$$\operatorname{dist}(H(f), \{x(t)\colon t \in [T, T+\ell]\}) \le \varepsilon_0 \quad \textit{for all } T \in [S+i_0\ell, S+(i_0+7)\ell].$$

Set

$$D_f = \sup\{|h|\colon\ h \in H(f)\}. \tag{4.3}$$

For each $g \in \mathfrak{A}$ denote by $\mathcal{A}(g)$ the set of all a.c. functions $v\colon [0,\infty) \to R^n$ such that for each $T_1 \in [0,\infty), T_2 \in (T_1,\infty)$

$$I^g(T_1,T_2,v) = (T_2-T_1)\mu(g) + \pi^g(v(T_1)) - \pi^g(v(T_2)). \tag{4.4}$$

For $K,\tau > 0$ and $g \in \mathfrak{A}$ we define

$$\ell(g,K,\tau) = \inf\{U^g(0,\tau,x,y) - \pi^f(x) + \pi^f(y)\colon x,y \in R^n, |x|,|y| \le K\}. \tag{4.5}$$

It follows from the representation formula (see (1.6), (1.7)), Proposition 3.11, that

$$\ell(f,K,\tau) = \mu(f)\tau, \qquad \tau > 0, \quad K > D_f. \tag{4.6}$$

Equations (4.5), (4.6), and Proposition 3.8 imply the following result.

LEMMA 4.5. *Let $K > D_f, 0 < \tau_1, \tau_2, \delta > 0$. Then there exists a neighborhood $\mathcal{U}$ of f in $\mathfrak{A}$ such that $|\ell(g,K,\tau) - \mu(f)\tau| \le \delta$ for each $g \in \mathcal{U}$ and each $\tau \in [\tau_1,\tau_2]$.*

LEMMA 4.6. *Let $h \in H(f)$. Then there exists an (f)-good function $v\colon [0,\infty) \to H(f)$ such that $v \in \mathcal{A}(f)$ and $v(0) = h$.*

Proof. By Proposition 3.14 there exists an (f)-good function $u \in \mathcal{A}(f)$. We may assume that

$$d(u(t),H(f)) \le 1 \ \text{ for all } \ t \in [0,\infty). \tag{4.7}$$

There exists a sequence of positive numbers $\{T_p\}_{p=0}^{\infty}$ such that

$$T_{p+1} \ge T_p + 1, \quad p = 0,1,\ldots, \quad u(T_p) \to h \quad \text{as } p \to \infty. \tag{4.8}$$

For every integer $p \ge 1$ we set

$$v_p(t) = u(t+T_p), \qquad t \in [0,\infty). \tag{4.9}$$

By Proposition 3.4, (4.9), and (4.7) there exist a subsequence $\{v_{p_j}\}_{j=1}^{\infty}$ and an a.c. function $v\colon [0,\infty) \to R^n$ such that for every integer $N \ge 1$

$$v_{p_j}(t) \to v(t) \text{ as } j \to \infty \text{ uniformly in } [0,N], \tag{4.10}$$

$$I^f(0,N,v) \le \liminf_{j\to\infty} I^f(0,N,v_{p_j}). \tag{4.11}$$

Since $\Omega(u) = H(f)$, it follows from (4.9) and (4.10) that $v(t) \in H(f)$ for all $t \in [0,\infty)$. By (4.8)–(4.10) $v(0) = h$. Since $u \in \mathcal{A}(f)$, it follows from (4.9)–(4.11) that for each integer $N \ge 1$

$$\begin{aligned} I^f(0,N,v) &\le \liminf_{j\to\infty} I^f(T_{p_j}, T_{p_j}+N, u) \\ &\le \liminf_{j\to\infty} [N\mu(f) + \pi^f(u(T_{p_j})) - \pi^f(u(T_{p_j}+N))] \\ &= N\mu(f) + \pi^f(v(0)) - \pi^f(v(N)). \end{aligned}$$

Together with the representation formula (see (1.6), (1.7)) this implies that $v \in \mathcal{A}(f)$. The lemma is proven.

LEMMA 4.7. *Let $\varepsilon \in (0,1)$ and $K > D_f + 1$. Then there exists a neighborhood $\mathcal{U}$ of f in $\mathfrak{A}$ such that for each $g \in \mathcal{U}$ and each number $T \ge 1$ there exists an a.c. function $x\colon [0,T] \to R^n$ such that $x(0), x(T) \in H(f)$,*

$$I^g(0,T,x) - \pi^f(x(0)) + \pi^f(x(T)) \le \ell(g,K,T) + \varepsilon. \tag{4.12}$$

Proof. By Lemma 4.6 there exists an (f)-good function

$$v_0 : [0,\infty) \to H(f) \text{ such that } v_0 \in \mathcal{A}(f). \tag{4.13}$$

By Proposition 3.2 there exists a neighborhood $\mathcal{U}_1$ of f in $\mathfrak{A}$ and $M_1 > 0$ such that for each $g \in \mathcal{U}_1$, each $T_1 \in [0,\infty), T_2 \in [T_1 + 8^{-1}, \infty)$, and each a.c. function x: $[T_1, T_2] \to R^n$ which satisfies

$$|x(T_i)| \le K, \quad i = 1,2, \quad I^g(T_1, T_2, x) \le U^g(T_1, T_2, x(T_1), x(T_2)) + 1 \tag{4.14}$$

the following relation holds:

$$|x(t)| \le M_1, \qquad t \in [T_1, T_2]. \tag{4.15}$$

By Proposition 3.10 there exists a number

$$\delta \in (0, 8^{-1}\varepsilon) \tag{4.16}$$

such that for each $y_1, y_2, z_1, z_2 \in R^n$ which satisfy

$$|y_i|, |z_i| \le M_1 + 2K + 1, \quad i = 1,2, \quad |y_i - z_i| \le 4\delta, \quad i = 1,2, \tag{4.17}$$

the following relations hold:

$$|U^f(0,1,y_1,y_2) - U^f(0,1,z_1,z_2)| \le 2^{-6}\varepsilon, \quad |\pi^f(y_i) - \pi^f(z_i)| \le 2^{-6}\varepsilon, \quad i = 1,2. \tag{4.18}$$

By Proposition 3.9 there exists an integer $L \ge 1$ such that for each (f)-good function v: $[0,\infty) \to R^n$

$$\operatorname{dist}(H(f), \{v(t):\ t \in [T, T+L]\}) \le 8^{-1}\delta \tag{4.19}$$

for all large T.

By Lemma 4.4 and the definition of L there exist an integer $N_1 \ge 10$ and a neighborhood $\mathcal{U}_2$ of f in $\mathfrak{A}$ such that for each $g \in \mathcal{U}_2$, each $T_1 \ge 0$, $T_2 \ge T_1 + N_1 L$, each a.c. function x: $[T_1, T_2] \to R^n$ which satisfies (4.14), and each $S \in [T_1, T_2 - N_1 L]$ there exists an integer $i_0 \in [0, N_1 - 8]$ such that for all $T \in [S + i_0 L, S + (i_0 + 7)L]$

$$\operatorname{dist}(H(f), \{x(t):\ t \in [T, T+L]\}) \le \delta. \tag{4.20}$$

Since $\Omega(v_0) = H(f)$ it follows from (4.13) that there exist integers

$$N_2 \ge 4N_1 L + 4, \qquad N_3 \ge 8 \tag{4.21}$$

such that

$$\begin{aligned} &\operatorname{dist}(H(f), (v_0(t):\ t \in [0, N_2 L]\}) \le 8^{-1}\delta, \\ &\operatorname{dist}(H(f), \{v_0(t):\ t \in [8(N_2+1)L, 8(N_2+1)L + N_3 L]\}) \le 8^{-1}\delta. \end{aligned} \tag{4.22}$$

Fix an integer

$$N_0 \ge 8L(N_1 + N_2 + N_3 + 4). \tag{4.23}$$

By Proposition 3.8 there exists a neighborhood $\mathcal{U}_3$ of f in $\mathfrak{A}$ such that for each $g \in \mathcal{U}_3$ and each z, $y \in R^n$ which satisfy $|y|$, $|z| \le 4M_1 + 4K + 4$ the following relation holds:

$$|U^f(0,1,y,z) - U^g(0,1,y,z)| \le 2^{-6}\varepsilon. \tag{4.24}$$

By Proposition 3.7 there exists a neighborhood $\mathcal{U}_4$ of f in $\mathfrak{A}$ such that for each $g \in \mathcal{U}_4$, each $T_1 \geq 0$, $T_2 \in [T_1+1, T_1+N_0+1]$, and each a.c. function $x\colon [T_1, T_2] \to R^n$ which satisfies

$$\inf\{I^f(T_1,T_2,x), I^g(T_1,T_2,x)\} \leq N_0(|\mu(f)|+1)+2\sup\{|\pi^f(h)|\colon\ h \in H(f)\} \tag{4.25}$$

the following relation holds:

$$|I^f(T_1,T_2,x) - I^g(T_1,T_2,x)| \leq 2^{-6}\varepsilon. \tag{4.26}$$

By Lemma 4.5 there exists a neighborhood $\mathcal{U}_5$ of f in $\mathfrak{A}$ such that for each $g \in \mathcal{U}_5$ and each $\tau \in [1, 8(N_0+1)]$

$$|\ell(g,K,\tau) - \mu(f)\tau| \leq 2^{-6}\varepsilon. \tag{4.27}$$

Set

$$\mathcal{U} = \bigcap_{i=1}^{5} \mathcal{U}_i. \tag{4.28}$$

Let $g \in \mathcal{U}$ and a number $T \geq 1$. There are two cases: (i) $T \leq N_0$; (ii) $T > N_0$.

Consider the case (i). Set

$$x(t) = v_0(t), \qquad t \in [0,T]. \tag{4.29}$$

It follows from (4.13) and the definition of $\mathcal{A}(f)$ (see (4.4)) that

$$I^f(0,T,x) \leq N_0|\mu(f)| + 2\sup\{|\pi^f(h)|\colon\ h \in H(f)\}.$$

By this relation and the definition of $\mathcal{U}_4$ (see (4.26))

$$|I^f(0,T,x) - I^g(0,T,x)| \leq 2^{-6}\varepsilon. \tag{4.30}$$

By the definition of $\mathcal{U}_5$ (see (4.27))

$$|\ell(g,K,T) - \mu(f)T| \leq 2^{-6}\varepsilon. \tag{4.31}$$

Combining (4.30), (4.29), and (4.13) we obtain that

$$\begin{aligned} I^g(0,T,x) - \pi^f(x(0)) + \pi^f(x(T)) &\leq I^f(0,T,v_0) - \pi^f(v_0(0)) + \pi^f(v_0(T)) + 2^{-6}\varepsilon \\ &\leq \mu(f)T + 2^{-6}\varepsilon \leq \ell(g,K,T) + 2^{-5}\varepsilon. \end{aligned}$$

Therefore in the case (i) the assertion of the lemma is valid.

Consider the case (ii). Then

$$T > N_0. \tag{4.32}$$

There exists an a.c. function $y\colon [0,T] \to R^n$ such that

$$|y(0)|, |y(T)| \leq K, \qquad I^g(0,T,y) - \pi^f(y(0)) + \pi^f(y(T)) \leq \ell(g,K,T) + 16^{-1}\delta. \tag{4.33}$$

Clearly

$$I^g(0,T,y) \leq U^g(0,T,y(0),y(T)) + 16^{-1}\delta.$$

By this relation, (4.33), (4.32), (4.23), and the definition of N_1, $\mathcal{U}_2$ (see (4.20)) there exist integers i_1, $i_2 \in [0, N_1 - 8]$ such that

$$\operatorname{dist}(H(f), \{y(t)\colon t \in [S, S+L]\}) \le \delta \tag{4.34}$$

for each

$$S \in [i_1 L, (i_1+7)L] \cup [T - 2N_1 L + i_2 L, \quad T - 2N_1 L + (i_2+7)L]. \tag{4.35}$$

It follows from (4.34), (4.35), and (4.22) that there exist

$$t_1 \in [8(N_2+1)L, 8(N_2+1)L + N_3 L], \qquad t_2 \in [0, N_2 L] \tag{4.36}$$

for which

$$|y(i_1 L+1) - v_0(t_1)| \le \delta + 4^{-1}\delta, \quad |y(T - 2N_1 L + i_2 L + 1) - v_0(t_2)| \le \delta + 4^{-1}\delta. \tag{4.37}$$

By Proposition 3.5, (4.13), (4.36), (4.21), (4.32), and (4.23) there exists an a.c. function $x\colon [0,T] \to R^n$ such that

$$x(t) = v_0(t + t_1 - i_1 L - 1), \quad t \in [0, i_1 L + 1), \quad x(t) = y(t), \quad t \in [i_1 L + 2, T - 2N_1 L + i_2 L],$$

$$x(t) = v_0(t + t_2 - (T - 2N_1 L + i_2 L + 1)), \qquad t \in [T - 2N_1 L + i_2 L + 1, T],$$

$$I^g(\tau, \tau+1, x) = U^g(0, 1, x(\tau), x(\tau+1)), \quad \tau = i_1 L + 1, \quad T - 2N_1 L + i_2 L. \tag{4.38}$$

For each a.c. function $u\colon [0,T] \to R^n$ and each $r_1, r_2 \in [0,T]$ satisfying $r_1 < r_2$ we set

$$\sigma(r_1, r_2, u) = I^g(r_1, r_2, u) - \pi^f(u(r_1)) + \pi^f(u(r_2)). \tag{4.39}$$

Set

$$\tau_0 = 0, \quad \tau_1 = i_1 L + 1, \quad \tau_2 = i_1 L + 2, \quad \tau_3 = T - 2N_1 L + i_2 L,$$
$$\tau_4 = T - 2N_1 L + i_2 L + 1, \qquad \tau_5 = T. \tag{4.40}$$

It follows from (4.38)–(4.40) that

$$\sigma(0, T, x) - \sigma(0, T, y) = \sum_{i=0}^{1} [\sigma(\tau_i, \tau_{i+1}, x) - \sigma(\tau_i, \tau_{i+1}, y)]$$
$$+ \sum_{i=3}^{4} [\sigma(\tau_i, \tau_{i+1}, x) - \sigma(\tau_i, \tau_{i+1}, y)]. \tag{4.41}$$

Analogously to the case (i) we can show that

$$\sigma(t_1 - i_1 L - 1, t_1, v_0) \le \ell(g, K, i_1 L + 1) + 2^{-5}\varepsilon,$$
$$\sigma(t_2, t_2 + 2N_1 L + i_2 L + 1, v_0) \le \ell(g, K, 2N_1 L + i_2 L + 1) + 2^{-5}\varepsilon.$$

Together with (4.38)–(4.40) this implies that

$$\sigma(\tau_i, \tau_{i+1}, x) \le \ell(g, K, \tau_{i+1} - \tau_i) + 2^{-5}\varepsilon, \qquad i = 0, 4. \tag{4.42}$$

By (4.40), (4.33), (4.37), (4.16), and (4.13)

$$\sigma(\tau_i, \tau_{i+1}, y) \geq \ell(g, K, \tau_{i+1} - \tau_i), \qquad i = 0, 4. \tag{4.43}$$

It follows from (4.38)–(4.40) that for $i = 1, 3$

$$\begin{aligned}\sigma(\tau_i, \tau_{i+1}, x) - \sigma(\tau_i, \tau_{i+1}, y) &\leq U^g(0, 1, x(\tau_i), x(\tau_{i+1})) - \pi^f(x(\tau_i)) + \pi^f(x(\tau_{i+1}))\\ &\quad -[U^g(0, 1, y(\tau_i), y(\tau_{i+1})) - \pi^f(y(\tau_i)) + \pi^f(y(\tau_{i+1}))].\end{aligned} \tag{4.44}$$

It follows from (4.33) and the definition of $\mathcal{U}_1, M_1$ (see (4.14), (4.15)) that

$$|y(t)| \leq M_1, \qquad t \in [0, T]. \tag{4.45}$$

By (4.40), (4.38), (4.37), (4.13), and the definition of δ (see (4.16)–(4.18)) for $i = 1, 3$

$$|\pi^f(x(\tau_{i+1})) - \pi^f(x(\tau_i)) - (\pi^f(y(\tau_{i+1})) - \pi^f(y(\tau_i)))| \leq 2^{-6}\varepsilon. \tag{4.46}$$

It follows from (4.40), (4.38), (4.45), (4.13), the definition of $\mathcal{U}_3$ (see (4.24)), the definition of δ (see (4.16)–(4.18)), and (4.37) that for $i = 1, 3$

$$\begin{aligned}&|U^g(0, 1, x(\tau_i), x(\tau_{i+1})) - U^g(0, 1, y(\tau_i), y(\tau_{i+1}))|\\ &\quad \leq\ 2^{-5}\varepsilon + |U^f(0, 1, x(\tau_i), x(\tau_{i+1})) - U^f(0, 1, y(\tau_i), y(\tau_{i+1}))| \leq 2^{-4}\varepsilon.\end{aligned}$$

Together with (4.44) and (4.46) this implies that for $i = 1, 3$

$$\sigma(\tau_i, \tau_{i+1}, x) - \sigma(\tau_i, \tau_{i+1}, y) \leq 2^{-3}\varepsilon.$$

By this relation and (4.41)–(4.43), $\sigma(0, T, x) - \sigma(0, T, y) \leq 2^{-1}\varepsilon$. Together with (4.39), (4.33), and (4.16) this implies (4.12). This completes the proof of the lemma.

LEMMA 4.8. *Let $\varepsilon \in (0, 1)$ and $K > D_f + 1$. Then there exists a neighborhood $\mathcal{U}$ of f in $\mathfrak{A}$ such that for each $g \in \mathcal{U}$, each $h \in H(f)$, and each number $T \geq 1$ there exists an a.c. function $x\colon [0, T] \to R^n$ for which*

$$x(0) = h, \qquad x(T) \in H(f), \tag{4.47}$$

$$I^g(0, T, x) - \pi^f(x(0)) + \pi^f(x(T)) \leq \ell(g, K, T) + \varepsilon. \tag{4.48}$$

Proof. By Proposition 3.2 there exist a neighborhood $\mathcal{U}_1$ of f in $\mathfrak{A}$ and $M_1 > 2K + 1$ such that for each $g \in \mathcal{U}_1$, each $T_1 \in [0, \infty), T_2 \in [T_1 + 8^{-1}, \infty)$, and each a.c. function $x\colon [T_1, T_2] \to R^n$ which satisfies

$$|x(T_i)| \leq K, \quad i = 1, 2, \quad I^g(T_1, T_2, x) \leq U^g(T_1, T_2, x(T_1), x(T_2)) + 1 \tag{4.49}$$

the following relation holds:

$$|x(t)| \leq M_1, \qquad t \in [T_1, T_2]. \tag{4.50}$$

By Lemma 4.7 there exists a neighborhood $\mathcal{U}_2$ of f in $\mathfrak{A}$ such that for each $g \in \mathcal{U}_2$ and each number $T \geq 1$ there exists an a.c. function $x\colon [0, T] \to R^n$ such that

$$\begin{aligned}&x(0), \quad x(T) \in H(f),\\ &I^g(0, T, x) - \pi^f(x(0)) + \pi^f(x(T)) \leq \ell(g, M_1, T) + 2^{-6}\varepsilon.\end{aligned} \tag{4.51}$$

By Proposition 3.10 there exists a number

$$\delta \in (0, 8^{-1}\varepsilon) \tag{4.52}$$

such that for each $y_1, y_2, z_1, z_2 \in R^n$ which satisfy

$$|y_i|, |z_i| \le 2M_1 + 4, \quad i = 1, 2, \quad |y_i - z_i| \le 4\delta, \quad i = 1, 2 \tag{4.53}$$

the following relations hold:

$$|U^f(0, 1, y_1, y_2) - U^f(0, 1, z_1, z_2)| \le 2^{-6}\varepsilon, \quad |\pi^f(y_i) - \pi^f(z_i)| \le 2^{-6}\varepsilon, \quad i = 1, 2. \tag{4.54}$$

By Proposition 3.8 there exists a neighborhood $\mathcal{U}_3$ of f in $\mathfrak{A}$ such that for each $g \in \mathcal{U}_3$ and each z, $y \in R^n$ which satisfy $|y|$, $|z| \le 2M_1 + 4$ the following relation holds:

$$|U^f(0, 1, y, z) - U^g(0, 1, y, z)| \le 2^{-6}\varepsilon. \tag{4.55}$$

By Lemma 4.6 there exists an (f)-good function

$$v_0\colon [0, \infty) \to H(f) \quad \text{such that } v_0 \in \mathcal{A}(f). \tag{4.56}$$

Since $\Omega(v_0) = H(f)$, it follows from (4.56) that there exist integers N_1, $N_2 \ge 8$ for which

$$\begin{aligned} &\operatorname{dist}(H(f), \{v_0(t)\colon\ t \in [4, N_1 + 4]\}) \le 8^{-1}\delta, \\ &\operatorname{dist}(H(f), \{v_0(t)\colon\ t \in [2N_1 + 16, 2N_1 + 16 + N_2]\}) \le 8^{-1}\delta. \end{aligned} \tag{4.57}$$

Fix an integer

$$N_0 \ge 8(N_1 + N_2 + 20). \tag{4.58}$$

By Proposition 3.7 there exists a neighborhood $\mathcal{U}_4$ of f in $\mathfrak{A}$ such that for each $g \in \mathcal{U}_4$, each $T_1 \ge 0$, $T_2 \in [T_1 + 1, T_1 + N_0]$, and each a.c. function $x\colon [T_1, T_2] \to R^n$ which satisfies

$$\inf\{I^f(T_1, T_2, x), I^g(T_1, T_2, x)\} \le N_0|\mu(f)| + 2\sup\{|\pi^f(z)|\colon\ z \in H(f)\} \tag{4.59}$$

the following relation holds:

$$|I^f(T_1, T_2, x) - I^g(T_1, T_2, x)| \le 2^{-6}\varepsilon. \tag{4.60}$$

By Lemma 4.5 there exists a neighborhood $\mathcal{U}_5$ of f in $\mathfrak{A}$ such that for each $g \in \mathcal{U}_5$ and each $T \in [1, 2N_0 + 8]$

$$|\ell(g, M_1, T) - \mu(f)T| \le 2^{-6}\varepsilon. \tag{4.61}$$

Set

$$\mathcal{U} = \bigcap_{i=1}^{5} \mathcal{U}_i. \tag{4.62}$$

Let $g \in \mathcal{U}$, $h \in H(f)$, and a number $T \ge 1$. There are two cases: (i) $T \le N_0$; (ii) $T > N_0$. Consider the case (i). By Lemma 4.6 there exists an (f)-good function

$$x\colon [0, \infty) \to H(f) \quad \text{such that } x \in \mathcal{A}(f), \quad x(0) = h.$$

Analogously to the case (i) in the proof of Lemma 4.7 we can show that relation (4.48) holds. Therefore in the case (i) the assertion of the lemma is valid.

Consider the case (ii). Then

$$T > N_0. \tag{4.63}$$

It follows from the definition of $\mathcal{U}_2$ (see (4.51)), (4.63), and (4.58) that there exists an a.c. function $u\colon [0, T - 4(N_1 + N_2 + 5)] \to R^n$ such that

$$u(0),\ u(T - 4(N_1 + N_2 + 5)) \in H(f), \tag{4.64}$$

$$\begin{aligned} &I^g(0, T - 4(N_1 + N_2 + 5), u) - \pi^f(u(0)) + \pi^f(u(T - 4(N_1 + N_2 + 5))) \\ &\le\ \ell(g, M_1, T - 4(N_1 + N_2 + 5)) + 2^{-6}\varepsilon. \end{aligned} \tag{4.65}$$

By (4.57) and (4.64) there exist

$$t_1,\ t_3 \in [4, N_1 + 4], \qquad t_2 \in [2N_1 + 16, 2N_1 + 16 + N_2] \tag{4.66}$$

for which

$$\begin{aligned} &|h - v_0(t_1)| \le 4^{-1}\delta, \quad |u(0) - v_0(t_2)| \le 4^{-1}\delta, \\ &|u(T - 4(N_1 + N_2 + 5)) - v_0(t_3)| \le 4^{-1}\delta. \end{aligned} \tag{4.67}$$

By Proposition 3.5, (4.66), (4.63), and (4.58) there exists an a.c. function $x\colon [0, T] \to R^n$ such that

$$\begin{aligned} x(0) &= h, \quad x(t) = v_0(t + t_1), \quad t \in [1, t_2 - t_1 - 1], \\ x(t) &= u(t - t_2 + t_1), \\ &\quad t \in [t_2 - t_1, T - 4(N_1 + N_2 + 5) + t_2 - t_1], \\ x(t) &= v_0(t + t_3 - (T - 4(N_1 + N_2 + 5) + t_2 - t_1)), \\ &\quad t \in [T - 4(N_1 + N_2 + 5) + t_2 - t_1 + 1, T], \\ I^g(\tau, \tau + 1, x) &= U^g(0, 1, x(\tau), x(\tau + 1)), \\ &\quad \tau = 0, t_2 - t_1 - 1, T - 4(N_1 + N_2 + 5) + t_2 - t_1. \end{aligned} \tag{4.68}$$

Clearly (4.47) holds. We will show that (4.48) holds.

For each a.c. function $v\colon [0, T] \to R^n$ and each $r_1, r_2 \in [0, T]$ satisfying $r_1 < r_2$ we set

$$\sigma(r_1, r_2, v) = I^g(r_1, r_2, v) - \pi^f(v(r_1)) + \pi^f(v(r_2)). \tag{4.69}$$

Set

$$\begin{aligned} &\tau_0 = 0, \quad \tau_1 = 1, \quad \tau_2 = t_2 - t_1 - 1, \quad \tau_3 = t_2 - t_1, \\ &\tau_4 = T - 4(N_1 + N_2 + 5) + t_2 - t_1, \\ &\tau_5 = T - 4(N_1 + N_2 + 5) + t_2 - t_1 + 1, \qquad \tau_6 = T. \end{aligned} \tag{4.70}$$

There exists an a.c. function $y\colon [0, T] \to R^n$ such that

$$|y(0)|,\ |y(T)| \le K, \qquad \sigma(0, T, y) \le \ell(g, K, T) + 2^{-6}\varepsilon. \tag{4.71}$$

It follows from (4.71) and the definition of $\mathcal{U}_1$, M_1 (see (4.49), (4.50)) that

$$|y(t)| \le M_1, \qquad t \in [0,T]. \tag{4.72}$$

Equations (4.69) and (4.70) imply that

$$\sigma(0,T,x) - \sigma(0,T,y) = \sum_{i=0}^{5} [\sigma(\tau_i, \tau_{i+1}, x) - \sigma(\tau_i, \tau_{i+1}, y)]. \tag{4.73}$$

By (4.72) and (4.69)

$$\sigma(\tau_i, \tau_{i+1}, y) \ge \ell(g, M_1, \tau_{i+1} - \tau_i), \qquad i = 0, \dots, 5. \tag{4.74}$$

It follows from (4.70), (4.68), (4.69), and (4.65) that

$$\sigma(\tau_3, \tau_4, x) \le \ell(g, M_1, \tau_4 - \tau_3) + 2^{-6}\varepsilon. \tag{4.75}$$

Analogously to (4.42) (see the proof of Lemma 4.7) we can show that

$$\sigma(\tau_i, \tau_{i+1}, x) \le \ell(g, M_1, \tau_{i+1} - \tau_i) + 2^{-5}\varepsilon, \qquad i = 1, 5. \tag{4.76}$$

By (4.68)–(4.70), the definition of $\mathcal{U}_3$ (see (4.55)), (4.64), and (4.56) for $i = 0, 2, 4$

$$\begin{aligned} \sigma(\tau_i, \tau_{i+1}, x) &= U^g(0, 1, x(\tau_i), x(\tau_{i+1})) - \pi^f(x(\tau_i)) + \pi^f(x(\tau_{i+1})) \\ &\le U^f(0, 1, x(\tau_i), x(\tau_{i+1})) - \pi^f(x(\tau_i)) + \pi^f(x(\tau_{i+1})) + 2^{-6}\varepsilon. \end{aligned} \tag{4.77}$$

We set

$$\gamma_0 = t_1, \quad \gamma_2 = t_2 - 1, \quad \gamma_4 = t_3. \tag{4.78}$$

Equations (4.78), (4.70), (4.68), and (4.67) imply that for $i = 0, 2, 4$

$$|x(\tau_i) - v_0(\gamma_i)|, |x(\tau_i + 1) - v_0(\gamma_i + 1)| \le 4^{-1}\delta. \tag{4.79}$$

It follows from (4.77), (4.70), (4.79), and (4.56) and the definition of δ (see (4.52)–(4.54)) that for $i = 0, 2, 4$

$$\begin{aligned} \sigma(\tau_i, \tau_{i+1}, x) &\le U^f(0, 1, x(\tau_i), x(\tau_i + 1)) - \pi^f(x(\tau_i)) + \pi^f(x(\tau_i + 1)) + 2^{-6}\varepsilon \\ &\le U^f(0, 1, v_0(\gamma_i), v_0(\gamma_i + 1)) - \pi^f(v_0(\gamma_i)) + \pi^f(v_0(\gamma_i + 1)) \\ &\quad + 2^{-4}\varepsilon = \mu(f) + 2^{-4}\varepsilon. \end{aligned}$$

By this relation and the definition of $\mathcal{U}_5$ for $i = 0, 2, 4$

$$\sigma(\tau_i, \tau_{i+1}, x) \le \ell(g, M_1, \tau_{i+1} - \tau_i) + 2^{-4}\varepsilon + 2^{-6}\varepsilon.$$

Combining this relation and (4.73)–(4.76) we obtain that

$$\sigma(0,T,x) - \sigma(0,T,y) \le 2^{-6}\varepsilon + 2^{-4}\varepsilon + 3(2^{-6}\varepsilon + 2^{-4}\varepsilon) \le 2^{-1}\varepsilon.$$

Together with (4.69) and (4.71) this implies (4.48). This completes the proof of the lemma.

LEMMA 4.9. *Let $\varepsilon \in (0,1)$, $K > D_f + 1$, and let L be a positive integer such that for each (f)-good function $v\colon [0,\infty) \to R^n$*

$$\operatorname{dist}(H(f), \{v(t):\ t \in [T, T+L]\}) \le 8^{-1}\varepsilon \tag{4.80}$$

for all large T (the existence of L follows from Proposition 3.9).

Then there exist a neighborhood $\mathcal{U}$ of f in $\mathfrak{A}$ and $\delta \in (0,1)$ such that for each $g \in \mathcal{U}$, each $T \in [L, \infty)$, and each a.c. function v: $[0,T] \to R^n$ which satisfies

$$d(v(0), H(f)) \le \delta, \qquad d(v(T), H(f)) \le \delta, \tag{4.81}$$

$$I^g(0,T,v) - \pi^f(v(0)) + \pi^f(v(T)) \le \ell(g,K,T) + \delta \tag{4.82}$$

the relation

$$\operatorname{dist}(H(f), \{v(t):\ t \in [S, S+L]\}) \le \varepsilon \tag{4.83}$$

holds for every $S \in [0, T-L]$.

Proof. By Lemma 4.3 there exists

$$\delta_1 \in (0, \varepsilon) \tag{4.84}$$

such that for each $T \in [L, \infty)$ and each a.c. function v: $[0,T] \to R^n$ which satisfies

$$\begin{aligned} &d(v(0), H(f)),\ d(v(T), H(f)) \le \delta_1, \\ &I^f(0,T,v) - T\mu(f) - \pi^f(v(0)) + \pi^f(v(T)) \le \delta_1 \end{aligned} \tag{4.85}$$

relation (4.83) holds for every $S \in [0, T-L]$. Fix a number

$$\delta \in (0, 8^{-1}\delta_1). \tag{4.86}$$

By Proposition 3.9 there exists an integer $L_1 \ge 1$ such that for each (f)-good function v: $[0,\infty) \to R^n$

$$\operatorname{dist}(H(f),\ \{v(t):\ t \in [T, T+L_1]\}) \le 8^{-1}\delta \tag{4.87}$$

for all large T. We may assume that

$$L_1 \ge 10L + 24. \tag{4.88}$$

By Lemma 4.4 there exist an integer $N_1 \ge 10$ and a neighborhood $\mathcal{U}_1$ of f in $\mathfrak{A}$ such that for each $g \in \mathcal{U}_1$, each $T_1 \ge 0$, $T_2 \ge T_1 + N_1 L_1$, each a.c. function x: $[T_1, T_2] \to R^n$ which satisfies

$$|x(T_i)| \le K, \quad i = 1,2, \quad I^g(T_1, T_2, x) \le U^g(T_1, T_2, x(T_1), x(T_2)) + 1, \tag{4.89}$$

and each $S \in [T_1, T_2 - N_1 L_1]$ there exists an integer $i_0 \in [0, N_1 - 8]$ such that for all $T \in [S + i_0 L_1, S + (i_0 + 7)L_1]$

$$\operatorname{dist}(H(f), \{x(t):\ t \in [T, T+L_1]\}) \le \delta. \tag{4.90}$$

By Lemma 4.8 there exists a neighborhood $\mathcal{U}_2$ of f in $\mathfrak{A}$ such that for each $g \in \mathcal{U}_2$, each $h \in H(f)$, and each $T \ge 1$ there exists an a.c. function x: $[0,T] \to R^n$ for which

$$\begin{aligned} &x(0) = h, \qquad x(T) \in H(f), \\ &I^g(0,T,x) - \pi^f(x(0)) + \pi^f(x(T)) \le \ell(g,K,T) + 8^{-1}\delta. \end{aligned} \tag{4.91}$$

Choose an integer

$$N_0 \geq 100 L_1 N_1. \tag{4.92}$$

By Lemma 4.5 there exists a neighborhood $\mathcal{U}_3$ of f in $\mathfrak{A}$ such that for each $g \in \mathcal{U}_3$ and each $\tau \in [1, N_0]$

$$|\ell(g, K, \tau) - \mu(f)\tau| \leq 4^{-1}\delta. \tag{4.93}$$

By Proposition 3.7 there exists a neighborhood $\mathcal{U}_4$ of f in $\mathfrak{A}$ such that for each $g \in \mathcal{U}_4$, each $T_1 \geq 0, T_2 \in [T_1+1, T_1+2N_0+1]$, and each a.c. function $x\colon [T_1, T_2] \to R^n$ which satisfies

$$\begin{aligned} &\inf\{I^f(T_1, T_2, x), I^g(T_1, T_2, x)\} \\ &\quad \leq 2N_0|\mu(f)| + 4 + 2\sup\{|\pi^f(h)|\colon\ h \in R^n,\ |h| \leq K+2\} \end{aligned} \tag{4.94}$$

the following relation holds:

$$|I^f(T_1, T_2, x) - I^g(T_1, T_2, x)| \leq \delta. \tag{4.95}$$

Set

$$\mathcal{U} = \bigcap_{i=1}^{4} \mathcal{U}_i. \tag{4.96}$$

Assume that $g \in \mathcal{U}$, $T \geq L$, and an a.c. function $v\colon\ [0, T] \to R^n$ satisfies (4.81) and (4.82). There are two cases: (i) $T \leq N_0$; (ii) $T > N_0$.

Consider case (i). It follows from (4.82), the definition of $\mathcal{U}_3$ (see (4.93)) that

$$I^g(0, T, v) - \pi^f(v(0)) + \pi^f(v(T)) \leq |\mu(f)|T + 2\delta. \tag{4.97}$$

By this relation and the definition of $\mathcal{U}_4$ (see (4.94), (4.95)) and (4.81)

$$|I^f(0, T, v) - I^g(0, T, v)| \leq \delta.$$

Together with (4.97), (4.81), and (4.86) this implies (4.85). It follows from (4.85) and the definition of δ_1 (see (4.84)) that (4.83) holds for all $S \in [0, T-L]$. Therefore in case (i) the assertion of the lemma is valid.

Consider case (ii). Then

$$T > N_0. \tag{4.98}$$

It follows from (4.81), (4.82), (4.98), (4.91) and the definition of N_1 and $\mathcal{U}_1$ (see (4.89), (4.90)) that there exists a sequence of numbers $\{T_i\}_{i=0}^q$ such that

$$\begin{aligned} &T_0 = 0, \quad T_q = T, \quad T_{i+1} - T_i \in [2L_1, 2(2N_1 - 6)L_1], \\ &\qquad i = 0, \dots, q-1, \\ &d(v(T_i), H(f)) \leq \delta, \qquad i = 0, \dots, q. \end{aligned} \tag{4.99}$$

For each a.c. function $y\colon [0, T] \to R^n$ we define

$$\sigma(r_1, r_2, y) = I^g(r_1, r_2, y) - \pi^f(y(r_1)) + \pi^f(y(r_2)) \tag{4.100}$$

for each $r_1, r_2 \in [0,T]$ satisfying $r_1 < r_2$ and set

$$\sigma(r,r,y) = 0 \tag{4.101}$$

for each $r \in [0,T]$. We set

$$\ell(g,K,0) = 0. \tag{4.102}$$

Let integers $j, p \in [0,q], j < p$. We will estimate

$$\sigma(T_j, T_p, v) - \ell(g, K, T_p - T_j).$$

By the definition of $\mathcal{U}_2$ (see 4.91)) and (4.99)–(4.102) there exists an a.c. function y: $[0,T] \to R^n$ such that

$$y(T_i) \in H(f), \quad i = 0, j, p, q, \quad \sigma(0, T_j, y) \le \ell(g, K, T_j) + 8^{-1}\delta,$$
$$\sigma(T_j, T_p, y) \le \ell(g, K, T_p - T_j) + 8^{-1}\delta, \qquad \sigma(T_p, T_q, y) \le \ell(g, K, T_q - T_p) + 8^{-1}\delta.$$

It follows from this relation, (4.82), and (4.99)–(4.102) that

$$\begin{aligned}\delta \ge \sigma(0,T,v) - \sigma(0,T,y) &= [\sigma(0,T_j,v) - \sigma(0,T_j,y)] + [\sigma(T_j,T_p,v) - \sigma(T_j,T_p,y)] \\ &\quad + [\sigma(T_p,T_q,v) - \sigma(T_p,T_q,y)] \\ &\ge \sigma(T_j,T_p,v) - \sigma(T_j,T_p,y) - 4^{-1}\delta, \\ \sigma(T_j,T_p,v) &\le \delta + 4^{-1}\delta + \ell(g,K,T_p - T_j) + 8^{-1}\delta. \end{aligned} \tag{4.103}$$

We have shown that (4.103) holds for each integer $j, p \in [0,q]$ satisfying $j < p$.

Let $S \in [0, T-L]$. By (4.99) and (4.88) there exist integers $j, p \in [0,q]$ such that

$$j < p, \quad S \in [T_j, T_p - L], \quad T_p - T_j \in [2L_1, 8N_1L_1]. \tag{4.104}$$

Evidently (4.103) holds. By (4.103), (4.100), (4.104), (4.92), and the definition of $\mathcal{U}_3$ (see (4.93))

$$I^g(T_j, T_p, v) - \pi^f(v(T_j)) + \pi^f(v(T_p)) \le \mu(f)(T_p - T_j) + 2\delta.$$

By this relation, the definition of $\mathcal{U}_4$ (see (4.94), (4.95)), (4.99), (4.104), and (4.92)

$$|I^f(T_j, T_p, v) - I^g(T_j, T_p, v)| \le \delta,$$
$$I^f(T_j, T_p, v) - \pi^f(v(T_j)) + \pi^f(v(T_p)) \le \mu(f)(T_p - T_j) + 3\delta.$$

It follows from these relations, (4.99), (4.88), (4.86), and the definition of δ_1 (see (4.84)) that (4.83) holds. This completes the proof of the lemma.

LEMMA 4.10. *Let $\varepsilon \in (0,1), K > D_f + 4$, and let $L \ge 1$ be an integer such that for each (f)-good function v: $[0,\infty) \to R^n$*

$$\operatorname{dist}(H(f), \{v(t):\ t \in [T, T+L]\}) \le 8^{-1}\varepsilon$$

for all large T (the existence of L follows from Proposition 3.9).

Then there exists a neighborhood $\mathcal{U}$ of f in $\mathfrak{A}$ such that for each $g \in \mathcal{U}$ and each $h \in H(f)$ there exists an a.c. function x: $[0,\infty) \to R^n$ such that $x(0) = h$;

$$\operatorname{dist}(H(f), \{x(t):\ t \in [S, S+L]\}) \le \varepsilon \tag{4.105}$$

for all $S \in [0, \infty)$*;*

$$I^g(0,T,x) - \pi^f(x(0)) + \pi^f(x(T)) \le \ell(g,K,T) + \varepsilon \tag{4.106}$$

for each $T \ge 1$*;*

$$I^g(t_1,t_2,x) - \pi^f(x(t_1)) + \pi^f(x(t_2)) \le \ell(g,K,t_2-t_1) + \varepsilon \tag{4.107}$$

for each $t_1 \ge 1, t_2 \ge t_1 + 1$.

Proof. By Lemma 4.9 there exist a neighborhood $\mathcal{U}_1$ of f in $\mathfrak{A}$ and

$$\delta_1 \in (0, 8^{-1}\varepsilon) \tag{4.108}$$

such that for each $g \in \mathcal{U}_1$, each $T \in [L, \infty)$, and each a.c. function $v\colon [0,T] \to R^n$ which satisfies

$$\begin{aligned} d(v(0), H(f)) \le \delta_1, \qquad d(v(T), H(f)) \le \delta_1, \\ I^g(0,T,v) - \pi^f(v(0)) + \pi^f(v(T)) \le \ell(g,K,T) + \delta_1 \end{aligned} \tag{4.109}$$

relation (4.105) holds for all $S \in [0, T-L]$.

By Lemma 4.8 there exists a neighborhood $\mathcal{U}_2$ of f in $\mathfrak{A}$ such that for each $g \in \mathcal{U}_2$, each $h \in H(f)$, and each $T \ge 1$ there exists an a.c. function $v\colon [0,T] \to R^n$ for which

$$\begin{aligned} v(0) = h, \qquad v(T) \in H(f), \\ I^g(0,T,v) - \pi^f(v(0)) + \pi^f(v(T)) \le \ell(g,K,T) + 8^{-1}\delta_1. \end{aligned} \tag{4.110}$$

Set

$$\mathcal{U} = \mathcal{U}_1 \cap \mathcal{U}_2. \tag{4.111}$$

Assume that $g \in \mathcal{U}$ and $h \in H(f)$. By the definition of $\mathcal{U}_2$ (see (4.110)) for each $N \ge 1$ there exists an a.c. function $x_N\colon [0,N] \to R^n$ such that (4.110) holds with $T = N$, $v = x_N$. It follows from the definition of $\mathcal{U}_1$ and δ_1 (see (4.109)) that for each integer $N \ge L$ and each number $S \in [0, N-L]$

$$\operatorname{dist}(H(f), \{x_N(t)\colon\ t \in [S, S+L]\}) \le \varepsilon. \tag{4.112}$$

Let $N \ge L$ be an integer. For each a.c. function $y\colon [0,N] \to R^n$ and each number $r_1, r_2 \in [0,N]$ satisfying $r_1 < r_2$ we define $\sigma(r_1,r_2,y)$ by (4.100). Assume that an integer $q \in \{2,3,4\}$, $\{t_i\}_{i=0}^q \subset [0,N]$, $t_0 = 0$, $t_q = N$, $t_{i+1} - t_i \ge 1$, $i = 0, \dots, q-1$. Equation (4.112), which holds for each $S \in [0, N-L]$, implies that

$$\sigma(t_i, t_{i+1}, x_N) \ge \ell(g,K,t_{i+1}-t_i), \qquad i = 0, \dots, q-1. \tag{4.113}$$

By the definition of $\mathcal{U}_2$ (see (4.110)) there exists an a.c. function $y\colon [0,N] \to R^n$ for which

$$y(t_i) \in H(f), \quad i = 0, \dots, q, \quad \sigma(t_i, t_{i+1}, y) \le \ell(g,K,t_{i+1}-t_i) + 8^{-1}\delta_1, \quad i = 0, \dots, q-1.$$

Together with (4.110), which holds with $T = N$, $v = x_N$, and (4.113), this implies that for each $j \in [0, q-1]$

$$\begin{aligned} 8^{-1}\delta_1 \ge \sigma(0,N,x_N) - \sigma(0,N,y) &= \sum_{i=0}^{q-1} [\sigma(t_i,t_{i+1},x_N) - \sigma(t_i,t_{i+1},y)] \\ &\ge \sigma(t_j,t_{j+1},x_N) - \ell(g,K,t_{j+1}-t_j) - 8^{-1}\delta_1 q. \end{aligned}$$

This implies that for each integer $N \geq L+3$ and each $\tau_1 \in \{0\} \cup [1, N-2]$, $\tau_2 \in [\tau_1 + 1, N-1]$

$$I^g(\tau_1, \tau_2, x_N) - \pi^f(x_N(\tau_1)) + \pi^f(x_N(\tau_2)) \leq \ell(g, K, \tau_2 - \tau_1) + 3 \cdot 4^{-1}\delta_1. \tag{4.114}$$

By this relation, (4.112), which holds for each $N \geq L$ and each number $S \in [0, N-L]$, and by Proposition 3.4 there exist a subsequence $\{x_{N_p}\}_{p=1}^{\infty}$ and an a.c. function $x\colon [0,\infty) \to R^n$ such that for each integer $N \geq 1$

$$x_{N_p}(t) \to x(t) \quad \text{as } p \to \infty \text{ uniformly in } [0, N] \tag{4.115}$$

and

$$I^g(T_1, T_2, x) \leq \liminf_{p\to\infty} I^g(T_1, T_2, x_{N_p}) \tag{4.116}$$

for each $T_1 \in [1,\infty) \cup \{0\}$, $T_2 \geq T_1 + 1$.

Equation (4.110), which holds with $T = N$, $v = x_N$, and (4.115), implies that $x(0) = h$. Equations (4.112) and (4.115) imply (4.105) for all $S \in [0,\infty)$. Equations (4.114)–(4.116) and (4.108) imply (4.106) for each $T \geq 1$ and (4.107) for each $t_1 \geq 1$, $t_2 \geq t_1 + 1$. This completes the proof of the lemma.

LEMMA 4.11. $\sup\{\pi^f(h)\colon\ h \in H(f)\} = 0$.

Proof. There exists $h_0 \in H(f)$ for which

$$\pi^f(h_0) \geq \pi^f(h), \qquad h \in H(f). \tag{4.117}$$

Let $v\colon [0,\infty) \to R^n$ be an a.c. function,

$$v(0) = h_0. \tag{4.118}$$

We will show that $\liminf_{T\to\infty}[I^f(0, T, v) - T\mu(f)] \geq 0$.

By Proposition 1.1 we may assume that v is an (f)-good function. Then $\Omega(v) = H(f)$. It follows from this relation, the representation formula (see (1.6)), (4.117), and (4.118) that

$$\liminf_{T\to\infty}[I^f(0, T, v) - T\mu(f)] \geq \liminf_{T\to\infty}[\pi^f(v(0)) - \pi^f(v(T))] \geq 0.$$

This implies that $\pi^f(h_0) \geq 0$. By Proposition 3.11 there exists an (f)-good function $u \in \mathcal{A}(f)$ satisfying $u(0) = h_0$. It is easy to see that $\Omega(u) = H(f)$,

$$\liminf_{T\to\infty}[I^f(0, T, u) - T\mu(f)] = \liminf_{T\to\infty}[\pi^f(u(0)) - \pi^f(u(T))] = 0.$$

This completes the proof of the lemma.

LEMMA 4.12. *Let $\varepsilon \in (0,1)$. Then there exists a neighborhood $\mathcal{U}$ of f in $\mathfrak{A}$ such that $|\mu(f) - \mu(g)| \leq \varepsilon$ for each $g \in \mathcal{U}$.*

Proof. By Proposition 3.1 there exists a neighborhood $\mathcal{U}_1$ of f in $\mathfrak{A}$ and $M_0 > 0$ such that for each $g \in \mathcal{U}_1$ and each (g)-good function $x\colon [0,\infty) \to R^n$

$$\limsup_{t\to\infty} |x(t)| < M_0. \tag{4.119}$$

Set

$$M_1 = \sup\{|U^f(0, 1, x, y)|\colon\ x, y \in R^n,\ |x|, |y| \leq 2M_0 + 2\}. \tag{4.120}$$

By Proposition 3.7 there is a neighborhood $\mathcal{U}_2$ of f in $\mathfrak{A}$ such that for each $g \in \mathcal{U}_2$, each $T \geq 0$, and each a.c. function y: $[T, T+1] \to R^n$ satisfying

$$\inf\{I^f(T, T+1, y), I^g(T, T+1, y)\} \leq 2M_1 + 4$$

the following relation holds:

$$|I^f(T, T+1, y) - I^g(T, T+1, y)| \leq 8^{-1}\varepsilon. \tag{4.121}$$

By Proposition 3.8 there exists a neighborhood $\mathcal{U}_3$ of f in $\mathfrak{A}$ such that for each $g \in \mathcal{U}_3$ and each $x, y \in R^n$ satisfying $|y|$, $|x| \leq 2M_0 + 2$ the following relation holds:

$$|U^f(0, 1, x, y) - U^g(0, 1, x, y)| \leq 16^{-1}\varepsilon. \tag{4.122}$$

Set

$$\mathcal{U} = \bigcap_{i=1}^{3} \mathcal{U}_i. \tag{4.123}$$

Let $g_1, g_2 \in \mathcal{U}$. Consider a (g_1)-good function x: $[0, \infty) \to R^n$. We have

$$\sup\{|I^{g_1}(0, T, x) - T\mu(g_1)|\colon\ T \in (0, \infty)\} < \infty. \tag{4.124}$$

By the definition of $\mathcal{U}_1, M_0$ (see (4.119)) we may assume that

$$|x(t)| \leq M_0, \qquad t \in [0, \infty). \tag{4.125}$$

By Proposition 1.1 we may assume without loss of generality that

$$I^{g_1}(T, T+1, x) \leq U^{g_1}(0, 1, x(T), x(T+1)) + 4^{-1} \quad \text{for all } T \in [0, \infty).$$

It follows from this relation, (4.125), the definition of $\mathcal{U}_3$ (see (4.122)), and (4.120) that

$$I^{g_1}(T, T+1, x) \leq M_1 + 2^{-1} \quad \text{for all } T \in [0, \infty).$$

By this relation and the definition of $\mathcal{U}_2$, for each $T \in [0, \infty)$,

$$\begin{aligned} &|I^f(T, T+1, x) - I^{g_1}(T, T+1, x)| \leq 8^{-1}\varepsilon, \\ &\quad I^f(T, T+1, x) \leq M_1 + 1, \\ &|I^f(T, T+1, x) - I^{g_2}(T, T+1, x)| \leq 8^{-1}\varepsilon, \\ &\quad |I^{g_1}(T, T+1, x) - I^{g_2}(T, T+1, x)| \leq 4^{-1}\varepsilon. \end{aligned} \tag{4.126}$$

Equations (4.124) and (4.126) imply that

$$\sup\{|I^{g_2}(0, N, x) - N4^{-1}\varepsilon - N\mu(g_1)|\colon\ N = 1, 2, \dots, \} < \infty.$$

Together with Proposition 1.1, this implies that $\mu(g_2) \leq \mu(g_1) + 4^{-1}\varepsilon$. This completes the proof of the lemma.

LEMMA 4.13. *Let $K > D_f + 1$. Then there exists a neighborhood $\mathcal{U}$ of f in $\mathfrak{A}$ such that $\ell(g, K, \tau) \leq \tau\mu(g)$ for each $g \in \mathcal{U}$ and each $\tau > 0$.*

Proof. By Proposition 3.13 there exists a neighborhood $\mathcal{U}$ of f in $\mathfrak{A}$ such that for each $g \in \mathcal{U}$ and each (g)-good function $x\colon [0,\infty) \to R^n$

$$\limsup_{t\to\infty} |x(t)| < K.$$

Let $g \in \mathcal{U}$, $\tau > 0$. Consider any (g)-good function $x\colon [0,\infty) \to R^n$. We may assume that

$$|x(t)| \le K, \qquad t \in [0,\infty). \tag{4.127}$$

It follows from (4.127) that for each integer $N \ge 1$

$$\begin{aligned} I^g(0,N\tau,x) - \pi^f(x(0)) + \pi^f(x(N\tau)) &= \sum_{i=0}^{N-1} [I^g(i\tau,(i+1)\tau,x) \\ &\qquad - \pi^f(x(i\tau)) + \pi^f(x((i+1)\tau))] \\ &\ge N\ell(g,K,\tau), \\ I^g(0,N\tau,x) &\ge N\ell(g,K,\tau) + 2\sup\{|\pi^f(z)|\colon\ z \in R^n,\ |z| \le K\}. \end{aligned}$$

Since x is a (g)-good function, we have

$$\sup\{N\ell(g,K,\tau) - N\tau\mu(g)\colon\ N = 1,2,\ldots\} < \infty.$$

This completes the proof of the lemma.

There exists

$$h_* \in H(f) \quad \text{such that } \pi^f(h_*) \ge \pi^f(h),\ h \in H(f). \tag{4.128}$$

LEMMA 4.14. *Let $\varepsilon \in (0,1)$ and $K > D_f + 4$. Then there exist a neighborhood $\mathcal{U}$ of f in $\mathfrak{A}$ and an integer $L \ge 1$ such that for each $g \in \mathcal{U}$ and each $h \in H(f)$ there exists a (g)-good function $v\colon [0,\infty) \to R^n$ such that*

$$v(0) = h; \tag{4.129}$$

$$\operatorname{dist}(H(f), \{v(t)\colon\ t \in [T,T+L]\}) \le \varepsilon \tag{4.130}$$

for all $T \in [0,\infty)$;

$$I^g(T_1,T_2,v) - \pi^f(v(T_1)) + \pi^f(v(T_2)) \le \ell(g,K,T_2-T_1) + \varepsilon \tag{4.131}$$

for each $T_1 \in \{0\} \cup [1,\infty), T_2 \ge T_1 + 1$;

$$|I^g(0,T,v) - \pi^f(v(0)) + \pi^f(v(T)) - T\mu(g)| \le \varepsilon \tag{4.132}$$

for each $T \in [1,\infty)$;

$$|\liminf_{T\to\infty} [I^g(0,T,v) - T\mu(g)] - \pi^f(h)| \le 2\varepsilon. \tag{4.133}$$

Proof. By Lemma 4.6 and (4.128) there exists an (f)-good function $v_0\colon [0,\infty) \to H(f)$ for which

$$v_0 \in \mathcal{A}(f), \qquad v_0(0) = h_*. \tag{4.134}$$

By Proposition 3.10 there exists a number

$$\delta \in (0, 2^{-8}\varepsilon) \tag{4.135}$$

such that for each $y_1, y_2, z_1, z_2 \in R^n$ which satisfy $|y_i|, |z_i| \le 2K+8, i = 1, 2, |y_i - z_i| \le 4\delta, i = 1, 2$, the following relations hold:

(4.136)

$$|U^f(0, 1, y_1, y_2) - U^f(0, 1, z_1, z_2)| \le 2^{-8}\varepsilon, \quad |\pi^f(y_i) - \pi^f(z_i)| \le 2^{-8}\varepsilon, \quad i = 1, 2.$$

By Proposition 3.8 there exists a neighborhood $\mathcal{U}_1$ of f in $\mathfrak{A}$ such that for each $g \in \mathcal{U}_1$ and each $z, y \in R^n$ which satisfy $|y|, |z| \le 2K + 8$ the following relation holds:

$$|U^f(0, 1, y, z) - U^g(0, 1, y, z)| \le 2^{-4}\delta. \tag{4.137}$$

Since $\Omega(v_0) = H(f)$, there exist integers $N_1, N_2 \ge 10$ for which

$$\begin{aligned} &\operatorname{dist}(H(f), \{v_0(t):\ t \in [0, N_1]\}) \le 2^{-4}\delta, \\ &\operatorname{dist}(H(f), \{v_0(t):\ t \in [4N_1, 4N_1 + N_2]\}) \le 2^{-4}\delta. \end{aligned} \tag{4.138}$$

By Proposition 3.9 there exists an integer $L \ge 1$ such that for each (f)-good function $v\colon [0, \infty) \to R^n$

$$\operatorname{dist}(H(f), \{v(t):\ t \in [T, T + L]\}) \le 2^{-4}\delta \tag{4.139}$$

for all large T.

By Lemma 4.10 there exists a neighborhood $\mathcal{U}_2$ of f in $\mathfrak{A}$ such that for each $g \in \mathcal{U}_2$ and each $h \in H(f)$ there exists an a.c. function $v\colon [0, \infty) \to R^n$ satisfying (4.129) and such that for each $T \in [0, \infty)$

$$\operatorname{dist}(H(f), \{v(t):\ t \in [T, T + L]\}) \le \delta, \tag{4.140}$$

$$I^g(T_1, T_2, v) - \pi^f(v(T_1)) + \pi^f(v(T_2)) \le \ell(g, K, T_2 - T_1) + \delta \tag{4.141}$$

for each $T_1 \in \{0\} \cup [1, \infty), T_2 \ge T_1 + 1$.

By Lemma 4.13 there exists a neighborhood $\mathcal{U}_3$ of f in $\mathfrak{A}$ such that for each $g \in \mathcal{U}_3$ and each $T > 0$

$$\ell(g, K, T) \le T\mu(g). \tag{4.142}$$

By Lemma 4.12 there exists a neighborhood $\mathcal{U}_4$ of f in $\mathfrak{A}$ such that for each $g \in \mathcal{U}_4$

$$|\mu(g) - \mu(f)| \le 2^{-8}\delta(8N_1 + 8N_2)^{-1}. \tag{4.143}$$

By Proposition 3.7 there exists a neighborhood $\mathcal{U}_5$ of f in $\mathfrak{A}$ such that for each $g \in \mathcal{U}_5$, each $T_1 \ge 0$, $T_2 \in [T_1 + 1, T_1 + 8(N_1 + N_2)]$, and each a.c. function $v\colon [T_1, T_2] \to R^n$ which satisfies

$$\begin{aligned} &\inf\{I^f(T_1, T_2, v), I^g(T_1, T_2, v)\} \\ &\quad \le (8N_1 + 8N_2)|\mu(f)| + 4 + 2\sup\{|\pi^f(z)|:\ z \in R^n, |z| \le 2K + 2\} \end{aligned} \tag{4.144}$$

the following relation holds:

$$|I^f(T_1, T_2, v) - I^g(T_1, T_2, v)| \le 2^{-8}\delta. \tag{4.145}$$

Set

$$\mathcal{U} = \bigcap_{i=1}^{5} \mathcal{U}_i. \tag{4.146}$$

Let $g \in \mathcal{U}$ and $h \in H(f)$. It follows from the definition of $\mathcal{U}_2$ that there exists an a.c. function $v\colon [0,\infty) \to R^n$ such that (4.129) holds, relation (4.140) holds for each $T \in [0,\infty)$, and relation (4.141) holds for each $T_1 \in \{0\} \cup [1,\infty), T_2 \ge T_1 + 1$. By (4.141), which holds for $T_1 = 0$ and each $T_2 \ge 1$, and the definition of $\mathcal{U}_3$ (see (4.142)) v is a (g)-good function.

Fix a number $T \ge 1$. We will establish (4.132). It follows from (4.140) and (4.138) that there exist

$$t_1 \in [0, N_1], \qquad t_2 \in [4N_1, 4N_1 + N_2] \tag{4.147}$$

for which

$$|v(T) - v_0(t_1)| \le \delta + 8^{-1}\delta, \qquad |h - v_0(t_2)| \le \delta + 8^{-1}\delta. \tag{4.148}$$

By Proposition 3.5 there exists an a.c. function $x\colon [0, T + t_2 - t_1] \to R^n$ such that

$$\begin{aligned}
&x(t) = v(t), \quad t \in [0,T], \quad x(t) = v_0(t + t_1 - T),\\
&t \in [T+1, T + t_2 - t_1 - 1],\\
&x(T + t_2 - t_1) = h, \quad I^g(S, S+1, x) = U^g(0, 1, x(S), x(S+1)),\\
&S = T, \quad T + t_2 - t_1 - 1.
\end{aligned} \tag{4.149}$$

Equations (4.149) and (4.129) imply that

$$I^g(0, T + t_2 - t_1, x) \ge \mu(g)(T + t_2 - t_1). \tag{4.150}$$

Equation (4.149) implies that

$$\begin{aligned}
&I^g(T, T + t_2 - t_1, x) - \pi^f(x(T)) + \pi^f(x(T + t_2 - t_1))\\
&\quad = U^g(0, 1, x(T), x(T+1)) - \pi^f(x(T))\\
&\qquad + \pi^f(x(T+1)) + I^g(t_1 + 1, t_2 - 1, v_0)\\
&\qquad - \pi^f(v_0(t_1 + 1)) + \pi^f(v_0(t_2 - 1))\\
&\qquad + U^g(0, 1, x(T + t_2 - t_1 - 1), x(T + t_2 - t_1))\\
&\qquad - \pi^f(x(T + t_2 - t_1 - 1)) + \pi^f(x(T + t_2 - t_1)).
\end{aligned} \tag{4.151}$$

Analogously to the proof of the case (i) in Lemma 4.7 (see (4.30) and (4.31)) we can show by using (4.134), (4.147), and the definition of $\mathcal{U}_5$ (see (4.145)) that

$$I^g(t_1+1, t_2-1, v_0) - \pi^f(v_0(t_1+1)) + \pi^f(v_0(t_2-1)) \le 2^{-8}\delta + \mu(f)(t_2 - t_1 - 2). \tag{4.152}$$

Set

$$S_1 = T, \quad S_2 = T + t_2 - t_1 - 1, \quad r_1 = t_1, \quad r_2 = t_2 - 1. \tag{4.153}$$

It follows from this relation; (4.149); the definition of $\mathcal{U}_1, \delta$ (see (4.135)–(4.137)); (4.140), which holds for each $T \geq 0$; and (4.134) that for $i = 1, 2$

$$\begin{aligned} &U^g(0,1,x(S_i),x(S_i+1)) - \pi^f(x(S_i)) + \pi^f(x(S_i+1)) \\ &\quad \leq U^f(0,1,x(S_i),x(S_i+1)) - \pi^f(x(S_i)) + \pi^f(x(S_i+1)) + 2^{-4}\delta \\ &\quad \leq U^f(0,1,v_0(r_i),v_0(r_i+1)) - \pi^f(v_0(r_i)) + \pi^f(v_0(r_i+1)) + 2^{-6}\varepsilon \\ &\quad \leq \mu(f) + 2^{-6}\varepsilon. \end{aligned}$$

By this relation, (4.151)–(4.153), the definition of $\mathcal{U}_4$ (see (4.143)), and (4.147)

$$\begin{aligned} &I^g(T, T+t_2-t_1, x) - \pi^f(x(T)) + \pi^f(x(T+t_2-t_1)) \\ &\quad = 2\mu(f) + 2^{-5}\varepsilon + \mu(f)(t_2-t_1-2) + 2^{-8}\varepsilon \\ &\quad \leq 2^{-5}\varepsilon + \mu(g)(t_2-t_1) + 2^{-7}\varepsilon. \end{aligned} \tag{4.154}$$

It follows from this relation, (4.150), (4.149), and (4.129) that

$$I^g(0,T,v) - \pi^f(v(0)) + \pi^f(v(T)) \geq \mu(g)T - 2^{-4}\varepsilon.$$

This relation; (4.141), which holds with $T_1 = 0, T_2 = T$; and (4.142) imply (4.132). Therefore we have shown that (4.132) holds for each $T \in [1, \infty)$. Together with (4.129), this implies that

$$\varepsilon \geq |\liminf_{T\to\infty}[I^g(0,T,v) - T\mu(g)] - \liminf_{T\to\infty}[\pi^f(h) - \pi^f(v(T))]|. \tag{4.155}$$

By (4.140), which holds for each $T \in [0, \infty)$, and the definition of δ (see (4.135), (4.136))

$$|\liminf_{T\to\infty}[\pi^f(h) - \pi^f(v(T))] - [\pi^f(h) - \sup\{\pi^f(z):\ z \in H(f)\}]| \leq 2^{-8}\varepsilon.$$

Equation (4.133) now follows from this relation, (4.155), and Lemma 4.11. The lemma is proven.

LEMMA 4.15. *Let $\varepsilon \in (0,1)$. Then there exist $\delta \in (0, \varepsilon)$ and a neighborhood $\mathcal{U}$ of f in $\mathfrak{A}$ such that for each $g \in \mathcal{U}$, each $h \in H(f)$, and each $y \in R^n$ satisfying $|y - h| \leq \delta$ the relation $|\pi^g(y) - \pi^f(h)| \leq \varepsilon$ holds.*

Proof. Fix

$$K > D_f + 4. \tag{4.156}$$

By Proposition 3.10 there exists a number

$$\delta \in (0, 8^{-1}\varepsilon) \tag{4.157}$$

such that for each $x_1, x_2, y_1, y_2 \in R^n$ which satisfy

$$|x_i|, |y_i| \leq K+2, \quad i = 1,2, \quad |x_i - y_i| \leq 8\delta, \quad i = 1,2, \tag{4.158}$$

the following relations hold:

$$|U^f(0,1,x_1,x_2) - U^f(0,1,y_1,y_2)| \leq 2^{-8}\varepsilon, \ \ |\pi^f(x_i) - \pi^f(y_i)| \leq 2^{-8}\varepsilon, \ \ i = 1,2. \tag{4.159}$$

By Proposition 3.8 there exists a neighborhood $\mathcal{U}_1$ of f in $\mathfrak{A}$ such that for each $g \in \mathcal{U}_1$ and each $y_1, y_2 \in R^n$ which satisfy $|y_i| \leq 2K+2, i = 1,2$,

$$|U^f(0,1,y_1,y_2) - U^g(0,1,y_1,y_2)| \leq 2^{-8}\varepsilon. \tag{4.160}$$

By Lemma 4.14 there exist a neighborhood $\mathcal{U}_2$ of f in $\mathfrak{A}$ and an integer $L \geq 1$ such that for each $g \in U_2$ and each $h \in H(f)$ there exists a (g)-good function v: $[0,\infty) \to R^n$ such that $v(0) = h$;

$$\operatorname{dist}(H(f), \{v(t):\ t \in [T, T+L]\}) \leq 2^{-6}\delta \tag{4.161}$$

for all $T \in [0,\infty)$;

$$I^g(T_1, T_2, v) - \pi^f(v(T_1)) + \pi^f(v(T_2)) \leq \ell(g, K, T_2 - T_1) + 2^{-6}\delta \tag{4.162}$$

for each $T_1 \in \{0\} \cup [1,\infty), T_2 \geq T_1 + 1$;

$$|I^g(0, T, v) - \pi^f(v(0)) + \pi^f(v(T)) - T\mu(g)| \leq 2^{-6}\delta \tag{4.163}$$

for each $T \in [1,\infty)$;

$$|\liminf_{T\to\infty}[I^g(0, T, v) - T\mu(g)] - \pi^f(h)| \leq 2^{-6}\delta. \tag{4.164}$$

By Lemmas 4.13 and 4.12 there exist a neighborhood $\mathcal{U}_3$ of f in $\mathfrak{A}$ such that for each $g \in \mathcal{U}_3$ and each $T \in (0,\infty)$

$$|\mu(g) - \mu(f)| \leq 2^{-6}\varepsilon, \qquad \ell(g, K, T) \leq T\mu(g). \tag{4.165}$$

By Proposition 3.13 there exist an integer $L_1 \geq 1$ and a neighborhood $\mathcal{U}_4$ of f in $\mathfrak{A}$ such that for each $g \in \mathcal{U}_4$ and each (g)-good function v: $[0,\infty) \to R^n$

$$\operatorname{dist}(H(f), \quad \{v(t):\quad t \in [T, T+L_1]\}) \leq 2^{-8}\delta \tag{4.166}$$

for all large T. Set

$$\mathcal{U} = \bigcap_{i=1}^{4} \mathcal{U}_i. \tag{4.167}$$

Assume that

$$g \in \mathcal{U}, \quad h \in H(f), \quad y \in R^n, \quad |y - h| \leq \delta. \tag{4.168}$$

By the definition of $\mathcal{U}_2, L$ and (4.168) there exists a (g)-good function v: $[0,\infty) \to R^n$ such that $v(0) = h$, (4.161) holds for each $T \geq 0$, (4.162) holds for each $T_1 \in \{0\} \cup [1,\infty)$ and each $T_2 \geq T_1 + 1$, and (4.163) holds for each $T \in [1,\infty)$. Together with (4.165) this implies that

$$0 \leq T\mu(g) - \ell(g, K, T) \leq 2^{-5}\delta \quad \text{for all } T \in [1,\infty). \tag{4.169}$$

Consider any (g)-good function u: $[0,\infty) \to R^n$ for which $u(0) = y$. By the definition of $\mathcal{U}_4, L_1$, (4.168) holds with $v = u$ for all large T. Together with (4.169) and (4.168), this implies that

$$\begin{aligned} \liminf_{T\to\infty}[I^g(0, T, u) - T\mu(g)] &\geq \liminf_{T\to\infty}[I^g(0, T, u) - \ell(g, K, T)] - 2^{-5}\delta \\ &\geq \liminf_{T\to\infty}[\pi^f(y) - \pi^f(u(T))] - 2^{-5}\delta. \end{aligned} \tag{4.170}$$

It follows from (4.166), which holds with $v = u$ for all large T; (4.170); (4.168); the definition of δ (see (4.157)–(4.159)); and Lemma 4.1 that

$$\begin{aligned}\liminf_{T\to\infty}[\pi^f(y) - \pi^f(u(T))] &\geq \pi^f(h) - 2^{-8}\varepsilon - \limsup_{T\to\infty}\pi^f(u(T))\\ &\geq \pi^f(h) - 2^{-8}\varepsilon - \sup\{\pi^f(z)\colon z \in H(f)\} - 2^{-8}\varepsilon\\ &\geq \pi^f(h) - 2^{-7}\varepsilon.\end{aligned}$$

Together with (4.170) and (4.157) this implies that

$$\liminf_{T\to\infty}[I^g(0,T,u) - T\mu(g)] \geq \pi^f(h) - \varepsilon.$$

Therefore

$$\pi^g(y) \geq \pi^f(h) - \varepsilon. \tag{4.171}$$

We will show that $\pi^g(y) \leq \pi^f(h) + \varepsilon$. By Lemma 4.6 there exists an (f)-good function

$$v_0\colon [0,\infty) \to H(f) \quad \text{such that } v_0 \in \mathcal{A}(f),\ v_0(0) = h. \tag{4.172}$$

By the definition of $\mathcal{U}_2$ and L and (4.172) there exists a (g)-good function $v_1\colon [0,\infty) \to R^n$ such that

$$v_1(0) = v_0(1), \tag{4.173}$$

(4.161) holds with $v = v_1$ for each $T \in [0,\infty)$; (4.162) holds with $v = v_1$ for each $T_1 \in \{0\} \cup [1,\infty)$, $T_2 \geq T_1 + 1$; (4.163) holds with $v = v_1$ for each $T \in [1,\infty)$; and (4.164) holds with $v = v_1$, $h = v_0(1)$.

By Proposition 3.5 there exists an a.c. function $w\colon [0,\infty) \to R^n$ such that

(4.174)

$$w(0) = y, \quad w(t) = v_1(t-1), \quad t \in [1,\infty), \quad I^g(0,1,w) = U^g(0,1,w(0),w(1)).$$

Equations (4.173), (4.174), and (4.164), which holds with $v = v_1$, $h = v_0(1)$, imply that

$$\begin{aligned}\pi^g(y) &\leq \liminf_{T\to\infty}[I^g(0,T,w) - T\mu(g)] = U^g(0,1,y,v_0(1)) - \mu(g)\\ &\quad + \liminf_{T\to\infty}[I^g(0,T,v_1) - T\mu(g)]\\ &= U^g(0,1,y,v_0(1)) - \mu(g) + \pi^f(v_0(1)) + 2^{-6}\delta. \qquad (4.175)\end{aligned}$$

It follows from (4.168), (4.156), (4.172), the definition of $\mathcal{U}_1$ (see (4.160)), and the definition of δ (see (4.157)–(4.159)) that

$$\begin{aligned}U^g(0,1,y,v_0(1)) &\leq U^g(0,1,y,v_0(1)) + 2^{-8}\varepsilon \leq 2^{-8}\varepsilon + U^f(0,1,v_0(0),v_0(1)) + 2^{-8}\varepsilon\\ &= 2^{-7}\varepsilon + \pi^f(h) - \pi^f(v_0(1)) + \mu(f).\end{aligned}$$

Together with (4.175) and (4.165) this implies that $\pi^g(y) \leq 2^{-7}\varepsilon + 2^{-5}\varepsilon + \pi^f(h) \leq \pi^f(h) + \varepsilon$. This completes the proof of the lemma.

There exists $h_f \in H(f)$ such that

$$\pi^f(h_f) \geq \pi^f(h), \qquad h \in H(f). \tag{4.176}$$

LEMMA 4.16. *Let $\varepsilon \in (0,1)$, $K > D_f + 4$. Then there exist a neighborhood $\mathcal{U}$ of f in $\mathfrak{A}$ and integers $Q_1 \geq 8$, $Q_2 \geq 8 + Q_1$ such that, for each $g \in \mathcal{U}$ and each $x \in R^n$ satisfying $|x| \leq K$,*

$$\pi^g(x) = \inf\{\liminf_{T\to\infty}[I^g(0,T,v) - T\mu(g)]\colon\ v\colon\ [0,\infty) \to R^n \text{ is an a.c. function},$$

$$v(0) = x,\ \inf\{|v(t) - h_f|\colon\ t \in [Q_1, Q_2]\} \leq \varepsilon\}. \tag{4.177}$$

Proof. By Proposition 3.13 there exist an integer $L \geq 1$ and a neighborhood $\mathcal{U}_1$ of f in $\mathfrak{A}$ such that, for each $g \in \mathcal{U}_1$ and each (g)-good function $v\colon\ [0,\infty) \to R^n$,

$$\operatorname{dist}(H(f),\ \{v(t)\colon\ t \in [T, T+L]\}) \leq 16^{-1}\varepsilon \tag{4.178}$$

for all large T.

By Lemma 4.4 and the definition of L there exist an integer $N \geq 10$ and a neighborhood $\mathcal{U}_2$ of f in $\mathfrak{A}$ such that for each $g \in \mathcal{U}_2$, each $T_1 \geq 0$, $T_2 \geq T_1 + NL$, each a.c. function $v\colon\ [T_1, T_2] \to R^n$ which satisfies

$$|v(T_i)| \leq K, \quad i = 1,2, \quad I^g(T_1, T_2, v) \leq U^g(T_1, T_2, v(T_1), v(T_2)) + 4, \tag{4.179}$$

and each $S \in [T_1, T_2 - NL]$ there exists an integer $i_0 \in [0, N-8]$ such that, for all $T \in [S + i_0 L, S + (i_0 + 7)L]$,

$$\operatorname{dist}(H(f), \{v(t)\colon\ t \in [T, T+L]\}) \leq 2^{-1}\varepsilon. \tag{4.180}$$

Set

$$\mathcal{U} = \mathcal{U}_1 \cap \mathcal{U}_2, \quad Q_1 = NL, \quad Q_2 = 2NL. \tag{4.181}$$

Assume that

$$g \in \mathcal{U}, \quad x \in R^n, \quad |x| \leq K.$$

Denote by E the set of all (g)-good functions $v\colon\ [0,\infty) \to R^n$ for which

$$v(0) = x, \qquad \liminf_{T\to\infty}[I^g(0,T,v) - T\mu(g)] \leq \pi^g(x) + 1. \tag{4.182}$$

It is easy to see that

$$\pi^g(x) = \inf\{\liminf_{T\to\infty}[I^g(0,T,v) - T\mu(g)]\colon\ v \in E\}. \tag{4.183}$$

Consider any $v \in E$. By the definition of $\mathcal{U}_1, L$

$$|v(t)| \leq K \text{ for all large } t. \tag{4.184}$$

Equation (4.182) implies that

$$I^g(0,T,v) \leq U^g(0,T,v(0),v(T)) + 2, \quad \text{for all } T \in [1,\infty).$$

It follows from this relation, (4.184), (4.181), and the definition of $N, \mathcal{U}_2$ (see (4.179), (4.180)) that $\inf\{|v(t) - h_f|\colon t \in [Q_1, Q_2]\} \le \varepsilon$. This completes the proof of the lemma.

Proof of Theorem 2.1. By Lemma 4.12 f is a continuity point of the mapping $g \to \mu(g)$, $g \in \mathfrak{A}$. We will show that f is a continuity point of the mapping $g \to \pi^g$, $g \in \mathfrak{A}$.

Assume that $\varepsilon \in (0,1), K > D_f + 4$. By Lemma 4.15 there exist a neighborhood $\mathcal{U}_1$ of f in $\mathfrak{A}$ and

$$\delta \in (0, 16^{-1}\varepsilon) \tag{4.185}$$

such that, for each $g \in \mathcal{U}_1$, each $h \in H(f)$, and each $y \in R^n$ satisfying $|y - h| \le \delta$, the following relation holds:

$$|\pi^g(y) - \pi^f(h)| \le 16^{-1}\varepsilon. \tag{4.186}$$

By Lemma 4.16 there exist a neighborhood $\mathcal{U}_2$ of f in $\mathfrak{A}$ and integers $Q_1 \ge 8$, $Q_2 \ge 8 + Q_1$ such that for each $g \in \mathcal{U}_2$ and each $x \in R^n$ satisfying $|x| \le K$, relation (4.177) holds with $\varepsilon = 8^{-1}\delta$. By Lemma 4.12 and Proposition 3.8 there exists a neighborhood $\mathcal{U}_3$ of f in $\mathfrak{A}$ such that

$$|\mu(g) - \mu(f)| \le (16(Q_1 + Q_2))^{-1}\varepsilon \quad \text{for each } g \in \mathcal{U}_3, \tag{4.187}$$

$$|U^f(0,\tau,x,y) - U^g(0,\tau,x,y)| \le 16^{-1}\varepsilon \tag{4.188}$$

for each $g \in \mathcal{U}_3$, each $\tau \in [1, 2Q]$, and each $x, y \in R^n$ which satisfy $|x|, |y| \le 2K + 2$. Set

$$\mathcal{U} = \bigcap_{i=1}^{3} \mathcal{U}_i.$$

It follows from the definition of $\mathcal{U}_2$, Q_1, Q_2, and (4.177) that for each $g \in \mathcal{U}$ and each $x \in R^n$ satisfying $|x| \le K$

$$\pi^g(x) = \inf\{U^g(0,T,x,y) - T\mu(g) + \pi^g(y)\colon\ t \in [Q_1, Q_2], y \in R^n, |y - h_f| \le 8^{-1}\delta\}.$$

By this relation, (4.187), (4.188), and the definition of $\mathcal{U}_1, \delta$ (see (4.185) and (4.186)), for each $g \in \mathcal{U}$ and each $x \in R^n$ satisfying $|x| \le K$,

$$\begin{aligned}|\pi^g(x) - \inf\{U^f(0,T,x,y) - T\mu(f) + \pi^f(h_f)\colon\ T \in [Q_1, Q_2],\ y \in R^n,\ |y - h_f| \le 8^{-1}\delta\}| \\ \le\ 16^{-1}\varepsilon + 8^{-1}\varepsilon + 16^{-1}\varepsilon \le 4^{-1}\varepsilon.\end{aligned}$$

This implies that for each g_1, $g_2 \in \mathcal{U}$ and each $x \in R^n$ satisfying $|x| \le K$

$$|\pi^{g_1}(x) - \pi^{g_2}(x)| \le 2^{-1}\varepsilon.$$

Therefore f is a continuity point of the mapping $g \to \pi^g$, $g \in \mathfrak{A}$. The theorem is proven.

Proof of Theorem 2.2. Let $\varepsilon \in (0,1)$. By Proposition 3.9 there exists an integer $L \ge 1$ such that for each (f)-good function $v\colon\ [0,\infty) \to R^n$

$$\operatorname{dist}(H(f), \{v(t)\colon\ t \in [T, T + L]\}) \le 2^{-8}\varepsilon \tag{4.189}$$

for all large T. By Lemma 4.9 there exist a neighborhood $\mathcal{U}_1$ of f in $\mathfrak{A}$ and

$$\delta_0 \in (0, 8^{-1}\varepsilon) \tag{4.190}$$

such that for each $g \in \mathcal{U}_1$, each $T \in [L, \infty)$, and each a.c. function $v\colon [0,T] \to R^n$ which satisfies

$$\begin{aligned} &d(v(0), H(f)) \le \delta_0, \qquad d(v(T), H(f)) \le \delta_0, \\ &I^g(0,T,v) - \pi^f(v(0)) + \pi^f(v(T)) \le \ell(g, D_f + 4, T) + \delta_0 \end{aligned} \tag{4.191}$$

the relation

$$\operatorname{dist}(H(f), \{v(t)\colon\ t \in [S, S+L]\}) \le \varepsilon \tag{4.192}$$

holds for all $S \in [0, T-L]$.

By Lemma 4.8 there exists a neighborhood $\mathcal{U}_2$ of f in $\mathfrak{A}$ such that for each $g \in \mathcal{U}_2$, each $h \in H(f)$, and each $T \ge 1$ there exists an a.c. function $v\colon [0,T] \to R^n$ for which

(4.193)

$$v(0) = h, \quad v(T) \in H(f), \quad I^g(0,T,v) - \pi^f(v(0)) + \pi^f(v(T)) \le \ell(g, D_f + 4, T) + 8^{-1}\delta_0.$$

Clearly there exists

$$\delta \in (0, 2^{-6}\delta_0) \tag{4.194}$$

such that, for each $x, y \in R^n$ satisfying $|x - y| \le \delta$ and $|x|, |y| \le D_f + 4$,

$$|\pi^f(x) - \pi^f(y)| \le 2^{-4}\delta_0. \tag{4.195}$$

By Theorem 2.1 there exists a neighborhood $\mathcal{U}_3$ of f in $\mathfrak{A}$ such that for each $g \in \mathcal{U}_3$ and each $x \in R^n$ satisfying $|x| \le D_f + 4$

$$|\pi^f(x) - \pi^g(x)| \le 2^{-4}\delta_0. \tag{4.196}$$

By Proposition 3.13 there exist an integer $L_0 \ge 1$ and a neighborhood $\mathcal{U}_4$ of f in $\mathfrak{A}$ such that for each $g \in \mathcal{U}_4$ and each (g)-good function $v\colon [0, \infty) \to R^n$

$$\operatorname{dist}(H(f),\ \{v(t)\colon\ t \in [T, T + L_0]\}) \le \delta \tag{4.197}$$

for all large T.

We may assume that $L_0 \ge L$. Set

$$\mathcal{U} = \bigcap_{i=1}^{4} \mathcal{U}_i.$$

Assume that

$$g \in \mathcal{U}, \quad v \in \mathcal{A}(g), \quad d(v(0), H(f)) \le \delta. \tag{4.198}$$

It follows from (4.198) and Proposition 3.14 that v is a (g)-good function. By the definition of $\mathcal{U}_4$ and L_0 there exists a number $T_0 > 0$ such that (4.197) holds for each $T \ge T_0$.

Let $T \geq T_0 + L$. There exists

$$\tau \in [T + L_0, T + 2L_0] \quad \text{such that } |v(\tau) - h_f| \leq \delta \tag{4.199}$$

(recall h_f in (4.176)). We will show that

$$I^g(0, \tau, v) - \pi^f(v(0)) + \pi^f(v(\tau)) \leq \ell(g, D_f + 4, \tau) + \delta_0. \tag{4.200}$$

It follows from (4.198), (4.199), and the definition of $\mathcal{U}_3$ (see (4.196)) that

$$\begin{aligned} \ell(g, D_f + 4, \tau) &\leq I^g(0, \tau, v) - \pi^f(v(0)) + \pi^f(v(\tau)) \\ &\leq I^g(0, \tau, v) - \pi^g(v(0)) + \pi^g(v(\tau)) + 2^{-3}\delta_0 \\ &= \tau\mu(g) + 2^{-3}\delta_0. \end{aligned} \tag{4.201}$$

By the definition of $\mathcal{U}_2$ (see (4.193)) there exists an a.c. function $u\colon [0, \tau] \to R^n$ such that

$$u(0), u(\tau) \in H(f), \qquad I^g(0, \tau, u) - \pi^f(u(0)) + \pi^f(u(\tau)) \leq \ell(g, D_f + 4, \tau) + 8^{-1}\delta_0.$$

It follows from these relations, the definition of $\mathcal{U}_3$ (see (4.196)), and the representation formula (1.6) that

$$\ell(g, D_f + 4, \tau) + 8^{-1}\delta_0 \geq I^g(0, \tau, u) - \pi^g(u(0)) + \pi^g(u(\tau)) - 2^{-3}\delta_0 \geq \tau\mu(g) - 2^{-3}\delta_0.$$

Together with (4.201) this implies (4.200). By (4.198)–(4.200) and the definition of $\mathcal{U}_1$, (4.192) holds for all $S \in [0, \tau - L]$. This completes the proof of the theorem.

Proof of Theorem 2.3. Let $\varepsilon \in (0, 1)$ and $K > D_f + 4$. By Theorem 2.2 there exist $\delta \in (0, \varepsilon)$, $L > 0$, and a neighborhood $\mathcal{U}_1$ of f in $\mathfrak{A}$ such that, for each $g \in \mathcal{U}_1$, each $v \in \mathcal{A}(g)$ satisfying $d(v(0), H(f)) \leq \delta$, and each $T \geq 0$,

$$\operatorname{dist}(H(f),\ \{v(t)\colon\ t \in [T, T + L]\}) \leq \varepsilon. \tag{4.202}$$

By Proposition 3.13 there exist an integer $L_0 \geq 1$ and a neighborhood $\mathcal{U}_2$ of f in $\mathfrak{A}$ such that for each $g \in \mathcal{U}_2$ and each (g)-good function $v\colon [0, \infty) \to R^n$

$$\operatorname{dist}(H(f),\ \{v(t)\colon\ t \in [T, T + L_0]\}) \leq 8^{-1}\delta \tag{4.203}$$

for all large T.

By Lemma 4.4 there exists an integer $N \geq 10$ and a neighborhood $\mathcal{U}_3$ of f in $\mathfrak{A}$ such that, for each $g \in \mathcal{U}_3$; each $T_1 \geq 0$, $T_2 \geq T_1 + NL_0$; each a.c. function $v\colon [T_1, T_2] \to R^n$ which satisfies

$$|v(T_i)| \leq K + 8, \quad i = 1, 2, \quad I^g(T_1, T_2, v) \leq U^g(T_1, T_2, v(T_1), v(T_2)) + 4; \tag{4.204}$$

and each $S \in [T_1, T_2 - NL_0]$, there exists an integer $i_0 \in [0, N - 8]$ such that for all $T \in [S + i_0L_0, S + (i_0 + 7)L_0]$

$$\operatorname{dist}(H(f), \{v(t)\colon\ t \in [T, T + L_0]\}) \leq \delta. \tag{4.205}$$

Set

$$\mathcal{U} = \bigcap_{i=1}^{3} \mathcal{U}_i, \qquad Q = NL_0. \tag{4.206}$$

Assume that

$$g \in \mathcal{U}, \quad v \in \mathcal{A}(g), \quad |v(0)| \le K. \tag{4.207}$$

Equation (4.207) and Proposition 3.14 imply that v is a (g)-good function. Therefore by the definition of $\mathcal{U}_2$

$$|v(t)| \le K + 1 \quad \text{for all large } t.$$

It follows from this relation, (4.207), (4.206), and the definition of $\mathcal{U}_3$, N that there exists $\tau \in [0, Q]$ for which $d(v(\tau), H(f)) \le \delta$. By this relation, (4.207), and the definition of $\mathcal{U}_1$, δ, L, relation (4.202) holds for each $T \ge \tau$. This completes the proof of the theorem.

5. Proof of Theorem 2.4.

LEMMA 5.1. *Assume that $f \in \mathfrak{A}, x\colon [0,\infty) \to R^n$ is an (f)-good function and $h \in \Omega(x)$. Then there exists an a.c. function $v\colon R^1 \to \Omega(x)$ such that $v \in \mathcal{B}(f)$, $v(0) = h$.*

Proof. By Proposition 3.1 the function x is bounded. It is easy to see that the following property holds:

(a) for each $\varepsilon > 0$ there exists $T(\varepsilon) > 0$ such that for each $T_1 \ge T(\varepsilon), T_2 > T_1$

$$I^f(T_1, T_2, x) - \pi^f(x(T_1)) + \pi^f(x(T_2)) - (T_2 - T_1)\mu(f) \le \varepsilon.$$

There exists a sequence of numbers $\{T_p\}_{p=0}^{\infty}$ such that

$$T_{p+1} \ge T_p + 1, \quad p = 0, 1, \ldots, \quad x(T_p) \to h \quad \text{as } p \to \infty. \tag{5.1}$$

For every integer $p \ge 1$ we set

$$v_p(t) = x(t + T_p), \qquad t \in [-T_p, \infty). \tag{5.2}$$

By Proposition 3.4, the boundness of x, (5.1), and (5.2) there exists a subsequence $\{v_{p_j}\}_{j=1}^{\infty}$ and an a.c. function $v\colon R^1 \to R^n$ such that for each integer $N \ge 1$

$$v_{p_j}(t) \to v(t) \quad \text{as } j \to \infty \text{ uniformly in } [-N, N],$$

$$I^f(-N, N, v) \le \liminf_{j \to \infty} I^f(-N, N, v_{p_j}). \tag{5.3}$$

Equations (5.1)–(5.3) imply that $v(0) = h$ and $v(t) \in \Omega(x), t \in R^1$. It follows from property (a), (5.3), and (5.2) that $v \in \mathcal{B}(f)$. The lemma is proven.

Propositions 3.1, 3.2, and 3.14 imply the following result.

LEMMA 5.2. *Assume that $f \in \mathfrak{A}$ and $v \in \mathcal{B}(f)$. Then* $\sup\{|v(t)|\colon t \in R^1\} < \infty$.

Assertion (1) of Theorem 2.4 follows from Lemma 5.1. Assertion (2) of Theorem 2.4 follows from Lemma 5.2 and Theorem 2.3. Assertion (3) of Theorem 2.4 follows from assertion (2) and Theorem 2.4.

Lemma 5.1 implies the following result.

PROPOSITION 5.1. *Assume that $f \in \mathfrak{A}$ and there exists a compact set $H(f) \subset R^n$ such that for each $v \in \mathcal{B}(f)$ the following relations hold:*

$$v(t) \in H(f), \qquad t \in R^1,$$

$$\{y \in R^n\colon \text{ there exists a sequence } \{t_i\}_{i=0}^{\infty} \subset [0,\infty) \text{ for which } t_i \to \infty, v(t_i) \to y \text{ as } i \to \infty\} = H(f).$$

Then $\Omega(u) = H(f)$ for each (f)-good function $u\colon [0,\infty) \to R^n$.

6. Examples. Fix a constant $a > 0$ and set $\psi(t) = t (t \in [0, \infty))$. Consider the complete metric space $\mathfrak{A}$ of integrands $f\colon R^n \times R^n \to R^1$ defined in the introduction.

Example 1. Consider an integrand $f(x, u) = |x|^2 + |u|^2, x, u \in R^n$. It is easy to see that $f \in \mathfrak{A}$. We can show (see [25, sect. 14]) that $\Omega(v) = \{0\}$ for every (f)-good function $v\colon [0, \infty) \to R^n$.

Example 2. Fix a number $q > 0$, and consider an integrand $g(x, u) = q|x|^2|x - e|^2 + |u|^2 (x, u \in R^n)$, where $e = (1, 1, \ldots, 1) \in R^n$. It is easy to see that $g \in \mathfrak{A}$ if the constant a is large enough. Clearly the functions $v_1(t) = 0$ and $v_2(t) = e$ $(t \in [0, \infty))$ are (g)-good and g does not have property B.

Acknowledgment. The author thanks A. Leizarowitz and M. Marcus for helpful discussions.

REFERENCES

[1] B.D.O. Anderson and J.B. Moore, *Linear Optimal Control*, Prentice-Hall, Englewood Cliffs, NJ, 1971.
[2] Z. Artstein and A. Leizarowitz, *Tracking periodic signals with overtaking criterion*, IEEE Trans. Automat. Control, AC-30 (1985), pp. 1122–1126.
[3] W.A. Brock and A. Haurie, *On existence of overtaking optimal trajectories over an infinite horizon*, Math. Oper. Res., 1 (1976), pp. 337–346.
[4] D.A. Carlson, *On the existence of sporadically catching-up optimal solutions for infinite horizon optimal control problems*, J. Optim. Theory Appl., 53 (1987), pp. 219–235.
[5] D.A. Carlson, *The existence of catching-up optimal solutions for a class of infinite horizon optimal control problems with time delay*, SIAM J. Control Optim., 28 (1990), pp. 402–422.
[6] D.A. Carlson, A. Haurie, and A. Jabrane, *Existence of overtaking optimal solutions to infinite dimension control problems on unbounded time intervals*, SIAM J. Control Optim., 25 (1987), pp. 1517–1541.
[7] D.A. Carlson, A. Haurie, and A. Leizarowitz, *Infinite Horizon Optimal Control*, Springer-Verlag, Berlin, 1991.
[8] B.D. Coleman, M. Marcus, and V.J. Mizel, *On the thermodynamics of periodic phases*, Arch. Rational Mech. Anal., 117 (1992), pp. 321–347.
[9] D. Gale, *On optimal development in a multisector economy*, Rev. Econom. Stud., 34 (1967), pp. 1–19.
[10] A. Haurie, *Optimal control on an infinite time horizon: The turnpike approach*, J. Math. Econom., 3 (1976), pp. 81–102.
[11] J.L. Kelley, *General Topology*, D. Van Nostrand, New York, 1955.
[12] A. Leizarowitz, *Existence of overtaking optimal trajectories for problems with convex integrands*, Math. Oper. Res., 10 (1985), pp. 450–461.
[13] A. Leizarowitz, *Infinite horizon autonomous systems with unbounded cost*, Appl. Math. Optim., 13 (1985), pp. 19–43.
[14] A. Leizarowitz, *Tracking nonperiodic trajectories with the overtaking criterion*, Appl. Math. Optim., 14 (1986), pp. 155–171.
[15] A. Leizarowitz, *Optimal trajectories of infinite horizon deterministic control systems*, Appl. Math. Optim., 19 (1989), pp. 11–32.
[16] A. Leizarowitz and V.J. Mizel, *One dimensional infinite-horizon variational problems arising in continuum mechanics*, Arch. Rational Mech. Anal., 106 (1989), pp. 161–194.
[17] V.L. Makarov and A.M. Rubinov, *Mathematical Theory of Economic Dynamics and Equilibria*, Nauka, Moscow, 1973 (in Russian). English translation, Springer-Verlag, 1977.
[18] R.T. Rockafellar, *Saddle points of Hamiltonian systems in convex problem of Lagrange*, J. Optim. Theory Appl., 12 (1973), pp. 367–389.
[19] A.M. Rubinov, *Economic dynamics*, in Itogy Nauki, Sovremennye problemy mat. 19, VINITI, Moscow, 1982, pp. 59–110. English translation in J. Soviet Math., 26 (1984).
[20] C.C. von Weizsacker, *Existence of optimal programs of accumulation for an infinite horizon*, Rev. Econom. Studies, 32 (1965), pp. 85–104.
[21] A.J. Zaslavski, *Optimal programs on infinite horizon*, SIAM J. Control Optim., 33 (1995), pp. 1643–1660.
[22] A.J. Zaslavski, *Optimal programs on infinite horizon* 2, SIAM J. Control Optim., 33 (1995), pp. 1661–1686.

[23] A.J. Zaslavski, *The existence of periodic minimal energy configurations for one-dimensional infinite horizon variational problems arising in continuum mechanics*, J. Math. Anal. Appl., 194 (1995), pp. 459–476.

[24] A.J. Zaslavski, *The existence and structure of extremals for a class of second order infinite horizon variational problems*, J. Math. Anal. Appl., 194 (1995), pp. 660–696.

[25] A.J. Zaslavski, *Dynamics properties of optimal solutions of variational problems*, Nonlinear Anal., 27 (1996), pp. 895–931.

SIAM J. CONTROL OPTIM.
Vol. 35, No. 4, pp. 1204–1216, July 1997

006

AN ABSTRACT BANG-BANG PRINCIPLE AND TIME-OPTIMAL BOUNDARY CONTROL OF THE HEAT EQUATION*

VICTOR J. MIZEL† AND THOMAS I. SEIDMAN‡

Abstract. A principal technical result of this paper is that the one-dimensional heat equation with boundary control is exactly null-controllable with control restricted to an arbitrary set $\mathcal{E} \subset [0,T]$ of positive measure. A general abstract argument is presented to show that, in contrast to previous results, this implies the bang-bang property for time-optimal controls—i.e., such a control can take only extreme values of (the hull of) the constraint set—without imposing any condition regarding the target state.

Key words. partial differential equation, bang-bang control, nullcontrol, reachability

AMS subject classifications. 49K20, 49K30

PII. S0363012996265470

1. Introduction. Our principal concern will be with the bang-bang property for time-optimal boundary control of the one-dimensional heat equation

$$\begin{aligned} &u_t = u_{xx} \qquad (0 < t < T,\ 0 < x < 1),\\ &u(\cdot,0) = \varphi = \text{ control}, \qquad u(\cdot,1) = 0,\\ &u(0,\cdot) = \omega_0 \in \mathcal{X}_0 = L^2(0,1). \end{aligned} \tag{1.1}$$

(Here we will assume a pointwise control constraint and then will say that φ has the bang-bang property if it takes only extremal values.) We note at this point (cf. Remark 4.1) that the previously known results [12], [8] on the bang-bang property for time-optimal control of (1.1) are incomplete in that their hypotheses impose conditions on the target state which turn out to be extraneous for the bang-bang property per se; our focal goal will be the removal of such restrictions.

To this end, we will introduce an abstract formulation of the problem, following [14] in spirit if not quite in detail, and will prove a general abstract result (Theorem 2) which, in the context of (1.1), reduces the problem to a question of independent interest: *exact nullcontrollability for* (1.1) *when* φ *is restricted to* $L^\infty(\mathcal{E})$ *for an arbitrary set* $\mathcal{E}$ *of positive measure in* $[0,T]$. The argument for the latter (Theorem 5) is based on a recent result in nonharmonic analysis by Borwein and Erdélyi [2], [3].

We begin with the observation that there are really two quite distinct versions of the time-optimality problem in control theory:

- immediately initiate control so as to reach the goal as early as possible;
- reach the goal by a fixed time T while delaying initiation of active control to as late as possible.

The first of these is usually taken as the standard statement of the problem, but, much as in [14], it will be more convenient here to use the second version for our

*Received by the editors January 22, 1996; accepted for publication (in revised form) April 30, 1996.

http://www.siam.org/journals/sicon/35-4/26547.html

†Department of Mathematics, Carnegie Mellon University, Pittsburgh, PA 15217 (vm09+@andrew.cmu.edu). The research of this author was supported in part by National Science Foundation grants DMS 9201221 and DMS 9500915.

‡Department of Mathematics and Statistics, University of Maryland–Baltimore County, Baltimore, MD 21250 (seidman@math.umbc.edu).

abstract formulation. We do observe that the two versions are clearly equivalent when the problem is autonomous (with suitable initial conditions). It is also worth emphasizing that (still when the problem is autonomous) the construction used in the proof of the abstract bang-bang principle, Theorem 2 below, could be used equally well directly for a proof of the bang-bang property in the context of version 1, with no restriction on the initial data.

In the context of version 2, we adjoin to (1.1) the target condition that the profile at time T belongs to a prescribed set

$$u(T,\cdot) \in \mathcal{S}_T \tag{1.2}$$

and formulate the time-optimality problem as finding a pair (φ, τ) which maximizes τ subject to the *admissibility constraints* that, for a given set-valued function $A : [0,T] \to 2^{\mathbb{R}}$, one has

$$\begin{aligned} &\text{(i)} \quad \varphi(t) = \varphi_*(t) = \text{given} \qquad \text{for } 0 \le t < \tau, \\ &\text{(ii)} \quad \varphi(t) \in A(t) \qquad \text{for } \tau \le t \le T, \\ &\text{(iii)} \quad \text{the solution } u \text{ of (1.1) using this control } \varphi \text{ satisfies (1.2).} \end{aligned} \tag{1.3}$$

We interpret φ_* as a trivial or passive control (e.g., $\varphi_* \equiv 0$), so τ represents the time at which we initiate active control and maximizing τ is just minimizing the duration $(T-\tau)$ of the actively controlled interval. (If $\omega_0 = 0$, $\varphi_* \equiv 0$, then (1.1) gives $u(\tau,\cdot) = 0$. So, if $A(t)$ were also independent of t, we could translate $[\tau, T]$ to $[0, T-\tau]$ to get the usual earliest arrival (version 1) for this autonomous problem.) Concerning the set function $A(\cdot)$, we assume that

$$\begin{gathered} a(t), b(t) \in A(t) \text{ and } a(\cdot), b(\cdot) \in L^\infty(0,T) \\ \text{for } a(t) := \min\{A(t)\}, \quad b(t) := \max\{A(t)\}. \end{gathered} \tag{1.4}$$

We emphasize that to obtain the bang-bang property *we need impose no hypotheses whatsoever on the data* $\{\omega_0, T, \mathcal{S}_T, \varphi_*\}$ *beyond the implicit assumption that such a time-optimal control does exist.* We do note that the usual argument gives existence of an optimizer in the setting above with $\mathcal{S}_T = \{\omega_T\}$, provided only that the data are *compatible* (i.e., there is some control satisfying (1.3)), and this reachability is the *only* restriction to be imposed regarding the target state ω_T in $\mathcal{X}_T = L^2(0,1)$.

For (1.1), (1.2) with $\mathcal{S}_T = \{\omega_T\}$, (1.3ii) with (1.4), we will show that *pointwise almost everywhere (a.e.) on* $[\tau, T]$*, the values of any time-optimal control* φ *must be either* $\varphi(t) = a(t)$ *or* $\varphi(t) = b(t)$*, with* a, b *as in* (1.4) and that this time-optimal control φ is unique. Note that in this situation the control φ is a *scalar* function of t and this strongly affects the ease with which we can use Theorems 1 and 2 to obtain such a bang-bang property. We will, however, comment in section 4 on the related situation in which one has control at both ends of the interval so

$$u\Big|_{x=0} = \varphi_1, \qquad u\Big|_{x=1} = \varphi_2, \tag{1.5}$$

and the control $\varphi := (\varphi_1, \varphi_2)$ is then an $\mathbb{R}^2$-valued function on $[0,T]$.

2. Evolutionary abstract control systems. We begin by recalling from [14], in slightly modified form, the notion of an *evolutionary abstract control system.* Consider $\mathcal{I}_0 := [0,T]$ as an order category; i.e., writing $\mathcal{I} = \mathcal{I}'\mathcal{I}''$ for $\mathcal{I} = [r,t]$ means

$\mathcal{I}'' = [r,s]$, $\mathcal{I}' = [s,t]$ with $0 \le r \le s \le t \le T$. Let $(\mathcal{X}, \mathbf{E})$ be a functor from $\mathcal{I}_0$ to the category of Banach spaces and continuous linear maps so $\mathcal{I} = [r,t]$ gives $\mathbf{E}_{\mathcal{I}} : \mathcal{X}_r \to \mathcal{X}_t$ with $\mathcal{I} = \mathcal{I}'\mathcal{I}''$ implying $\mathbf{E}_{\mathcal{I}} = \mathbf{E}_{\mathcal{I}'} \circ \mathbf{E}_{\mathcal{I}''}$; i.e.,

$$\mathbf{E}_{r,t} = \mathbf{E}_{s,t}\mathbf{E}_{r,s} : \mathcal{X}_r \to \mathcal{X}_s \to \mathcal{X}_t \quad \text{for } r \le s \le t. \tag{2.1}$$

In general, $\mathbf{E}_{\cdot\cdot}$ represents uncontrolled system evolution for a possibly nonautonomous well-posed problem. Next we associate *control spaces* $\mathcal{U}_{\mathcal{I}} = \mathcal{U}_{r,t}$ with the intervals $\mathcal{I} = [r,t] \subset \mathcal{I}_0$. We always think of each $\mathcal{U}_{\mathcal{I}}$ as a space of functions defined on $\mathcal{I}$, e.g., $\mathcal{U}_{\mathcal{I}} = L^2(\mathcal{I})$, so when $\mathcal{I} = \mathcal{I}'\mathcal{I}''$ we may decompose $\varphi \in \mathcal{U}_{\mathcal{I}}$ into a pair of functions (φ', φ'') defined on $\mathcal{I}'$ and $\mathcal{I}''$, respectively, by restriction maps $\mathbf{\Omega}'$, $\mathbf{\Omega}''$. We then ask that $\varphi' = \mathbf{\Omega}'\varphi \in \mathcal{U}_{\mathcal{I}'}$ and $\varphi'' = \mathbf{\Omega}''\varphi \in \mathcal{U}_{\mathcal{I}''}$. The control maps $\mathbf{C}_{r,t} : \mathcal{U}_{r,t} \to \mathcal{X}_t$ must satisfy the obvious identity

$$\mathbf{C}_{r,t} = \mathbf{C}_{s,t}\mathbf{\Omega}'_s + \mathbf{E}_{s,t}\mathbf{C}_{r,s}\mathbf{\Omega}''_s \tag{2.2}$$

for any such decomposition (any choice of $s \in [r,t]$).

Given any Banach space $\mathcal{V}$ with an injection $\mathbf{I}_{\mathcal{V}} : \mathcal{V} \to \mathcal{U}_{\mathcal{I}}$ — so we may think of $\mathcal{V}$ as consisting of functions with support in (some specified subset of) $\mathcal{I} = [t,T]$ — we say that $\mathcal{V}$ has the *nullcontrollability property* and write $\mathcal{V} \in \mathcal{NC}_{t,T}$ if there is a nullcontrol $v \in \mathcal{V}$ for each initial state in $\mathcal{X}_t$; i.e., if

$$\begin{gathered} \textit{for each } x = x_t \in \mathcal{X}_t \ \textit{ there is some } v \in \mathcal{V} \textit{ such that} \\ \mathbf{E}_{t,T}x + \mathbf{C}_{\mathcal{V}}v = 0 \quad (\mathbf{C}_{\mathcal{V}} := \mathbf{C}_{t,T}\mathbf{I}_{\mathcal{V}})\,. \end{gathered} \tag{2.3}$$

(Clearly, if $\mathcal{V} \in \mathcal{NC}_{t,T}$ for some $\mathcal{V} \hookrightarrow \mathcal{U}_{\mathcal{I}}$, then $\mathcal{U}_{\mathcal{I}} \in \mathcal{NC}_{t,T}$. We recall Theorem 1 of [14]: If $\mathcal{U}_{t,T} \in \mathcal{NC}_{t,T}$, then $\mathcal{K}_s = \mathcal{K}_t^0$ for all $s \le t$, where $\mathcal{K}_s := \mathcal{R}(\mathbf{E}_{s,T}) + \mathcal{R}(\mathbf{C}_{s,T})$ and $\mathcal{K}_s^0 := \mathcal{R}(\mathbf{C}_{s,T})$.) Slightly more delicate than (2.3), but useful later, is the *restricted nullcontrollability property*: we write $\mathcal{V} \in \mathcal{NC}_{t,T}^r$ if

$$\begin{gathered} \textit{for each } x \in \overline{\mathcal{R}(\mathbf{C}_{0,t})} \subset \mathcal{X}_t \ \textit{ there is some } v \in \mathcal{V} \textit{ such that} \\ \mathbf{E}_{t,T}x + \mathbf{C}_{\mathcal{V}}v = 0; \end{gathered} \tag{2.4}$$

i.e., we are restricting initial states in (2.3) to $\overline{\mathcal{R}(\mathbf{C}_{0,t})}$.

We will need the following result, which we present in full although the argument is already known in somewhat different contexts.

THEOREM 1. *If $\mathcal{V} \in \mathcal{NC}_{t,T}$ (respectively, $\mathcal{V} \in \mathcal{NC}_{t,T}^r$), then there is a constant $K_{\mathcal{V}}$ such that v in* (2.3) (*respectively,* (2.4)) *may be chosen with $\|v\|_{\mathcal{V}} \le K\|x\|_{\mathcal{X}_t}$ for any $K > K_{\mathcal{V}}$. Dually, if $\mathcal{V} \in \mathcal{NC}_{t,T}$, one has*

$$\|\mathbf{E}_{t,T}^*\,\xi\|_{\mathcal{X}_t^*} \le K_{\mathcal{V}}\|\mathbf{C}_{\mathcal{V}}^*\,\xi\| \tag{2.5}$$

for $\xi \in \mathcal{X}_T^*$ with the $\mathcal{V}^*$-norm on the right. Conversely, if $\mathcal{V}$ contains a dual space $(\mathcal{W}^* \subset \mathcal{V})$ and $\mathbf{C}_{\mathcal{V}}^*\,\xi \in \mathcal{W}$ for a dense set $\mathcal{D}$ of $\xi \in \mathcal{X}_T^*$, then (2.5)*, using the $\mathcal{W}$-norm on the right for $\xi \in \mathcal{D}$, implies that $\mathcal{V} \in \mathcal{NC}_{t,T}$.*

Proof. For brevity we now simply write $\mathbf{E}$ for $\mathbf{E}_{t,T}$ and $\mathbf{C}$ for $\mathbf{C}_{\mathcal{V}} := \mathbf{C}_{t,T}\mathbf{I}_{\mathcal{V}}$. Clearly, $\mathcal{V} \in \mathcal{NC}_{t,T}$ is equivalent to range containment:

$$\mathcal{R}(\mathbf{E}) \subset \mathcal{R}(\mathbf{C}) =: \mathcal{K}^0(\mathcal{V}) \subset \mathcal{X}_T.$$

Set $\hat{\mathcal{V}} := \mathcal{V}/\mathcal{N}(\mathbf{C})$ with an injective induced map $\hat{\mathbf{C}} : \hat{\mathcal{V}} \to \mathcal{X}_T$ (i.e., $\hat{\mathbf{C}}v = \mathbf{C}_{\mathcal{V}}\hat{v}$ for $v \in \hat{v} \in \hat{\mathcal{V}}$), and let

$$\Gamma := \{(x, \hat{v}) : \mathbf{E}x + \hat{\mathbf{C}}\hat{v} = 0\} \ \subset \ \mathcal{X}_t \times \hat{\mathcal{V}}.$$

Note that Γ is a subspace and is the graph of a linear map $\mathbf{L} = \mathbf{L}_{\mathcal{V}} : \mathcal{X}_t \to \hat{\mathcal{V}}$ which is well defined on all of $\mathcal{X}_t$ by (2.3) and the injectivity of $\hat{\mathbf{C}}$. Since Γ is closed (as $\mathbf{E}, \hat{\mathbf{C}}$ are continuous), it follows that $\mathbf{L}_{\mathcal{V}}$ is bounded, by the closed graph theorem, and the bound on $\|v\|$ with $K_{\mathcal{V}} = \|\mathbf{L}\|$ follows from the definition of the quotient space norm on $\hat{\mathcal{V}}$. Simply replacing $\mathcal{X}_t$ by $\overline{\mathcal{R}(\mathbf{C}_{0,t})}$ in the argument above now gives the bound when $\mathcal{V} \in \mathit{NC}^r_{t,T}$. To obtain (2.5) when $\mathcal{V} \in \mathit{NC}_{t,T}$, we note that the construction of $\mathbf{L}$ gives

$$\mathbf{E} = -\hat{\mathbf{C}}\mathbf{L} \quad \text{so, dually,} \quad \mathbf{E}^* = -\mathbf{L}^*\hat{\mathbf{C}}^*$$

with $\mathbf{L}^* : \hat{\mathcal{V}}^* \to {\mathcal{X}_t}^*$. We then have $\|\mathbf{L}^*\| = \|\mathbf{L}\| =: K_{\mathcal{V}}$ and, since $\langle \hat{\mathbf{C}}^*\xi, \hat{v}\rangle = \langle \mathbf{C}^*\xi, v\rangle$ for $v \in \hat{v} \in \hat{\mathcal{V}}$ and $\xi \in {\mathcal{X}_T}^*$, this gives (2.5).

For the converse, consider any $\eta \in \mathbf{C}^*\mathcal{D}$, i.e., $\eta = \mathbf{C}^*\,\xi$ for some $\xi \in \mathcal{D} \subset \mathcal{X}_T^*$, we can set $\zeta := -\mathbf{E}^*\xi$, noting that if ξ is nonunique (so also $\eta = \mathbf{C}\xi'$ with $\xi' \in \mathcal{D}$), then (2.5) ensures that $\|\mathbf{E}^*(\xi - \xi')\| = 0$, so ζ is well defined. Now, arbitrarily fixing $x = x_t \in \mathcal{X}_t$, we consider

$$\Phi : \eta \longmapsto \langle x, \zeta\rangle = -\langle x, \mathbf{E}^*\xi\rangle$$

for such η. It is clear that the functional Φ is linear on $\mathbf{C}^*\mathcal{D} \subset \mathcal{W}$ and that

$$|\langle \Phi, \eta\rangle| = |\langle x, \zeta\rangle| \le \|x\|\|\zeta\| \le \|x\|K\|\eta\|.$$

Thus, Φ extends by continuity to the $\mathcal{W}$-closure $\overline{\mathbf{C}^*\mathcal{D}}$ and then, by the Hahn–Banach theorem, to a linear functional v on $\mathcal{W}$ (i.e., $v \in \mathcal{W}^* \subset \mathcal{V}$) without increase of norm, so $\|v\| \le K_{\mathcal{V}}\|x\|$. Since

$$\langle \mathbf{C}v, \xi\rangle = \langle v, \mathbf{C}^*\xi\rangle = -\langle x, \mathbf{E}^*\xi\rangle = \langle -\mathbf{E}x, \xi\rangle$$

for ξ dense in ${\mathcal{X}_T}^*$, it follows that $\mathbf{E}x + \mathbf{C}v = 0$ and $v \in \mathcal{V}$ is a nullcontrol for x. As $x \in \mathcal{X}_t$ was arbitrary, we have (2.3), so $\mathcal{V} \in \mathit{NC}_{t,T}$ as asserted. ∎

3. Time-optimality. We now turn to formulation of the abstract time-optimality problem. It will be convenient here to abuse notation slightly by thinking of $\mathcal{U} = \mathcal{U}_{0,T}$ as the common domain of the control maps $\mathbf{C}_{s,t} : \mathcal{U} \to \mathcal{X}_t$, omitting explicit indication of the $\mathbf{\Omega}$ operators; note that we think of $\mathbf{C}_{s,t}\varphi$ as depending only on the part of φ between s and t, so $\mathcal{N}(\mathbf{C}_{s,t}) \supset \mathcal{N}(\mathbf{\Omega}_{[s,t]})$, where, in the obvious notation,

$$\mathbf{\Omega}_{[s,t]} := \mathbf{\Omega}'_{s,[0,t]}\mathbf{\Omega}''_{t,[0,T]} = \mathbf{\Omega}''_{t,[s,T]}\mathbf{\Omega}'_{s,[0,T]}.$$

Fixing the *passive control* $\varphi_* \in \mathcal{U}$, a basic assumption is that for each $s \in (0,T)$ we have

$$\varphi \in \mathcal{U} \Rightarrow \mathbf{P}_s\varphi := \begin{cases} \varphi_* & \text{on } [0,s) \\ \varphi & \text{on } [s,T] \end{cases} \in \mathcal{U} \tag{3.1}$$

or, more formally, $\mathbf{\Omega}'_s\mathbf{P}_s\varphi = \mathbf{\Omega}'_s\varphi$ and $\mathbf{\Omega}''_s\mathbf{P}_s\varphi = \mathbf{\Omega}''_s\varphi_*$; note that $\mathbf{P}_s$ will not generally be linear unless $\varphi_* \equiv 0$. We impose the continuity condition that

$$\mathbf{C}_{r,t}\mathbf{P}_s\varphi \to \mathbf{C}_{r,t}\varphi \text{ as } s \searrow r \tag{3.2}$$

for $0 \le r < t \le T$, which just says that changing φ on the vanishingly small interval $[r,s]$ has vanishingly small control effect at any $t > r$.

The data for the time-optimality problem will be

$$x_0 \in \mathcal{X}_0, \qquad \varphi_* \in \mathcal{U}, \qquad \mathcal{A} \subset \mathcal{U}, \qquad \mathcal{S}_T \subset \mathcal{X}_T, \tag{3.3}$$

where x_0 is an *initial state*, φ_* is the passive control, $\mathcal{A}$ is a *constraint set*, and $\mathcal{S}_T$ is the *target set*. We will require, to simplify our statement rather than to restrict $\mathcal{A}$, that $\varphi \in \mathcal{A}$ implies $\mathbf{P}_s\varphi \in \mathcal{A}$ for each s. The set of *admissible pairs* $\mathcal{P} = \mathcal{P}(\mathcal{A}, \mathcal{S}_T; x_0, \varphi_*)$ is then defined as

$$\mathcal{P} := \{(\varphi, \tau) \in \mathcal{A} \times [0,T] : \varphi = \mathbf{P}_\tau\varphi, [\mathbf{E}_{0,T}x_0 + \mathbf{C}_{0,T}\varphi] \in \mathcal{S}_T\}, \tag{3.4}$$

and we say that a control $\bar{\varphi}$ or, more precisely, an admissible pair $(\bar{\varphi}, \bar{\tau}) \in \mathcal{P}$ is *time-optimal* (with respect to these data) if it maximizes τ over $(\varphi, \tau) \in \mathcal{P}$.

We will say that $\varphi \in \mathcal{U}$ is *slack with respect to* $(\mathcal{A}, \mathcal{V})$ (for a Banach space $\mathcal{V}$ with $\mathbf{I}_\mathcal{V} : \mathcal{V} \to \mathcal{U}$) if there is some $\varepsilon > 0$ such that

$$[\varphi + \mathbf{I}_\mathcal{V} v] \in \mathcal{A} \ \textit{ for all } v \in \mathcal{V} \textit{ with } \|v\|_\mathcal{V} < \varepsilon. \tag{3.5}$$

At this point we may state and prove our abstract bang-bang principle.

THEOREM 2. *Suppose that $\varphi \in \mathcal{U}$ is slack with respect to $(\mathcal{A}, \mathcal{V})$ for some $\mathcal{V} \in \mathcal{NC}^r_{t,T}$. Then (φ, τ) with $\tau < t$ cannot be time-optimal with respect to any data set $(\mathcal{A}, \mathcal{S}_T; x_0, \varphi_*)$ involving this $\mathcal{A}$.*

Proof. Note that, while we have written simply $\mathbf{I}_\mathcal{V} : \mathcal{V} \to \mathcal{U}$, the condition $\mathcal{V} \in \mathcal{NC}^r_{t,T}$ includes the implication that $\mathcal{R}(\mathbf{I}_\mathcal{V})$ is actually in $\mathcal{U}_{t,T}$, so for $s \le t$ one has

$$\mathbf{P}_s\varphi_s = \varphi_s \ \text{ for } \varphi_s := \mathbf{P}_s(\varphi + \mathbf{I}_\mathcal{V} v) \qquad (\text{any } v \in \mathcal{V}). \tag{3.6}$$

Now let K^r be as $K_\mathcal{V}$ in Theorem 1 applied to this $\mathcal{V} \in \mathcal{NC}^r_{t,T}$ and let $\varepsilon > 0$ be as in (3.5). In view of (3.2) with $r = \tau < t$, we may choose $s =: \hat{\tau}$ close enough to τ (with $\tau < \hat{\tau} < t$) that

$$\tilde{x} := \mathbf{C}_{\tau,t}\left[\mathbf{P}_{\hat{\tau}}\varphi - \varphi\right] \ \textit{ gives } \ \|\tilde{x}\| < \varepsilon / K^r, \tag{3.7}$$

noting that $\tilde{x} \in \mathcal{R}(\mathbf{C}_{0,t}) \subset \mathcal{X}_t$. By Theorem 1 we may then choose $v \in \mathcal{V}$ such that

$$\mathbf{E}_{t,T}\tilde{x} + \mathbf{C}_{t,T}v = 0 \text{ and } \|v\|_\mathcal{V} < \varepsilon. \tag{3.8}$$

Now set

$$\hat{\varphi} := \mathbf{P}_{\hat{\tau}}\left(\varphi + \mathbf{I}_\mathcal{V} v\right); \quad \text{i.e., } \ \hat{\varphi} = \begin{cases} \varphi_* & \text{on } [0,\tau), \\ \mathbf{P}_{\hat{\tau}}\varphi & \text{on } [\tau,t), \\ \varphi + \mathbf{I}_\mathcal{V} v & \text{on } [t,T]. \end{cases} \tag{3.9}$$

Since $\|v\|_\mathcal{V} < \varepsilon$, we have $[\varphi + \mathbf{I}_\mathcal{V} v] \in \mathcal{A}$ by (3.5) and also $\hat{\varphi} \in \mathcal{A}$; we have $\mathbf{P}_{\hat{\tau}}\hat{\varphi} = \hat{\varphi}$ by (3.6). Using (2.2) twice (splitting $[0,T]$ at τ and at t), we have

$$\begin{aligned} \mathbf{C}_{0,T}\varphi &= \mathbf{E}_{\tau,T}\mathbf{C}_{0,\tau}\varphi_* + \mathbf{C}_{\tau,T}\varphi \\ &\qquad\qquad \text{as } \varphi = \varphi_* \text{ on } [0,\tau) \\ &= \mathbf{E}_{\tau,T}\mathbf{C}_{0,\tau}\varphi_* + \mathbf{E}_{t,T}\mathbf{C}_{\tau,t}\varphi + \mathbf{C}_{t,T}\varphi, \end{aligned} \tag{3.10}$$

and, similarly, we have

$$\mathbf{C}_{0,T}\hat{\varphi} = \mathbf{E}_{\tau,T}\mathbf{C}_{0,\tau}\varphi_* + \mathbf{E}_{t,T}\mathbf{C}_{\tau,t}\mathbf{P}_{\hat{\tau}}\varphi + \mathbf{C}_{t,T}\left[\varphi + \mathbf{I}_\mathcal{V} v\right] \tag{3.11}$$

using (3.9). Comparing (3.11) with (3.10) gives (with $\mathbf{C}_{\mathcal{V}} := \mathbf{C}_{t,T}\mathbf{I}_{\mathcal{V}}$ as before)

$$\begin{aligned} \mathbf{C}_{0,T}\hat{\varphi} - \mathbf{C}_{0,T}\varphi &= \mathbf{E}_{t,T}\mathbf{C}_{\tau,t}\left[\mathbf{P}_{\hat{\tau}}\varphi - \varphi\right] + \mathbf{C}_{\mathcal{V}}v \\ &= \mathbf{E}_{t,T}\tilde{x} + \mathbf{C}_{\mathcal{V}}v = 0 \end{aligned} \tag{3.12}$$

by (3.7) and (3.8).

It follows that $(\hat{\varphi}, \hat{\tau}) \in \mathcal{P} = \mathcal{P}(\mathcal{A}, \mathcal{S}_T; x_0, \varphi_*)$ for any data, which gives $(\varphi, \tau) \in \mathcal{P}$: if $\mathbf{E}_{0,T}x_0 + \mathbf{C}_{0,T}\varphi =: x_T \in \mathcal{S}_T$, then also $\mathbf{E}_{0,T}x_0 + \mathbf{C}_{0,T}\hat{\varphi} = x_T$ for the very same $x_T \in \mathcal{S}_T$. Since $\hat{\tau} > \tau$, it would then be impossible for τ to be maximal and φ could not be a time-optimal control. □

To see why we refer to Theorem 2 as an abstract bang-bang principle, we note our motivating consequence. Observe, first, that in considering scalar controls with a uniform pointwise bound as in (1.4), there is some arbitrariness about the specification of the control space $\mathcal{U}$. We will, somewhat arbitrarily, take $\mathcal{U} := L^p(0,T)$ for some finite $p > 1$ (so, in particular, $\mathcal{U}$ is reflexive) and assume that each of the operators $\mathbf{E}_{s,t}$, $\mathbf{C}_{s,t}$ is continuous for this choice of $\mathcal{U}$.

THEOREM 3. *Consider a time-optimality problem, as above, with $\mathcal{S}_T$ closed and convex in $\mathcal{X}_T$, scalar control (say, $\mathcal{U} = L^p(0,T)$ for some $p \geq 1$), and $\mathcal{A}$ of the form*

$$\mathcal{A} := \{\varphi \in \mathcal{U} : \varphi(t) \in A(t) \text{ a.e. on } [0,T]\} \tag{3.13}$$

with $A(\cdot)$ as in (1.4) *and $\varphi_* \in \mathcal{A}$. Assume*

$$\begin{gathered} \text{for each } t \in (0,T), \text{ each set } \mathcal{E} \subset (t,T) \text{ of positive measure,} \\ \text{one has } L^\infty(\mathcal{E}) \in \mathcal{NC}^r_{t,T}. \end{gathered} \tag{3.14}$$

Then there is a unique time-optimal control $\bar{\varphi}$, and this necessarily has the bang-bang property:

$$[\varphi(t) = a(t) \text{ or } \varphi(t) = b(t)] \quad \text{a.e. on } [\tau, T], \tag{3.15}$$

with a, b as in (1.4).

The key to this is that for (3.15) to fail one must have

$$a(t) + \varepsilon \leq \varphi(t) \leq b(t) - \varepsilon \quad \text{for } t \in \mathcal{E} \tag{3.16}$$

for some $\varepsilon > 0$ and some set $\mathcal{E}$ of positive measure in $[\tau, T]$, perhaps restricting to an intersection. We may assume that this set $\mathcal{E}$ is actually contained in some $[\bar{t}, T]$ with $\bar{t} > \tau$. We do note that the very existence of a time-optimal control is not immediately clear at this point, since we have not even assumed that $A(t)$ should be a closed set.

Proof. We first consider the situation with $\mathcal{A}$ replaced by $\mathcal{A}_*$, where

$$\mathcal{A}_* := \{\varphi \in \mathcal{U} : \varphi(t) \in [a(t), b(t)] =: A_*(t) \text{ a.e. on } [0,T]\}.$$

As $\mathcal{A}_*$ is bounded, closed, and convex (hence weakly compact in $\mathcal{U} = L^p(0,T)$), the usual argument gives existence of a time-optimal control: let (φ_ν) be an optimizing sequence so we may assume $\varphi_\nu \rightharpoonup \bar{\varphi}$ with $\tau_\nu \nearrow \tau$. Noting that $\mathbf{C}_{0,T}\varphi_\nu \rightharpoonup \mathbf{C}_{0,T}\bar{\varphi}$, we must have $\mathbf{E}_{0,T}x_0 + \mathbf{C}_{0,T}\bar{\varphi} \in \mathcal{S}_T$, whence $(\bar{\varphi}, \tau)$ is admissible and so is time-optimal. By (3.14) and Theorem 2, we see that $\bar{\varphi}$ cannot be slack with respect to $(\mathcal{A}_*, \mathcal{V})$ for any $\mathcal{V} = L^\infty(\mathcal{E})$ with $\mathcal{E}$ of positive measure in $(\bar{t}, T)$, $\bar{t} > \tau$. On the other hand, we have already noted that a failure of (3.15) would give (3.16), which would imply

such slackness and give a contradiction. Hence, $\bar{\varphi}$ must satisfy (3.15). So by (1.4) we have $\bar{\varphi} \in \mathcal{A}$, and this pair $(\bar{\varphi}, \tau)$ is also admissible for the original problem. Since the problem using $\mathcal{A}_*$ is a relaxed version of that, $(\bar{\varphi}, \tau)$ must be time-optimal for the original problem.

To see uniqueness, note that if $(\hat{\varphi}, \tau)$ were a different time-optimal pair for the original problem (necessarily with the same τ), then we could set $\tilde{\varphi} := (\bar{\varphi} + \hat{\varphi})/2$. Note also that $(\tilde{\varphi}, \tau)$ is an admissible pair for the problem using $\mathcal{A}_*$, since the system is linear and $\mathcal{A}_*, \mathcal{S}_T$ are convex. Whether or not $\hat{\varphi}$ satisfies (3.15), it is clear that (3.15) cannot hold for $\tilde{\varphi}$ on the (assumed nonnull) set, where $\hat{\varphi} \neq \bar{\varphi}$. As above, we then see that $(\tilde{\varphi}, \tau)$ cannot be time-optimal for that problem, contradicting the assumed maximality of τ. Thus, $\bar{\varphi}$ is the unique optimal control for the original problem. ∎

For the finite-dimensional case (state space $\mathbb{R}^n$) we see that the hypotheses above are easily established for control systems governed by

$$\dot{x} = Ax + \varphi \mathbf{b} \qquad x(0) = x_0. \tag{3.17}$$

COROLLARY 4. *The results of Theorem* 3 *apply to finite-dimensional time-optimality problems of the indicated form for* (3.17)*, provided that* $A(\cdot)\,, \mathbf{b}(\cdot)$ *are real-analytic on* $[0, T]$ *when this is nonautonomous.*

Proof. We need only verify the hypothesis (3.14), and for this it is convenient to take $\mathcal{X}_t := \mathcal{R}(\mathbf{C}_{0,t})$ for $t \in [0, T]$, so, in particular, $\mathbf{C} = \mathbf{C}_{0,T}$ is surjective to $\mathcal{X}_T$. The choice of control space $\mathcal{U}$ is not very significant, and we take, e.g., $\mathcal{U} := L^2(0, T)$. One easily verifies that the adjoint map $\mathbf{C}^*$ is given for $\eta \in \mathcal{X}_T^*$ $(\subset \mathbb{R}^n)$ by $\mathbf{C}^* : \eta \mapsto \langle \mathbf{b}, y \rangle \in L^2(0, T)$, where

$$-\dot{y} = A^* y, \qquad y(T) = \eta. \tag{3.18}$$

The range $\mathcal{R}(\mathbf{C}^*) = \{\langle \mathbf{b}, y \rangle\}$ is then finite dimensional: indeed, as $\mathbf{C}$ is surjective, it follows that $\mathbf{C}^*$ is injective and $\dim \mathcal{R}(\mathbf{C}^*) = \dim \mathcal{X}_T^* = \dim \mathcal{X}_T \leq n$. The analyticity assumptions on $A(\cdot)$, $\mathbf{b}(\cdot)$ ensure that y and $\langle \mathbf{b}, y \rangle$ are real analytic on $[0, T]$. Hence, if $\langle \mathbf{b}, y \rangle = 0$ on any set $\mathcal{E}$ of positive measure, one must have $\langle \mathbf{b}, y \rangle \equiv 0$ on $[0, T]$. Thus, the map $\mathbf{L}_{\mathcal{E}} : \eta \mapsto \langle \mathbf{b}, y \rangle|_{\mathcal{E}} : \mathcal{X}_T^* \to \mathcal{R}(\mathbf{C}) \to \hat{\mathcal{W}}$ (where $\hat{\mathcal{W}}$ consists of the restrictions to $\mathcal{E}$ of functions in $\mathcal{R}(\mathbf{C}^*)$) is injective and so invertible. Since $\hat{\mathcal{W}}$ is finite dimensional, $[\mathbf{L}_{\mathcal{E}}]^{-1}$ is continuous, with $\hat{\mathcal{W}}$ normed as a subspace of $\mathcal{W} := L^1(\mathcal{E})$ (so $\mathcal{V} := L^\infty(0, T)$ is just $\mathcal{W}^*$), and (2.5) holds, giving (3.14) by Theorem 1. The conclusion is now immediate from Theorem 3. ∎

This argument seems new, even for the finite-dimensional case; we do note that it does not seem to be usefully related to the usual characterization of time-optimal controls as in the Pontryagin maximum principle.

4. Boundary control of the heat equation. In this section we return to consideration of (1.1) as an example of the abstract formulation of sections 2 and 3. Our principal new result is exact boundary nullcontrollability from measurable sets—more precisely, that $L^\infty(\mathcal{E}) \in \mathcal{NC}_{t,T}$ for any set $\mathcal{E}$ of positive measure in $[t, T]$. This is just (3.14)—one notes that $\mathcal{NC}_{t,T}$ and $\mathcal{NC}_{t,T}^r$ are equivalent here—so Theorem 3 then gives the desired bang-bang property for time-optimal boundary control of (1.1).

We will take $\mathcal{X}_t = \mathcal{X} := L^2(0, 1)$ for each $t \in [0, T]$ and will, e.g., take $\mathcal{U} = L^2(0, T)$, so $\mathcal{U}_{\mathcal{I}} = L^2(\mathcal{I})$ with the obvious interpretations of the $\mathbf{\Omega}$ operators by restriction. For this autonomous situation one has $\mathbf{E}_{r,t} = \mathbf{S}(t - r)$, where $\mathbf{S}(\cdot)$ is the semigroup on $L^2(0, 1)$ corresponding to (1.1) with homogeneous boundary conditions.

Then $\mathbf{C}_{s,t}$ is the control effect (so $\mathbf{C}_{s,t} : \varphi \mapsto u(t,\cdot)$, where u satisfies the first two lines of (1.1) with $u(s,\cdot) = 0$), and it is standard (cf., e.g., [10]) that each $\mathbf{C}_{s,t}$ is continuous, indeed compact, from $L^2(s,t)$ to $\mathcal{X} = L^2(0,1)$. (We note in passing that there is a well-known explicit representation for this control mapping associated with (1.1), using convolution with a fundamental solution, expressible in terms of a theta function; cf., e.g., [5, p. 171].) The identities (2.1), (2.2) are clear in this context. For this $\mathcal{U}$ there is no difficulty in defining $\mathbf{P}_s$, and the continuity condition (3.2) here follows a fortiori from the stronger fact that $\mathbf{P}_s \to \mathbf{P}_r$ (strongly on $\mathcal{U} = L^2(0,T)$) as $s \to r$.

To compute the adjoint maps $\mathbf{E}^*_{\bar{t},T}$, $\mathbf{C}^*_{\bar{t},T}$ we consider u satisfying (1.1) for $\bar{t} < t \le T$ with $u(\bar{t},\cdot) \equiv 0$ and y satisfying

$$\begin{aligned} &-y_t = y_{xx} \qquad (0 < t < T,\ 0 < x < 1),\\ &y(T,\cdot) = \eta \in \mathcal{X}^*_T = L^2(0,1),\\ &y(\cdot,0) \equiv 0 \equiv y(\cdot,1). \end{aligned} \tag{4.1}$$

A simple computation involving (1.1) with $u(\bar{t},\cdot) = 0$, (4.1), and an integration by parts gives the identity

$$\int_0^1 uy\,dx\Big|_{t=T} = \int_{\bar{t}}^T \varphi\,[y_x(\cdot,0)]\,dt,$$

and, since $u(T,\cdot) = \mathbf{C}_{\bar{t},T}\varphi$ here, this gives

$$\begin{aligned} \mathbf{C}^*_{\bar{t},T} :&\quad \mathcal{X}^*_T \to L^2(\bar{t},T) \subset L^2(0,T)\\ :&\quad \eta \longmapsto \psi := y_x(\cdot,0)\Big|_{[\bar{t},T]}. \end{aligned} \tag{4.2}$$

Even more simply, (4.1) gives

$$\mathbf{E}^*_{\bar{t},T} : \mathcal{X}^*_T \to \mathcal{X}^*_{\bar{t}} = L^2(0,1) : \eta \longmapsto y(\bar{t},\cdot). \tag{4.3}$$

It will be necessary to represent y in terms of the eigenfunctions and eigenvalues

$$\eta_k(x) := \sqrt{2}\sin\sqrt{\lambda_k}x, \qquad \lambda_k := k^2\pi^2, \tag{4.4}$$

so that

$$\eta = \sum_k c_k\eta_k \ \text{ gives } \ \begin{cases} y = \displaystyle\sum_k c_k e^{-\lambda_k(T-t)}\eta_k,\\ \psi = \displaystyle\sum_k \left[\sqrt{2\lambda_k}\,c_k\right] e^{-\lambda_k(T-t)}. \end{cases} \tag{4.5}$$

Our immediate observation is that

$$\eta \in \mathcal{D} := \text{ span}\,\{\eta_k\} \quad \Rightarrow \quad \mathbf{C}^*_{\bar{t},T}\eta = \psi \in \mathcal{M} = \mathcal{M}(\Lambda) := \text{ span}\,\{e^{-\lambda_k(T-t)}\},$$

where $\Lambda := \{\lambda_k : k = 1,2,\ldots\}$ with, looking to a somewhat more general setting, $0 < \lambda_1 < \lambda_2 < \cdots$ such that $\Sigma_k 1/\lambda_k$ is convergent, as is obviously the case here.

Our starting point will be an inequality

$$\|y(0,\cdot)\|_{L^2(0,1)} \le M_{\bar{t}}\|y_x(\cdot,0)\|_{L^2(0,\bar{t})} \tag{4.6}$$

for solutions of (4.1); it is sufficient to consider this only for $\eta \in \mathcal{D}$. We recognize this as (2.5), giving (2.3) by Theorem 1, corresponding to having $\mathcal{U}_{0,\bar{t}} \in \mathit{NC}_{0,\bar{t}}$, replacing T by $\bar{t}$ here. We will take this nullcontrollability as well known, but note that essentially this inequality (with time reversed and an interchange of Dirichlet and Neumann conditions) was the principal result of [11], with the nullcontrollability form given in [4]). From (4.6) with time reversed, one sees clearly the interpretation of (2.5) as asserting *well-posed observability*: predicting the terminal state from (boundary) observations without knowing the initial state.

Our major new resource is an inequality recently obtained by Borwein and Erdélyi; this is Theorem 5.6 of [2], but see also [1], [3].

THEOREM (BE). *Assume* $\Sigma_k 1/\lambda_k < \infty$, *etc. Then, for every* $q > 0$, $s > 0$, $\rho \in (0,1)$, *there is a constant* $c = c_q(s,\rho,\Lambda)$ *such that*

$$\text{for every set } \mathcal{A} \subset [\rho,1] \text{ with } \operatorname{meas} \mathcal{A} \geq s \text{ one has}$$

$$\|p\|_{L^\infty(0,\rho)} \leq c\|p\|_{L^q(\mathcal{A})} \tag{4.7}$$

$$\text{for every polynomial } p \in \mathcal{M}_0 = \mathcal{M}_0(\Lambda) := \{\Sigma_k a_k x^{\lambda_k}\}.$$

For our present purposes, we make the substitution $x = e^{-(T-t)}$ and set $\rho = e^{-(T-\bar{t})}$ so $t \in [0,\bar{t}]$, $[\bar{t},T]$, $\mathcal{E}$ correspond, respectively, to $x \in [e^{-T},\rho] \subset [0,\rho]$, $[\rho,1]$, $\mathcal{A}$ and $\mathcal{M}$ corresponds to $\mathcal{M}_0$. Noting that meas $\mathcal{A} \geq \rho \operatorname{meas} \mathcal{E}$ for $\mathcal{E} \subset [\bar{t},T]$, one easily sees that, specializing to $q = 1$, (4.7) gives just the inequality we will need

$$\|\tilde{\psi}\|_{L^2(0,\bar{t})} \leq \tilde{c}\|\tilde{\psi}\|_{L^1(\mathcal{E})} \qquad \text{for } \tilde{\psi} \in \mathcal{M} \tag{4.8}$$

with $\tilde{c} = \sqrt{\bar{t}}\, c_1(\rho \operatorname{meas} \mathcal{E}, \rho, \Lambda)$ for any set $\mathcal{E}$ of positive measure in $[\bar{t},T]$.

At this point we are in a position to state and prove our second principal result, on exact boundary nullcontrollability of the one-dimensional heat equation from arbitrary sets of positive measure.

THEOREM 5. *Let* $T > 0$ *and suppose that* $\mathcal{E} \subset [0,T]$ *has positive measure. Then there is a constant* K *such that*

for every $\omega_0 \in \mathcal{X} = L^2(0,1)$ *there is a control* φ *such that*

$|\varphi(t)| \leq K\|\omega_0\|_{\mathcal{X}}$ *for* $t \in \mathcal{E}$, $\qquad \varphi(t) = 0$ *for* $t \notin \mathcal{E}$,

and the solution u *of* (1.1), *using* φ, *has* $u(T,\cdot) = 0$.

Proof. This follows directly from the results we have already developed. Choose any $\bar{t} > 0$ such that $\hat{\mathcal{E}} \cap [\bar{t},T]$ has positive measure, set $\mathcal{W} := L^1(\hat{\mathcal{E}})$, and $\mathcal{V} := \mathcal{W}^* = L^\infty(\hat{\mathcal{E}})$. Consider $y(0,\cdot) = \mathbf{E}^*_{0,T}\eta$ and $\tilde{\psi} = \psi = \mathbf{C}^*_{0,T}\eta$ for $\eta \in \mathcal{D}$. Then (4.6) and (4.8) with $\mathcal{E}$ replaced by $\hat{\mathcal{E}}$ give $\|y(0,\cdot)\|_{L^2(0,1)} \leq M_{\bar{t}}\tilde{c}\|\psi\|_{L^1(\hat{\mathcal{E}})}$ or, equivalently,

$$\|\mathbf{E}^*_{0,T}\eta\|_{\mathcal{X}^*} \leq K_{\mathcal{V}}\|\mathbf{C}^*_{\mathcal{V}}\eta\|_{\mathcal{W}},$$

which we recognize as (2.5). The second part of Theorem 1 then gives $\mathcal{V} \in \mathit{NC}_{\bar{t},T}$ which, since $\hat{\mathcal{E}} \subset \mathcal{E}$ so $\mathcal{V} \hookrightarrow L^\infty(\mathcal{E})$, gives precisely the conclusion of the present theorem. ∎

COROLLARY 6. *The results of Theorem* 3 *apply to the time-optimality problem for* (1.1).

Proof. Theorem 5 just gives the hypothesis (3.14) in this context so Theorem 3 applies. ∎

The argument in Theorem 5 establishing that for each $\mathcal{E}$ in $[t,T]$ of positive measure one has $L^\infty(\mathcal{E})$ in NC^r and hence that Theorem 3 applies shows (cf. Theorem V 1.1 of [7]) that the vector measure

$$m : B[0,T] \to \mathcal{X}_T = L^2(0,1) : \mathcal{E} \mapsto C_{0,T}(\chi_{\mathcal{E}})$$

is a *Liapunov measure*; i.e., for each Borel set $\mathcal{F} \subset [0,T]$ of positive measure, the set $\{m(\mathcal{E}) : \mathcal{E} \subset \mathcal{F}\}$ is a convex, weakly compact subset of L^2. The control-theoretic implications of this property of m are discussed in Chapters V and IX of [7].

Remark 4.1. We remark that the bang-bang property for time-optimal controls is classical for the finite-dimensional case but has previously been shown in the context of boundary controls of the heat equation only with the imposition of a slackness condition on the target state: the control constraint has the form $|\varphi| \le M$, where it is to be known that the target is actually reachable (in *some* time) subject to $|\varphi| \le M'$ with the slackness consisting of asking that $M > M'$. Some years ago, when [12] appeared, we felt that this condition might be an artifact of the proof technique, and we attempted to demonstrate the bang-bang property without it, i.e., for arbitrary (reachable) targets. We failed at that time. The gap in our argument was the need for an estimate such as (4.7), and it is the recent availability of the result by Borwein and Erdélyi [1] which has enabled us now to return successfully to the problem at least for one dimension.

It should be noted that a newer proof of the bang-bang property was presented in Krabs's book [8], but this proof also imposes an auxiliary condition on the target state ω_T. The result, Theorem 2.4.13 of [8], is formulated in terms of a moment problem, so some translation is necessary for comparison. He requires that $\mathbf{c} \in W$, where $\mathbf{c} = (c_k)$ is the sequence of Fourier coefficients of the target u_* and the space W is such that this requirement is equivalent to asking that u_* is a limit—in the sense that differences are reachable by controls with L^∞-norm approaching zero—of targets of the special form $\tilde{u}(\varepsilon, \cdot)$ for $\varepsilon > 0$ and $\tilde{u}$ satisfying the equation $\tilde{u}_t = \tilde{u}_{xx}$ with control vanishing on $[T-\varepsilon, T]$. Certainly the special targets then have $\tilde{u}(\varepsilon, x) = 0$ at $x = 0, 1$, so this, in particular, will also be true in the limit, i.e., for the targets to which Krabs's Theorem 2.4.13 would apply. Krabs also provides Theorem 2.4.14, explicitly following ideas of [12], giving the conclusion with essentially the same slackness condition mentioned earlier; this condition certainly implies that $|\tilde{u}(0)| \le M' < M$. Thus, neither of these theorems would apply to use as target, e.g., the trivially reachable state obtained by taking $\varphi \equiv M$ on $[0, T_*]$. In comparison, we emphasize that we have imposed *no* requirement on the target to get the bang-bang property for a time-optimal control except as is implicit in the very existence of such a control.

The paper [12] considers the n-dimensional case (a bounded spatial region $\Omega \subset \mathbb{R}^n$ with control φ on $[0,T] \times \partial\Omega$) subject to a constraint of the form

$$|\varphi(t,x)| \le M \qquad \text{a.e. for } 0 \le t \le T,\ x \in \partial\Omega. \tag{4.9}$$

To use our present approach to prove the strong form of the bang-bang property, that $|\varphi^*| = M$ a.e. on $[0,T^*] \times \partial\Omega$, would require an n-dimensional form of Theorem 5, showing exact nullcontrollability with controls in $L^\infty(\mathcal{E})$, where $\mathcal{E}$ is now an arbitrary subset of positive measure in $[0,T^*] \times \partial\Omega$. This seems well out of reach by currently available ideas; indeed, even the nullcontrollability from a patch ($\mathcal{E} = [0,T^*] \times \mathcal{P}$ with $\mathcal{P} \subset \partial\Omega$ open but small) has only recently been demonstrated ([9], cf. [13]). On the other hand, it seems to be a tractable open problem to show the weaker bang-bang property that $\|\varphi^*(t,\cdot)\|_{L^\infty(\partial\Omega)} = M$ a.e. on $[0,T^*]$ by showing nullcontrollability from $L^\infty(\mathcal{E} \times \partial\Omega)$ with $\mathcal{E}$ of positive measure in $[0,T^*]$ as earlier.

Note that each of the results above obtains the bang-bang property by way of the adjoint characterization: $\varphi = \{M \text{ where } v_x \ge M; -M \text{ where } v_x \le -M\}$ for some solution v of the adjoint problem. A plausible conjecture is that the additional

restriction on the target state might be significant to ensure this characterization (so there might conceivably be examples for which this characterization fails in the absence of some such slackness condition; this could be a subject for future investigation), although we have seen that it is not necessary for the bang-bang property itself.

Remark 4.2. An essentially identical argument works if we replace the heat equation in (1.1) by

$$u_t = (pu_x)_x - qu \tag{4.10}$$

and/or replace the Dirichlet boundary conditions there by some alternative type of boundary control. For this case we let $\{\lambda_k, z_k\}$ be the eigenvalues and eigenfunctions of the Sturm–Liouville operator $\mathbf{A} : z \mapsto -(pz')' + qz$ whose (homogeneous) boundary conditions are those of the new form of boundary control.

Similarly, one could consider the problem with scalar control in the equation itself:

$$u_t = (pu_x)_x - qu + \varphi(t)b \tag{4.11}$$

for some specified $b(\cdot) \in \mathcal{X}$, using homogeneous boundary condition. In this connection one might note Henry's example [6] of a problem with time-optimal control not of bang-bang form, as in (4.10), but effectively considering version 1 of the time-optimality problem with time-dependent constraints, so it does not correspond to the situation we have analyzed.

Remark 4.3. We may consider the problem with a nonscalar control: $\boldsymbol{\varphi} = [\varphi_0, \varphi_1]$ so the boundary conditions in (1.1) are replaced by

$$u(\cdot, 0) = \varphi_1 \qquad u(\cdot, 1) = \varphi_2 \tag{4.12}$$

and in (1.3) we take $A(t) = \mathcal{K} \subset \mathbb{R}^2$, where $\mathcal{K}$ is a closed, bounded, convex set. Here we may distinguish two forms of the bang-bang property:

weak: *a.e. on* $[0, T^*]$ *one has* $\boldsymbol{\varphi}(t) \in \partial\mathcal{K}$,

strong: *a.e. on* $[0, T^*]$ *one has* $\boldsymbol{\varphi}(t)$ *an extreme point of* $\mathcal{K}$.

The weak form is immediate from the previous arguments: if there were $\mathcal{E} \subset [0, T]$ with positive measure for which $\boldsymbol{\varphi}$ remained in the interior, then we could obtain a contradiction as in the proof of Theorem 2, perturbing only the component φ_0 as there. For the strong form one needs a modification of this to avoid the possibility that $\boldsymbol{\varphi}$ might remain interior to some face within $\partial\mathcal{K}$ so that one must consider perturbations with a linear restriction: $\tilde{\varphi}(t) = \hat{\varphi}(t)\mathbf{c}$ for some nonzero $\mathbf{c} \in \mathbb{R}^2$. What would be needed then is the appropriate modification of the inequality (4.6), obtainable along similar lines.

To illustrate the situation, consider first $\mathcal{K} = [0, 1] \times [0, 1]$. Any case where $\boldsymbol{\varphi}$ is weakly optimal but not strongly optimal can be reduced to the following: for a set E of positive measure in $[\tau, T]$ and some $\varepsilon > 0$, one has $\varphi_1(t) \in (\varepsilon, 1-\varepsilon)$ with $\varphi_2(t) \in \{0, 1\}$ for all $t \in \mathcal{E}$. By selecting $\hat{\tau} > \tau$ but close, we can ensure that $\mathcal{E}' = E \cap [\tau, T]$ has positive measure and that the state $\omega'_{\hat{\tau}}$ produced at $\hat{\tau}$ by use of the modified controls $\varphi'_i(t) = \{\varphi_i(t)$ for $t < \tau$; $= 0$ for $t \in [\tau, \hat{\tau}]\}$ differs from the state $\omega_{\hat{\tau}}$ produced by the original $\boldsymbol{\varphi} = (\varphi_1, \varphi_2)$ by less than ε/K^r, i.e., $\|\omega'_{\hat{\tau}} - \omega_{\hat{\tau}}\| < \varepsilon/K^r$. Consequently, by modifying φ'_1 by v supported on $\mathcal{E}'$ and of sup norm$< \varepsilon$, we obtain, as in the proof of Theorem 2, that $(\varphi'_1 + v, \varphi'_2)$ attains the same target ω_T as $\boldsymbol{\varphi}$ yet with τ replaced by

the larger $\hat{\tau}$, contradicting the assumed optimality of τ. On the other hand, if we take $\mathcal{K} = \{(x, y) : x, y \geq 0,\ x + y \leq 1\}$, it is clear that such an argument is only available if one knows that pairs with $\varphi_1 + \varphi_2 = 0$ on $\mathcal{E}$ are available as nullcontrols for the state perturbation. Such "odd" control pairs only produce corresponding odd states and so can only compensate for odd state perturbations. Hence our argument cannot be expected to work in this setting, although we cannot on this basis conclude that the bang-bang property fails.

Similar considerations apply if one would generalize (4.11) to

$$u_t = (pu_x)_x - qu + \Sigma_j \varphi_j(t) b_j \tag{4.13}$$

with pointwise constraints imposed on the vector control $\boldsymbol{\varphi} = [\varphi_1, \ldots, \varphi_J]$.

Remark 4.4. There is little difficulty in generalizing the abstract Theorem 3 to treat state-dependent constraints. It is convenient to take a space $\boldsymbol{\mathcal{X}} = \{x(\cdot)\}$ of "controlled trajectories," where the state trajectory is defined by $x(t) := \mathbf{C}_{0,t}\varphi$ for $t \in [0, T]$, $\varphi \in \mathcal{U}$; we assume the topology imposed on $\boldsymbol{\mathcal{X}}$ is such that the linear map $\mathbf{X} : \varphi \mapsto x(\cdot) : \mathcal{U} \to \boldsymbol{\mathcal{X}}$ is continuous. By a state-dependent constraint we mean a set-valued function

$$(t, x) \longmapsto A(t, x) \subset \mathbb{R} \qquad \text{for } t \in [0, T], x \in \boldsymbol{\mathcal{X}}, \tag{4.14}$$

so the control restriction (1.3ii) becomes

$$\varphi \in \mathcal{A} := \{\varphi \in \mathcal{U} : \varphi(t) \in A(t, \mathbf{X}\varphi) \text{ a.e. on } [0, T]\}. \tag{4.15}$$

We continue to take $\mathcal{U} = L^p(0, T)$ and to assume (1.4), now also writing $a(t) = a(t, x)$, $b(t) = b(t, x)$; we will further assume that one has uniform bounds: $a \leq a(t, x) \leq b(t, x) \leq b$ for all $x \in \boldsymbol{\mathcal{X}}$. Finally, we need a mild continuity condition[1]

$$\begin{aligned} &\varphi_n \rightharpoonup \bar{\varphi} \textit{(weak convergence in } \mathcal{U}\textit{) with } (4.15) \textit{ for each } n \\ &\textit{implies} \qquad a(t, \mathbf{X}\bar{\varphi}) \leq \bar{\varphi}(t) \leq b(t, \mathbf{X}\bar{\varphi}) \textit{ a.e. on } [0, T]. \end{aligned} \tag{4.16}$$

We may then argue much as in the proof of Theorem 3. If φ_n is an optimizing sequence for the time-optimality problem given by (4.15), we have $\varphi_n \rightharpoonup \bar{\varphi}$, using our assumptions on $A(\cdot\cdot)$ and extracting a subsequence if necessary, so we may set $A_*(t) := [a(t, \mathbf{X}\bar{\varphi}), b(t, \mathbf{X}\bar{\varphi})]$ and have $\bar{\varphi}(t) \in A_*(t)$ a.e. on $[0, T]$. As in the proof of Theorem 3, we consider the time-optimality problem using A_* for $\mathcal{A}$ to obtain a (unique) time-optimal control $\hat{\varphi}$. As there, $\hat{\varphi}$ has the bang-bang property and so is also admissible for the original problem, whence the control times τ are the same and we can conclude that $\hat{\varphi} = \bar{\varphi}$ so that this is the unique time-optimal control for the original problem.

Acknowledgments. The first author wishes to express his appreciation to Peter Borwein for informing him of the recent results achieved by Borwein and Erdélyi in [1] and to Greg Knowles for very stimulating discussions on this topic several years ago.

[1]For example, it is not hard to see that (4.16) will hold if one can take $\boldsymbol{\mathcal{X}}$ compact in $C([0, T] \to \mathcal{X})$ and if, with $A(t) = A(t, x(t))$ so $A : [0, T] \times \mathcal{X} \to 2^{\mathbb{R}}$, one has

$$r_n \to \bar{r}, z_n \to \bar{z}, r_n \in A(t, z_n) \qquad \Rightarrow \qquad a(t, \bar{z}) \leq \bar{r} \leq b(t, \bar{z}).$$

REFERENCES

[1] P. Borwein and T. Erdélyi, *Müntz spaces and Remez inequalities*, Bull. Amer. Math. Soc., 32 (1994), pp. 38–42.

[2] P. Borwein and T. Erdélyi, *Generalizations of Müntz's Theorem via a Remez-type Inequality for Müntz Spaces*, Simon Fraser University, Burnaby, BC, 1994, preprint.

[3] P. Borwein and T. Erdélyi, *Polynomials and Polynomial Inequalities*, Springer-Verlag, New York, 1995.

[4] H. O. Fattorini and D. L. Russell, *Exact controllability theorems for linear parabolic equations in one space dimension*, Arch. Rational Mech. Anal., 43 (1971), pp. 271–292.

[5] G. Hellwig, *Partial Differential Equations*, Blaisdell, New York, 1964.

[6] J. Henry, *Un contre-exemple en théorie de la commande en temps minimal des systèmes paraboliques*, C.R. Acad. Sci. Paris Sér. A 289 (1979), pp. 87–89.

[7] I. Kluvanek and G. Knowles, *Vector Measures and Control Systems*, North-Holland Mathematics Studies 20, North-Holland, Amsterdam, 1975.

[8] W. Krabs, *On Moment Theory and Controllability of One-Dimensional Vibrating Systems and Heating Processes*, Lecture Notes in Control and Information Sciences 173, Springer-Verlag, New York, 1992.

[9] G. Lebeau and L. Robbiano, *Contrôle exact de l'équation de la chaleur*, Comm. Partial Differential Equations, 20 (1995), pp. 335–356.

[10] J. L. Lions and E. Magenes, *Non-Homogeneous Boundary Value Problems and Applications*, vol. II, Springer-Verlag, Berlin, 1972.

[11] V. J. Mizel and T. I. Seidman, *Observation and prediction for the heat equation*, I, J. Math. Anal. Appl., 28 (1969), pp. 303–312.

[12] E. J. P. G. Schmidt, *The 'bang-bang' principle for the time-optimal problem in boundary control of the heat equation*, SIAM J. Control Optim., 18 (1980), pp. 101–107.

[13] T. I. Seidman, *Observation and prediction for the heat equation,* IV: *Patch observability and controllability*, SIAM J. Control Optim., 15 (1977), pp. 412–427.

[14] T. I. Seidman, *Time-invariance of the reachable set for linear control problems*, J. Math. Anal. Appl., 72 (1979), pp. 17–20.

SIAM J. CONTROL OPTIM.
Vol. 35, No. 4, pp. 1217–1243, July 1997

007

$\mathcal{H}_\infty$ CONTROL AND ESTIMATION PROBLEMS WITH DELAYED MEASUREMENTS: STATE-SPACE SOLUTIONS*

KRISHAN M. NAGPAL† AND R. RAVI‡

Abstract. Most physical processes exhibit transport delay in the measured output, and it is well known that this can have disastrous effects on system stability and performance if it is not accounted for. In this paper, we give necessary and sufficient conditions for existence of estimators and controllers that achieve the desired $\mathcal{H}_\infty$ performance criterion when such a measurement delay is present. We also give the complete characterization of all controllers and estimators that achieve the desired performance criterion. The necessary and sufficient conditions are easy to check and are given in terms of the familiar pair of algebraic Riccati equations that appear in the nondelay versions of the corresponding $\mathcal{H}_\infty$ problems, along with an additional Riccati differential equation. Explicit state-space formulas for the controllers and estimators are also obtained. They have a linear periodic structure and are easily implementable. To obtain these results, we first obtain state-space results for a "modified" Nehari problem, which may be of independent interest (see Problem 5 in section 2).

Key words. $\mathcal{H}_\infty$ control, $\mathcal{H}_\infty$ estimation, delay systems, optimal control

AMS subject classifications. 93, 49

PII. S0363012994277499

1. Introduction.

The problem of controller and estimator design for finite-dimensional systems with the $\mathcal{H}_\infty$ performance criterion is by now fairly well understood (see, for example, [3] and [8] and the references therein). In several practical situations, one encounters systems whose models are distributed or infinite dimensional. Flexible beams or systems involving delay are examples of such systems. An interesting area of work in control has been the extension of finite-dimensional results to infinite-dimensional systems. The aim of the present paper is to provide state space solutions to several commonly encountered control and estimation problems for finite-dimensional linear systems when there is a delay present in the measurements. Our optimality criterion is the minimization of the $\mathcal{L}_2$-induced norm (or the $\mathcal{H}_\infty$ norm) and the approach is based on time domain techniques.

Some early results obtained for $\mathcal{H}_\infty$ control of certain classes of distributed plants using various operator-theoretic ideas are, for example, [4], [11], [10], [12], and [17]. These approaches are usually based on the commutant lifting methods or the skew Toeplitz theory. The approach we adopt here is time domain in nature and follows such recent approaches as [2], [13], [14], and [15]. Tadmor in [15] presents state-space solutions to the $\mathcal{H}_\infty$ problem for general linear system. However, because of the presence of delay, direct application of his results would involve Riccati equations that are in operator form and to which solutions may not be easy to compute. *Since the underlying system is finite dimensional, we have tried as much as possible to stick to finite-dimensional techniques in the paper and obtain solutions that would be*

*Received by the editors November 21, 1994; accepted for publication (in revised form) May 2, 1996. The results of this paper were presented in K. Nagpal and R. Ravi, *Proceedings of the* 1994 *American Control Conference*, Baltimore, MD.

http://www.siam.org/journals/sicon/35-4/27749.html

†Scientific Systems Co. Inc., 500 W. Cummings Pk., Suite 3000, Woburn, MA 01801 (nagpal@ssci.com). The research of this author was supported in part by the University of Iowa.

‡Control Systems & Electronic Technologies Laboratory, General Electric Research and Development Center, P. O. Box 8, Schenectady, NY 12301 (ravi@crd.ge.com). The research of this author was supported in part by the General Electric Research and Development Center, Schenectady, NY.

similar and comparable with the solutions when there is no delay in the measurements. Consequently, our results show that the extra price one has to pay for tolerating the delay, as compared with the case where there is no delay, is the solution of an *additional* Riccati differential equation defined over the duration of the delay. The approach adopted here uses lifting ideas that have also been used for some sampled data control problems (see, for example, [1]). Some aspects of two of the problems considered here—namely, the problems of state feedback control and output feedback control with measurement delay—have also been addressed in [2] using game theory ideas. The problem of prediction with an $\mathcal{H}_\infty$ criterion has not been previously addressed in the literature. Crucial to the results outlined here are the results of a "modified" Nehari problem which may be of independent interest.

The paper is organized as follows. In section 2 we describe the problems considered in this paper, and the main results are given in section 3. Some preliminary results are presented in section 4, and section 5 contains the proofs of the main results. A summary and some concluding remarks are contained in section 6.

We end this introduction with some remarks on the notation. Let $\mathcal{R}$ denote the set of real numbers, $\mathcal{R}^n$ denote the n-dimensional Euclidean space (identified with $n \times 1$ vectors of real numbers), and $\mathcal{R}^{n\times m}$ be the set of all $n \times m$ real matrices. We will use A' (or v') to denote the transpose of the matrix A (or vector v) and $\rho(A) := \max_i |\lambda_i(A)|$ to denote the spectral radius. A square matrix A is called stable if all its eigenvalues are in the open left half plane. The following norms will be used: $\|\eta\| := (\sum_1^n |\eta_i|^2)^{1/2}$ for $\eta \in \mathcal{R}^n$; $\|z\|_{[a,b]} := (\int_a^b \|z(t)\|^2 dt)^{1/2}$, and $\langle z_1, z_2\rangle_{[a,b]}$ will denote the norm and the inner product in $\mathcal{L}_2^n[a,b]$. Whenever there is no ambiguity about the interval of interest, we ignore the subscript $[a,b]$ in the definition of the norm and the inner product. The map from w to z is denoted by T_{zw}, and its $\mathcal{L}_2$-induced norm is denoted as $\|T_{zw}\|$ and its adjoint as T_{zw}^*. The map T_{zw} is said to be causal if $z(t) = f(w(s)),\ 0 \le s \le t$, and anticausal if $z(t) = f(w(s)),\ s \ge t$. If T_{zw} admits a state-space representation as follows,

$$\begin{aligned} \dot{x} &= A(t)x + B(t)w, \\ z &= C(t)x + D(t)w, \end{aligned}$$

then we will use the following *packed matrix* abbreviation to describe T_{zw}:

$$T_{zw} = \left[\begin{array}{c|c} A & B \\ \hline C & D \end{array}\right].$$

We now briefly describe the notation used for lifting. Let $l_{\mathcal{L}_2[0,h]}$ be the space of sequences where each element is in $\mathcal{L}_2[0,h]$; i.e.,

$$l_{\mathcal{L}_2[0,h]} := \{\{q_i\} : q_i \in \mathcal{L}_2[0,h]\ \forall i \ge 0\},$$

with the norm defined as

$$\|\{q_i\}\| = \left[\sum_0^\infty \|q_i\|^2\right]^{\frac{1}{2}},$$

where $\|q_i\|$ is the $\mathcal{L}_2[0,h]$ norm of q_i. The lifting operation is defined as $W : \mathcal{L}_2[0,\infty) \to l_{\mathcal{L}_2[0,h]}$, where

$$\{q_i\} = Wq, \quad q_i(t) = q(ih+t) \ \text{ for } 0 \le t \le h. \tag{1}$$

The lifting map W can be visualized as breaking up a signal q defined for $t \geq 0$ into an infinite number of pieces where each piece is an identical copy of q restricted to an interval of length h. It is easily seen that W is a one-to-one, invertible isometry between $\mathcal{L}_2[0,\infty)$ and $l_{\mathcal{L}_2[0,h]}$. A map $\Theta : l_{\mathcal{L}_2[0,h]} \to l_{\mathcal{L}_2[0,h]}$ is called causal if $W^{-1}\Theta W : \mathcal{L}_2[0,\infty) \to \mathcal{L}_2[0,\infty)$ is causal.

2. Problem definitions. In this section we describe the problems that are addressed in this paper. The first four concern delay systems and the fifth one is a "modified" Nehari problem. In Problems 1 to 4, we will limit ourselves to finite-dimensional linear time-invariant systems. Because of the time-domain techniques employed in the proofs, most of the results are easily generalizable to linear time-varying systems as well. Some of the issues in Problem 1 have also been discussed in [17], and some aspects of Problems 2 and 4 also appear in [2].

Problem 1 (*the basic delay problem*). Given are a real number $h \geq 0$ (representing delay) and a finite-dimensional linear system G (not necessarily stable) with a realization as follows:

$$G \left\{ \begin{array}{rcl} \dot{x} & = & Ax + Bw,\ t \in [0,\infty),\ x(0) = 0, \\ z & = & Cx. \end{array} \right. \tag{2}$$

The estimate of $z(t)$ denoted by $\hat{z}(t)$ is generated by a causal operator $f(\cdot)$ that has available to it all $w(s)$ with $s \leq t - h$; i.e.,

$$\hat{z}(t) = f(w(s)),\ s \leq t - h.$$

The problem is then one of determining conditions for a causal operator $f(\cdot)$ to exist so that

$$\sup_w \frac{\|z - \hat{z}\|^2}{\|w\|^2} < 1.$$

Moreover, if the above problem is solvable, we would like to characterize all such causal operators $f(\cdot)$ that achieve the above performance bound.

(In the frequency domain terminology, if the system G is stable ($G \in \mathcal{H}_\infty$), the above problem is equivalent to determining conditions for existence of an operator $Q \in \mathcal{H}_\infty$ such that

$$\inf_{Q \in \mathcal{H}_\infty} \|G - Qe^{-hs}\| < 1,$$

and if such a system exists, to obtain the set of all such Q.)

Problem 2 (*full information control problem with delay*). Given are a real number $h \geq 0$ and a finite-dimensional linear time-invariant system G_{fi}:

$$G_{fi} \left\{ \begin{array}{rcl} \dot{x} & = & Ax + B_1 w + B_2 u,\ t \in [0,\infty),\ x(0) = 0, \\ z & = & C_1 x + D_{12} u, \\ y & = & \begin{pmatrix} x \\ w \end{pmatrix}. \end{array} \right. \tag{3}$$

We would like to determine conditions for a controller $f(\cdot)$ to exist such that with

$$u(t) = f(y(s)) \text{ with } s \leq t - h,$$

$\|T_{zw}\| < 1$ and the feedback system is internally stable. Note that the measurements y are available to the controller only after a time delay of h units. Again, we seek a parameterization of all controllers $f(\cdot)$ that achieve $\|T_{zw}\| < 1$.

Problem 3 (*prediction problem*). Given are a real number $h \geq 0$ and a finite-dimensional linear time-invariant system G_p:

$$G_p \begin{cases} \dot{x} &= Ax + B_1 w,\ t \in [0,\infty),\ x(0) = 0, \\ z &= C_1 x, \\ y &= C_2 x + D_{21} w. \end{cases} \tag{4}$$

We would like to determine conditions under which we can construct an estimator for z to achieve a desired $\mathcal{H}_\infty$ bound, given measurements delayed by h time units. In other words we would like to find a function $f(\cdot)$ so that with

$$\hat{z}(t) = f(y(s)),\ s \leq t - h,$$

$\|T_{ew}\| < 1$, where $e := z - \hat{z}$ is the estimation error. We also seek to parameterize the set of all predictors $f(\cdot)$ that achieve $\|T_{ew}\| < 1$.

Problem 4 (*output feedback control problem with delay*). Given are a real number $h \geq 0$ and a finite-dimensional linear time-invariant system G_{of}:

$$G_{of} \begin{cases} \dot{x} &= Ax + B_1 w + B_2 u,\ t \in [0,\infty),\ x(0) = 0, \\ z &= C_1 x + D_{12} u, \\ y &= C_2 x + D_{21} w. \end{cases} \tag{5}$$

We would like to determine conditions for a controller $f(\cdot)$ to exist so that with

$$u(t) = f(y(s)) \text{ with } s \leq t - h,$$

$\|T_{zw}\| < 1$ and the feedback system is internally stable. Again, we seek a parameterization of all $f(\cdot)$ that achieve this performance bound.

Problem 5 (*modified Nehari problem*). Given are the following two finite-dimensional linear (possibly time-varying) *anti-causal systems* $\Sigma_1 : w_1 \to y_1$ and $\Sigma_2 : w_2 \to y_2$:

$$\Sigma_1 \begin{cases} \dot{x}_1 &= F_1 x_1 + G_1 w_1,\ t \in [0,T],\ x_1(T) = 0, \\ y_1 &= H_1 x_1; \end{cases} \tag{6}$$

$$\Sigma_2 \begin{cases} \dot{x}_2 &= F_2 x_2 + G_2 w_2,\ t \in [0,T],\ x_2(T) = 0, \\ y_2 &= H_2 x_2 + w_2. \end{cases} \tag{7}$$

We would like to determine conditions for a *causal function* $f(\cdot)$ to exist so that with

$$\hat{y}_1(t) = f(w(s)) \text{ with } 0 \leq s \leq t$$

the following holds:

$$\sup_{w_1} \frac{\|\Sigma_2(\Sigma_1 w_1 - \hat{y}_1)\|^2}{\|w_1\|^2} < 1.$$

Again, we seek a parameterization of all $f(\cdot)$ that achieve this performance bound.

The last problem described above reduces to the standard (finite-horizon version) Nehari problem when $\Sigma_2 = I$, where I is the identity operator. State-space results for the standard Nehari problem have been obtained for both time-invariant and time-varying systems by various authors (see, for example, [6] and [16]). Problem 5 stated above can be viewed as a generalization of the standard Nehari problem because it allows one to also include a "weighting filter" Σ_2. The results that we obtain can be easily generalized to the infinite-horizon version of the above problem, but we consider the finite-horizon version simply because of its relevance to the first four delay problems presented in this section.

3. Main results. We will first present the results of the modified Nehari problem since the results of the other problems rely quite critically on its solution.

THEOREM 3.1. *Given are two anticausal linear systems Σ_1 and Σ_2 as described in equations* (6) *and* (7). *There exists a causal system $f(\cdot)$ such that with $\hat{y}_1(t) = f(w(s))$, where $0 \le s \le t$,*

$$\sup_{w_1} \frac{\|\Sigma_2(\Sigma_1 w_1 - \hat{y}_1)\|^2}{\|w_1\|^2} < 1$$

if and only if

$$\lambda_{max}(P(t)Q(t)) < 1 \quad \forall t \in [0,T],$$

where λ_{max} denotes the largest eigenvalue and

$$\begin{aligned} -\dot{Q} &= \bar{F}'Q + Q\bar{F} - \bar{H}'\bar{H}, \quad Q(0) = 0, \\ \dot{P}_1 &= F_1P_1 + P_1F_1' - G_1G_1', \quad P_1(T) = 0, \\ \dot{P}_2 &= (F_2 - G_2H_2)P_2 + P_2(F_2 - G_2H_2)' - G_2G_2', \quad P_2(T) = 0, \\ P &= \begin{bmatrix} P_1 & 0 \\ 0 & -P_2 \end{bmatrix}, \end{aligned} \tag{8}$$

where $\bar{F}$ and $\bar{H}$ are defined as

$$\bar{F} = \begin{bmatrix} F_1 & 0 \\ G_2H_1 & F_2 \end{bmatrix}, \qquad \bar{H} = \begin{bmatrix} H_1 & H_2 \end{bmatrix}.$$

Moreover, if $\lambda_{max}(P(t)Q(t)) < 1$ then the set of all causal operators $f : w_1 \to \hat{y}_1$ that achieve the desired bound are given by

$$\begin{aligned} \dot{q} &= -(\tilde{F}' + N(t)Q(t)\tilde{G}_1\tilde{G}_1')q + N(t)Q(t)\tilde{G}_1w_1 - N(t)[\bar{H}' - Q(t)\tilde{G}_2]v, \ q(0) = 0, \\ \hat{y}_1(t) &= -[\bar{H}P(t) - \tilde{G}_2']q - v, \\ \eta(t) &= w_1(t) - \tilde{G}_1'q, \\ v(t) &= \Theta\eta, \ \textit{where } \Theta \textit{ is causal, and } \|\Theta\| < 1, \end{aligned} \tag{9}$$

where

$$\tilde{F} = \begin{bmatrix} F_1 & 0 \\ 0 & F_2 - G_2H_2 \end{bmatrix}, \ \tilde{G}_1 = \begin{bmatrix} G_1 \\ 0 \end{bmatrix}, \ \tilde{G}_2 = \begin{bmatrix} 0 \\ G_2 \end{bmatrix}, \quad N(t) = (I - Q(t)P(t))^{-1}.$$

For convenience in describing the results of the other problems, we will use the following notation to describe the system given in (9):

$$\mathcal{N}^1_{[\Sigma_1,\Sigma_2,T,\Theta]} : w_1 \to \hat{y}_1 := \text{ the corresponding map described by (9)}, \tag{10}$$

$$\mathcal{N}^2_{[\Sigma_1,\Sigma_2,T,\Theta]} : w_1 \to \eta := \text{ the corresponding map described by (9)}. \tag{11}$$

The next theorem gives the result for the basic delay problem or Problem 1 of the previous section. Before presenting the result, we will describe some notation that appears in the statement of the next theorem. Let $X(t)$ be the solution to the Riccati

differential equation defined in (22). Also, let $\Psi(\cdot,\cdot)$ denote the transition matrix of $A + BB'X$, and let $\bar{B}(\tau)$ be defined as follows:

(12)
$$\frac{d\Psi(\tau,s)}{d\tau} = (A + BB'X(\tau))\Psi(\tau,s), \quad \Psi(s,s) = I, \quad \bar{B}(\tau) = e^{A\tau}\Psi(h,\tau)B \text{ for } \tau \in [0,h].$$

Let $S(t)$ be the solution of the following Riccati equation defined over the interval $[0,h]$:

$$-\dot{S} = -(A + BB'X)S - S(A + BB'X)' - SC'CS + BB', \quad S(0) = 0. \tag{13}$$

We now define four operators from $\mathcal{L}_2[0,h]$ to $\mathcal{L}_2[0,h]$, where G_0 and G_2 are causal while G_1 and G_3 are anticausal. With the notation $G_i : f \to g$ for $0 \le i \le 3$, the state-space realizations of these systems are

$$G_0 \left\{ \begin{array}{rcl} \dot{x}_0(\tau) &=& Ax_0(\tau) + Bf(\tau),\ \tau \in [0,h],\ x_0(0) = 0, \\ g(\tau) &=& f(\tau) - B'X(\tau)x_0(\tau); \end{array} \right. \tag{14}$$

$$G_1 \left\{ \begin{array}{rcl} \dot{x}_1(\tau) &=& Ax_1(\tau) + \bar{B}(\tau)f(\tau),\ \tau \in [0,h],\ x_1(h) = 0, \\ g(\tau) &=& -Cx_1(\tau); \end{array} \right. \tag{15}$$

$$G_2 \left\{ \begin{array}{rcl} \dot{x}_2(\tau) &=& (A + BB'X(\tau))x_2(\tau) + Bf(\tau),\ \tau \in [0,h],\ x_2(0) = 0, \\ g(\tau) &=& Cx_2(\tau); \end{array} \right. \tag{16}$$

$$G_3 \left\{ \begin{array}{rcl} \dot{x}_3(\tau) &=& -(A + BB'X(\tau))'x_3 - C'f(\tau),\ \tau \in [0,h],\ x_3(h) = 0, \\ g(\tau) &=& -CS(\tau)x_3(\tau) + f(\tau). \end{array} \right. \tag{17}$$

Let $\{w_i\}, \{r_i\}, \{z_{1i}\}, \{z_{2i}\} \in l_{\mathcal{L}_2[0,h]}$ be defined as follows (recall that W is the lifting operator defined in (1)):

$$\{w_i\} = Ww, \tag{18}$$

$$r_i = G_0 w_i, \quad i \ge 0, \tag{19}$$

$$z_{10} = 0, \quad z_{1i} = G_1 r_{i-1} \text{ for } i \ge 1, \tag{20}$$

$$z_{20} = 0, \quad z_{2i} = G_2 r_i \text{ for } i \ge 1, \tag{21}$$

where the systems G_0, G_1, and G_2 are as defined in (14), (15), and (16). Also define

$$\tilde{F}_d(t) := \begin{bmatrix} A & 0 \\ 0 & -(A + BB'X(t))' - C'CS(t) \end{bmatrix}, \quad \tilde{G}_{1d}(t) := \begin{bmatrix} \bar{B}(t) \\ 0 \end{bmatrix},$$

$$\tilde{G}_{2d} := \begin{bmatrix} 0 \\ -C' \end{bmatrix}, \quad t \in [0,h];$$

$$\bar{F}_d(t) = \begin{bmatrix} A & 0 \\ C'C & -(A + BB'X(t))' \end{bmatrix}, \ \bar{H}_d(t) = -\begin{bmatrix} C & CS(t) \end{bmatrix}, \quad t \in [0,h].$$

$$\begin{array}{rcl} -\dot{Q}_d(t) &=& \bar{F}_d'(t)Q(t) + Q(t)\bar{F}_d(t) - \bar{H}_d'(t)\bar{H}_d(t), \quad t \in [0,h],\ Q_d(0) = 0, \\ \dot{P}_{1d}(t) &=& AP_1(t) + P_1(t)A' - \bar{B}'(t)\bar{B}(t), \quad t \in [0,h],\ P_{1d}(h) = 0, \\ \dot{P}_{2d} &=& -(A' + X(t)BB' + C'CS(t))P_{2d}(t) - P_{2d}(t)(A' + X(t)BB' + C'CS(t))' \\ && -C'C, \quad t \in [0,h],\ P_{2d}(h) = 0, \\ P_d(t) &=& \begin{bmatrix} P_{1d} & 0 \\ 0 & -P_{2d} \end{bmatrix}, \\ N_d(t) &=& (I - Q_d(t)P_d(t))^{-1}, \quad t \in [0,h]. \end{array}$$

We now present the main result for the basic delay problem.

THEOREM 3.2. *Given are a causal linear system G described by equation* (2) *and a positive real number h representing the delay. There exists a causal system $f(\cdot)$ such that with*

$$\hat{z}(t) = f(w(s)), \quad 0 \le s \le t-h,$$

the following holds:

$$\sup_w \frac{\|z-\hat{z}\|^2}{\|w\|^2} < 1$$

if and only if there exists a matrix function $X(t)$ for $t \in [0,h]$ that satisfies the following (Riccati) differential equation:

$$-\dot{X} = A'X + XA + XBB'X + C'C, \quad X(h) = 0. \tag{22}$$

Moreover, if the above condition holds, then the set of all $f : w \to \hat{z}$ that achieve the desired performance is of the form

$$\begin{aligned} \hat{z}(t) &= 0, \quad 0 \le t \le h, \\ \hat{z}(ih+t) &= C\{e^{A(t+h)}x((i-1)h) + e^{At}\textstyle\int_0^t \Psi(h,s)Br_{i-1}(s)ds\} + \hat{\hat{z}}_i(t), \\ &\quad 0 \le t \le h, \ i \ge 1, \end{aligned} \tag{23}$$

where r_i and $\Psi(\cdot,\cdot)$ are defined in the equations (19) *and* (12)*, respectively, and $\hat{\hat{z}}_i \in l_{\mathcal{L}_2[0,h]}$ is obtained as follows:*

$$\begin{aligned} w_{10} &= r_0; \quad w_{1i} = r_i - G_2^*(z_{1i} - \hat{\hat{z}}_i) \ \text{ for } i \ge 1, \\ \dot{q}_i(t) &= -[\tilde{F}_d' + N_d(t)Q_d(t)\tilde{G}_{1d}(t)\tilde{G}_{1d}'(t)]q_i(t) + N_d(t)Q_d(t)\tilde{G}_{1d}(t)w_{1i}(t) \\ &\quad -N_d(t)[\bar{H}_d'(t) - Q_d(t)\tilde{G}_{2d}]v_i(t), \quad q_i(0) = 0, \ t \in [0,h], \ i \ge 0, \\ \hat{y}_{i+1}(t) &= -[\bar{H}_d P_d(t) - \tilde{G}_{2d}']q_i(t) - v_i(t), \quad t \in [0,h], \ i \ge 0, \\ \eta_i(t) &= w_{1i}(t) - \tilde{G}_{1d}'q_i(t), \ t \in [0,h], \quad i \ge 0, \\ \{v_i\} &= \Theta\{\eta_i\}, \ \text{where } \Theta \text{ is causal and } \|\Theta\| < 1, \\ \hat{\hat{z}}_0 &= 0, \quad \hat{\hat{z}}_{i+1}(t) = (G_1G_2^*(z_{1i} - \hat{\hat{z}}_i))(t) + \hat{y}_{i+1}(t), \quad t \in [0,h], \ i \ge 0. \end{aligned} \tag{24}$$

Remark 1. To emphasize the connection between the above result and the modified Nehari problem and for the reader's convenience, the above equations are also summarized below:

$$\begin{aligned} \{w_i\} &= Ww, \\ r_i &= G_0 w_i, \quad i \ge 0, \\ z_{10} &= 0, \quad z_{1i} = G_1 r_{i-1} \text{ for } i \ge 1, \\ w_{10} &= 0, w_{1i} = r_i - G_2^*(z_{1i} - \hat{\hat{z}}_i), \\ \hat{y}_{i+1} &= \mathcal{N}^1_{[G_1,G_3,h,\Theta]}(w_{1i}), \quad i \ge 0, \\ \eta_i &= \mathcal{N}^2_{[G_1,G_3,h,\Theta]} : (w_{1i}), \quad i \ge 0, \\ \hat{\hat{z}}_0 &= 0, \quad \hat{\hat{z}}_{i+1} = \hat{y}_{i+1} + G_1G_2^*(z_{1i} - \hat{\hat{z}}_i), \quad i \ge 0, \\ \hat{z}(ih+t) &= C\{e^{A(t+h)}x((i-1)h) + e^{At}\textstyle\int_0^t \Psi(h,s)Br_{i-1}(s)ds\} + \hat{\hat{z}}_i(t), \\ &\quad 0 \le t \le h, \ i \ge 1, \end{aligned}$$

where the maps $\mathcal{N}^1$ and $\mathcal{N}^2$ are as defined in (10) and (11).

For convenience in describing the results of Problems 2 through 5, we will use the following notation to characterize the set $\Omega : w \to \hat{z}$ of all admissible functions that solve the basic delay problem (Problem 1) described in the Theorem 3.2; i.e.,

(25)
$$\mathcal{D}_{[A,B,C,h]} := \{\Omega : \text{ where } \Omega : w \to \hat{z} \text{ admits a realization of the form given by (23) and (24)}\}.$$

For the full information control problem, we will make the following standard assumptions:

A1: (A, B_1, C_1) is stabilizable and detectable; (A, B_2) is stabilizable;
A2: $D_{12}'[C_1 \ D_{12}] = [0 \ I]$.

THEOREM 3.3. *For the linear system G_{fi} with a realization as described in (3), there exists a controller $f(\cdot)$ so that with $u(t) = f(y(s))$, $0 \le s \le t - h$, the feedback system is internally stable and*

$$\sup_w \frac{\|z\|^2}{\|w\|^2} < 1$$

if and only if the following conditions hold.

(1) *There exists a positive semidefinite matrix $X \ge 0$ that is a stabilizing solution of the following algebraic Riccati equation:*

$$A'X + XA + X(B_1B_1' - B_2B_2')X + C_1'C_1 = 0. \tag{26}$$

(2) *There exists a positive semidefinite matrix function $S_1(t)$ for $t \in [0, h]$ that satisfies the following Riccati equation:*

(27)
$$-\dot{S}_1 = (A + B_1B_1'X)'S_1 + S_1(A + B_1B_1'X) + S_1B_1B_1'S_1 + XB_2B_2'X, \quad S_1(h) = 0.$$

If the above conditions hold, then the set of all controllers that achieve the desired performance can be represented as

$$u(t) = 0, \quad t \in [0, h),$$

$$u(t) = -B_2'Xx_2(t) + (\Omega \ w)(t), \quad t \ge h,$$

where $\Omega \in \mathcal{D}_{[A+B_1B_1'X, B_1, -B_2'X, h]}$ (where this set is as defined in (25)) and x_2 is obtained as follows:

$$\dot{x}_2 = (A + B_1B_1'X)x_2 + B_2u, \quad x_2(0) = 0.$$

Remark 2. Condition 2 of the above theorem can equivalently be written as follows: there exists positive semidefinite matrix function $Q(t)$ defined over $[0, h]$ that satisfies

$$-\dot{Q} = A'Q + QA + QB_1B_1'Q + C_1'C_1, \ Q(h) = X.$$

This is easily verified by noting that $Q(t) := S_1(t) + X$ satisfies the above equation.

Next we turn to the prediction problem, for which we will make the following assumptions:

A3: (A, B_1, C_1) is stabilizable and detectable; (A, C_2) is detectable;
A4: $D_{21}[B_1' \ D_{21}'] = [0 \ I]$.

THEOREM 3.4. *Given the linear system G_p described in* (4)*, there exists a predictor $\hat{z}(t) = f(y(s))$ with $0 \le s \le t-h$ that achieves*

$$\sup_w \frac{\|z-\hat{z}\|^2}{\|w\|^2} < 1$$

if and only if the following conditions hold.

(1) *There exists a positive semidefinite matrix $Y \ge 0$ that is a stabilizing solution of the following algebraic Riccati equation:*

$$AY + YA' + Y(C_1'C_1 - C_2'C_2)Y + B_1B_1' = 0. \tag{28}$$

(2) *There exists a positive semidefinite matrix function $S_2(t)$ for $t \in [0,h]$ that satisfies the following Riccati equation:*

$$-\dot{S}_2 = (A + YC_1'C_1)'S_2 + S_2(A + YC_1'C_1) + S_2YC_2'C_2Y_2S_2 + C_1'C_1, \quad S_2(h) = 0. \tag{29}$$

Let p be defined as an output of the following system:

$$\begin{aligned} \dot{x}_1 &= (A + YC_1'C_1 - YC_2'C_2)x_1 + YC_2'y - YC_1'\hat{z}, \quad x_1(0) = 0, \\ p &= -C_2x_1 + y. \end{aligned}$$

Then the set of all predictors that achieve the desired performance can be represented as

$$\hat{z}(t) = 0, \quad t \in [0,h),$$
$$\hat{z}(t) = C_1x_2(t) - (\Omega\ p)(t), \quad t \ge h,$$

where $\Omega \in \mathcal{D}_{[(A+YC_1'C_1),YC_2',-C_1,h]}$ (where this set is as defined in (25)*) and x_2 is obtained as follows:*

$$\dot{x}_2 = (A + YC_1'C_1)x_2 - YC_1'\hat{z}, \quad x_2(0) = 0.$$

For the output feedback control problem with delay, we will make the following assumptions:

A5: (A, B_1, C_1) and (A, B_2, C_2) are stabilizable and detectable,
A6: $D_{21}[B_1'\ D_{21}'] = [0\ I]$,
A7: $D_{12}'[C_1\ D_{12}] = [0\ I]$.

THEOREM 3.5. *For the linear system G_{of} described in* (5)*, there exists a controller $f(\cdot)$ so that with $u(t) = f(y(s)), 0 \le s \le t-h$, the feedback system is internally stable and*

$$\sup_w \frac{\|z\|^2}{\|w\|^2} < 1$$

if and only if the following conditions hold.

(1) *There exists a positive semidefinite matrix $X \ge 0$ that is a stabilizing solution of the following algebraic Riccati equation:*

$$A'X + XA + X(B_1B_1' - B_2B_2')X + C'C = 0. \tag{30}$$

(2) *There exists a positive semidefinite matrix $Y \ge 0$ that is a stabilizing solution of the following algebraic Riccati equation:*

$$AY + YA' + Y(C_1'C_1 - C_2'C_2)Y + B_1B_1' = 0. \tag{31}$$

(3)

$$\rho(YX) < 1. \tag{32}$$

(4) *With Z defined as $Z := X(I - YX)^{-1}$, there exists a positive semidefinite matrix function $S_3(t)$ for $t \in [0, h]$ that satisfies the following Riccati equation:*

$$\begin{aligned} -\dot{S}_3 &= (A + YC_1'C_1 + YC_2C_2'YZ)'S_3 + S_3(A + YC_1'C_1 + YC_2C_2'YZ) \\ &\quad + S_3YC_2'C_2YS_3 + ZB_2B_2'Z, \quad S_3(h) = 0. \end{aligned} \tag{33}$$

Let r be defined as an output of the following system

$$\begin{aligned} \dot{x}_1 &= (A + YC_1'C_1 - YC_2C_2')x_1 + YC_2'y + B_2u, \quad x_1(0) = 0, \\ r &= -C_2(I + YZ)x_1 + y. \end{aligned}$$

Then the set of all controllers that achieve the desired performance can be represented as

$$u(t) = 0, \quad t \in [0, h),$$

$$u(t) = -B_2'Zx_2(t) + (\Omega r)(t), \quad t \geq h,$$

where $\Omega \in \mathcal{D}_{[(A+YC_1'C_1+YC_2C_2'YZ),YC_2',-B_2'Z,h]}$ (where this set is as defined in (25)*) and x_2 is given by*

$$\dot{x}_2 = (A + YC_1'C_1 + YC_2C_2'YZ)x_2 + B_2u, \quad x_2(0) = 0.$$

Remark 3. Condition 4 of the necessity in the above equation is equivalent to the existence of positive semidefinite matrix function $Q(t)$ defined over $[0, h]$ that satisfies

$$-\dot{Q} = A'Q + QA + QB_1B_1'Q + C_1'C_1, \; Q(h) = X$$

with an additional spectral radius condition of $\rho(YQ(0)) < 1$.

4. Preliminary results. In this section we present some preliminary results that will be used in the proofs of the main results considered here. All the results that are well known in the literature are presented without proofs.

LEMMA 4.1. *Define $P(t)$ to be the solution of*

$$\dot{P} = AP + PA' + BB', \quad P(0) = 0. \tag{34}$$

If (A, B) is controllable, then for the system described in equation (2)

$$\inf_w \left\{ \|w\|^2_{[0,\tau]} : x(0) = 0, x(\tau) = x_\tau \right\} = x_\tau' P(\tau)^{-1} x_\tau. \tag{35}$$

When (A, B) is not controllable, the statement of the above lemma is still valid for all x_τ that are reachable with the given initial condition with $P(\tau)^{-1}$ replaced by the pseudoinverse of $P(\tau)$. The following lemma follows from the standard factorization results for finite horizon linear regulator theory, but for completeness we give a brief proof for it.

LEMMA 4.2. *Let G_2 and G_3 be the systems defined in* (16) *and* (17)*, respectively. Then,*

$$I + G_2G_2^* = G_3^*G_3.$$

Proof. The system $G_2^* : f_a \to g_a$, the adjoint of the system G_2, has a state-space realization

$$\begin{aligned} \dot{x}_{2a}(\tau) &= -(A + BB'X(\tau))'x_{2a}(\tau) - C'f_a(\tau),\ \tau \in [0,h],\ x_{2a}(h) = 0, \\ g_a(\tau) &= B'x_{2a}(\tau). \end{aligned}$$

If S satisfies the differential equation described in (13), then it is easily verified that

$$\frac{d}{dt}(x'_{2a}(t)S(t)x_{2a}(t)) = -x'_{2a}(t)BB'x_{2a}(t) - f'_a(t)f_a + [f_a - CS(t)x_{2a}(t)]'[f_a - CS(t)x_{2a}(t)].$$

Integrating the above from 0 to h and noting that the boundary conditions drop out because of the boundary conditions on S and x_a, one obtains

$$\|f_a\|^2_{[0,h]} + \|g_a\|^2_{[0,h]} = \|f_a - CSx_{2a}\|^2_{[0,h]}.$$

Thus, for any $f_a \in \mathcal{L}_2[0,h]$,

$$\langle (I + G_2G_2^*)f_a, f_a\rangle_{[0,h]} = \|f_a\|^2_{[0,h]} + \|g_a\|^2_{[0,h]} = \|f_a - CSx_{2a}\|^2_{[0,h]} = \|G_3f_a\|^2_{[0,h]} \qquad \square$$

The next lemma is the so-called Redheffer's lemma [3, Lemma 15] and plays a crucial role in several $\mathcal{H}_\infty$ optimization problems. Most of the subsequent results are from [3] (or easily obtained from the results therein), where they are proven for finite-dimensional linear time-invariant systems. However, they can be generalized to infinite-dimensional systems, as for example in [15].

LEMMA 4.3. *Let P be a system partitioned as*

$$P = \begin{bmatrix} P_{11} & P_{12} \\ P_{21} & P_{22} \end{bmatrix},$$

and let Q be another system connected to P as in Figure 1. *Suppose that P has an exponentially stable realization and that Q is any linear operator. In addition, suppose that the closed loop system is also stabilizable and detectable (from w and z, respectively). Let P be isometric (i.e., $\|w\|^2 + \|v\|^2 = \|z\|^2 + \|r\|^2\ \forall v, w \in \mathcal{L}_2$), and let P_{21}^{-1} exist and be stable. Then the closed loop system is exponentially stable and $||T_{zw}|| < 1$ if and only if Q is exponentially stable and $||Q|| < 1$, where T_{zw} is the closed loop input-output operator mapping w to z.*

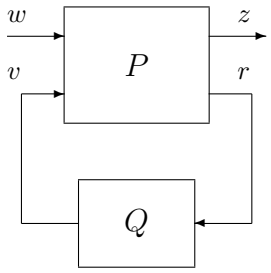
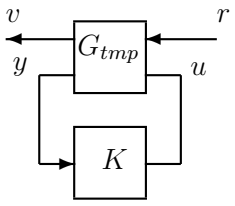


FIG. 1.

The next result introduces a transformation that is by now standard in this field (see for example [3, Lemma 9]).

LEMMA 4.4. *Let assumptions* A1–A2 *hold and suppose that X as defined in* (26) *exists. A controller K stabilizes G_{of} defined in* (5) *and achieves $\|T_{zw}\| < 1$ if and*

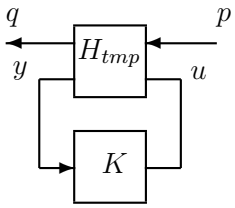


$$G_{tmp} \sim \left[\begin{array}{c|cc} A + B_1 B_1' X & B_1 & B_2 \\ \hline B_2' X & 0 & I \\ C_2 & D_{21} & 0 \end{array}\right]$$

FIG. 2.

only if K stabilizes the system G_{tmp} and achieves $\|T_{vr}\| < 1$ for the system G_{tmp} described in Figure 2.

The next result that we will state requires an additional auxiliary system, H_{tmp}, which we interconnect with a controller K as in Figure 3. This result is the dual of Lemma 9 in [3] or, equivalently, the dual of the lemma stated above.



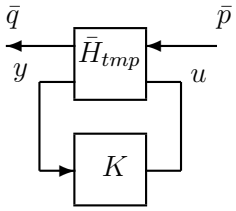
$$H_{tmp} \sim \left[\begin{array}{c|cc} A + Y C_1' C_1 & Y C_2' & B_2 \\ \hline C_1 & 0 & D_{12} \\ C_2 & I & 0 \end{array}\right]$$

FIG. 3.

LEMMA 4.5. *Let assumptions* A5–A7 *hold and suppose that Y as defined in* (31) *exists. A controller K stabilizes the system G_{of} defined in* (5) *and achieves $\|T_{zw}\| < 1$ if and only if K stabilizes H_{tmp} and achieves $\|T_{qp}\| < 1$ for the system H_{tmp}.*

The next result is obtained by combining the previous two lemmas. Even though its proof is straightforward, we provide a brief sketch of it since this transformation is not often used in the literature (a closely related observation appears in the work of Tadmor [15, Proposition 3.5.1]).

LEMMA 4.6. *Let assumptions* A5–A7 *hold. Let X and Y satisfy equations* (30), (31), *and* (32), *and let $Z := X(I - YX)^{-1}$. A controller K internally stabilizes the system G_{of} defined in* (5) *and achieves $\|T_{zw}\| < 1$ if and only if K stabilizes $\bar{H}_{tmp}$ (as defined in Figure* 4) *and achieves $\|T_{\bar{q}\bar{p}}\| < 1$.*



$$\bar{H}_{tmp} \sim \left[\begin{array}{c|cc} A + Y C_1' C_1 + Y C_2' C_2 Y Z & Y C_2' & B_2 \\ \hline B_2' Z & 0 & I \\ C_2(I + YZ) & I & 0 \end{array}\right]$$

FIG. 4.

Proof. It can be easily verified that if X and Y satisfy the given conditions, then $Z := X(I - YX)^{-1}$ is positive semidefinite and satisfies the algebraic Riccati equation

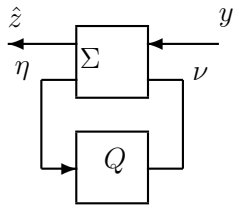
$$(A + Y C_1' C_1)' Z + Z(A + Y C_1' C_1) + Z(Y C_2' C_2 Y - B_2 B_2') Z + C_1' C_1 = 0,$$

where $A + Y C_1' C_1 + (Y C_2' C_2 Y - B_2 B_2') Z$ is stable. From Lemma 4.5, a controller

K solves the $\mathcal{H}_\infty$ control problem for the system G_{of} defined in (5) if and only if K stabilizes H_{tmp} and achieves $\|T_{qp}\|_\infty < 1$. Now treating the system H_{tmp} as the system G_{of} and applying the Lemma 4.4 to it, one obtains the above Lemma 4.6. □

The following result is from [5] and gives a parameterization of all causal estimators that achieve $\|T_{ew}\| < 1$ for the estimation problem.

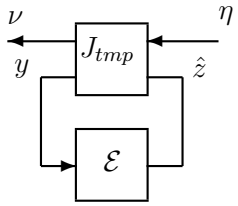
LEMMA 4.7. *Let assumptions* A3, A4 *hold and let* Y *satisfy* (28). *Any causal estimator* $\mathcal{E} : y \to \hat{z}$ *that solves the estimation problem for the system* G_p *defined in* (4) *and achieves* $\|T_{ew}\| < 1$ *(where* $e = z - \hat{z}$*) can be represented as in Figure* 5, *where* Q *is a causal operator with* $\|Q\| < 1$.



$$\Sigma \sim \left[\begin{array}{c|cc} A - YC_2'C_2 & YC_2' & YC_1' \\ \hline C_1 & 0 & -I \\ -C_2 & I & 0 \end{array}\right]$$

FIG. 5.

The following lemma states that a causal estimator for the system G_p defined in (4) achieves $\|T_{ew}\| < 1$ if and only if it achieves $\|T_{\nu\eta}\| < 1$ for the system described in Figure 6. This result follows immediately from the above parameterization of causal estimators.



$$J_{tmp} \sim \left[\begin{array}{c|cc} A + YC_1'C_1 & YC_2' & -YC_1' \\ \hline C_1 & 0 & -I \\ C_2 & I & 0 \end{array}\right]$$

FIG. 6.

LEMMA 4.8. *Let assumptions* A3, A4 *hold and let* Y *satisfy* (28). *A causal estimator* $\mathcal{E} : y \to \hat{z}$ *solves the estimation problem for the system* G_p *defined in* (4) *and achieves* $\|T_{ew}\| < 1$ *if and only if* $\|T_{\nu\eta}\|_\infty < 1$ *for the system described in Figure* 6.

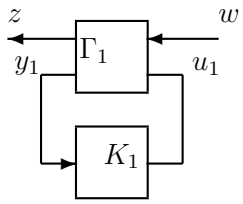
Proof. Let $\hat{x}$ be the state of the system Σ described in the Figure 5. Then the system Σ can be redescribed as

$$\dot{\hat{x}} = (A + YC_1'C_1)\hat{x} + YC_2\eta - YC_1'\hat{z},$$

$$y = C_2\hat{x} + \eta, \quad \hat{z} = \mathcal{E}y, \quad \nu = C_1\hat{x} - \hat{z}.$$

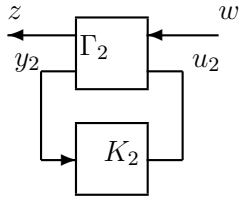
The above is nothing but the description of the system J_{tmp}. Now the conclusion of the above lemma follows immediately from Lemma 4.7 since the estimator $\mathcal{E}$ achieves the desired performance if and only if $\|T_{\nu\eta}\| < 1$. □

The following lemma, a proof of which is contained in [3, Propositions 3 and 4], is fairly straightforward. To state the result, we need to describe the feedback connection of systems and controllers shown below in Figures 7 and 8.



$$\Gamma_1 \sim \left[\begin{array}{c|cc} A & B_1 & B_2 \\ \hline C_1 & 0 & D_{12} \\ 0 & I & 0 \end{array}\right]$$

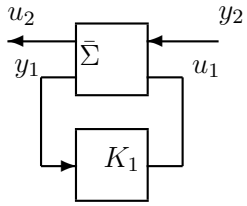
FIG. 7.



$$\Gamma_2 \sim \left[\begin{array}{c|cc} A & B_1 & B_2 \\ \hline C_1 & 0 & D_{12} \\ C_2 & I & 0 \end{array}\right]$$

FIG. 8.

LEMMA 4.9. *Let $A-B_1C_2$ be stable and the controller K_1 be such that it internally stabilizes the feedback system shown in Figure* 7*. If the controller $K_2 : y_2 \to u_2$ can be represented as in Figure* 9*, then the feedback system represented in Figure* 8 *is internally stable. Moreover, if K_2 can be represented as in Figure* 9*, then the closed loop map from w to z for the two feedback systems described in Figures* 7 *and* 8 *is identical.*



$$\bar{\Sigma} \sim \left[\begin{array}{c|cc} A - B_1C_2 & B_1 & B_2 \\ \hline 0 & 0 & I \\ -C_2 & I & 0 \end{array}\right]$$

FIG. 9.

Let $\mathcal{K}_1$ be the set of all controllers that internally stabilize the feedback system shown in Figure 7 *and render $\|T_{zw}\| < 1$. If $A - B_1C_2$ is stable, then any controller K_2 that internally stabilizes the feedback system shown in the Figure* 8 *and renders $\|T_{zw}\| < 1$ can be represented as in Figure* 9*, where $K_1 \in \mathcal{K}_1$.*

5. Proofs.

Proof of Theorem 3.1: *The modified Nehari problem.*

Necessity. Here we show that if there exists a causal system f that achieves the desired performance, then $\lambda_{max}(P(t)Q(t)) < 1 \, \forall t \in [0, T]$. For clarity of exposition, we will make a simplifying assumption that (F_1, G_1) and (F_2, G_2) are controllable (all the subsequent arguments go through if this is not the case by replacing matrix inverses by their pseudoinverses). Let the observability grammian Q defined in (8) be partitioned as

$$Q := \begin{bmatrix} Q_{11} & Q_{12} \\ Q_{12}' & Q_{22} \end{bmatrix}. \tag{36}$$

We will show that if the problem is solvable then with P_1 and P_2 as defined in (8),

$$P_1^{-1}(t) - Q_{11}(t) + Q_{12}(t)(P_2^{-1}(t) + Q_{22}(t))^{-1}Q_{12}'(t) > 0 \quad \forall t \in (0,T). \tag{37}$$

Let x_1 be the state of Σ_1, x_2 be the state of Σ_2, and $\hat{y}_1 = f(w_1)$, where $f(\cdot)$ is a given causal function. Then the system can be written in state-space form as

$$\begin{bmatrix} \dot{x}_1 \\ \dot{x}_2 \end{bmatrix} = \begin{bmatrix} F_1 & 0 \\ G_2H_1 & F_2 \end{bmatrix} \begin{bmatrix} x_1 \\ x_2 \end{bmatrix} + \begin{bmatrix} G_1 \\ 0 \end{bmatrix} w_1 - \begin{bmatrix} 0 \\ G_2 \end{bmatrix} \hat{y}_1, \quad x_1(T) = x_2(T) = 0,$$

$$e = \begin{bmatrix} H_1 & H_2 \end{bmatrix} \begin{bmatrix} x_1 \\ x_2 \end{bmatrix} - \hat{y}_1, \quad \text{where } e := \Sigma_2(\Sigma_1 w_1 - \hat{y}_1).$$

Fix a $\tau \in (0,T)$ and a nonzero vector $x_{1\tau}$. Let w_1 be such that $w_1(s) = 0$ for all $s < \tau$, and for $t \in [\tau, T]$ let w_1 be the minimum norm input such that $x_1(\tau) = x_{1\tau}$ for the system (6). Then from Lemma 4.1 (note that the system here is anticausal),

$$\|w\|_{[0,T]}^2 = \|w\|_{[\tau,T]}^2 = x_{1\tau}' P_1^{-1}(\tau) x_{1\tau}.$$

Since $w_1(s) = 0$ for all $0 < s < \tau$ and $\hat{y}_1$ is causally generated from w_1, $\hat{y}_1(s) = 0$ for all $0 < s < \tau$. Thus for $0 < s < \tau$,

$$\begin{bmatrix} \dot{x}_1 \\ \dot{x}_2 \end{bmatrix} = \begin{bmatrix} F_1 & 0 \\ G_2H_1 & F_2 \end{bmatrix} \begin{bmatrix} x_1 \\ x_2 \end{bmatrix}, \quad e = \begin{bmatrix} H_1 & H_2 \end{bmatrix} \begin{bmatrix} x_1 \\ x_2 \end{bmatrix},$$
$$\text{and } \|e\|_{[0,\tau]}^2 = x'(\tau) Q(\tau) x(\tau),$$

where

$$x := \begin{bmatrix} x_1 \\ x_2 \end{bmatrix}.$$

Since $e := \Sigma_2(\Sigma_1 w_1 - \hat{y}_1)$, the dynamics of system Σ_2 can also be written as (with x_2 as the state of Σ_2)

$$\dot{x}_2 = (F_2 - G_2H_2)x_2 + G_2 e, x_2(T) = 0.$$

Suppose there exists a causal $f(\cdot)$ that achieves the desired performance for the given problem. Then for this $f(\cdot)$ and the given choice of w_1, we have the following series of inequalities, where the fourth step follows from application of Lemma 4.1 to the above equation for x_2:

$$\begin{aligned}
0 &< \|w\|_{[0,T]}^2 - \|e\|_{[0,T]}^2 \\
&= \|w\|_{[\tau,T]}^2 - \|e\|_{[0,\tau]}^2 - \|e\|_{[\tau,T]}^2 \\
&= x_{1\tau}' P_1^{-1}(\tau) x_{1\tau} - x'(\tau) Q(\tau) x(\tau) - \|e\|_{[\tau,T]}^2 \\
&\le x_{1\tau}' P_1^{-1}(\tau) x_{1\tau} - \inf_{x_2(\tau)} \left\{ x'(\tau) Q(\tau) x(\tau) + \inf_e \left\{ \|e\|_{[\tau,T]}^2 \right\} \right\} \\
&= x_{1\tau}' P_1^{-1}(\tau) x_{1\tau} - \inf_{x_2(\tau)} \left\{ x'(\tau) Q(\tau) x(\tau) + x_2(\tau)' P_2^{-1}(\tau) x_2(\tau) \right\} \\
&= x_{1\tau}' P_1^{-1}(\tau) x_{1\tau} - x_{1\tau}' [Q_{11}(\tau) - Q_{12}(\tau)(P_2^{-1}(\tau) + Q_{22}(\tau))^{-1} Q_{12}'(\tau)] x_{1\tau},
\end{aligned}$$

where in obtaining the last step we have used the fact that if $M > 0$, then

$$\begin{aligned}\inf_{x_2}\left\{\begin{bmatrix} x_2' & x_1' \end{bmatrix}\begin{bmatrix} M & S \\ S' & N \end{bmatrix}\begin{bmatrix} x_2 \\ x_1 \end{bmatrix}\right\} \\ = \inf_{x_2}\left\{\|M^{1/2}(x_2 + M^{-1}Sx_1)\|^2 + x_1'(N - S'M^{-1}S)x_1\right\} \\ = x_1'(N - S'M^{-1}S)x_1.\end{aligned}$$

Since the above inequality is true for all $\tau \in (0,T)$ and for all $x_{1\tau}$ we see that (37) holds. This implies that the matrix $P^{-1}(\tau) - Q(\tau)$ is invertible for all $\tau \in (0,T)$, or equivalently that $\lambda_{max}(P(\tau)Q(\tau)) < 1 \ \ \forall \tau \in [0,T]$.

Sufficiency. Here we first show that any causal $f : w_1 \to \hat{y}_1$ that is of the form (9) achieves $\sup_{w_1} \frac{\|\Sigma_2(\Sigma_1 w_1 - \hat{y}_1)\|^2}{\|w_1\|^2} < 1$. With x as defined above, the composite state-space expression for the system Σ_1, Σ_2 and the operator $f(\cdot)$ defined in (9) can be written as

$$\begin{aligned}\dot{\eta} &= \check{F}\eta + \check{G}\nu, \\ \begin{bmatrix} e \\ \eta \end{bmatrix} &= \check{H}\eta + \begin{bmatrix} I & 0 \\ 0 & I \end{bmatrix}\nu,\end{aligned} \tag{38}$$

where

$$\eta = \begin{bmatrix} x \\ q \end{bmatrix}, \quad \nu = \begin{bmatrix} v \\ w_1 \end{bmatrix}, \quad \check{F} = \begin{bmatrix} \bar{F} & \tilde{G}_2(\bar{H}P - \tilde{G}_2') \\ 0 & -\tilde{F}' - NQ\tilde{G}_1\tilde{G}_1' \end{bmatrix}, \tag{39}$$

$$\check{G} = \begin{bmatrix} \tilde{G}_2 & \tilde{G}_1 \\ N(Q\tilde{G}_2 - \bar{H}') & NQ\tilde{G}_1 \end{bmatrix}, \quad \check{H} = \begin{bmatrix} \bar{H} & \bar{H}P - \tilde{G}_2' \\ 0 & -\tilde{G}_1' \end{bmatrix}. \tag{40}$$

With

$$\check{X} := \begin{bmatrix} Q & -(I - QP) \\ -(I - PQ) & -(I - PQ)P \end{bmatrix} = \begin{bmatrix} Q & -N^{-1} \\ -N^{-T} & -N^{-T}P \end{bmatrix}, \tag{41}$$

one can verify by straightforward though tedious algebra that the following hold:

$$\begin{aligned}-\frac{d\check{X}}{dt} &= \check{F}'\check{X} + \check{X}\check{F} - \check{H}'\check{H}, \\ \check{G}'\check{X} &= \check{H}.\end{aligned}$$

From (38) and above, one notes that

$$\frac{d(\eta'\check{X}\eta)}{dt} = \eta'\left[\check{F}'\check{X} + \check{X}\check{F} + \frac{d(\check{X})}{dt}\right]\eta + \eta'\check{X}\check{G}\nu + \nu'\check{G}'\check{X}\eta = \left[\nu + \check{H}\eta\right]'\left[\nu + \check{H}\eta\right] - \nu'\nu.$$

Integrating $\frac{d(\eta'\check{X}\eta)}{dt}$ from 0 to T one obtains (the boundary terms drop out because of boundary conditions on x, q, Q, and P)

$$\|w_1\|^2 + \|v\|^2 = \|\eta\|^2 + \|e\|^2. \tag{42}$$

Since $\|\Theta\| < 1$, there exists an $\epsilon > 0$ such that $\|\eta\|^2 - \|v\|^2 \geq \epsilon\|\eta\|^2$. Thus

$$\|w_1\|^2 - \|e\|^2 \geq \epsilon\|\eta\|^2 \geq \frac{\epsilon}{\|T_{w_1\eta}\|}\|w_1\|^2.$$

To complete the proof, it remains to be shown that any causal $f : w_1 \to \hat{y}_1$ that achieves the desired performance admits a representation as described in (9) for some Θ with $\|\Theta\| < 1$. This will now be shown using Lemma 4.3. Let $f : w_1 \to \hat{y}_1$ be a causal operator that achieves $\sup_{w_1} \frac{\|\Sigma_2(\Sigma_1 w_1 - \hat{y}_1)\|^2}{\|w_1\|^2} < 1$. Define $v := \hat{y}_1 - [\bar{H}P - \tilde{G}_2']q$, where q satisfies the differential equation described in (9). Then

$$P : \begin{bmatrix} w_1 \\ v \end{bmatrix} \to \begin{bmatrix} e \\ \eta \end{bmatrix}$$

is described by the system (38). As observed above, P is isometric (i.e., $\|w_1\|^2 + \|v\|^2 = \|\eta\|^2 + \|e\|^2$). It is also easily seen that the map from w_1 to η is invertible (i.e., P_{21}^{-1} exists), and since the problem is defined over a finite interval, P and P_{21}^{-1} are stable. Thus from Lemma 4.3 we conclude that $\|T_{ew_1}\| < 1$ if and only if $\|T_{v\eta}\| < 1$. Moreover, since the map $f(\cdot)$ is required to be causal, from (9) we observe that so must be $T_{v\eta}$. The required result follows by noting $\Theta := T_{v\eta}$. □

Before proceeding to the proof of the next theorem, we state the following lemma, the proof of which is implicit in the necessity proof of the modified Nehari problem, and thus we state it without proof. For any given $\tau \in [0, T]$, let $\mathcal{W}_\tau$ and $\mathcal{Y}_\tau$ be defined as

$$\mathcal{W}_\tau := \{w_1 \in L_2[0,T] : w_1(s) = 0 \ \forall\ 0 \le s < \tau\},$$
$$\mathcal{Y}_\tau := \{\hat{y}_1 \in L_2[0,T] : \hat{y}_1(s) = 0 \ \forall\ 0 \le s < \tau\}.$$

LEMMA 5.1. *Let Σ_1 and Σ_2 be anticausal systems described in equations* (6) *and* (7)*, respectively. The given modified Nehari problem* (*Problem* 5 *in section* 2) *is solvable* (*i.e., there exists a causal $f : w_1 \to \hat{y}_1$ that achieves* $\sup_{w_1} \frac{\|\Sigma_2(\Sigma_1 w_1 - \hat{y}_1)\|^2}{\|w_1\|^2} < 1$) *if and only if there exists an $\epsilon > 0$ so that*

$$\inf_{w_1 \in \mathcal{W}_\tau} \sup_{\hat{y}_1 \in \mathcal{Y}_\tau} \{\|w_1\|^2 - \|\Sigma_2(\Sigma_1 w_1 - \hat{y}_1)\|^2\} \ge \epsilon \|w_1\|^2$$

holds for all $\tau \in (0, T)$.

Proof of Theorem 3.2: *The basic delay problem.*

Necessity. Here we show that if there exists a causal operator $f(\cdot)$ that achieves the desired performance bound, then the Riccati differential equation described in (22) admits a solution. Since the information available to any admissible f about w is delayed by h units, for $t \in [0, h)$ it has no information and thus $\hat{z}(t) = 0$ for $t \in [0, h]$. Thus an existence of an admissible f so that $\sup_w \frac{\|z - \hat{z}\|^2}{\|w\|^2} < 1$ implies that, for some $\epsilon > 0$,

$$\|w\|_{[0,h]}^2 - \|Cx\|_{[0,h]}^2 \ge \epsilon \|w\|_{[0,h]}^2$$

for the system G defined in (2). It is well known (see, for example, [7]) that the above requirement is equivalent to the existence of a solution to the Riccati differential equation described by (22).

Sufficiency. The proof is somewhat long and hence we will break it down into smaller steps.

(1) Here we rewrite Cx, the signal to be estimated for the system (2), as a sum of three parts. The first part is predictable with the information available to the estimator, i.e., with knowledge of w with a delay of h units. For easier reading, we

rewrite some of the notation described previously.

$$\frac{d\Psi(\tau,s)}{d\tau} = (A + BB'X(\tau))\Psi(\tau,s), \quad \Psi(s,s) = I, \quad \bar{B}(\tau) = e^{A\tau}\Psi(h,\tau)B \text{ for } \tau \in [0,h],$$

$$\{w_i\} = Ww,$$

$$r_i = G_0 w_i, \quad i \geq 0,$$

$$z_{10} = 0, \quad z_{1i} = G_1 r_{i-1} \text{ for } i \geq 1,$$

$$z_{20} = 0, \quad z_{2i} = G_2 r_i \text{ for } i \geq 1,$$

where the systems G_0, G_1, and G_2 are as defined in (14), (15), and (16), respectively. From the above equations it is easily verified that, for all $\tau \in [0,h]$ and $i \geq 1$,

$$z_{2i}(\tau) = C\int_0^\tau \Psi(\tau,s)Br_i(s)ds = C\int_0^\tau e^{A(\tau-s)}Bw_i(s)ds,$$

$$\begin{aligned} z_{1i}(\tau) &= -C\int_h^\tau e^{A(\tau-s)}\bar{B}(s)r_{i-1}(s)ds \\ &= C\int_\tau^h e^{A(\tau-s)}e^{As}\Psi(h,s)Br_{i-1}(s)ds = Ce^{A\tau}\int_\tau^h \Psi(h,s)Br_{i-1}(s)ds, \end{aligned}$$

where the second equality of the first equation above follows from the fact that $G_2G_0 = G$. Again using the fact that $G_2G_0 = G$ and from (2), we have that, for any $\tau \in [0,h]$ and for any integer $i \geq 1$,

$$\begin{aligned} x(ih+\tau) &= e^{A\tau}x(ih) + \int_0^\tau e^{A(\tau-s)}Bw_i(s)ds \\ &= e^{A\tau}\left[e^{Ah}x((i-1)h) + \int_0^h \Psi(h,s)Br_{i-1}(s)ds\right] + \int_0^\tau e^{A(\tau-s)}Bw_i(s)ds \\ &= \left[e^{A(\tau+h)}x((i-1)h) + e^{A\tau}\int_0^\tau \Psi(h,s)Br_{i-1}(s)ds \ \right] \\ &\quad + e^{A\tau}\int_\tau^h \Psi(h,s)Br_{i-1}(s)ds + \int_0^\tau \Psi(\tau,s)Br_i(s)ds. \end{aligned}$$

From the above and definitions of z_{1i} and z_{2i} as defined in (20) and (21), one obtains

$$\begin{aligned} z(ih+\tau) &= Cx(ih+\tau) \\ &= C\left[e^{A(\tau+h)}x((i-1)h) + e^{A\tau}\int_0^\tau \Psi(h,s)Br_{i-1}(s)ds\right] \quad (\text{“term 1”}) \\ &\quad + Ce^{A\tau}\int_\tau^h \Psi(h,s)Br_{i-1}(s)ds + C\int_0^\tau \Psi(\tau,s)Br_i(s)ds \\ &= \text{“term 1”} + G_1 r_{i-1}(\tau) + G_2 r_i(\tau) = \text{“term 1”} + z_{1i}(\tau) + z_{2i}(\tau). \end{aligned} \tag{43}$$

The “term 1” depends only on $w(s)$ for $s \leq (i-1)h + \tau$ and thus is completely obtainable at $t = ih + \tau$ with knowledge of w with h units' delay. Note also that the

first term in the estimate $\hat{z}$ of z in the statement of Theorem 3.2 is nothing but the "term 1." With the following notation of $\{\hat{\hat{z}}_i\} \in l_{\mathcal{L}_2[0,h]}$,

$$\begin{aligned}
&\hat{\hat{z}}_0 = 0,\\
&\hat{\hat{z}}_i(\tau) := \hat{z}(ih+\tau) - C\left\{e^{A(\tau+h)}x((i-1)h) + e^{A\tau}\int_0^\tau \Psi(h,s)Br_{i-1}(s)ds\right\} \quad \text{for } i \geq 1\\
&(44) \quad = \hat{z}(ih+\tau) - \text{"term 1,"}
\end{aligned}$$

and the above decomposition of Cx, it follows that to show that the $\hat{z} = f(w)$ achieves the desired performance, we have to show that

$$\sup_w \frac{\|Cx\|^2_{[0,h]} + \|\{z_{1i} + z_{2i} - \hat{\hat{z}}_i\}\|^2}{\|w\|^2} < 1, \tag{45}$$

where $\hat{\hat{z}}_i$ is as defined in (44). Thus the given problem is solvable if and only if there exists a function $\hat{f} : \mathcal{L}_2 \to l_{\mathcal{L}_2[0,h]}$ so that with $\hat{\hat{z}}_0 = 0$, $\hat{\hat{z}}_i(\tau) = \hat{f}(w(s))(\tau)$, where $0 \leq s \leq (i-1)h + \tau$, $\tau \in [0,h]$, the above inequality holds.

(2) Here we obtain an expression for

$$\|w\|^2 - \|Cx\|^2_{[0,h]} - \|\{z_{1i} + z_{2i} - \hat{\hat{z}}_i\}\|^2$$

for any given $\{\hat{\hat{z}}_i\} \in l_{\mathcal{L}_2[0,h]}$ such that $\hat{\hat{z}}_0 = 0$. The expression we obtain not only is an important step in the proof but also may provide an insight into the structure of the problem and how $\hat{\hat{z}}_i$ must be chosen so as to satisfy (45).

Noting that $G_2G_0 = G$, one can write the expressions for r_i and z_{2i} defined in (19) and (21) using a single differential equation as

$$\begin{aligned}
\dot{x}_0(\tau) &= Ax_0(\tau) + Bw_i(\tau),\ \tau \in [0,h],\ x_0(0) = 0,\\
r_i(\tau) &= w_i(\tau) - B'X(\tau)x_0(\tau),\\
z_{2i}(\tau) &= Cx_0(\tau).
\end{aligned}$$

Integrating $\frac{d(x_0'Xx_0)}{dt}$ from 0 to h, where X satisfies (22), one obtains that

$$\|w_i\|^2 - \|z_{2i}\|^2 = \|r_i\|^2.$$

Similarly one can show that $\|w_0\|^2 - \|Cx\|^2_{[0,h]} = \|r_0\|^2$. From the above equation we have that for any $\{\hat{\hat{z}}_i\}$ such that $\hat{\hat{z}}_0 = 0$,

$$\begin{aligned}
&(46)\quad \|w\|^2 - \|Cx\|^2_{[0,h]} - \|\{z_{1i} + z_{2i} - \hat{\hat{z}}_i\}\|^2\\
&= \|r_0\|^2 + \textstyle\sum_{i\geq 1}\left[\{\|w_i\|^2 - \|z_{2i}\|^2\} - 2\langle z_{1i} - \hat{\hat{z}}_i, z_{2i}\rangle_{[0,h]} - \|z_{1i} - \hat{\hat{z}}_i\|^2\right]\\
&= \|r_0\|^2 + \textstyle\sum_{i\geq 1}\left[\|r_i\|^2 - 2\langle z_{1i} - \hat{\hat{z}}_i, G_2r_i\rangle_{[0,h]} - \|z_{1i} - \hat{\hat{z}}_i\|^2\right]\\
&= \|r_0\|^2 + \textstyle\sum_{i\geq 1}\left[\|r_i\|^2 - 2\langle G_2^*(z_{1i} - \hat{\hat{z}}_i), r_i\rangle_{[0,h]} - \|z_{1i} - \hat{\hat{z}}_i\|^2\right]\\
&= \|r_0\|^2 + \textstyle\sum_{i\geq 1}\left[\|r_i - G_2^*(z_{1i} - \hat{\hat{z}}_i)\|^2 - \langle (I + G_2G_2^*)(z_{1i} - \hat{\hat{z}}_i), (z_{1i} - \hat{\hat{z}}_i)\rangle_{[0,h]}\right]\\
&= \|r_0\|^2 + \textstyle\sum_{i\geq 1}\left[\|r_i - G_2^*(z_{1i} - \hat{\hat{z}}_i)\|^2 - \|G_3(z_{1i} - \hat{\hat{z}}_i)\|^2\right].
\end{aligned}$$

The last equality stems from Lemma 4.2. Noticing the way $\{w_{1i}\} \in l_{\mathcal{L}_2[0,h]}$ is defined in the theorem statement ($w_{10} = r_0$ and $w_{1i} = r_i - G_2^*(z_{1i} - \hat{\hat{z}}_i)$ for $i \geq 1$) and using the definition of z_{1i} ($z_{1i} = G_1 r_{i-1}$), the above equation can be rewritten as

$$\begin{aligned} &\|w\|^2 - \|Cx\|^2_{[0,h]} - \textstyle\sum_{i\geq 1} \|z_{1i} + z_{2i} - \hat{\hat{z}}_i\|^2 \\ &= \textstyle\sum_{i\geq 0} \left[\|w_{1i}\|^2 - \|G_3(z_{1(i+1)} - \hat{\hat{z}}_{i+1})\|^2\right] \\ &= \textstyle\sum_{i\geq 0} \left[\|w_{1i}\|^2 - \|G_3(G_1 w_{1i} + \nu_i - \hat{\hat{z}}_{i+1})\|^2\right], \end{aligned} \tag{47}$$

where $\{\nu_i\} \in l_{\mathcal{L}_2[0,h]}$ is defined as

$$\nu_i := G_1 G_2^*(z_{1i} - \hat{\hat{z}}_i), \quad i \geq 0.$$

(3) In this part we will show that a certain modified Nehari problem admits a solution. Based on this modified Nehari problem and (47), we will qualitatively outline the structure of all admissible $f : w \to \hat{z}$ (or equivalently, $\hat{f} : w \to \hat{\hat{z}}$) that achieve the desired performance bound.

The signal z_{1i} ($z_{1i} = G_{1i} r_{i-1}$) depends only on $w(s)$, $0 \leq s \leq ih$. Because of the presence of delay, for all admissible $\hat{f} : w \to \hat{\hat{z}}$, $\hat{\hat{z}}_i$ also depends only on $w(s)$, $0 \leq s \leq ih$. Thus both of the signals z_{1i} and $\hat{\hat{z}}_i$ are known if one knows $w(s)$ for all $0 \leq s \leq t = ih$. Thus the signal $\nu_i = G_1 G_2^*(z_{1i} - \hat{\hat{z}}_i)$ is also known if one knows $w(s)$ for all $0 \leq s \leq t = ih$. Treating ν_i as a known signal, each of the terms in the summation in (47) can be viewed as a modified Nehari problem (problem 5 in the problem definition section) if one makes the following association with the terminology of the modified Nehari problem (note that both G_1 and G_3 are anticausal systems):

$$\Sigma_1 \leftarrow G_1, \quad \Sigma_2 \leftarrow G_3, \quad w_1 \leftarrow w_{1i}, \quad \hat{y}_1 \leftarrow \hat{\hat{z}}_{i+1} - \nu_i, \quad T \leftarrow h.$$

It is shown in the appendix that if the Riccati differential equation (22) admits a solution, then indeed the above modified Nehari problem is solvable.

We now briefly outline how $\hat{\hat{z}}_{i+1}$ is chosen from the solution of the above modified Nehari problem. Let us make the following association with the notation adopted in the modified Nehari problem:

$$\Sigma_1 \leftarrow G_1, \quad \Sigma_2 \leftarrow G_3, \quad w_1 \leftarrow w_{1i}, \quad T \leftarrow h, \; \hat{y}_1 \leftarrow \hat{y}_{i+1}, \quad v \leftarrow v_i, \quad \eta \leftarrow \eta_i.$$

It follows from equation (42) of the modified Nehari problem proof that for all integers $i \geq 0$

$$\|w_{1i}\|^2 - \|G_3(G_1 w_{1i} - \hat{y}_{i+1})\|^2 = \|\eta_i\|^2 - \|v_i\|^2.$$

If $\hat{\hat{z}}_{i+1}$ is chosen as $\hat{\hat{z}}_{i+1} = \hat{y}_{i+1} + G_1 G_2^*(z_{1i} - \hat{\hat{z}}_i) = \hat{y}_{i+1} + \nu_i$, then it follows from the above that

$$\|w_{1i}\|^2 - \|G_3(G_1 w_{1i} + G_1 G_2^*(z_{1i} - \hat{\hat{z}}_i) - \hat{\hat{z}}_{i+1})\|^2 = \|\eta_i\|^2 - \|v_i\|^2 \quad \forall i \geq 0.$$

Summing the above for all $i \geq 0$, one observes that

$$P : \begin{bmatrix} \{w_{1i}\} \\ \{v_i\} \end{bmatrix} \to \begin{bmatrix} \{G_3(G_1 w_{1i} + G_1 G_2^*(z_{1i} - \hat{\hat{z}}_i) - \hat{\hat{z}}_{i+1})\} \\ \{\eta_i\} \end{bmatrix} \text{ is isometric.} \tag{48}$$

Thus $\|\{w_1\}\|^2 + \|\{v_i\}\|^2 = \|\{\eta_i\}\|^2 + \|\{G_3(G_1 w_{1i} + G_1 G_2^*(z_{1i} - \hat{\hat{z}}_i) - \hat{\hat{z}}_{i+1})\}\|^2$. If additionally $\|\{\eta_i\}\|^2 > \|\{v_i\}\|^2$, then the right-hand side of (47) would be greater than or equal to zero and thus the required performance bound would be achieved.

Because of the presence of delay, one also needs to show that for any $\tau \in [0, h]$, $\hat{\hat{z}}_{i+1}(\tau)$ chosen as above depends only on $w(s)$ for $0 \le s \le ih + \tau$. Let $i \ge 1$ be any given integer. We have already established that ν_i depends only on $w(s)$ for $s \le ih$. Also, $\hat{\hat{z}}_{i+1} = \hat{y}_{i+1} + \nu_i$ where $\hat{y}_{i+1}$ is the estimate for the modified Nehari problem. From the definition of modified Nehari problem, for any $\tau \in [0, h]$, $\hat{y}_{i+1}(\tau)$ is generated causally from w_{1i} and thus depends only on $w_{1i}(s)$ for $0 \le s \le \tau$. From the definition of w_{1i} it then follows that $\hat{y}_{i+1}(\tau)$ depends only on $w(s)$ for $0 \le s \le ih + \tau$. Thus $\hat{\hat{z}}_{i+1}(\tau) = y_{i+1}(\tau) + \nu_i(\tau)$ depends only on $w(s)$ for $0 \le s \le ih + \tau$.

(4) In this part we complete the proof of the "sufficiency" along the lines outlined in the previous subsection. For convenience, the set of all estimators described in (24) is reproduced below:

$$\begin{aligned}
\hat{y}_{i+1} &= \mathcal{N}^1_{[G_1, G_3, h, \Theta]}(w_{1i}), \quad i \ge 0, \\
\eta_i &= \mathcal{N}^2_{[G_1, G_3, h, \Theta]} : (w_{1i}), \quad i \ge 0, \\
\hat{\hat{z}}_{i+1} &= \hat{y}_{i+1} + G_1 G_2^*(z_{1i} - \hat{\hat{z}}_i), \quad i \ge 0, \\
\{v_i\} &= \Theta\{\eta_i\}, \quad \text{where } \|\Theta\| < 1.
\end{aligned}$$

From (48) and (47) it follows that

$$\begin{aligned}
& \|w\|^2 - \|Cx\|^2_{[0,h]} - \textstyle\sum_{i \ge 1} \|z_{1i} + z_{2i} - \hat{\hat{z}}_i\|^2 \\
= \;& \textstyle\sum_{i \ge 0} \left[\|\eta_i\|^2 - \|v_i\|^2 \right] \\
\ge \;& \gamma \|\{\eta_i\}\|^2 \quad (\text{for some } \gamma > 0 \text{ since } \|\Theta\| < 1) \qquad (49) \\
\ge \;& \epsilon_1 \gamma \|\{w_{1i}\}\|^2 \quad (\text{for some } \epsilon_1 > 0) \\
\ge \;& \epsilon_2 \epsilon_1 \gamma \|\{w_i\}\|^2 = \epsilon \|w\|^2 \quad (\text{for some } \epsilon_2 > 0, \text{ and thus } \epsilon = \epsilon_2 \epsilon_1 \gamma > 0),
\end{aligned}$$

where the third from last inequality follows from the fact that $\|\Theta\| < 1$, where $\Theta : \{\eta_i\} \to \{v_i\}$. The second from last inequality follows from the fact that with $\{v_i\} = \Theta\{\eta_i\}$, the map from $\{\eta_i\}$ to $\{w_{1i}\}$ is bounded. The last inequality is a consequence of the fact that the map from $\{w_{1i}\}$ to $\{w_i\}$ is bounded.

The above inequality, together with the fact that it is enough to show (45), leads to the conclusion that the maps $f : w \to \hat{z}$ described in (24) achieve the desired performance bound.

(5) Next we show that any $f : w \to \hat{z}$ that is admissible and achieves the desired performance bound, admits a realization as in the theorem statement with some causal Θ such that $\|\Theta\| < 1$. This is shown as in the proof of the modified Nehari problem. Let $f : w \to \hat{z}$ be admissible and achieve the desired performance bound. For this f, let the operator $\hat{f}$ be defined as $\hat{f} : w \to \hat{\hat{z}}$, where $\hat{\hat{z}}_i$ is as defined in (44). Then from (45) one notes that there exists an $\epsilon > 0$ such that

$$\|w\|^2 - \|Cx\|^2_{[0,h]} - \|\{z_{1i} + z_{2i} - \hat{\hat{z}}_i\}\|^2 \ge \epsilon \|w\|^2.$$

Since $z_{1i} = G_1 G_0 w_{i-1}$ and $z_{2i} = G_2 G_0 w_i$, $\|\{z_{1i}\}\| \le \|G_1 G_0\| \|w\|$ and $\|\{z_{2i}\}\| \le \|G_2 G_0\| \|w\|$, where both $\|G_1 G_0\|$ and $\|G_2 G_0\|$ are bounded, since these operators are defined over a finite interval. From this together with the fact that the above inequality

holds, one concludes that there exists an $M < \infty$ such that $\|\{\hat{\hat{z}}_i\}\| \le M\|w\|$. Since $w_{1i} = r_i - G_2^*(z_{1i} - \hat{\hat{z}}_i)$, $\|\{w_{1i}\}\| < \gamma\|w\|$ for some $\gamma < \infty$. Thus for this $\hat{f} : w \to \hat{\hat{z}}$ that is admissible and achieves the desired performance, from the above equation and (47),

$$\begin{aligned}&\|w\|^2 - \|Cx\|^2_{[0,h]} - \textstyle\sum_{i\ge1} \|z_{1i} + z_{2i} - \hat{\hat{z}}_i\|^2\\ &= \textstyle\sum_{i\ge0} \left[\|w_{1i}\|^2 - \|e_i\|^2\right] \ge \epsilon_1 \|\{w_{1i}\}\|^2,\end{aligned}$$

where $\epsilon_1 = \frac{\epsilon}{\gamma} > 0$ and

$$e_i := G_3(G_1 w_{1i} + G_1 G_2^*(z_{1i} - \hat{\hat{z}}_i) - \hat{\hat{z}}_{i+1}).$$

For this function $\hat{f}$, let $\{v_i\} \in l_{\mathcal{L}_2[0,h]}$ be defined as $v_i := G_1 G_2^*(z_{1i} - \hat{\hat{z}}_i) - [\bar{H}_d P_d - \tilde{G}'_{2d}]q_i - \hat{\hat{z}}_{i+1}$, where q_i is the solution of the differential equation described in (24). Also, let η_i be as defined in (24). The map from w_{1i} to η_i is invertible and bounded (since the system is defined over a finite interval). Now from (48) and Lemma 4.3 we conclude that $\|T_{\{e_i\}\{w_{1i}\}}\| < 1$ if and only if $\|T_{\{v_i\}\{\eta_i\}}\| =: \|\Theta\| < 1$. Next we argue that Θ must also be causal. Because of the delay, for any $\tau \in [0,h]$, $\hat{\hat{z}}_{i+1}(\tau)$ should depend only on $w_{1j},\ j < i$, and $w_{1i}(s),\ 0 \le s \le \tau$. It is easily observed from (24) that this is the case if and only if $v_i(s)$ depends only on $w_{1j},\ j < i$, and $w_{1i}(s),\ 0 \le s \le \tau$. This is the case if and only if $\Theta = T_{\{v_i\}\{\eta_i\}}$ is causal. □

Proof of Theorem 3.3: *Full information control problem with delay.*

The proof of the Theorem 3.3 proceeds by converting the original problem to one that is equivalent to a basic delay problem using Lemma 4.4 adapted to the system G_{fi}.

Necessity. Here we show that conditions (1) and (2) of Theorem 3.3 must be satisfied for there to exist a controller which stabilizes the system and achieves $\|T_{zw}\| < 1$. Condition 1 is a necessary condition for the existence of controllers that achieve the desired performance even in the absence of delay (see, for example, [3]). Thus it must also be a necessary condition when the measurements are available with a delay.

From the full information (full information is used to denote complete information about the state and the exogenous signal) equivalent to Lemma 4.4 (see, for example, [3]), one concludes that if there exists a controller K that achieves $\|T_{zw}\| < 1$ for the system G_{fi}, then for the same controller $\|T_{vr}\| < 1$ for the following system:

$$\begin{aligned}\dot{x} &= (A + B_1 B_1' X)x + B_1 r + B_2 u,\ t \in [0,\infty),\ x(0) = 0,\\ v &= B_2' X x + u,\\ y &= \begin{bmatrix} x \\ r \end{bmatrix},\\ u(t) &= f(y(s)) \text{ with } s \le t - h,\end{aligned} \tag{50}$$

where the function $f(\cdot)$ defines the controller. Since for $t \in [0,h)$ the controller has no information, $u(t) = 0$ for $t \in [0,h)$. Thus if $\|T_{vr}\| < 1$, there exists an $\epsilon > 0$ such that

$$\|r\|^2_{[0,h]} - \|B_2' X x\|^2_{[0,h]} \ge \epsilon \|r\|^2_{[0,h]}$$

for the system given by the equation (50) with $u(t) = 0$ for $t \in [0,h)$. A necessary and sufficient condition for this to hold is nothing but condition (2) of Theorem 3.3.

Sufficiency. Here we show that all the controllers that achieve the desired performance are as described in the statement of Theorem 3.3. Because of Lemma 4.4, it is enough to show that all such controllers are the ones that achieve $\|T_{vr}\| < 1$ for the system described by (50).

The state x of the system described in equation (50) can be written as the sum of two states, one driven by the input u and the other by the exogenous signal r:

$$\begin{aligned} \dot{x_1} &= (A + B_1 B_1' X)x_1 + B_1 r,\ t \in [0,\infty),\ x_1(0) = 0, \\ \dot{x_2} &= (A + B_1 B_1' X)x_2 + B_2 u,\ t \in [0,\infty),\ x_2(0) = 0, \\ x &= x_1 + x_2, \\ v &= B_2' X(x_1 + x_2) + u. \end{aligned}$$

At any time t, $x_2(t)$ is known since it is generated by past inputs which are known. Thus x_2 can be treated as a deterministic/known signal. Let $\bar{u} := u + B_2' X x_2$. Thus a given controller $K : r \to u$ achieves the desired performance if and only if with $\bar{u} := u + B_2' X x_2$, $\|T_{vr}\| < 1$ for the following system:

$$\begin{aligned} \dot{x_1} &= (A + B_1 B_1' X)x_1 + B_1 r,\ t \in [0,\infty),\ x_1(0) = 0, \\ v &= B_2' X x_1 + \bar{u}, \\ \bar{u}(t) &= \bar{f}(r(s)) \text{ with } 0 \le s \le t - h. \end{aligned} \tag{51}$$

We now note that for the above system, the problem of determining $\bar{u}$ so as to achieve $\|T_{vr}\| < 1$ is nothing but a basic delay problem (Problem 1) if one makes the following association with the terminology of the basic delay problem:

$$A \leftarrow A + B_1 B_1' X, \quad B \leftarrow B_1, \quad C \leftarrow -B_2' X, \quad \hat{z} \leftarrow \bar{u}.$$

From Theorem 3.2 one concludes that the above basic delay problem is solvable because of condition (2) of Theorem 3.3. Let $\Omega : r \to \bar{u}$, where $\Omega \in \mathcal{D}_{[A+B_1B_1'X, B_1, -B_2'X, h]}$; i.e., Ω solves the basic delay problem for the system described in (51) and achieves $\|T_{vr}\| < 1$. Then from the above arguments it follows that $u = \bar{u} - B_2' X x_2$ achieves $\|T_{vr}\| < 1$ for the system given by (50). It is easily verified that all controllers presented in Theorem 3.3 are of this form.

Next we show that any controller that achieves $\|T_{vr}\| < 1$ for the system described in equation (50) admits a realization as described in Theorem 3.3. Let the controller $K : r \to u$ be admissible and achieve $\|T_{vr}\| < 1$ for the system described by (50). Define $\bar{u} := u + B_2' X x_2$, where x_2 is as defined above and in the theorem statement. For this controller K, let $\bar{K} : r \to \bar{u}$ be the corresponding map from r to $\bar{u}$. Then using the above decomposition of x, one concludes that $\bar{K}$ solves the basic delay problem for the system described in (51) and achieves $\|T_{vr}\| < 1$ or, equivalently, $\bar{K} \in \mathcal{D}_{[A+B_1B_1'X, B_1, -B_2'X, h]}$. □

Proof of Theorem 3.5: *Output feedback control problem with delay.*

The proof of Theorem 3.5 proceeds by converting the original problem to one that is equivalent to a basic delay problem using Lemmas 4.6 and 4.9.

Necessity. Here we show that conditions (30), (31), (32), and (33) of Theorem 3.5 must be satisfied for there to exist a controller which stabilizes the system and achieves $\|T_{zw}\| < 1$. Conditions (30), (31), and (32) are necessary conditions for existence of controllers that achieve the desired performance even in the absence of delay (see, for example, [3]). Thus they must also necessarily hold when the measurements are available with a delay.

From Lemma 4.6 one concludes that if there exists a controller K that achieves $\|T_{zw}\| < 1$, then for the same controller, $\|T_{\bar{q}\bar{p}}\| < 1$ for the following system:

$$
\begin{array}{rcl}
\dot{x} & = & (A + YC_1'C_1 + YC_2'C_2YZ)x + YC_2'\bar{p} + B_2u,\ t \in [0,\infty),\ x(0) = 0, \\
\bar{q} & = & B_2'Zx + u, \\
y & = & C_2(I + YZ)x + \bar{p}, \\
u(t) & = & f(y(s)) \text{ with } 0 \le s \le t - h,
\end{array} \tag{52}
$$

where the function $f(\cdot)$ defines the controller. Since for $t \in [0, h)$ the controller has no information, $u(t) = 0$ for $t \in [0, h)$. Thus if $\|T_{\bar{q}\bar{p}}\| < 1$, there exists an $\epsilon > 0$ such that

$$\|\bar{p}\|^2_{[0,h]} - \|B_2'Zx\|^2_{[0,h]} \ge \epsilon\|\bar{p}\|^2_{[0,h]}$$

for the system described by (52) with $u(t) = 0$ for $t \in [0, h)$. A necessary and sufficient condition for this to hold is nothing but the condition (33) of the theorem statement.

Sufficiency. Here we show that all the controllers that achieve the desired performance are as described in the statement of Theorem 3.5. From the Lemma 4.6, all controllers that achieve the desired performance are the ones that achieve $\|T_{\bar{q}\bar{p}}\| < 1$ for the system described by (52).

Consider the following system, which has the same dynamics as the system (52), but now the controller has delayed measurements of $\bar{p}$ instead of $C_2(I + YZ)x + \bar{p}$; i.e.,

$$
\begin{array}{rcl}
\dot{x} & = & (A + YC_1'C_1 + YC_2'C_2YZ)x + YC_2'\bar{p} + B_2u,\ t \in [0,\infty),\ x(0) = 0, \\
\bar{q} & = & B_2'Zx + u, \\
y_2 & = & \bar{p}, \\
u(t) & = & g(y_2(s)) \text{ with } 0 \le s \le t - h.
\end{array} \tag{53}
$$

For the system described by (53), the problem of obtaining $g(\cdot)$ so that $\|T_{\bar{q}\bar{p}}\| < 1$ is of the same type as was encountered in equation (50) in the proof of Theorem 3.3. As in the proof of Theorem 3.3, one concludes that the above problem is solvable because of condition 4 of Theorem 3.5. One now proceeds exactly as before in the proof of Theorem 3.3 to conclude that any controller that achieves the desired performance $\|T_{\bar{q}\bar{p}}\| < 1$ for the system (53) has a realization of the following form:

$$
K_\Omega \left\{ \begin{array}{rcl}
\dot{x_2} & = & (A + YC_1'C_1 + YC_2'C_2YZ)x_2 + B_2u,\ t \in [0,\infty),\ x_2(0) = 0, \\
u(t) & = & 0, \quad t \in [0, h), \\
u(t) & = & -B_2'Zx_2 + (\Omega\bar{p})(t) \quad t \ge h,
\end{array} \right.
$$

where $\Omega \in \mathcal{D}_{[(A+YC_1'C_1+YC_2C_2'YZ),YC_2',-B_2'Z,h]}$. Let us define the above controller as $K_\Omega : \bar{p} \to u$.

Note that from condition 2 of Theorem 3.5, $A + YC_1'C_1 - YC_2C_2'$ is stable. From Lemma 4.9 it follows that all controllers that achieve $\|T_{\bar{q}\bar{p}}\| < 1$ for system (52) are of the form

$$
\begin{array}{rcl}
\dot{x_1} & = & (A + YC_1'C_1 - YC_2'C_2)x_1 + YC_2'y + B_2u,\ t \in [0,\infty),\ x_1(0) = 0, \\
r & = & -C_2(I + YZ)x_1 + y, \\
u & = & K_\Omega r.
\end{array}
$$

It is easily verified that the controllers described in Theorem 3.5 that achieve the desired performance are of the above form. □

Proof of Theorem 3.4: *Prediction problem.*

The proof of Theorem 3.4 proceeds by converting the original problem to one that is equivalent to a basic delay problem using Lemma 4.8. From Lemma 4.8 one concludes that if there exists a predictor $\mathcal{E}$ that achieves $\|T_{ew}\| < 1$, then for the same predictor $\|T_{\nu\eta}\| < 1$ for the following system:

$$\begin{aligned} \dot{x} &= (A + YC_1'C_1)x + YC_2'\eta - YC_1'\hat{z},\ t \in [0,\infty),\ x(0) = 0, \\ \nu &= C_1x - \hat{z}, \\ y &= C_2x + \eta, \\ \hat{z}(t) &= f(y(s)) \text{ with } 0 \le s \le t - h, \end{aligned} \tag{54}$$

where the function $f(\cdot)$ defines the predictor. This problem is completely analogous to (52), discussed in the proof of Theorem 3.5, and the proof proceeds along the same lines if one makes the following association with the terminology in (52):

$$(A + YC_1'C_1 + YC_2'C_2YZ) \leftarrow (A + YC_1'C_1),\quad YC_2' \leftarrow YC_2',\quad B_2 \leftarrow YC_1',$$

$$B_2'Z \leftarrow C_1,\quad C_2(I + YZ) \leftarrow C_2,\quad \bar{p} \leftarrow \eta,\quad \bar{q} \leftarrow \nu,\quad u \leftarrow -\hat{z}. \qquad \square$$

6. Summary. In this paper we have presented state-space solutions to $\mathcal{H}_\infty$ control and estimation problems when there is a delay present in the measurements or implementation of control. The necessary and sufficient conditions are given in terms of finite-dimensional algebraic and differential Riccati equations. Using the solutions to these Riccati equations, one can obtain explicit state-space realizations for the controllers and predictors. The "central" solution for the controllers/estimators is a linear periodic system, with the period being the amount of delay present. Though the approach presented here may provide some new insight about how "lifting" type techniques can be used for such problems, it is not clear if these ideas are generalizable to a wider class of problems, such as ones involving multiple delays in dynamics or measurements.

Appendix. Here we show that if the Riccati differential equation (22) admits a solution, then the modified Nehari problem with the following association with the terminology of the modified Nehari problem is solvable:

$$\Sigma_1 \leftarrow G_1, \quad \Sigma_2 \leftarrow G_3, \quad T \leftarrow h.$$

In the discussions that follow, τ is a fixed number in the interval $\tau \in (0, h)$. Recall that $\{w_i\} = Ww$ and $r_i = G_0w_i$. Let $\mathcal{W}_\tau$, $\mathcal{Z}_\tau$, and $\mathcal{R}_\tau$ be defined as follows:

$$\begin{aligned} \mathcal{W}_\tau &:= \{w_0 \in L_2[0,h] : w_0(s) = 0\ \forall\ 0 \le s < \tau\}, \\ \mathcal{Z}_\tau &:= \{\hat{z}_1 \in L_2[0,h] : \hat{z}_1(s) = 0\ \forall\ 0 \le s < \tau\}, \\ \mathcal{R}_\tau &:= \{r_0 \in L_2[0,h] : r_0(s) = 0\ \forall\ 0 \le s < \tau\}. \end{aligned} \tag{55}$$

Since $r_0 = G_0w_0$ and G_0 is causal and invertible, $r_0 \in \mathcal{R}_\tau$ if and only if $w_0 \in \mathcal{W}_\tau$. From (43) one notes that for any $w_0 \in \mathcal{W}_\tau$, $z = Cx$ for the system (2) in the interval $[h, 2h]$ can be written as

$$Cx(h + t) = z_{11}(t) + z_{12}(t),\ t \in [0,h], \tag{56}$$

where $z_{11}(t) = G_1 r_0(t)$ and $z_{21}(t) = G_2 r_1(t)$, since the "term 1" here is 0 because $x(0) = 0$ and $r_0(s) = 0,\ 0 \le s < \tau$.

It is well known that existence of a solution to the Riccati differential equation described by (22) implies that, for some $\epsilon > 0$,

$$\|w\|^2_{[0,h]} - \|Cx\|^2_{[0,h]} \ge \epsilon \|w\|^2_{[0,h]} \quad \forall w \in L_2[0,h] \tag{57}$$

for the system G defined in (2). The following observation follows immediately from the above inequality: for any $w \in L_2[0,2h]$ for which $w_0 \in \mathcal{W}_\tau$,

$$\begin{aligned} &\sup_q \{\|w\|^2_{[\tau,2h]} - \|Cx\|^2_{[\tau,\tau+h]} - \|Cx - q\|^2_{[\tau+h,2h]}\} \\ &= \{\|w\|^2_{[\tau,\tau+h]} - \|Cx\|^2_{[\tau,\tau+h]}\} + \|w\|^2_{[\tau+h,2h]} \\ &\quad \text{(with } q \text{ chosen as } q(t) = Cx(t),\ t \in [\tau+h, 2h]) \\ &\ge \epsilon \|w\|^2_{[\tau,\tau+h]} \quad \text{(from (57) and the fact that } x(\tau) = 0 \text{ since } w_0 \in \mathcal{W}_\tau). \end{aligned} \tag{58}$$

Moreover, if $w_0 \in \mathcal{W}_\tau$, the following holds for any choice of $\hat{\hat{z}}_1$:

$$\begin{aligned} &\|w\|^2_{[0,2h]} - \|Cx\|^2_{[0,h]} - \|Cx - \hat{\hat{z}}_1\|^2_{[h,2h]} \\ &= \|w_0\|^2_{[\tau,h]} + \|w_1\|^2 - \|Cx\|^2_{[\tau,h]} - \|Cx - \hat{\hat{z}}_1\|^2_{[h,2h]} \quad (\text{since } w_0 \in \mathcal{W}_\tau) \\ &= \|w_0\|^2_{[\tau,h]} + \|w_1\|^2 - \|Cx\|^2_{[\tau,h]} - \|z_{11} + z_{21} - \hat{\hat{z}}_1\|^2 \quad (\text{from (56)}) \\ &= \|r_0\|^2_{[\tau,h]} + \|r_1 - G^*_{2i}(z_{11} - \hat{\hat{z}}_1)\|^2 - \|G_3(z_{11} - \hat{\hat{z}}_1)\|^2 \quad (\text{from (46)}). \end{aligned}$$

From the above identity and (58) it follows that for any $r_1 \in L_2[0,h]$ and $r_0 \in \mathcal{R}_\tau$,

$$\sup_{\hat{\hat{z}}_1 \in \mathcal{Z}_\tau} \{\|r_0\|^2_{[\tau,h]} + \|r_1 - G^*_{2i}(z_{11} - \hat{\hat{z}}_1)\|^2 - \|G_3(z_{11} - \hat{\hat{z}}_1)\|^2\} \ge \epsilon \|w\|^2_{[\tau,\tau+h]} \ge \rho \|r_0\|^2_{[\tau,h]}$$

for some $\rho > 0$. Since the above holds for all r_1, it also holds for $r_1 = G^*_{2i}(z_{11} - \hat{\hat{z}}_1)$. Thus

$$\inf_{r_0 \in \mathcal{R}_\tau} \sup_{\hat{\hat{z}}_1 \in \mathcal{Z}_\tau} \{\|r_0\|^2_{[\tau,h]} - \|G_3(G_1 r_0 - \hat{\hat{z}}_1)\|^2\} \ge \rho \|r_0\|^2_{[\tau,h]},$$

where we have used the fact that $z_{11} = G_1 r_0$. Since the above holds for any $\tau \in (0,h)$, we conclude from Lemma 5.1 that the given modified Nehari problem is solvable.

REFERENCES

[1] B. A. Bamieh and J. B. Pearson, Jr., *A general framework for linear periodic systems with applications to $\mathcal{H}_\infty$ sampled-data control*, IEEE Trans. Automat. Control, 37 (1992), pp. 418–435.

[2] T. Başar and P. Bernhard, *H^∞-Optimal Control and Related Minimax Problems: A Dynamic Games Approach*, Birkhäuser Boston, Cambridge, MA, 1991.

[3] J. C. Doyle, K. Glover, P. P. Khargonekar, and B. A. Francis, *State-space solutions to standard H_2 and H_∞ control problems*, IEEE Trans. Automat. Control, 34 (1989), pp. 831–847.

[4] D. Flamm and S. Mitter, *H_∞ sensitivity minimization for delay systems*, Systems Control Lett., 9 (1987), pp. 17–24.

[5] D. J. Limebeer and U. Shaked, *New results in H_∞ filtering*, in Proc. 9th International Symposium on MTNS, Kobe, Japan, 1991, pp. 317–322.

[6] M. A. Kaashoek and J. Kos, *The Nehari-Takagi problem for input-output operators of time varying continuous time systems*, Integral Equations Operator Theory, 18 (1994), pp. 435–467.

[7] P. P. Khargonekar, K. M. Nagpal, and K. Poolla, *H_∞ control with transients*, SIAM J. Control Optim., 29 (1991), pp. 1373–1393.

[8] K. Nagpal and P. P. Khargonekar, *Filtering and smoothing in an H_∞ setting*, IEEE Trans. Automat. Control, 36 (1991), pp. 152–166.

[9] K. Nagpal and R. Ravi, *H_∞ control and estimation problems with delayed measurements: state space solutions*, in Proc. 1994 American Control Conference, Baltimore, MD, 1994, pp. 2379–2383.

[10] H. Ozbay, M. C. Smith, and A. Tannenbaum, *Controller design for unstable distributed plants*, in American Control Conference, San Diego, 1990, pp. 1583–1588.

[11] H. Ozbay and A. Tannenbaum, *A skew Toeplitz approach to the H_∞ optimal control of multivariable distributed systems*, SIAM J. Control Optim., 28 (1990), pp. 653–670.

[12] J. R. Partington, K. Glover, H. J. Zwart, and R. F. Curtain, *L_∞ approximation and nuclearity of delay systems*, Systems Control Lett., 10 (1988), pp. 59–65.

[13] R. Ravi, K. M. Nagpal, and P. P. Khargonekar, *H_∞ control of linear time-varying systems: A state-space approach*, SIAM J. Control Optim., 29 (1991), pp. 1394–1413.

[14] G. Tadmor, *Worst-case design in the time domain: The maximum principle and the standard H_∞ problem*, Math. Control Signals Systems, 3 (1990), pp. 301–324.

[15] G. Tadmor, *The standard H_∞ problem and the maximum principle—The general linear case*, SIAM J. Control Optim., 31 (1993), pp. 831–846.

[16] G. Tadmor and M. Verma, *Factorization and the Nehari theorem in time varying systems*, Math. Control Signals Systems, 5 (1992), pp. 418–452.

[17] K. Zhou and P. P. Khargonekar, *On the weighted sensitivity minimization problem for delay systems*, Systems Control Lett., 8 (1987), pp. 307–312.

SIAM J. CONTROL OPTIM.
Vol. 35, No. 4, pp. 1244–1262, July 1997

008

CONSTRAINED H_∞ OPTIMAL CONTROL OVER AN INFINITE HORIZON*

ATHANASIOS SIDERIS† AND HÉCTOR ROTSTEIN‡

Abstract. It has been recently shown in [*Automatica J. IFAC*, 29 (1993), pp. 969–983; *IEEE Trans. Automat. Control*, 39 (1994), pp. 762–779] that time-domain constraints can be incorporated *explicitly* into H_∞ optimal control problems over a finite horizon. In this paper, we construct a sequence of such finite-horizon constrained H_∞ optimization problems and show that their solutions converge to the solution of the infinite-horizon H_∞ optimization. We prove that convergence is uniform in the H_∞ norm and that the optimal solution to the infinite-horizon problem is unique.

Key words. H_∞ optimal control, time-domain constraints

AMS subject classifications. 93C05, 93C80, 49N05

PII. S0363012995256898

1. Introduction.

It has been recently shown that time-domain constraints can be incorporated *explicitly* into the H_∞ optimal control problems over a finite horizon. More specifically in [10, 8], the problem of minimizing the H_∞ norm of a closed-loop transfer function of a discrete-time system subject to convex constraints on the time responses of several closed-loop responses to given test signals such as steps, impulses, etc. imposed over a *finite* horizon is solved exactly.

These results allow significantly more insight into the properties of the constrained optimal systems and offer certain computational advantages over pure convex programming approaches for constrained control problems [4, 6, 2, 5]. These results are outlined in section 2.

In the approach in [10, 8], one perhaps expects that the decay of the time-domain responses implied by the asymptotic stability of the closed-loop system, along with the constraints over the finite horizon, produce a desirable time-domain behavior over the infinite horizon. However, constrained H_∞ optimal solutions may exhibit undesirable time-domain behavior *immediately* after the finite horizon, regardless of how long this horizon is extended.

This behavior is essentially caused because constrained H_∞ optimal solutions over a finite horizon are allpass, while such solutions over an infinite horizon are not necessarily so. Section 3 illustrates this point by a simple design example.

Given this situation, the following questions arise. What is the optimal constrained H_∞ norm, μ_∞, for the infinite-horizon problem? Is μ_∞ achieved, and if yes, what are the properties of the constrained optimal solution t^*? How can this solution be obtained or approximated? It is shown in [9] that under certain assumptions t^* exists, and a sequence of finite-horizon solutions t^n obtained by extending the horizon converges to t^* in the so-called normal sense; that is, there exists a subsequence t^{n_k} of t^n that converges to t^* uniformly on compact sets of $\mathcal{D}^c \doteq \{z \in \mathcal{C} \ / \ |z| > 1\}$ [1]. However, in the approach of [9] an approximating sequence is difficult to construct.

*Received by the editors January 19, 1995; accepted for publication (in revised form) May 2, 1996.

http://www.siam.org/journals/sicon/35-4/25689.html

†Department of Mechanical and Aerospace Engineering, University of California, Irvine, Irvine, CA 92717 (sideris@euclid.eng.uci.edu). The research of this author was supported in part by NSF contract ECS-9214993.

‡Department of Electrical Engineering, Technion-Israel Institute of Technology, Haifa 32000, Israel (hector@ee.technion.ac.il).

In this note, we take a different approach to the infinite-horizon constrained H_∞ problem. As is observed in [10], one can sacrifice optimality in the finite-horizon H_∞ optimization problems in order to obtain good behavior of the tail response. This is achieved in [10] by placing the closed-loop poles inside a disk $\mathcal{D}_{1/\rho}$ of radius $1/\rho$, $\rho > 1$, instead of the unit disk by means of a scaling $z \longleftarrow \rho \cdot z$, which is undone once the transfer function of the controller has been obtained. A similar device, although in a different setting, has been recently used in [11]. Examples in [10] show that extending the horizon coupled with such a ρ-scaling is very effective in producing a response that satisfies the constraints over the infinite horizon. However, no analysis has been given relating the solutions obtained in this manner to the optimal constrained H_∞ solution over the infinite horizon. The main result of this paper is to show that by taking the limit $\rho \to 1$ and by extending the time horizon *appropriately*, a sequence of real rational solutions is obtained that converges *uniformly* to the constrained optimal solution which is shown to be *unique*. Thus, we obtain complete answers to the questions raised above and most importantly a constructive procedure to approximate the optimal constrained H_∞ norm and solution by rational transfer functions. These results are obtained under certain assumptions on the given time-domain constraints which are easily satisfied in practice and can be readily checked. Section 4 contains our main results, while the most technical proofs are relegated to the appendix.

Our results are illustrated by a simple design example in section 5, and section 6 contains the conclusions.

Notation. L_∞ denotes the Lebesgue space of complex-valued functions which are essentially bounded on the unit circle, with norm $\|g\|_\infty \doteq \text{ ess } \sup_{\theta\in[0,2\pi]} |g(e^{j\theta})|$. By H_∞ we denote the set of functions $g(z) \in L_\infty$ that are analytic outside the closed unit disk and bounded on the unit circle. The H_∞ norm is defined as

$$\|g\|_\infty \doteq \text{ ess } \sup_{|z|>1} |g(z)|.$$

Under this definition, z^{-1} represents the unit delay operator. Finally, we define $g^\sim(z) \doteq g(1/z)$.

2. Problem formulation. Let $p(z)$ be the nominal plant, $k(z)$ be the controller to be designed, and $t(z) \doteq p(z)k(z)[1 + p(z)k(z)]^{-1}$ be the complementary sensitivity transfer function which represents the command response transfer function in the closed-loop system of $p(z)$ and $k(z)$. As $t(z)$ must be stable, it has an expansion

$$t(z) = \sum_{i=0}^{\infty} t_i z^{-i}.$$

The problem studied in this paper is the following.

PROBLEM 1 (H_∞ optimization with time-domain constraints over an infinite horizon). *Given two sequences* $\{ub_i\} = \{ub_0, ub_1, ub_2, \ldots\}$ *and* $\{lb_i\} = \{lb_0, lb_1, lb_2, \ldots\}$, *design a controller* $k(z)$ *such that*

(a) $k(z)$ *is internally stabilizing;*

(b) *the constraints*

$$lb_i \le t_i \le ub_i \qquad \forall\ i \ge 0$$

are satisfied; and

(c) $\|t\|_\infty$ *is minimized.* □

In [10, 8], the following problem is solved instead.

PROBLEM 2 (H_∞ optimization with time-domain constraints over a finite horizon). *Given two finite sequences* $\{ub_i\} = \{ub_0, ub_1, ub_2, \dots, ub_{n-1}\}$ *and* $\{lb_i\} = \{lb_0, lb_1, lb_2, \dots, lb_{n-1}\}$, *design a controller* $k(z)$ *such that*

(a) $k(z)$ *is internally stabilizing;*

(b) *the constraints*

$$lb_i \le t_i \le ub_i \qquad \forall\ i = 0, \dots, n-1$$

are satisfied; and

(c) $\|t\|_\infty$ *is minimized.* ∎

As observed in [10] and the introduction, as the horizon length $n \to \infty$, then the solutions will not in general converge in the H_∞-norm sense to a solution of the infinite-horizon problem [9]. Also we should point out that this form of the time-domain constrained H_∞ optimal control problem is considered here mainly for simplicity; the general version of the theory which requires a more elaborated notation and is technically more involved is treated in [8].

3. The finite-horizon case. In this section we summarize the solution of Problem 2 from [10]. By the well-known Youla parametrization lemma (see, e.g. [3]), the set of all admissible (i.e., resulting from internally stabilizing controllers) closed-loop transfer functions $t(z)$ can be expressed in the form

$$t(z) = u(z) - v(z)q(z),$$

where u is stable and v can be selected to be inner, i.e., stable and such that $v^\sim(z)v(z) = 1$ for $|z| = 1$ and both are determined from the problem data, while $q(z)$ is any stable transfer function $q(z)$ (the Youla parameter). Then the minimization of $\|t\|_\infty$ is easily seen to be equivalent to the minimization of $\|r - q\|_\infty$, where $r(z) = v^\sim(z)u(z)$ is real-rational and antistable (i.e., all its poles lie outside the unit disk). Let $u(z) = \sum_{i=0}^{\infty} u_i z^{-i}$, $v(z) = \sum_{i=0}^{\infty} v_i z^{-i}$, and $q(z) = \sum_{i=0}^{\infty} q_i z^{-i}$. Let $\mathbf{t_n} \doteq [t_0\ \cdots\ t_{n-1}]$, $\mathbf{u_n} \doteq [u_0\ \cdots\ u_{n-1}]$, $\mathbf{v_n} \doteq [v_0\ \cdots\ v_{n-1}]$, and

$$\mathbf{q_n} = [q_0 \cdots q_{n-1}].$$

Then it is easy to see that

$$t_i = u_i - \sum_{j=0}^{i} v_{i-j} q_j,$$

or in matrix notation,

$$\mathbf{t_n} = \mathbf{u_n} - \mathbf{q_n} V_n,$$

where

$$V_n = \begin{bmatrix} v_0 & v_1 & \cdots & v_{n-1} \\ 0 & v_0 & \cdots & v_{n-2} \\ \vdots & \vdots & \ddots & \vdots \\ 0 & \cdots & \cdots & v_0 \end{bmatrix}.$$

Introducing the additional notation

$$\mathbf{lb_n} \doteq [lb_0\ lb_1\ \cdots\ lb_{n-1}],$$
$$\mathbf{ub_n} \doteq [ub_0\ ub_1\ \cdots\ ub_{n-1}],$$

it is clear that the constraints up to the nth sample may be written as

$$\mathbf{lb_n} - \mathbf{u_n} \leq -\mathbf{q_n} V_n \leq \mathbf{ub_n} - \mathbf{u_n}.$$

Therefore, under the assumption that constraints are enforced only over a finite horizon, the time-domain constrained H_∞ problem may be formulated as

$$\mu_n = \min_{\substack{q \in H_\infty \\ \mathbf{lb_n} - \mathbf{u_n} \leq -\mathbf{q_n} V_n \leq \mathbf{ub_n} - \mathbf{u_n}}} \|r - q\|_\infty. \tag{1}$$

3.1. Summary of the solution in the finite-horizon case. It is convenient to replace r by its conjugate $g \doteq r^\sim$. The transfer function g is then stable and can be assumed to be strictly proper without loss of generality, by redefining q_0 if necessary. $g(z)$ has a minimal state-space realization

$$g = \left(\begin{array}{c|c} A & b \\ \hline c & 0 \end{array} \right).$$

Let W_c and W_o denote the controllability and observability Gramians of g, respectively, satisfying the discrete-time Lyapunov equations

$$W_c = A W_c A^t + b b^t,$$

$$W_o = A^t W_o A + c^t c,$$

and let w_c and w_o denote positive square roots of W_c and W_o. Then we have the following result.

THEOREM 1. *The solution to Problem* 1 *is obtained as*

$$q^*(z) = \sum_{i=0}^{n-1} q_i^* z^{-i} + z^{-n} q_{tail}^*(z),$$

where $\mathbf{q_n^*} \doteq [q_0^* \cdots q_{n-1}^*]$ *solves the convex minimization problem*

$$\mu_n = \min_{\substack{\mathbf{q_n} \\ \mathbf{lb_n} - \mathbf{u_n} \leq -\mathbf{q_n} V_n \leq \mathbf{ub_n} - \mathbf{u_n}}} \overline{\sigma}\left[W_n(\mathbf{q_n})\right]$$

with

$$W_n(\mathbf{q_n}) = \begin{bmatrix} w_o A^n w_c & w_o A^{n-1} b & \cdots & w_o A b & w_o b \\ c A^{n-1} w_c & c A^{n-2} b & \cdots & w_o b & -q_0 \\ \vdots & \vdots & & \vdots & \vdots \\ c A w_c & c w_c & \cdots & -q_{n-3} & -q_{n-2} \\ c w_c & -q_0 & \cdots & -q_{n-2} & -q_{n-1} \end{bmatrix}$$

and q_{tail}^* *the solution of the standard* H_∞ *optimization problem*

$$\|z^n \left[r - \sum_{i=0}^{n-1} q_i^* z^{-i} \right] - q_{tail}\|_\infty.$$

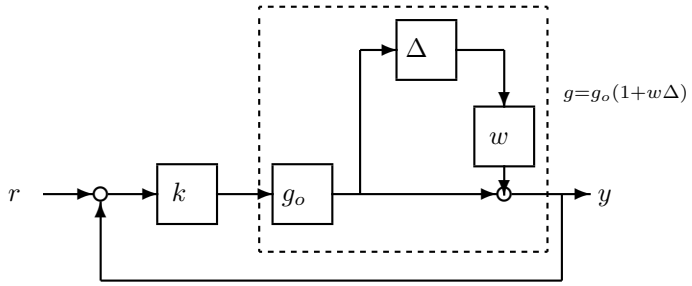


FIG. 1. *A simple design example.*

Proof. See [10]. ∎

A state-space procedure for computing $q(z)$ is given in [8]. It is a well-known fact that the solution to (1) is unique if no constraint is active at the optimum, since in that case the solution is given by the Nehari theorem (see, e.g., [3]).

It can be shown that this is the case even when a finite number of constraints are active, and therefore, the following result is true.

THEOREM 2 (see [9]). *Problem* 2 *has a unique solution.* ∎

3.2. A simple design example. Consider the system of Fig. 1. The nominal plant is given by $g_0(z) = \frac{z+0.2}{z^2-0.6z-1.12}$. We assume multiplicative model uncertainty of the form $g = g_o(1 + w\Delta)$, where $\Delta \in \mathcal{H}_\infty$ and the weighting function $w(z) = 0.3705\frac{z+0.986}{z+0.4682}$. We consider the following design specifications.

(1) The nominal closed-loop system should be asymptotically stable.
(2) For r a unit impulse at $i = 0$, y has a settling time of 10 samples for the nominal system, with a maximum peak of 0.6.
(3) The robustness of the closed-loop system is maximized for the multiplicative model uncertainty considered, i.e.,

$$\min_{k(z)\ \text{stab.}} \|w\ t\|_\infty$$

with $t(z) = \frac{k(z)g_o(z)}{1+k(z)g_o(z)} \equiv \sum_{i=0}^{\infty} t_i z^{-i}$.

As remarked previously, the formulation of Problems 1 and 2 can be easily extended to allow the use of weighting functions such as w above or even to impose time domain constraints on more than one, possibly different transfer functions than the ones involved in the $\mathcal{H}_\infty$ optimization [10, 8].

Standard H_∞ theory gives an optimal H_∞ norm of 0.66 but an impulse response that violates the second specification (Fig. 2). When we apply the method of constrained H_∞ optimization over finite horizons of lengths 30, 59, and 250 samples, we obtain the results depicted in Figs. 3, 4, and 5, respectively. These results clearly show that the oscillations immediately after the horizon cannot be controlled by simply extending the horizon.

4. Main results. In this section, we obtain a sequence of suboptimal solutions to finite-horizon constrained H_∞ optimal control problems that converges *uniformly* to the solution of the infinite horizon constrained H_∞ optimal control problem (Problem 1).

In proving the results of the paper, we make certain assumptions on the time-domain constraints. We collect these assumptions in the following subsection.

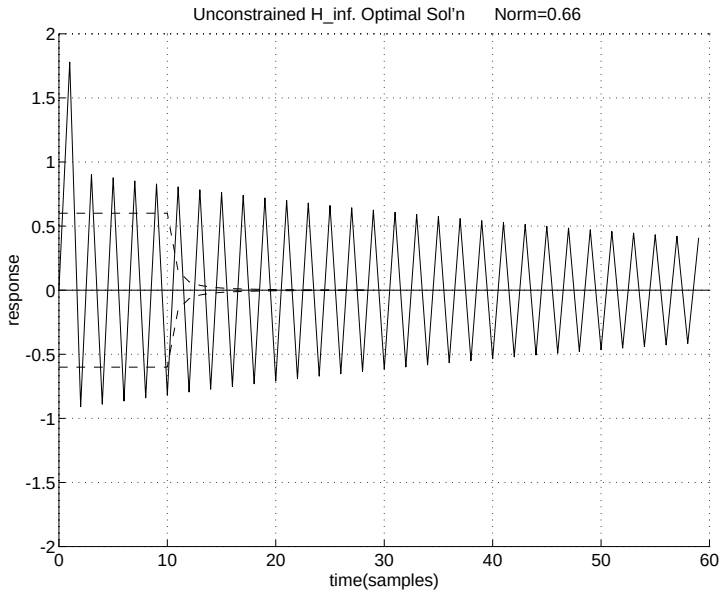


FIG. 2.

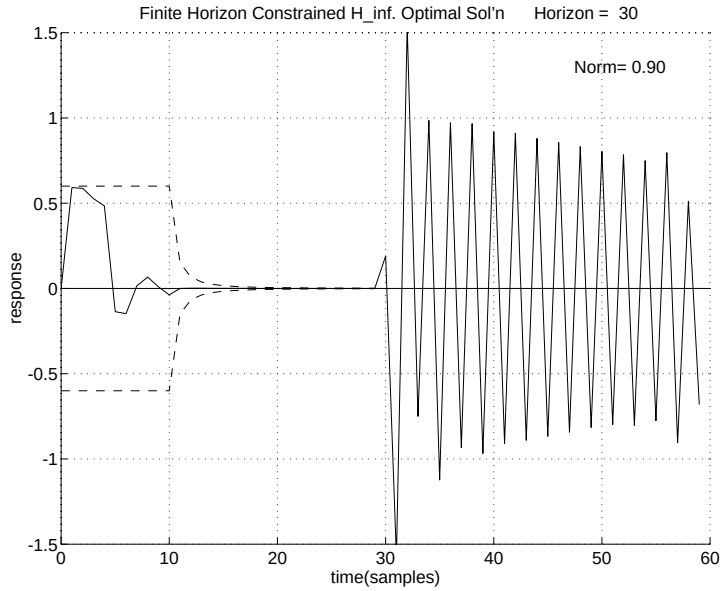


FIG. 3.

4.1. Assumptions on the time-domain constraints.

Assumption 1. The sequences $\{lb_i\}$ and $\{ub_i\}$ that define the time-domain constraints satisfy

$$\sum_{i=0}^{\infty} |lb_i - t_\infty| < \infty, \qquad \sum_{i=0}^{\infty} |ub_i - t_\infty| < \infty,$$

where $t_\infty \doteq \lim_{i\to\infty} t_i$ is the desired asymptotic value for t_i. ∎

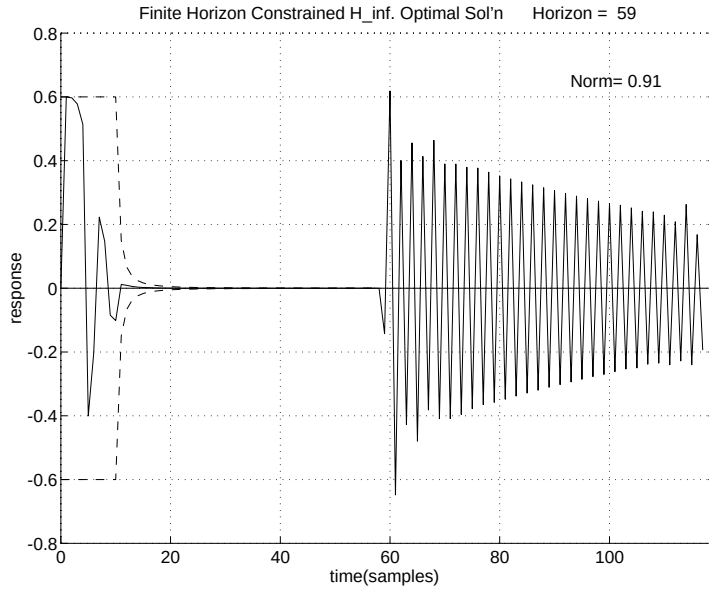


FIG. 4.

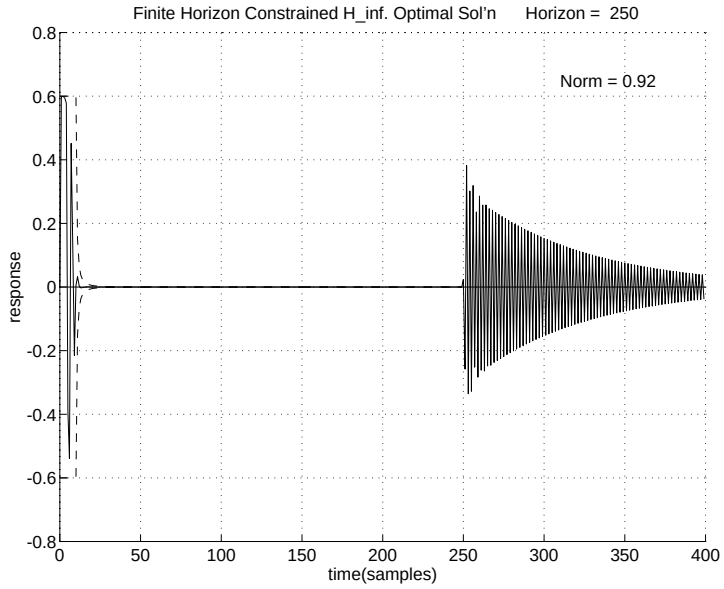


FIG. 5.

Assumption 2. There exists N_1 so that the time-domain constraints satisfy

$$lb_i \leq t_\infty \leq ub_i, \quad i \geq N_1,$$

where t_∞ is the desired asymptotic value for t_i. □

Assumption 3. The sequences $\{lb_i\}$ and $\{ub_i\}$ that define the time-domain constraints satisfy

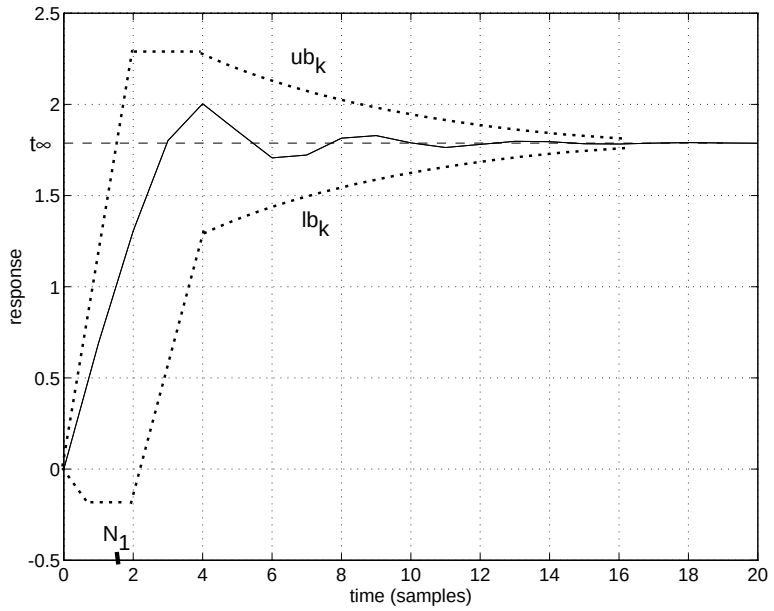


FIG. 6.

$$\sum_{i=0}^{\infty} \alpha^{-i} |lb_i - t_\infty| = \infty, \quad \sum_{i=0}^{\infty} \alpha^{-i} |ub_i - t_\infty| = \infty$$

for all $\alpha \in (0\ 1)$ and where t_∞ is the desired asymptotic value for t_i. □

The above assumptions are easily satisfied in all cases of practical interest.

This is illustrated by the typical step response bounds depicted in Fig. 6.

We also remark that Assumptions 1 to 3 are expressed in terms of the *given time response bounds*, that is, in terms of the problem data. Assumption 1 combined with Assumption 3 ensures that the sequences $\{ub_i\}$ and $\{lb_i\}$ converge to t_∞ but not exponentially fast. Assumption 2 requires that $\{ub_i\}$ and $\{lb_i\}$ converge to t_∞ from above and below, respectively.

Next, let Ω^n be the set of H_∞ functions that satisfy the time-domain constraints over the finite horizon of length n *and* the interpolation constraints necessary for nominal closed-loop stability. Each of the latter takes the form

$$\sum_{i=0}^{\infty} t_i w_k^{-i} = v_k, \quad k = 1, \ldots, r, \tag{2}$$

where w_k is plant pole or zero in $\mathcal{D}^c$ and $v_k = 1$ or 0, respectively (assuming for simplicity that w_k is of multiplicity one) [12]. Ω^n is then the feasible set for Problem 2. Similarly, Ω^∞ is the feasible set for Problem 1 and consists of the H_∞ functions satisfying the time-domain constraints over the infinite horizon and the stability interpolation constraints. We remark that Ω^∞ will generically either be empty or contain more than one point. Indeed, the case that Ω^∞ contains only one point is obtained when the hyperplane of the interpolation constraints (see (2)) intersects the set of the time-domain constraints only at one of its extreme points.

We also require the following assumption.

Assumption 4. *Ω^∞ contains a point $t^{FD}(z) = \sum_{i=0}^{L-1} t_i^{FD} z^{-i}$ of finite duration.*

Assumption 4 may appear at first to be restrictive. However, Lemma 1 of the next subsection shows that Assumption 4 is generically equivalent with the condition that Ω^∞ is not empty, that is, the existence of feasible solutions for Problem 1. Assumption 4 can be verified in practice by solving a linear programming problem. Consequently, Lemma 1 also gives a finite procedure to verify that Ω^∞ is not empty.

4.2. Preliminary results. In the remainder of the paper, we assume for simplicity that the asymptotic value $t_\infty = 0$ ($t_\infty = \lim_{i\to\infty} lb_i \equiv \lim_{i\to\infty} ub_i$ and, therefore, is a part of the problem data).

This is the case, for example, when t is the impulse response of a closed-loop rational transfer function. However, in case that $t_\infty \neq 0$, one should consider $t(z) - t_\infty$, $lb_k - t_\infty$, $ub_k - t_\infty$ in place of $t(z)$, lb_k, and ub_k, respectively, as well as $v_k - t_\infty$ in place of v_k in the interpolation constraints (2).

The following lemma provides justification for Assumption 4.

LEMMA 1. *Assume that the time-domain constraints satisfy Assumptions* 1–3 *and that t_∞, the desired asymptotic value for the solution, is zero. Then Ω^∞ is empty or it generically contains a point $t^{FD}(z) = \sum_{i=0}^{L-1} t_i^{FD} z^{-i}$ of finite duration.*

Proof. If Ω^∞ is empty, there is nothing to prove. Therefore, let us assume that Ω^∞ is not empty. By the previous discussion, Ω^∞ generically contains two functions $t(z)$, $s(z)$ such that $t(z) \neq s(z)$. Define

$$e(z) \doteq t(z) - s(z) = \sum_i e_i z^{-i},$$

and consider the stability interpolation constraints (2), assuming for ease of exposition that $w_k \in \mathbb{R}$, $k = 1, \ldots, r$. Then, e has at least $r+1$ nonzero coefficients. Indeed, let $e_i \neq 0$ for $i_1, \ldots, i_l$ (by hypothesis $l \geq 1$). It holds that

$$[e_{i_1} \;\; e_{i_2} \;\; \cdots \;\; e_{i_l}] \underbrace{\begin{bmatrix} 1 & 1 & \cdots & 1 \\ w_1^{-i_1} & w_2^{-i_1} & \cdots & w_r^{-i_1} \\ w_1^{-i_2} & w_2^{-i_2} & \cdots & w_r^{-i_2} \\ \vdots & \vdots & \vdots & \vdots \\ w_1^{-i_l} & w_2^{-i_l} & \cdots & w_r^{-i_l} \end{bmatrix}}_{\mathbf{W_r}} = [0 \;\; 0 \;\; \cdots \;\; 0].$$

Since $\mathbf{W_r}$ is full rank for $l < r+1$, it should be $l \geq r+1$.

Next, take

$$\hat{t}(z) \doteq \lambda t(z) + (1-\lambda) s(z)$$

for some $0 < \lambda < 1$. Clearly, $\hat{t}(z)$ satisfies the interpolation constraints, and moreover

$$lb_{i_j} \leq \min\{t_{i_j}, s_{i_j}\} < \hat{t}_{i_j} < \max\{t_{i_j}, s_{i_j}\} \leq ub_{i_j}.$$

Let $\epsilon = \min_{j=1,\ldots,r} \left\{ \hat{t}_{i_j} - lb_{i_j},\ ub_{i_j} - \hat{t}_{i_j} \right\} > 0$. This means that the coefficients $\hat{t}_{i_j}$, $j = 1, \ldots, r$ can be changed by ϵ without violating the time domain constraints. By Assumption 2, it holds that $lb_i \leq 0 \leq ub_i$ for some $i > N_1$. We next determine $L > N_1$ such that after dropping from $\hat{t}$ the tail

$$\hat{t}^{tail}(z) = \sum_{i=L}^{\infty} \hat{t}_i z^{-i} = z^{-L} \sum_{i=0}^{\infty} \hat{t}_{i+L} z^{-i},$$

$\hat{t}_{i_j},\ j=1,\dots,r$, in $\hat{t}^{FD}(z)=\sum_{i=0}^{L-1}\hat{t}_i z^{-i}$ can be modified by no more than ϵ to have $\hat{t}^{FD}(z)$ satisfy the stability interpolation constraints.

Since

$$\left|\hat{t}^{tail}(w_k)\right| = |w_k|^{-L}\left|\sum_{i=0}^{\infty}\hat{t}_{i+L}w_k^{-i}\right| \le |w_k|^{-L}\,\|\hat{t}\|_1,$$

choose L such that

$$|w_k|^{-L}\,\|\hat{t}\|_1 < \epsilon/\left\|\mathbf{W}\mathbf{r}^{-1}\right\|_1,$$

and compute

$$[\Delta_1\ \ \Delta_2\ \ \cdots\ \ \Delta_r] = [\hat{t}^{tail}(w_1)\ \ \hat{t}^{tail}(w_2)\ \ \cdots\ \ \hat{t}^{tail}(w_r)]\mathbf{W}\mathbf{r}^{-1}.$$

It follows that $|\Delta_i| \le \left\|\mathbf{W}\mathbf{r}^{-1}\right\|_1 \cdot \max_{j=1,\dots,r}\left\{|\hat{t}^{tail}(w_j)|\right\} < \epsilon$. Finally, define

$$t_i^{FD} = \begin{cases} \hat{t}_i, & 0\le i<L,\ i\ne i_1,\dots,i_r,\\ \hat{t}_i+\Delta_i, & i=i_1,\dots,i_r,\\ 0, & i\ge L,\end{cases}$$

and $t^{FD}=\sum_{i=0}^{\infty}t_i^{FD}z^{-i}$.

It is clear by construction that $t^{FD}\in\Omega^\infty$, and the proof is complete. ∎

We also have the following corollary to Lemma 1.

COROLLARY 1. *Consider Assumptions 1–4, and let $t\in\Omega^\infty$. Then there exists a sequence $\hat{t}^n\in\Omega^\infty$, $n=1,2,\dots$, of finite duration functions that converges uniformly to t.*

Proof. In the proof of Lemma 1, take $s(z)$ to be the finite duration point of Assumption 4, and let λ be sufficiently close to unity, so that $\|t-\hat{t}\|_\infty \le 1/(2n)$. Pick $\epsilon_1=\min\{\epsilon,1/(2nr)\}$ and L such that

$$|w_k|^{-L}\,\|\hat{t}\|_1 < \epsilon_1/\left\|\mathbf{W}\mathbf{r}^{-1}\right\|_1.$$

Then $|\Delta_i|<\epsilon$ and $\sum_{j=1}^{r}|\Delta_i|\le 1/(2n)$. It follows that $\left\|\hat{t}-t^{FD}\right\|_\infty \le \left\|\hat{t}-t^{FD}\right\|_1 \le 1/(2n)$ and $\|t-t^{FD}\|_\infty < 1/n$. Identify $\hat{t}^n$ with t^{FD}, and the proof is complete. ∎

Next, consider the disk of radius $1/\rho$

$$D_{1/\rho} = \{z\ /\ |z|<1/\rho\},\qquad \rho>1, \tag{3}$$

and let $t(z)$ have its poles inside $D_{1/\rho}$. We define the $\rho\infty$-norm of t as

$$\|t\|_{\rho\infty} \doteq \sup_{-\pi\le\theta\le\pi}\left|t\left(\frac{1}{\rho}e^{j\theta}\right)\right|. \tag{4}$$

LEMMA 2. *Let $t(z)=\sum_{k=0}^{\infty}t_k z^{-k}$ have its poles in $D_{1/\rho}$, and assume that the $\rho\infty$-norm of t satisfies*

$$\|t\|_{\rho\infty}\le M$$

for some constant M. Then

$$|t_k|\le M\rho^{-k}\quad\forall\ k.$$

Proof. It is well known that

$$t_k = \frac{1}{2\pi j} \oint_C t(z) z^{(k-1)} dz,$$

where C is a closed curve inside the region of convergence. Since all the singularities of $t(z)$ are confined inside a disk of radius $1/\rho$,

$$t_k = \frac{1}{2\pi} \int_{-\pi}^{\pi} t(e^{j\theta}/\rho) \rho^{-k} e^{jk\theta} d\theta.$$

This implies $|t_k| \le M\rho^{-k}$, and the result follows. ∎

4.3. Construction of suboptimal solutions in Ω^∞. We next describe a procedure to construct suboptimal solutions to Problem 1 which are in Ω^∞; that is, they satisfy the time-domain constraints over the infinite horizon. This is achieved by obtaining solutions to Problem 2 that have poles inside $D_{1/\rho}$ (see (3)) for $\rho > 1$ and for a horizon length n_ρ which is defined later. The construction is given in the following theorem.

THEOREM 3. *Assume that the time-domain constraints satisfy Assumptions* 1–4. *Fix* $\rho_o > 1$, *and let* $\rho_o > \rho > 1$.

Define

$$\hat{t}(z) \doteq t(\rho z) = \sum_{k=0}^{\infty} \underbrace{(t_k \rho^{-k})}_{\hat{t}_k} z^{-k}. \tag{5}$$

Also define an "equivalent" constraint set $\hat{\Omega}^n$ *for* $\hat{t}(z)$ *defined by bounds* $\hat{lb}_k$ *and* $\hat{ub}_k$ *derived from*

$$lb_k \le t_k \le ub_k \iff \underbrace{lb_k \rho^{-k}}_{\hat{lb}_k} \le \hat{t}_k \le \underbrace{ub_k \rho^{-k}}_{\hat{ub}_k}, \qquad k = 0, 1, \ldots, n-1,$$

and by the scaled interpolation constraints

$$\sum_{i=0}^{\infty} \hat{t}_i \underbrace{\left(\frac{w_k}{\rho}\right)}_{\hat{w}_k}^{-i} = v_k.$$

Let $\hat{t}^n$ *be the unique solution of the finite-horizon constrained* $\mathcal{H}_\infty$ *optimization problem*

$$\min_{\hat{t} \in \hat{\Omega}^n} \|\hat{t}\|_\infty,$$

and define

$$t^n(z) \doteq \hat{t}(z/\rho).$$

Then $t^n(z)$ *has the following properties.*

(1) t^n *is feasible over the finite horizon of length* n.

(2) t^n *has its poles inside* $D_{1/\rho} = \{z \ / \ |z| < 1/\rho\}$.

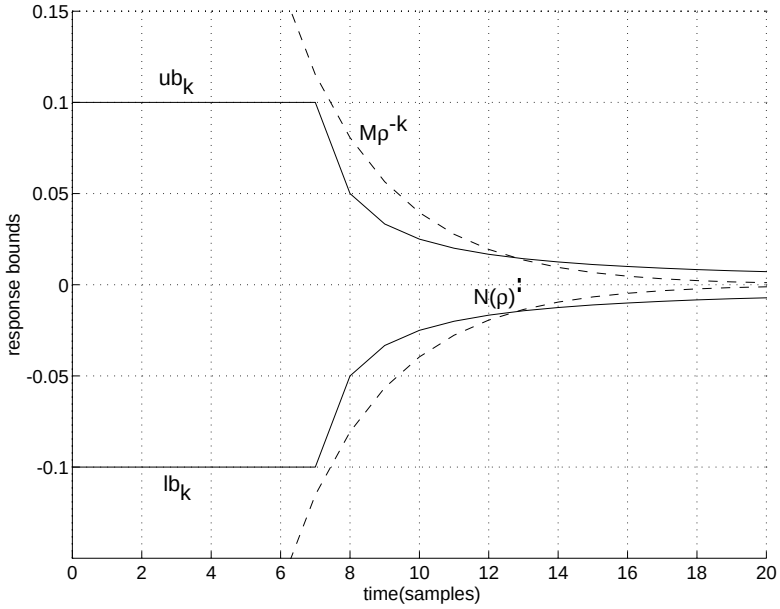


FIG. 7.

(3) *The $\rho\infty$-norm of t^n (see (4)) satisfies*

$$\|t\|_{\rho\infty} < M \quad \forall\ 1 < \rho < \rho_o,$$

where M is some constant depending only on the problem data and independent of both ρ and n.

(4) *There exists a horizon length n_ρ such that $t^{n_\rho} \in \Omega^\infty$; i.e., t^{n_ρ} is a feasible point for Problem* 1.

Proof. The first assertion follows from the construction of the set $\hat{\Omega}^n$. The second assertion is true because of the ρ-scaling in (5) and since $\hat{t}^n$ has its poles inside the unit disk. To show the third assertion consider the finite duration feasible point $t^{FD}(z)$ of Assumption 4 and let $M = \|t^{FD}\|_{\rho_o\infty}$. Then note that

$$\|t^n\|_{\rho\infty} \le \|t^{FD}\|_{\rho\infty} \le M \quad \forall\ \rho < \rho_o,$$

where the first inequality follows by the fact that t^{FD} is a feasible solution of Problem 2 for every ρ and t^n minimizes $\|t\|_{\rho\infty}$ over all $t \in \Omega^n$ by construction, and the second inequality follows by the maximum modulus principle. The last part of the proof is obtained by noting that t^n satisfies the assumptions of Lemma 2, and therefore it holds that

$$|t_k^n| \le M\rho^{-k} \quad \forall\ k \tag{6}$$

independently of the horizon length n. From Assumption 3, the bounds on the time-domain response converge to the steady-state value t_∞ less rapidly than the exponential decay prescribed by (6), and from Assumption 2, they remain on different sides of t_∞ after some $i > N_1$. Therefore, there is a time n_ρ after which the time-domain constraints are automatically satisfied by t^n (see Fig. 7). □

We are now in a position to state the main result of the paper.

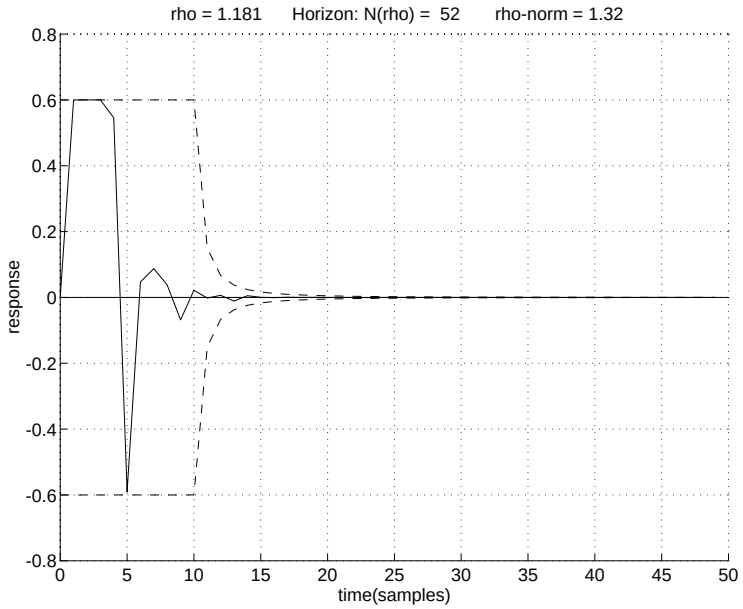


FIG. 8.

THEOREM 4. *Assume that the time-domain constraints satisfy Assumptions* 1–4. *Consider a sequence* ρ_n *such that* $\rho_n > 1 \ \forall \ n$ *and* $\rho_n \downarrow 1$ *as* $n \to \infty$ (*for example,* $\rho_n = 1 + \frac{1}{n}$). *Consider also* t^n *obtained from Theorem* 3 *for* $\rho = \rho_n$. *Then the following hold.*

(1) t^n *is feasible and unique for all* n.

(2) $\|t^n\|_{\rho_n \infty}$ *is monotonically decreasing and converges to* μ_∞, *the optimal constrained* H_∞ *norm over* Ω^∞.

(3) *The optimal norm in Problem* 1 *is achieved by a unique solution* t^*.

(4) $t^n \to t^*$ *uniformly.*

Proof. The proof of this result is rather technical and is included in the appendix. □

To illustrate Theorem 4 and the procedure of approximating the optimal solution, we apply it to the example of section 3.2. The results for different values of ρ are shown in Figs. 8–13. On top of each figure, we give the value of the ρ-scaling used, the horizon length n_ρ of Theorem 3, and the ρ-norm achieved by t_n—an upper bound on μ_∞.

We remark that it is also possible to derive a sequence of lower bounds converging to μ_∞ and thus obtain a meaningful stopping criterion based on the difference between the sequences of lower and upper bounds. Indeed, the H_∞-norm of the solution of Problem 2 for a horizon of length n_ρ and without the scaling of Theorem 3 gives such a sequence of lower bounds (see Figs. 3–5).

5. Conclusions. Theorem 4 resolves Problem 1 in a most satisfactory way. Under the mild Assumptions 1–4 on the time-domain constraints, Theorem 4 guarantees the existence of a unique optimal solution that can be approximated uniformly (that is, with approximation error measured in terms of the H_∞-norm) by rational functions. Furthermore, Theorem 4 gives a constructive procedure to obtain such approximations by solving finite-horizon problems. We remark that the horizon in each of these

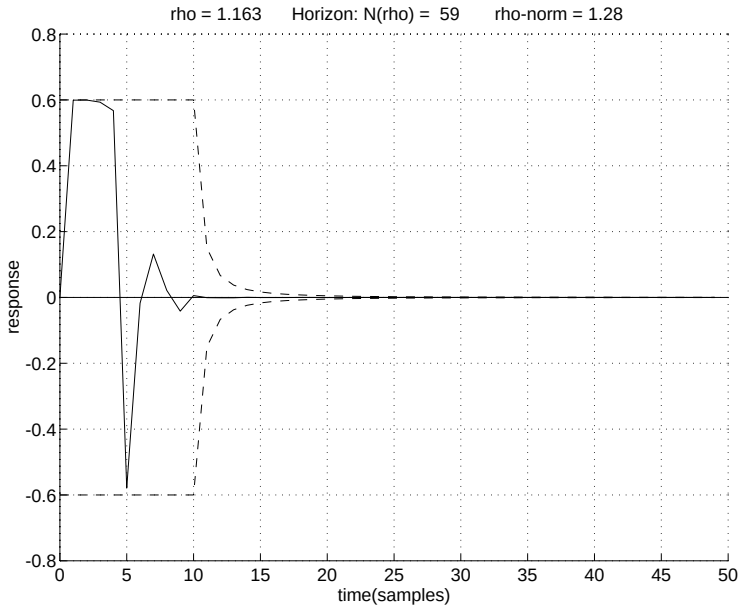


FIG. 9.

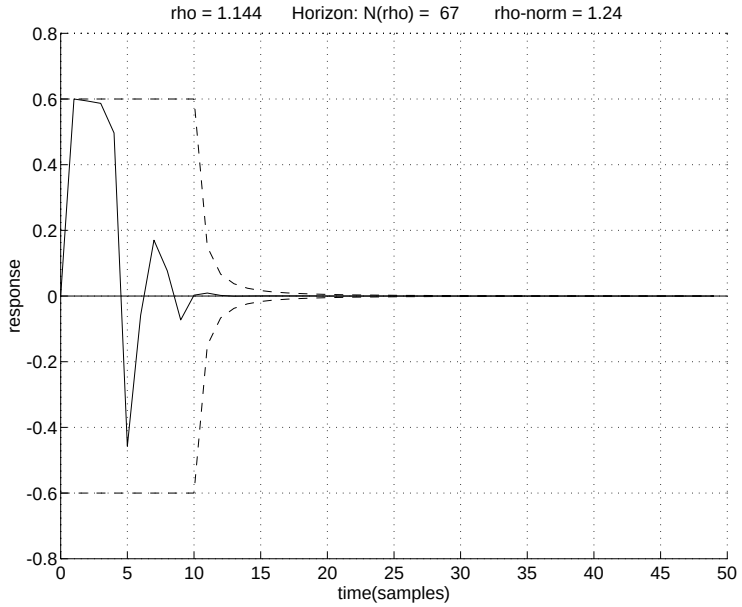


FIG. 10.

problems is determined from the given time-domain constraints and does not depend on the solution of the problem.

Although the basic concepts and some of the analysis used here carry through in the multivariable case, the latter apparently requires a different treatment. The main complications in the multivariable case arise from the nonuniqueness of the solution of

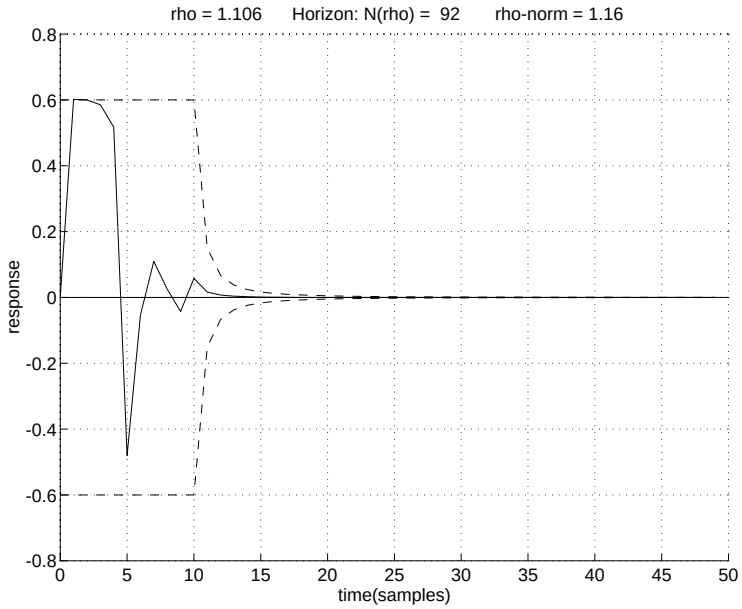


FIG. 11.

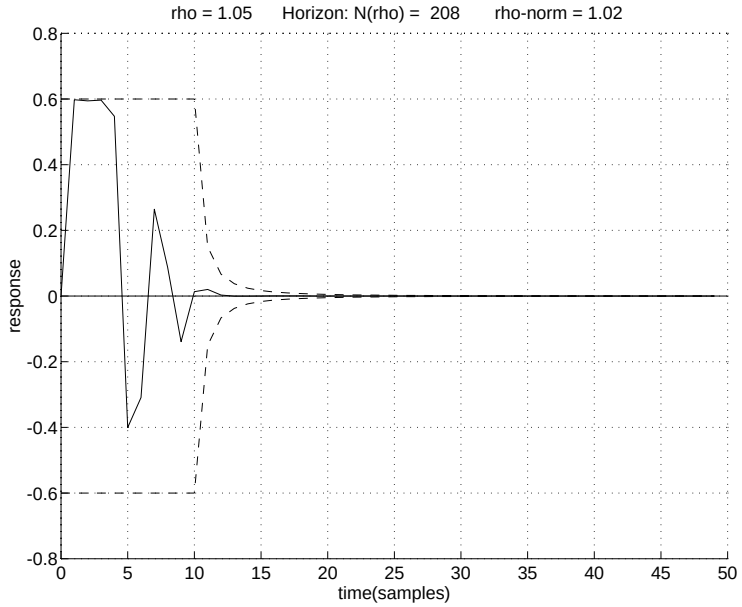


FIG. 12.

Problem 2 and the less straightforward representation of the interpolation constraints. One can equivalently pose the problem in terms of $Q \in H_\infty$, the Youla parameter of the Q-parametrization of all feedback stabilizing controllers. In this manner, the interpolation constraints are automatically satisfied but the time-domain constraints as reflected on Q are more complicated than the bound constraints assumed on t

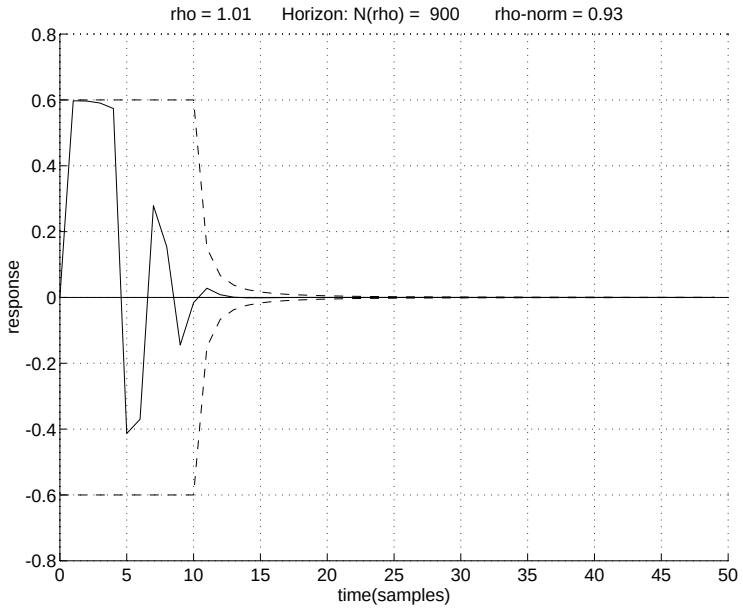


FIG. 13.

(see section 3 and [8]). However, experimentation with computer simulations in the multivariable case shows that the technique of ρ-scaling produces constrained solutions that satisfy the time-domain constraints over the infinite horizon and suggests that similar results with the ones obtained here for the SISO case are to be expected in the multivariable case also.

Appendix A. Proof of Theorem 4.

The proof of Theorem 4 is based on the following lemmas.

LEMMA 3. *Under Assumption* 1, Ω^∞ *is compact in the* ∞*-norm sense.*

Proof. Compactness of Ω^∞ is shown in [7]. We include the proof here for the reader's convenience. It is first shown that Ω^∞ is compact as a subset of the space l_1. Then Lemma [3] follows since the l_1 norm bounds the H_∞ norm.

Consider the set $\bar{\Omega}^\infty \subset l_1$ defined only by the time-domain constraints. The space l_1 is complete and $\bar{\Omega}^\infty$ is closed, and hence it suffices to show that for every $\epsilon > 0$ it is possible to cover $\bar{\Omega}^\infty$ with a finite number of balls in l_1 with radius ϵ. So let $\epsilon > 0$ be given. By Assumption 1, there exists N such that

$$\sum_{i=N}^{\infty} |lb_i - t_\infty| < \epsilon/4, \tag{7}$$

$$\sum_{i=N}^{\infty} |ub_i - t_\infty| < \epsilon/4, \tag{8}$$

implying $\sum_{i=N}^{\infty} |t_i - t_\infty| < \sum_{i=N}^{\infty} |lb_i - t_\infty| + \sum_{i=N}^{\infty} |ub_i - t_\infty| \le \epsilon/2 \ \forall\, t = \sum_{i=0}^{\infty} t_i z^{-i} \in \bar{\Omega}^\infty$. Now consider the set

$$\bar{\Omega}^N \doteq \{[t_0\ t_1\ \cdots\ t_{N-1}] \in \mathcal{R}^N \text{ s.t. } lb_i \le t_i \le ub_i\}.$$

$\bar{\Omega}^N$ is a closed and bounded set in $\mathcal{R}^N$ and hence there exists a finite number of balls

$\mathcal{B}(\hat{t}^k, \epsilon/2)$, centered at $\hat{t}^k \doteq [t_0^k \; t_1^k \; \cdots \; t_{N-1}^k]$ and of radius $\epsilon/2$ that cover $\bar{\Omega}^N$. Define

$$t^k \doteq [t_0^k \; t_1^k \; \cdots \; t_{N-1}^k \; t_\infty \; t_\infty \; \cdots],$$

and consider the balls $\mathcal{B}(t^k, \epsilon) \subset l_1$. Consider now any $t \in \bar{\Omega}^\infty$. By construction, there exist a k such that $\sum_{i=0}^{N-1} |t_i - t_i^k| < \epsilon/2$, and thus

$$\|t - t^k\|_1 \leq \sum_{i=0}^{N-1} |t_i - t_i^k| + \sum_{i=N}^{\infty} |t_i - t_\infty|, \tag{9}$$

$$\|t - t^k\|_1 < \epsilon/2 + \epsilon/2 = \epsilon. \tag{10}$$

Since the finite number of balls $\mathcal{B}(t^k, \epsilon)$ cover $\bar{\Omega}^\infty$, the set is compact.

Next, note that $\Omega^\infty \equiv \bar{\Omega}^\infty \cap \mathcal{I}$, where $\mathcal{I} \subset l_1$ is the set of functions satisfying the interpolation constraints (2). Since $\mathcal{I}$ is closed, it follows that Ω^∞ is also compact. □

The next lemma is instrumental in establishing uniqueness of the optimal solution in Problem 1.

LEMMA 4. *Let P be a continuous functional defined on a compact subset $\mathcal{B}$ of a Banach space, and consider the minimization*

$$\min_{x \in \mathcal{B}} P(x).$$

Suppose that x^* achieves the minimum $v^* = P(x^*)$ and that x^* is the unique minimizer. Then, given $\epsilon > 0$, there exists $\delta > 0$ such that if x_1, $x_2 \in \mathcal{B}$ with $v_1 = P(x_1)$, $v_2 = P(x_2)$ satisfying $|v_1 - v^*| \leq \delta$ and $|v_2 - v^*| \leq \delta$, it holds that $\|x_1 - x_2\| \leq \epsilon$.

Proof. Let us define the set $E_\delta = \{x \in \mathcal{B} / |P(x) - v^*| \leq \delta\}$, and let $d(\delta)$ be defined as

$$d(\delta) = \max_{x \in E_\delta} \|x - x^*\|.$$

Note that $d(\delta)$ is an increasing function of δ. We claim that $d(\delta) \to 0$ as $\delta \to 0$; that is, given ϵ, there exists δ such that $x \in E_\delta \implies d(\delta) \leq \epsilon$. Indeed, suppose that the last assertion is not true. Then for all δ, there exists $x_\delta \in \mathcal{B}$ such that $|P(x_\delta) - v^*| \leq \delta$ and $\|x_\delta - x^*\| > \epsilon$. Consider a sequence $\delta_n \to 0$ as $n \to \infty$ and the corresponding vectors $x_n \equiv x_{\delta_n}$. Since $\mathcal{B}$ is compact, $\{x_n\}$ has a convergence subsequence $y_n \equiv \{x_{n_k}\} \longrightarrow y^* \in \mathcal{B}$. Note that $\|y^* - x^*\| > \epsilon$ and the continuity of P implies $P(y^*) = \lim_{n \to \infty} P(y_n) = v^*$. This contradicts the hypothesis that x^* is the unique minimizer of P, and therefore, $d(\delta) \to 0$ as $\delta \to 0$ follows. Now given $\epsilon > 0$, let δ be such that $d(\delta) \leq \epsilon/2$, and assume that $|v_1 - v^*|$, $|v_2 - v^*| \leq \delta$. But then x_1, $x_2 \in E_\delta$ implying $\|x_1 - x_2\| \leq \|x_1 - x^*\| + \|x_2 - x^*\| \leq 2d(\delta) \leq \epsilon$, and the proof is complete. □

Next, we prove Theorem 4.

Proof of Part (1). The feasibility of t^n over the infinite horizon follows from Theorem 3, part (4). The uniqueness of t^n follows from Theorem 2. □

Proof of Part (2). Let $\mu_n \equiv \|t^n\|_{\rho_n \infty} \doteq \max_{|z| = \frac{1}{\rho_n}} |t^n(z)|$, and consider $1 < \rho_{n_2} < \rho_{n_1}$. Then it holds that

$$\mu_{n_2} = \max_{|z| = \frac{1}{\rho_{n_2}}} |t^{n_2}(z)| \leq \max_{|z| = \frac{1}{\rho_{n_2}}} |t^{n_1}(z)| \leq \max_{|z| = \frac{1}{\rho_{n_1}}} |t^{n_1}(z)| = \mu_{n_1},$$

where the first inequality above follows from the optimality of t^{n_2} and the fact that both t^{n_1} and t^{n_2} are feasible over the infinite horizon, and the second inequality follows from the maximum modulus principle. Since μ_n is monotone decreasing and bounded below, it converges to a number μ^*. Furthermore, since each t^n is feasible for the constrained H_∞ minimization over Ω^∞, it holds that

$$\mu_\infty \leq \|t^n\|_\infty \leq \mu_n.$$

Thus, we have

$$\mu_\infty \leq \mu^*. \tag{11}$$

Next, we show that the reverse inequality is true. Lemma 3 guarantees that μ_∞ in Problem 1 is achieved by some optimal solution $t^* = \sum_{i=0}^{\infty} t_i^* z^{-i}$.

Consider the sequence in Corollary 1, of finite-duration functions $\hat{t}^n \in \Omega^\infty$ that converges uniformly to t^*. By the optimality of t^n it holds that $\|\hat{t}^n\|_{\rho\infty} \geq \|t^n\|_{\rho\infty}\ \forall\ n$. Also by the maximum modulus principle, we have

$$\epsilon_n \doteq \|\hat{t}^n\|_{\rho\infty} - \|\hat{t}^n\|_\infty \geq 0. \tag{12}$$

Thus, we obtain

$$\|\hat{t}^n\|_\infty \geq \|t^n\|_{\rho\infty} - \epsilon_n. \tag{13}$$

Now note that $\epsilon_n \to 0$ as $n \to \infty$. (All functions $\hat{t}^n$ are uniformly bounded, and therefore they define an equicontinuous family of functions [1, p. 224]. Thus, given ϵ there exists δ such that

$$|z_1 - z_2| \leq \delta \implies |\hat{t}^n(z_1) - \hat{t}^n(z_2)| \leq \epsilon \quad \forall n. \tag{14}$$

Now take ϵ, let δ so that (14) holds, and consider N such that for $n \geq N$ it holds that $\rho_n \leq \frac{1}{1-\delta}$.

Then (14) implies

$$\left|\hat{t}^n\left(\frac{1}{\rho_n}e^{j\theta}\right) - \hat{t}^n(e^{j\theta})\right| \leq \epsilon \ \ \forall\theta \in [-\pi\ \pi] \quad n \geq N$$
$$\implies \left|\|\hat{t}^n\|_{\rho\infty} - \|\hat{t}^n\|_\infty\right| \leq \epsilon \quad n \geq N,$$

and $\epsilon_n \to 0$ follows from (12).)

By taking the limit in (13), we obtain

$$\mu_\infty \geq \mu^*, \tag{15}$$

and from (11) and (15) we obtain $\mu_\infty = \mu^*$. □

Proof of Part (3). To prove the uniqueness of t^*, let us assume that there exist t^{1*} and t^{2*} in Ω^∞ achieving μ_∞.

Next from Corollary 1, consider sequences $\hat{t}^{1n} \in \Omega^\infty$ and $\hat{t}^{2n} \in \Omega^\infty$ of finite-duration functions, converging uniformly to t^{1*} and t^{2*}, respectively Now note that $\|\cdot\|_{\rho\infty}$ is continuous, Ω^∞ is compact by Lemma 3, and the minimum of the finite-horizon constrained H_∞ optimization problem is unique by Theorem 2 and achieved by t^n. Therefore, Lemma 4 applies, and given ϵ, let δ be as in Lemma 4. From (12) we have $\|\hat{t}^{in}\|_{\rho\infty} - \|\hat{t}^{in}\|_\infty \to 0$, $i = 1, 2$, and from uniform convergence $\|\hat{t}^{in}\|_\infty \to$

$\|t^{i*}\|_\infty = \mu_\infty$, $i = 1, 2$. On the other hand, by parts (1) and (2), $\|t^n\|_{\rho\infty} \doteq \mu_n \to \mu^* \equiv \mu_\infty$. Therefore, there exists N_1 so that $\left|\|\hat{t}^{in}\|_{\rho\infty} - \mu_n\right| \leq \delta$, $i = 1, 2$ for $n \geq N_1$. From Lemma 4, we conclude

$$\|\hat{t}^{1n} - \hat{t}^{2n}\|_\infty \leq \epsilon \quad \forall\ n \geq N_1, \tag{16}$$

and from the uniform convergence of $\hat{t}^{1n}$ and $\hat{t}^{2n}$ to t^1 and t^2, respectively, there exists N_2 so that

$$\|\hat{t}^{in} - t^{i*}\|_\infty \leq \epsilon \quad \forall\ n \geq N_2, \quad i = 1, 2. \tag{17}$$

But then with $n \geq \max\{N_1,\ N_2\}$, (16) and (17) imply

$$\|t^{1*} - t^{2*}\|_\infty \leq 3\epsilon \quad \forall\ \epsilon;$$

i.e., $t^1 \equiv t^2$ and thus the optimal solution of Problem 1 over $\mathcal{S}$ is unique. □

Proof of Part (4). Let us apply Lemma 4 to $\hat{t}^n \doteq \hat{t}^{1n} \equiv \hat{t}^{2n}$ and t^n. With ϵ and δ as in Lemma 4, there exists N such that $|\|\hat{t}^n\|_\infty - \mu_\infty| \leq \delta$ and $|\|t^n\|_\infty - \mu_\infty| \leq \delta$ for $n \geq N_1$ (see the argument in the proof of part (3)). Since t^* is the unique minimizer of $\|\cdot\|_\infty$ over Ω^∞ by part (3), Lemma 4 implies

$$\|\hat{t}^n - t^n\|_\infty \leq \epsilon, \quad n \geq N_1. \tag{18}$$

On the other hand, $\hat{t}^n$ converges uniformly to t^*; therefore, there exists N_2 such that

$$\|\hat{t}^n - t^*\|_\infty \leq \epsilon, \quad n \geq N_2. \tag{19}$$

From (18) and (19), we conclude

$$\|t^n - t^*\|_\infty \leq 2\epsilon, \quad n \geq N = \max\{N_1,\ N_2\};$$

that is, t^n converges uniformly to t^*. □

REFERENCES

[1] L. Ahlfors, *Complex Analysis*, McGraw-Hill, New York, 1979.

[2] S. Boyd and C. Barratt, *Linear Controllers Design—Limits of Performance*, Prentice-Hall, Englewood Cliffs, NJ, 1990.

[3] B. Francis, *A Course in $\mathcal{H}_\infty$ Control Theory*, Lecture Notes in Control and Information Science, Springer-Verlag, New York, 1987.

[4] C. Gustafson and C. Desoer, *Controller design for linear multivariable feedback systems with stable plants, using optimization with inequality constraints*, Internat. J. Control, 37 (1983), pp. 881–907.

[5] J. W. Helton and A. Sideris, *Frequency response algorithms for $\mathcal{H}_\infty$ optimization with time domain constraints*, IEEE Trans. Automat. Control, 34 (1989), pp. 427–434.

[6] E. Polak and S. E. Salcudean, *On the design of linear multivariable feedback systems via constrained nondifferentiable optimization in $\mathcal{H}_\infty$ spaces*, IEEE Trans. Automat. Control, 34 (1989), pp. 268–276.

[7] H. Rotstein, *A Nevanlinna–Pick approach to time domain constrained $\mathcal{H}_\infty$ control*, SIAM J. Control Optim., 34 (1996), pp. 1329–1341.

[8] H. Rotstein and A. Sideris, *$\mathcal{H}_\infty$ optimization with time domain constraints*, IEEE Trans. Automat. Control, 39 (1994), pp. 762–779.

[9] H. Rotstein and A. Sideris, *$\mathcal{H}_\infty$-control with time domain constraints: The infinite horizon case*, Systems Control Lett., 24 (1995), pp. 251–258.

[10] A. Sideris and H. Rotstein, *Single input-single output $\mathcal{H}_\infty$-control with time domain constraints*, Automatica J. IFAC, 29 (1993), pp. 969–983.

[11] M. Sznaier, *An exact solution to general SISO mixed $\mathcal{H}_\infty/\mathcal{H}_2$ problems via convex optimization*, IEEE Trans. Automat. Control, 39 (1994), pp. 2511–2517.

[12] D. C. Youla and M. Saito, *Interpolation with positive-real functions*, J. Franklin Inst., 284 (1967), pp. 77–108.

SIAM J. CONTROL OPTIM.
Vol. 35, No. 4, pp. 1263–1296, July 1997

009

EXPERIMENTAL CONFIRMATION OF A PDE-BASED APPROACH TO DESIGN OF FEEDBACK CONTROLS*

H. T. BANKS†, RALPH C. SMITH‡, D. E. BROWN§, R. J. SILCOX¶, AND VERN L. METCALF‖

Abstract. Issues regarding the experimental implementation of PDE-based controllers are discussed in this work. While the motivating application involves the reduction of vibration levels for a circular plate through excitation of surface-mounted piezoceramic patches, the general techniques described here will extend to a variety of applications. The initial step is the development of a PDE model which accurately captures the physics of the underlying process. This model is then discretized to yield a vector-valued initial value problem. Optimal control theory is used to determine continuous-time voltages to the patches, and the approximations needed to facilitate discrete-time implementation are addressed. Finally, experimental results demonstrating the control of both transient and steady-state vibrations through these techniques are presented.

Key words. feedback control, piezoceramic actuators, PDE model

AMS subject classifications. Primary, 93C20; Secondary, 93B40

PII. S0363012995285909

1. Introduction.

An increasingly popular method for controlling structural vibrations is through the use of piezoceramic patches bonded to or embedded in the structure. These patches exhibit the piezoelectric property that inplane strains are generated in response to an applied voltage. Depending upon the geometry of patch placement, location with respect to the structure's neutral surface, and the method of excitation, this provides a mechanism for generating both inplane forces and/or bending moments in the underlying structure.

The advantage of using such patches as actuators in many applications is due to the fact that they are lightweight, space efficient, and relatively inexpensive and provide a means of obtaining structural control without significantly changing the passive structural dynamics. (They do not mass load the structure in the manner of a shaker or proof mass actuator.) Due to their ceramic nature, they can also be molded in a variety of shapes so as to fit the structure under consideration. Moreover, rigid body torques and spillover effects are minimized due to the fact that they are fully self-contained and distributed in nature. Finally, they also exhibit the inverse piezoelectric property and therefore generate a voltage in response to strains in the material. Hence a single patch or patch pair can be used for either sensing or actuation. This contributes to their efficiency and application in "smart material" structures.

*Received by the editors May 3, 1995; accepted for publication (in revised form) May 4, 1996. This research was supported by National Aeronautics and Space Administration contract NAS1-19480 while the first and second authors were visiting scientists at the Institute for Computer Applications in Science and Engineering (ICASE), NASA Langley Research Center, Hampton, VA 23681. Additional support was also provided in part under NASA grant NAG-1-1600.

http://www.siam.org/journals/sicon/35-4/28590.html

†Center for Research in Scientific Computation, North Carolina State University, Raleigh, NC 27695 (htbanks@eos.math.ncsu.edu). The research of this author was supported in part by Air Force Office of Scientific Research grant AFOSR-F49620-93-1-0198.

‡Department of Mathematics, Iowa State University, Ames, IA 50011 (rsmith@iastate.edu).

§Lockheed Engineering and Sciences Company, NASA Langley Research Center, Hampton, VA 23681.

¶Acoustics Division, NASA Langley Research Center, Hampton, VA 23681.

‖U.S. Army Research Laboratory, NASA Langley Research Center, Hampton, VA 23681.

In addition to their utility in purely structural applications, the patches are also finding increasing use in structural acoustic and fluid/structure applications. Again, their advantage lies in the fact that they provide an efficient means of controlling the structure without significantly altering its passive dynamics. Through consideration of the coupling between the structure and the adjacent media, this provides a means of controlling acoustic sound pressure levels or adjacent flow dynamics.

A great deal of research in the last several years has been directed at questions regarding the modeling of piezoceramic patch interactions with underlying structures (see [16] and the references therein), strategies for determination of optimal patch location and number, and control techniques which utilize the patches as sensors and actuators. Due to the steady-state periodic nature of the dynamics in many structural and structural acoustic systems, a large number of the current control methods are based upon frequency response input/output analysis. For example, Fuller, Gibbs, and Silcox have employed a feedforward filtered X version of an adaptive least mean squares algorithm to control flexural and extensional beam vibrations using piezoceramic actuators [25]. While such techniques have proven quite successful for controlling steady-state vibrations, they do not have the capability for direct control of transient responses. Other successfully implemented methods employing piezoelectric actuators to actively control structural vibrations include feedthrough techniques [24] and velocity feedback techniques [1, 32]. In general, these methods are based upon modal techniques and are designed to control purely steady-state responses. An exception to this is the experimental results reported in [32] in which transient plate vibrations, generated by an impact hammer, were reduced using an analog velocity feedback circuit.

Similar studies have demonstrated the experimental success of using surface-mounted piezoceramic patches to reduce structure-borne noise in structural acoustic systems [26, 27]. The emphasis in these studies was again on using frequency input/output analysis to control steady-state dynamics.

An alternative approach to controlling structural vibrations and sound pressure levels in structural acoustic systems is through the use of PDE-based feedback control methods. Analysis and numerical studies demonstrating these techniques for structural applications can be found in [7, 8, 9, 17] with corresponding results for structural acoustic systems given in [4, 6, 13]. These techniques start with an infinite-dimensional PDE model for the system under consideration. When developing such models, care should be taken to incorporate not only the contributions due to the piezoceramic patches but also dynamics due to inexact boundary conditions [18] coupling with adjacent acoustic fields [6], as well as any other physical phenomena which affect the dynamics of the structure. In this setting, mathematical issues such as model well-posedness and approximation issues concerning simulations, parameter estimation, and control can be addressed.

By approaching the problems in this manner, one can address the difficulties caused in purely modal methods by patch contributions, coupling between components, and inexact boundary conditions. Moreover, by combining the PDE model with appropriate time-dependent feedback control theory, one obtains a method which is equally applicable for controlling transient or steady-state vibrations.

There are several important features and benefits of a PDE-based approach. First, such an approach entails correct modeling based on physics. This facilitates treatment of actuator *loading* (passive) with respect to mass, stiffness, damping, geometry, etc., as well as the form of the actuator input. Of equal importance, the approach permits

the correct choice of an approximation framework so that computed quantities (which are suboptimal for the original infinite-dimensional system problem), such as states, feedback and compensator gains, and controls, actually converge to optimal quantities for the infinite-dimensional system, i.e., the actual physical model.

The choice of approximations is a *delicate* matter. It is known [11, 22] that failure to choose approximating elements "appropriately" in infinite-dimensional problems can yield poorly (weakly) convergent or even nonconvergent controls. "Appropriately" has been rather well documented in the PDE control research literature (e.g., see [3, 10, 28]) and involves concepts such as "costate or adjoint convergence" and "preservation of exponential stabilizability" (POES) or uniform stabilization under approximation (see Chapter 7 of [19] for a survey of some of the extensive theoretical literature on this subject). These theoretical concepts play a fundamental role in the practical choice of approximation schemes and associated computational algorithms.

For the problems and implementations addressed in this paper (including linear quadratic Gaussian (LQG) compensator and feedback design for unbounded input systems with periodic exogenous excitation), the theory is essentially complete if one combines and extends slightly the results in the literature (e.g., see [7, 23, 30, 35, 36]). A complete approximation and convergence analysis for the MinMax formulation is still under development, but the difficult points of the theory and their solution are basically resolved. We chose not to implement the MinMax design on the plate vibration problem described below since extensive computations [5] comparing the LQG and MinMax design suggested little advantage of the MinMax formulation for problems of the type considered in this paper.

In this paper, the experimental implementation of such a PDE-based control method is considered. While the motivating application involves the control of vibration levels for a circular plate through the excitation of surface-mounted piezoceramic patches, the general techniques described here will extend to a variety of applications. Following a brief discussion regarding the model and a Fourier–Galerkin scheme used to discretize it, relevant feedback control theory is discussed. In the discussion of the continuous- and discrete-time control results, two cases are considered; namely, the control of plate vibrations in the absence of a primary input force and the control of a plate driven by a periodic exogenous force. Implementation issues such as the effects of phase shifts and delays due to hardware are discussed, and the experimental setup is briefly described. Finally, experimental results demonstrating the transient and steady-state control results are presented. These demonstrate the effectiveness of the PDE-based controller for this system and indicate the potential of these control techniques for reducing transient and steady-state dynamics in other structural and structural acoustic systems.

Finally, we offer comments on the nature of our contributions here to the literature. This paper does *not* contain any new *theoretical* results; it reports on our successful use of PDE-based methods in experiments at NASA Langley Research Center. The methods (which *are* theoretically sound) for parameter estimation and feedback control are based on approximation theory developed (by us and many others in the PDE control community) during the past several decades. All necessary theorems on convergence of finite-dimensional parameter estimates, gains, filters, controls, observers, tracking variables, et cetera, needed for the examples treated here either have appeared or will soon appear in the research literature. Many of these theoretical results on PDE-based control have been largely viewed (especially in the engineering community) as nonimplementable and, hence, as somewhat irrelevant to

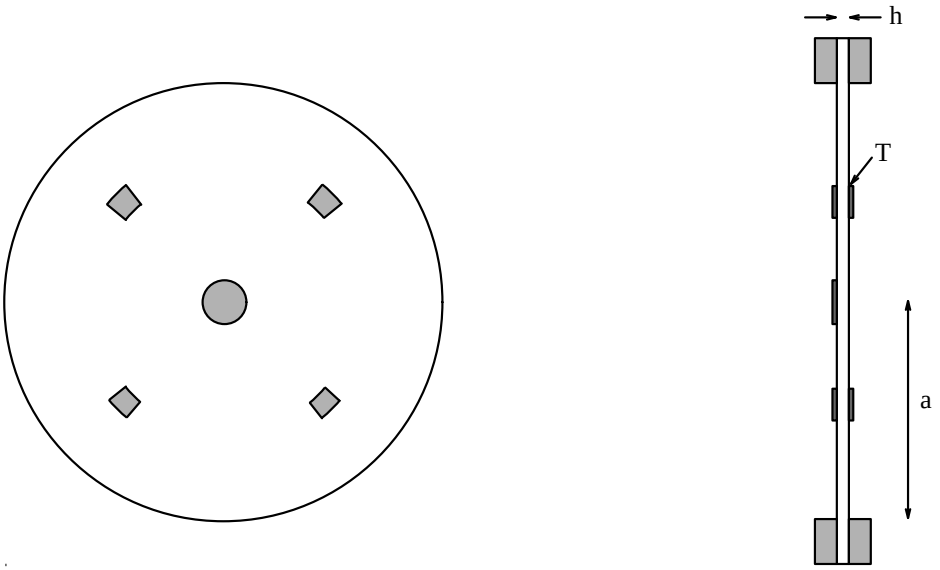


FIG. 2.1. *Thin circular plate with surface-mounted piezoceramic patches.*

applied scientists and engineers. The present manuscript refutes this notion and validates the practicality and importance of much of the theoretical efforts over the past years on PDE-based control. Our group is the first, to our knowledge, to provide substantial evidence that one can start from basic physical laws, derive careful infinite-dimensional distributed parameter or PDE control models, and successfully implement PDE control methods (with the approximations necessary to obtain finite-dimensional controls) which behave as theory and simulations predict.

2. Circular plate model. The structure under consideration is a thin circular plate with s sets of piezoceramic patches of thickness T bonded to the plate either singly or in pairs as depicted in Figure 2.1. Throughout this discussion, the radius and thickness of the plate are denoted by a and h, respectively. The density, Young's modulus, Poisson ratio, and Kelvin–Voigt damping parameters for the plate are given by $\rho_p, E_p, \nu_p, \hat{c}_{D_p}$, respectively, while similar parameters for the patches and bonding layer are denoted by $E_{pe}, \nu_{pe}, \nu_{pe}, \hat{c}_{D_{pe}}$ and $E_{b\ell}, \nu_{b\ell}, \nu_{b\ell}, \hat{c}_{D_{b\ell}}$, respectively. Moreover, μ, w, and $\tilde{g}$ are used to denote the air damping coefficient, transverse displacement, and external force on the plate. Finally, the region occupied by the unstrained neutral surface of the plate is indicated by Γ_0.

We point out that using current technology, piezoceramic patches can be cut to a variety of shapes. This includes circular and sectoral patches appropriate for the circular plate geometry considered here. From a modeling perspective, alignment of patch edges with coordinate axes is unnecessary when employing a weak form of the model; variations in shape simply affect the characteristic functions used to isolate the patch contributions. With regards to performance, experience indicates that while patch size, number, and placement strongly affect the ultimate control performance, slight variations in shape have minimal effect on the attenuation achieved using piezoceramic actuators.

Equations of motion for the plate can be determined from both Newtonian (force and moment balancing) and Hamiltonian (energy formulation) principles, and we

summarize both approaches here. The presentation will be for a general Kirchhoff plate with potentially nonaxisymmetric responses.

2.1. Strong form of plate model. Considering first the model which derives from Newtonian principles, we let $M_r, M_\theta, M_{r\theta}$ denote internal moments and $(M_r)_{pe}, (M_\theta)_{pe}$ denote external moments generated by the piezoceramic patches. As detailed in [12, 39], for a structure with s patch pairs, the internal moments are given by

$$\begin{aligned} M_r &= DK_r + \tilde{D}K_\theta + c_D\dot{K}_r + \tilde{c}_D\dot{K}_\theta, \\ M_\theta &= DK_\theta + \tilde{D}K_r + c_D\dot{K}_\theta + \tilde{c}_D\dot{K}_r, \\ M_{r\theta} &= M_{\theta r} = \frac{D}{2}\tau - \frac{\tilde{D}}{2}\tau + \frac{c_D}{2}\dot{\tau} - \frac{\tilde{c}_D}{2}\dot{\tau}, \end{aligned}$$

where

$$K_r = -\frac{\partial^2 w}{\partial r^2}, \quad K_\theta = -\frac{1}{r}\frac{\partial w}{\partial r} - \frac{1}{r^2}\frac{\partial^2 w}{\partial \theta^2}, \quad \tau = -\frac{2}{r}\frac{\partial^2 w}{\partial r \partial \theta} + \frac{2}{r^2}\frac{\partial w}{\partial \theta}.$$

The global flexural rigidity parameters D and $\tilde{D}$ and Kelvin–Voigt damping parameters c_D and $\tilde{c}_D$ are given by

$$\begin{aligned} D(r,\theta) &= \frac{E_p h^3}{12(1-\nu_p^2)} + \frac{2}{3}\sum_{i=1}^{s}\left[\frac{E_{pe}a_{3pe}}{1-\nu_{pe}^2} + \frac{E_{b\ell}a_{3b\ell}}{1-\nu_{b\ell}^2}\right]\chi_i(r,\theta), \\ \tilde{D}(r,\theta) &= \frac{E_p h^3 \nu_p}{12(1-\nu_p^2)} + \frac{2}{3}\sum_{i=1}^{s}\left[\frac{E_{pe}a_{3pe}\nu_{pe}}{1-\nu_{pe}^2} + \frac{E_{b\ell}a_{3b\ell}\nu_{b\ell}}{1-\nu_{b\ell}^2}\right]\chi_i(r,\theta), \\ c_D(r,\theta) &= \frac{\hat{c}_{D_p} h^3}{12(1-\nu_p^2)} + \frac{2}{3}\sum_{i=1}^{s}\left[\frac{\hat{c}_{D_{pe}}a_{3pe}}{1-\nu_{pe}^2} + \frac{\hat{c}_{D_{b\ell}}a_{3b\ell}}{1-\nu_{b\ell}^2}\right]\chi_i(r,\theta), \\ \tilde{c}_D(r,\theta) &= \frac{\hat{c}_{D_p} h^3 \nu_p}{12(1-\nu_p^2)} + \frac{2}{3}\sum_{i=1}^{s}\left[\frac{\hat{c}_{D_{pe}}a_{3pe}\nu_{pe}}{1-\nu_{pe}^2} + \frac{\hat{c}_{D_{b\ell}}a_{3b\ell}\nu_{b\ell}}{1-\nu_{b\ell}^2}\right]\chi_i(r,\theta). \end{aligned} \tag{2.1}$$

Here $a_{3b\ell} = (h/2+T_{b\ell})^3 - (h/2)^3$ and $a_{3pe} = (h/2+T_{b\ell}+T)^3 - (h/2+T_{b\ell})^3$ arise from integration through the bonding layer and patch thickness while $\chi_i(r,\theta)$ denotes the characteristic function which has a value of 1 in the region covered by the ith patch and is 0 elsewhere. A similar definition is used for the density which also exhibits a piecewise constant nature due to the presence of the patches. These definitions can be adapted to the case of a single patch that is bonded to the plate by replacing the 2/3 by 1/3. We point out that if the plate, patches, and bonding layers have the same Poisson ratios ($\nu_p = \nu_{pe} = \nu_{b\ell} = \nu$), then the internal moment expressions reduce to the familiar relations for a thin plate. For example, M_r in this case is given by

$$M_r = -D\left(\frac{\partial^2 w}{\partial r^2} + \frac{\nu}{r}\frac{\partial w}{\partial r} + \frac{\nu}{r^2}\frac{\partial^2 w}{\partial \theta^2}\right) - c_D\left(\frac{\partial^3 w}{\partial r^2 \partial t} + \frac{\nu}{r}\frac{\partial^2 w}{\partial r \partial t} + \frac{\nu}{r^2}\frac{\partial^3 w}{\partial \theta^2 \partial t}\right),$$

with D and c_D defined in (2.1).

The external moments generated by the patches in response to an applied voltage (out-of-phase for the patch pair) are given by

$$(M_r)_{pe} = (M_\theta)_{pe} = -\sum_{i=1}^{s} \mathcal{K}_i^B u_i(t)\chi_i(r,\theta),$$

where $u_i(t)$ is the voltage into the ith patch and $\mathcal{K}_i^B$ is a parameter which depends on the geometry, piezoceramic material properties, and piezoelectric strain constants (see [16] for details).

The internal and external moments can then be combined to yield the general plate moments

$$\begin{aligned} \mathcal{M}_r &= M_r - (M_r)_{pe}, \\ \mathcal{M}_\theta &= M_\theta - (M_\theta)_{pe}, \\ \mathcal{M}_{r\theta} &= M_{r\theta}. \end{aligned}$$

For a clamped plate with initial displacement $w_0(r,\theta)$ and velocity $w_1(r,\theta)$, force and moment balancing yield the equations

$$\begin{cases} \rho h \dfrac{\partial^2 w}{\partial t^2} + \mu \dfrac{\partial w}{\partial t} - \dfrac{\partial^2 \mathcal{M}_r}{\partial r^2} - \dfrac{2}{r}\dfrac{\partial \mathcal{M}_r}{\partial r} + \dfrac{1}{r}\dfrac{\partial \mathcal{M}_\theta}{\partial r} & 0 < \theta \le 2\pi, \\ \quad -\dfrac{2}{r}\dfrac{\partial^2 \mathcal{M}_{r\theta}}{\partial r \partial \theta} - \dfrac{2}{r^2}\dfrac{\partial \mathcal{M}_{r\theta}}{\partial \theta} - \dfrac{1}{r^2}\dfrac{\partial^2 \mathcal{M}_\theta}{\partial \theta^2} = \tilde{g}(t,r,\theta), & 0 \le r < a; \end{cases}$$

$$\text{(2.2)} \quad \begin{cases} w(t,a,\theta) = \dfrac{\partial w}{\partial r}(t,a,\theta) = 0; \end{cases}$$

$$\begin{cases} w(0,r,\theta) = w_0(r,\theta), \qquad \dfrac{\partial w}{\partial t}(0,r,\theta) = w_1(r,\theta) \end{cases}$$

as the strong form of the plate model (see [12, 37] for details). From an applications perspective, the following observations can be made regarding this model.

(1) It is first noted that in many applications, it is nearly impossible to maintain the truly fixed (zero displacement and slope) boundary conditions specified in the model (2.2). For the experimental plate which we used, parameter estimation results indicated minimal energy loss through the boundary conditions, and an adequate fit of the model (2.2) was obtained (see the identification results in [2, 15]). For plates in which the boundary clamping is less secure, a model for imperfectly clamped boundaries such as that presented in [18] should be used in order to obtain a model fit which is adequate for control applications.

(2) As discussed in [16] and noted in the model, the plate parameters ρ, D, ν, and c_D are discontinuous due to the presence and differing material properties of the patches. Moreover, while "handbook" values can be determined for ρ, D, and ν for a plate which is devoid of patches, those values usually cannot be used in the final system model with any accuracy due to nonuniformities in the plate and boundary conditions, variations in materials, and contributions due to the presence of the patches. In applications, such as that in this paper, these parameters are estimated using fit-to-data techniques (see [2, 15] for the parameter estimation results pertaining to the plate used here).

(3) The input parameters $\mathcal{K}_i^B$ are discontinuous since they are nonzero only over the regions of the patches. While expressions for these constants are derived in [16],

they too must be estimated in applications due to manufacturing variations in the patches.

It is readily noted in the strong form of the modeling equations that the discontinuous plate parameters and patch input terms are differentiated, thus leading to derivatives of the Dirac δ "function." The difficulties associated with this formulation are avoided in the weak form of the modeling equations which is presented in the next section.

2.2. Weak form of plate model. To provide a framework which facilitates analysis, approximation, and implementation, it is advantageous to consider a weak form of the modeling equations. Such a formulation can be determined directly from Hamiltonian (energy) principles and is equivalent to that obtained by integration by parts after multiplication of the strong form by suitably smooth test functions.

For the plate problem under consideration, a suitable space V of test functions is the subset of the Sobolev space $H^2(\Gamma_0)$ which satisfies the essential boundary conditions $w = \partial w/\partial r = 0$ at $r = a$. As detailed in [12, 39], a weak or variational form of the equations of motion for the plate is

$$\int_{\Gamma_0} \rho h \frac{\partial^2 w}{\partial t^2}\overline{\eta} d\gamma + \int_{\Gamma_0} \mu \frac{\partial w}{\partial t}\overline{\eta} d\gamma - \int_{\Gamma_0} M_r \overline{\frac{\partial^2 \eta}{\partial r^2}} d\gamma - \int_{\Gamma_0} \frac{1}{r^2} M_\theta \left[r\overline{\frac{\partial \eta}{\partial r}} + \overline{\frac{\partial^2 \eta}{\partial \theta^2}} \right] d\gamma$$
$$-2\int_{\Gamma_0} \frac{1}{r^2} M_{r\theta} \left[r\overline{\frac{\partial^2 \eta}{\partial r \partial \theta}} - \overline{\frac{\partial \eta}{\partial \theta}} \right] d\gamma = \int_{\Gamma_0} \sum_{i=1}^{s} \mathcal{K}_i^B u_i(t) \chi_i(r,\theta) \overline{\nabla^2 \eta} d\gamma + \int_{\Gamma_0} \tilde{g}\overline{\eta} d\gamma \tag{2.3}$$

for all $\eta \in V$. The overbar here denotes complex conjugation and the differential is $d\gamma = r d\theta dr$. It is easily noted that in this form, derivatives that were originally applied to moments have been transferred onto test functions. This eliminates the difficulties associated with differentiating the piecewise constant parameters ρ, D, ν, and c_D found in the internal moments as well as the discontinuous input parameters $\mathcal{K}_1^B, \ldots, \mathcal{K}_s^B$.

2.3. State approximation. As discussed in [12, 39], an appropriate choice for the basis and Fourier–Galerkin expansion of the plate displacement, when considering clamped boundary conditions, is $B_k^{\mathcal{N}}(r,\theta) = r^{|\hat{m}|} B_n^m(r) e^{im\theta}$ and

$$w^{\mathcal{N}}(t,r,\theta) = \sum_{m=-M}^{M} \sum_{n=1}^{N^m} w_{mn}^{\mathcal{N}}(t) r^{|\hat{m}|} B_n^m(r) e^{im\theta} = \sum_{k=1}^{\mathcal{N}} w_k^{\mathcal{N}}(t) B_k^{\mathcal{N}}(r,\theta). \tag{2.4}$$

Here $B_n^m(r)$ is the nth modified cubic spline satisfying $B_n^m(a) = \frac{dB_n^m(a)}{dr} = 0$, with the condition $\frac{dB_n^m(0)}{dr} = 0$ being enforced when $m = 0$ (this latter condition guarantees differentiability at the origin and implies that

$$N^m = \begin{cases} N, & m = 0, \\ N+1, & m \neq 0, \end{cases}$$

where N denotes the number of modified cubic splines). The total number of plate basis functions is $\mathcal{N} = (2M+1)(N+1) - 1$. As discussed in [12, 39], the inclusion of the weighting term $r^{|\hat{m}|}$ with

$$\hat{m} = \begin{cases} 0, & m = 0, \\ 1, & m \neq 0, \end{cases}$$

is motivated by the asymptotic behavior of the Bessel functions, which make up the analytic plate solution as $r \to 0$. It also serves to ensure the uniqueness of the solution at the origin. The Fourier coefficient in the weight is truncated to control the conditioning of the mass and stiffness matrices (see the examples in [12]).

To obtain a matrix system (again, see [12, 39] for a careful derivation with complete details), the $\mathcal{N}$-dimensional approximating subspace is taken to be $H^{\mathcal{N}} = \text{span}\{B_k^{\mathcal{N}}\}$ and the product space for the usual corresponding first-order vector system is $\mathcal{H}^{\mathcal{N}} \times \mathcal{H}^{\mathcal{N}}$. The restriction of the first-order form for the infinite-dimensional system (2.3) to the space $\mathcal{H}^{\mathcal{N}} \times \mathcal{H}^{\mathcal{N}}$ then yields the matrix equation

$$\begin{bmatrix} K_D^{\mathcal{N}} & 0 \\ 0 & M^{\mathcal{N}} \end{bmatrix} \begin{bmatrix} \dot{\vartheta}^{\mathcal{N}}(t) \\ \ddot{\vartheta}^{\mathcal{N}}(t) \end{bmatrix} = \begin{bmatrix} 0 & K_D^{\mathcal{N}} \\ -K_D^{\mathcal{N}} & -K_{c_D}^{\mathcal{N}} \end{bmatrix} \begin{bmatrix} \vartheta^{\mathcal{N}}(t) \\ \dot{\vartheta}^{\mathcal{N}}(t) \end{bmatrix} + \begin{bmatrix} 0 \\ \tilde{B}^{\mathcal{N}} \end{bmatrix} u(t) + \begin{bmatrix} 0 \\ \hat{g}^{\mathcal{N}}(t) \end{bmatrix},$$

$$\begin{bmatrix} K_D^{\mathcal{N}} & 0 \\ 0 & M^{\mathcal{N}} \end{bmatrix} \begin{bmatrix} \vartheta^{\mathcal{N}}(0) \\ \dot{\vartheta}^{\mathcal{N}}(0) \end{bmatrix} = \begin{bmatrix} g_1^{\mathcal{N}} \\ g_2^{\mathcal{N}} \end{bmatrix},$$

where $\vartheta^{\mathcal{N}}(t) = [w_1^{\mathcal{N}}(t), w_2^{\mathcal{N}}(t), \ldots, w_{\mathcal{N}}^{\mathcal{N}}(t)]^T$ denotes the column $\mathcal{N}$ vector containing the approximate state coefficients (see (2.4)). For the case in which $\nu_p = \nu_{pe} = \nu_{b\ell} = \nu$, the component matrices and vectors are given by

$$\begin{aligned}
&K_D^{\mathcal{N}} = K_{D1} + K_{D2} + K_{D3} + K_{D4} + K_{D5}, \\
&K_{c_D}^{\mathcal{N}} = K_{c_D1} + K_{c_D2} + K_{c_D3} + K_{c_D4} + K_{c_D5} + \int_{\Gamma_0} \mu B_k^{\mathcal{N}} \overline{B_\ell^{\mathcal{N}}} \, d\gamma, \\
&\left[M^{\mathcal{N}}\right]_{\ell,k} = \int_{\Gamma_0} \rho h B_k^{\mathcal{N}} \overline{B_\ell^{\mathcal{N}}} \, d\gamma, \qquad (2.5) \\
&\left[\hat{g}^{\mathcal{N}}(t)\right]_\ell = \int_{\Gamma_0} \tilde{g} \overline{B_\ell^{\mathcal{N}}} \, d\gamma, \quad \left[\tilde{B}^{\mathcal{N}}\right]_{\ell,j} = \int_{j^{th}patch} \mathcal{K}_j \overline{\nabla^2 B_\ell^{\mathcal{N}}} \, d\gamma, \\
&\left[g_1^{\mathcal{N}}\right]_\ell = \left\langle w_0, B_\ell^{\mathcal{N}} \right\rangle_V, \quad \left[g_2^{\mathcal{N}}\right]_\ell = \left\langle w_0, B_\ell^{\mathcal{N}} \right\rangle_H,
\end{aligned}$$

where

$$\begin{aligned}
[K_{D1}]_{\ell,k} &= \int_{\Gamma_0} D \left[\frac{\partial^2 B_k^{\mathcal{N}}}{\partial r^2} + \frac{\nu}{r} \frac{\partial B_k^{\mathcal{N}}}{\partial r} + \frac{\nu}{r^2} \frac{\partial^2 B_k^{\mathcal{N}}}{\partial \theta^2} \right] \frac{\partial^2 \overline{B_\ell^{\mathcal{N}}}}{\partial r^2} \, d\gamma, \\
[K_{D2}]_{\ell,k} &= \int_{\Gamma_0} D \left[\frac{1}{r^2} \frac{\partial B_k^{\mathcal{N}}}{\partial r} + \frac{1}{r^3} \frac{\partial^2 B_k^{\mathcal{N}}}{\partial \theta^2} + \frac{\nu}{r} \frac{\partial^2 B_k^{\mathcal{N}}}{\partial r^2} \right] \frac{\partial \overline{B_\ell^{\mathcal{N}}}}{\partial r} \, d\gamma, \\
[K_{D3}]_{\ell,k} &= \int_{\Gamma_0} D \left[\frac{1}{r^3} \frac{\partial B_k^{\mathcal{N}}}{\partial r} + \frac{1}{r^4} \frac{\partial^2 B_k^{\mathcal{N}}}{\partial \theta^2} + \frac{\nu}{r^2} \frac{\partial^2 B_k^{\mathcal{N}}}{\partial r^2} \right] \frac{\partial^2 \overline{B_\ell^{\mathcal{N}}}}{\partial \theta^2} \, d\gamma, \\
[K_{D4}]_{\ell,k} &= 2 \int_{\Gamma_0} D(1-\nu) \left[\frac{1}{r^2} \frac{\partial^2 B_k^{\mathcal{N}}}{\partial r \partial \theta} - \frac{1}{r^3} \frac{\partial B_k^{\mathcal{N}}}{\partial \theta} \right] \frac{\partial^2 \overline{B_\ell^{\mathcal{N}}}}{\partial r \partial \theta} \, d\gamma, \\
[K_{D5}]_{\ell,k} &= 2 \int_{\Gamma_0} D(1-\nu) \left[-\frac{1}{r^3} \frac{\partial^2 B_k^{\mathcal{N}}}{\partial r \partial \theta} + \frac{1}{r^4} \frac{\partial B_k^{\mathcal{N}}}{\partial \theta} \right] \frac{\partial \overline{B_\ell^{\mathcal{N}}}}{\partial \theta} \, d\gamma,
\end{aligned}$$

with D defined in (2.1) (similar expressions arise in the more general case of differing Poisson ratios). The index ranges here are $k, \ell = 1, \ldots, \mathcal{N}$. The matrices $K_{c_D1} - K_{c_D5}$ are defined similarly with the inclusion of the parameter c_D in the various integrals. Finally, we remind the reader that ρ, D, ν, and c_D are piecewise constant in these definitions due to the presence of the patches.

For application purposes, it is useful to note that the matrix system for the plate can thus be written as the Cauchy system

$$\begin{aligned} \dot{y}^{\mathcal{N}}(t) &= A^{\mathcal{N}} y^{\mathcal{N}}(t) + B^{\mathcal{N}} u(t) + g^{\mathcal{N}}(t), \\ y^{\mathcal{N}}(0) &= y_0^{\mathcal{N}}, \end{aligned} \tag{2.6}$$

where $y^{\mathcal{N}}(t) = [\vartheta^{\mathcal{N}}(t), \dot{\vartheta}^{\mathcal{N}}(t)]^T = [w_1^{\mathcal{N}}(t), \ldots, w_{\mathcal{N}}^{\mathcal{N}}(t), \dot{w}_1^{\mathcal{N}}(t), \ldots, \dot{w}_{\mathcal{N}}^{\mathcal{N}}(t)]^T$ denotes the column $2\mathcal{N}$ vector containing the generalized Fourier coefficients for the approximate displacement and velocity. In this form, the control problem can be readily discussed.

3. Continuous-time control problem. In the preceding discussion leading from the infinite-dimensional PDE model to the finite-dimensional matrix approximation, the superscript $\mathcal{N}$ was used to denote the level of discretization; i.e., the number of Fourier/spline basis elements used to approximate the state. This notation is standard in the theory of finite-element and spline approximations of infinite-dimensional systems. In finite-dimensional control theory, however, the level of discretization is typically fixed, and these superscript $\mathcal{N}$'s are usually omitted to simplify notation. We will do the same in this and subsequent sections so as to remain consistent with standard control notation.

3.1. Initial displacement and velocity—LQG control law. We consider first the $\mathcal{N}$-dimensional systems

$$\begin{aligned} \dot{y}(t) &= Ay(t) + Bu(t), \quad y(0) = y_0, \\ y_{ob}(t) &= Cy(t), \\ z(t) &= Hy(t) + Gu(t), \end{aligned} \tag{3.1}$$

where y_{ob} denotes observations in $\mathbb{R}^P$ and C is a $P \times \mathcal{N}$ observation matrix whose structure is determined by the manner and number of observations being used (the specific C matrix used in the plate experiments is described in section 4.2). Moreover, $z \in \mathbb{R}^r$ denotes the performance output obtained under the assumption that G and H are time-invariant matrices satisfying $H^T G = 0$. In the event that $P = \mathcal{N}$ and C is an identity, the optimal control u can be obtained from standard linear quadratic regulator (LQR) optimal control theory. The number of observations P is usually limited, however, and we concentrate instead on the case $P < \mathcal{N}$ which occurs when the full state is unavailable and must be reconstructed using a compensator (e.g., see [33]).

The general control problem for this case consists of determining the voltage u which minimizes the performance index (or cost functional)

$$\begin{aligned} J(u) &= \int_0^\infty |z(t)|^2 \, dt \\ &= \int_0^\infty \{\langle Qy(t), y(t)\rangle + \langle Ru(t), u(t)\rangle\} \, dt \end{aligned}$$

subject to (3.1). The $\mathcal{N} \times \mathcal{N}$ matrix Q can be chosen to satisfy various design criteria including frequency windowing, the weighting of various state components, or minimization of certain energy measures. The Q matrix used here was chosen using energy considerations, and construction details can be found in section 4.2. The $s \times s$ matrix R weights the voltages to the various patches or patch pairs.

Because full state information is not available in most applications, the state must be estimated or reconstructed from observations before a controlling voltage can be determined. We consider here a full order compensator or observer of Luenberger type [33] and refer the reader to [21, 30] for details on reduced order observers.

The compensator or reconstructed state satisfies the matrix system

$$\begin{aligned} \dot{y}_c(t) &= Ay_c(t) + Bu(t) + F\left[y_{ob}(t) - Cy_c(t)\right], \\ y_c(0) &= y_{c_0}, \end{aligned}$$

with the optimal voltage

$$u(t) = -Ky_c(t),$$

where F and K denote the compensator and feedback gains, respectively. We note that K and F are chosen so that the reconstruction error $|y(t) - y_c(t)| \to 0$ as $t \to \infty$. Under usual observability and controllability hypotheses (see [33]), the optimal feedback and compensator gains are given by

$$\begin{aligned} K &= R^{-1}B^T\Pi, \\ F &= PC^T\tilde{R}^{-1}, \end{aligned} \tag{3.2}$$

where Π and P are unique nonnegative-definite solutions to the following feedback (regulator) and compensator (observer) algebraic Riccati equations

$$\begin{aligned} \Pi A + A^T\Pi - \Pi BR^{-1}B^T\Pi + Q &= 0, \\ PA^T + AP - PC^T\tilde{R}^{-1}CP + \tilde{Q} &= 0, \end{aligned} \tag{3.3}$$

respectively. As was the case with the matrices Q and R, the matrices $\tilde{Q}$ and $\tilde{R}$ are design criteria for the specific control application under consideration (specific choices used in the plate experiments are summarized in section 4.2). We point out that in terms of the component matrices and control voltage, the compensator can be expressed as

$$\dot{y}_c(t) = \left[A - BR^{-1}B^T\Pi\right] y_c(t) + PC^T\tilde{R}^{-1}C\left[y(t) - y_c(t)\right].$$

The control law just described must be implemented in real time in order to be a viable method for reducing vibrations in physical structures. To facilitate implementation, it is prudent to calculate offline as many components as possible and then treat those precalculated components as filters when performing online computations. The method for continuous-time implementation is summarized, and offline and online components are categorized in Algorithm 3.1.

We note that the expensive (time-consuming) calculation of the component matrices A, B, Q, R, C, $\tilde{Q}$, and $\tilde{R}$ and Riccati solutions Π, P is performed offline, with the results loaded into the control code as datafiles. This leaves the integration of the system $\dot{y}_c(t) = A_c y_c(t) + F y_{ob}(t)$ as the primary computation to be performed during implementation. Issues regarding the numerical integration of the system as well as the effects of discrete-time calculations will be discussed in section 4, and a discrete version of Algorithm 3.1 is summarized in Algorithm 4.1.

While this compensator does provide the desired performance, it may lack robustness in some applications. In cases where added robustness with regard to certain types of system or observation noise and modeling errors is required, an H^∞/MinMax compensator of the type described in the next section can be used.

ALGORITHM 3.1. *Continuous time control of initial displacement and velocity.*

Offline	(i)	**Construct matrices** $A, B, C, Q, R, \tilde{Q}, \tilde{R}$
	(ii)	**Solve Riccati equations (3.3) for** Π **and** P
	(iii)	**Construct** $K = R^{-1}B^T\Pi$ **and** $F = PC^T\tilde{R}^{-1}$
Online	(i)	**Collect data** $y_{ob}(t) = Cy(t)$
	(ii)	**Solve the ODE system** $\dot{y}_c(t) = [A - BK - FC]\, y_c(t) + Fy_{ob}(t) = A_c y_c(t) + Fy_{ob}(t)$
	(iii)	**Calculate the voltage** $u(t) = -Ky_c(t)$

3.2. Initial displacement and velocity—H^∞/MinMax control law. We consider now the design of a dynamic compensator which is robust with respect to certain types of state and measurement uncertainties or disturbances (see [20] for details). To incorporate such uncertainties, we let $w(t) \in \mathbb{R}^q$ denote input and output disturbances. The system, with no exogenous force, is then given by

$$\begin{aligned}\dot{y}(t) &= Ay(t) + Bu(t) + Dw(t),\\ y_{ob}(t) &= Cy(t) + Ew(t),\\ z(t) &= Hy(t) + Gu(t).\end{aligned}$$

For this discussion, we assume that the input and output disturbances are independent and hence $DE^T = 0$. This is a matter of convenience, and the dependent case can be handled similarly after slight modifications are made (see [20]).

In this case, the MinMax optimization problem leading to the controller consists of finding a controller $u^* \in U \equiv L^2(0,\infty;\mathbb{R}^s)$ and disturbance $w^* \in W \equiv L^2(0,\infty;\mathbb{R}^q)$ such that

$$J_\gamma^* = \inf_{u\in U}\sup_{w\in W} J_\gamma(u,w) = J_\gamma(u^*,w^*)$$

for the disturbance-augmented functional

$$\begin{aligned}J_\gamma(u,w) &= \int_0^\infty \left\{|z(t)|^2 - \gamma^2\,|w(t)|^2\right\} dt\\ &= \int_0^\infty \left\{\langle Qy(t),y(t)\rangle + \langle Ru(t),u(t)\rangle - \gamma^2\,\langle w(t),w(t)\rangle\right\} dt.\end{aligned}$$

As noted in [17, 20], the results from this optimization problem yield a bound γ for the H^∞ norm of the transfer function from disturbance $\mathcal{L}(w)$ to the performance output $\mathcal{L}(z)$ where $\mathcal{L}$ denotes the Laplace transform.

Under the assumption that the pair (A,B) is stabilizable, (A,C) is detectable, (A,G) is controllable, and (A,H) is observable, one can prove the existence of (minimal) positive definite solutions Π and P to the algebraic Riccati equations

$$\begin{aligned}\Pi A + A^T\Pi - \Pi\left[BR^{-1}B^T - \gamma^{-2}\tilde{Q}\right]\Pi + Q = 0,\\ PA^T + AP - P\left[C^T\tilde{R}^{-1}C - \gamma^{-2}Q\right]P + \tilde{Q} = 0\end{aligned}$$

for a given attenuation $\gamma > 0$. Moreover, if the spectral radius ρ of $P\Pi$ satisfies the condition

$$\rho(P\Pi) < \gamma^2 \quad \text{or} \quad \Pi - \gamma^2 P^{-1} < 0,$$

then there exists a unique optimal controller

$$u^*(t) = -R^{-1}B^T\Pi y_c(t).$$

The state estimator $y_c(t) \in \mathbb{R}^{\mathcal{N}}$ satisfies

$$\begin{aligned}\dot{y}_c(t) &= A_c y_c(t) + F y_{ob}(t),\\ y_c(0) &= y_{c_0},\end{aligned}$$

where

$$\begin{aligned}A_c &= A - BK - FC + \gamma^{-2}\tilde{Q}\Pi,\\ F &= \left[I - \gamma^{-2}P\Pi\right]^{-1} PC^T\tilde{R}^{-1}.\end{aligned}$$

The implementation issues concerning the method are similar to those discussed in the algorithm for the LQG controller, but determination of the Riccati solutions Π, P and filter F are complicated by the fact that a suitable design value of γ must be determined before matrix calculations can proceed. Fortunately, these calculations can be performed offline and resulting matrices input as data files for the online computations. Hence the actual online controller can run at the same rate as that obtained using the LQG methodology.

3.3. Periodic primary excitation. For the case in which a periodic exogenous force drives the system, knowledge of that force can be used to extend previously discussed results to include the effects of periodicity. The $\mathcal{N}$-dimensional system in this case is

$$\begin{aligned}&\dot{y}(t) = Ay(t) + Bu(t) + g(t), \quad y(0) = y(\tau),\\ &y_{ob}(t) = Cy(t),\end{aligned} \tag{3.4}$$

where $g(t) \in \mathbb{R}^{\mathcal{N}}$ is periodic with period τ. That periodicity is then reflected in the performance index

$$J(u) = \int_0^\tau \left\{\langle Qy(t), y(t)\rangle + \langle Ru(t), u(t)\rangle\right\} dt$$

which is minimized subject to (3.4).

With K and F defined in (3.2), the reconstructed state in this case satisfies the system

$$\begin{aligned}&\dot{y}_c(t) = Ay_c(t) + Bu(t) + F\left[y_{ob}(t) - Cy_c(t)\right] + g(t),\\ &y_c(0) = y_c(\tau),\end{aligned} \tag{3.5}$$

with the optimal voltage given by

$$u(t) = -Ky_c(t) + R^{-1}B^T r(t). \tag{3.6}$$

Here r is a tracking variable defined by the system

$$\begin{aligned}&\dot{r}(t) = -\left[A - BK\right]^T r(t) + \Pi g(t),\\ &r(0) = r(\tau),\end{aligned} \tag{3.7}$$

where Π solves the first of the algebraic Riccati equations (3.3). We point out that in this case, the voltage contains two contributions. The first incorporates transient

ALGORITHM 3.2. *Continuous-time control with periodic exogenous force.*

Offline	(i)	**Construct matrices** $A, B, C, Q, R, \tilde{Q}, \tilde{R}$
	(ii)	**Solve Riccati equations (3.3) for** Π **and** P
	(iii)	**Construct** $K = R^{-1}B^T\Pi$ **and** $F = PC^T\tilde{R}^{-1}$
Online	(i)	**Collect data** $y_{ob}(t) = Cy(t)$ **and force measurements** $g(t)$
	(ii)	**Solve the ODE systems**
		$\dot{r}(t) = -[A - BK]^T r(t) + \Pi g(t)$
		$= A_{tr} r(t) + \hat{g}(t)$
		$\dot{y}_c(t) = [A - BK - FC]\, y_c(t) + F y_{ob}(t) + BR^{-1}B^T r(t) + g(t)$
		$= A_c y_c(t) + F y_{ob}(t) + A_r r(t) + g(t)$
	(iii)	**Calculate the voltage** $u(t) = -Ky_c(t) + R^{-1}B^T r(t)$

information by feeding back state estimates while the tracking component incorporates information regarding the periodic force.

Combining (3.5) and (3.6) yields the single expression

$$\begin{aligned} &\dot{y}_c(t) = [A - BK]\, y_c(t) + FC\,[y(t) - y_c(t)] + BR^{-1}B^T r(t) + g(t), \\ &y_c(0) = y_c(\tau) \end{aligned} \tag{3.8}$$

for the state estimator in terms of the tracking variable. The state estimate at time t is then obtained by integrating (3.7) and (3.8) after solving the necessary Riccati equations.

As was the case when considering control of the unforced system with initial displacement and velocity, the computations can be categorized with respect to those which can be performed offline and those which must be done online. Algorithm 3.2 summarizes the continuous-time control method for periodic excitation and categorizes the offline and online components. In this case, the tracking equation $\dot{r}(t) = A_{tr} r(t) + \hat{g}(t)$ must be solved before the state can be estimated by integrating the system $\dot{y}_c(t) = A_c y_c(t) + F y_{ob}(t) + A_r r(t) + g(t)$.

Expressions similar to those in section 3.2 arise when an H^∞/MinMax compensator is considered for the problem. For this case, the reader is referred to [4], where details concerning the design of an H^∞/MinMax compensator for a structural acoustic system that is subjected to a periodic exogenous force is considered.

4. Discrete-time control problem. The control laws summarized in the last section were derived under the assumption of continuous-time sampling of observed state and force data. Hence it was assumed that $y_{ob}(t)$ and $g(t)$ were available for all t within the temporal interval of interest. Moreover, it was assumed that $y(t), y_c(t)$, and $r(t)$ could be obtained through exact integration of the state, state estimator, and tracking equations. When implementing the method, however, one has available only discrete data values and the differential estimator and tracking equations must be numerically approximated at discrete-time values. The manner in which the control laws are implemented in discrete time and the influence of this discretization on the overall performance are discussed in this section.

In the last section, it was demonstrated that in the case with no primary exogenous force, the state equation, state estimator, and controlling voltage for the LQG and H^∞/MinMax problems had the form summarized in column 1 of Table 4.1. The corresponding quantities for a system subjected to a periodic exogenous force are summarized in the second column. Details regarding the component matrices can be found in the previous section.

TABLE 4.1
State, state estimator, and controlling voltage for the systems with no primary input and a periodic exogenous force.

	No primary input	Periodic exogenous force
State equation	$\dot{y}(t) = Ay(t) + Bu(t),\ y(0) = y_0,$ $y_{ob}(t) = Cy(t)$	$\dot{y}(t) = Ay(t) + Bu(t) + g(t),\ y(0) = y(\tau),$ $y_{ob}(t) = Cy(t)$
State estimator	$\dot{y}_c(t) = A_c y_c(t) + F y_{ob}(t),$ $y_c(0) = y_{c_0}$	$\dot{y}_c(t) = A_c y_c(t) + F y_{ob}(t) + A_r r(t) + g(t),$ $y_c(0) = y_c(\tau)$
Tracking equation		$\dot{r}(t) = A_{tr} r(t) + \hat{g}(t),$ $r(0) = r(\tau)$
Control voltage	$u(t) = -R^{-1} B^T \Pi y_c(t)$	$u(t) = -R^{-1} B^T \Pi y_c(t) + R^{-1} B^T r(t)$

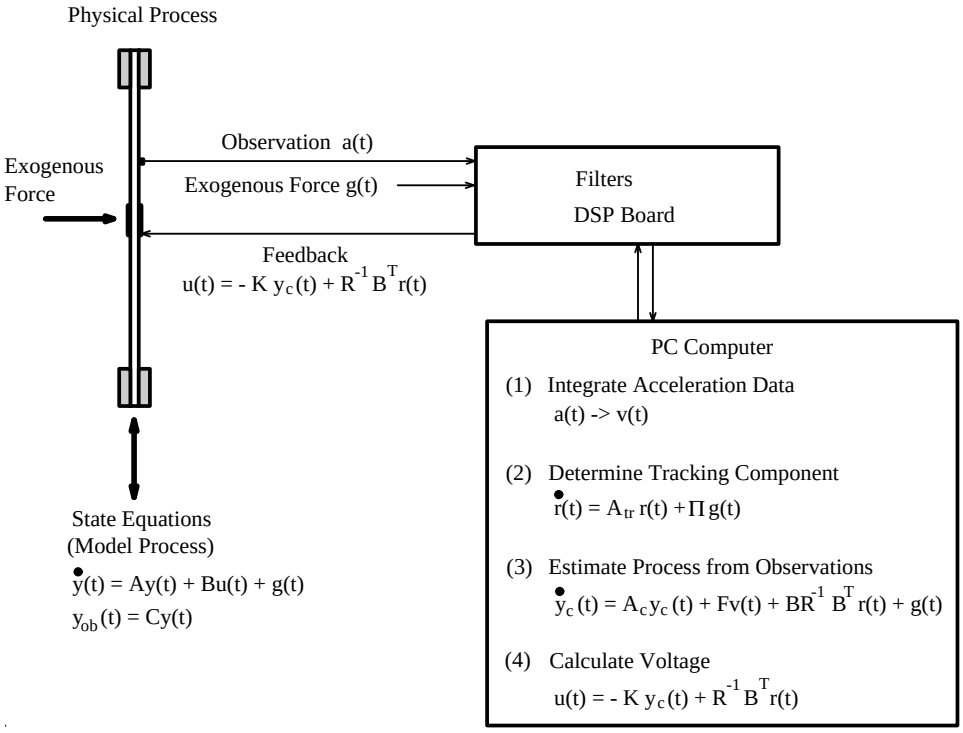


FIG. 4.1. *Experimental plate setup, modeling equations, and computations necessary for determination of controlling voltage.*

The relationship of these components to the experimental plate setup, driven by a periodic exogenous force, is illustrated in Figure 4.1. For that setup, the physical process consists of the clamped circular plate with attached piezoceramic patches. The state equations represent the spatial discretization of a PDE model of the process with physical parameters estimated using experimental data so that the model accurately captures the plate dynamics.

We point out that when modeling the process or performing simulations, the observations $y_{ob}(t)$ are obtained from the approximate state through the relation $y_{ob}(t) = Cy(t)$, where C is the matrix observation operator. When experimentally implementing the method, temporal data from the process are used to determine the observations $y_{ob}(t)$. This often requires processing either through digital or analog filters or in the software. In the circular plate experiments, accelerometer data $a(t)$ were integrated to yield velocity measurements $v(t)$ to be used for the observations $y_{ob}(t)$ (see the final subsection of this section for details). Once $y_{ob}(t)$ has been obtained, the state is estimated by numerically integrating the estimator equation and a controlling voltage is calculated. This voltage is then fed back to the process.

Because data is processed in a digital manner, the observations y_{ob} and force measurements g are obtained only at discrete times t_j. The rate at which these data can be sampled is governed by the data acquisition system, the software being used, and the number of calculations required between samples. In particular, the tracking component $r(t_j)$, state estimate $y_c(t_j)$, and voltage $u(t_j)$ must be calculated before the arrival of data at time t_{j+1}. While details regarding these calculations are postponed until section 4.1, it is noted here that the system sizes must be minimized in order to permit real-time approximation of the estimation equation. This is a major motivation for choosing appropriate, accurate approximation techniques to spatially discretize the modeling PDEs.

As indicated previously, the open and closed loop process dynamics can be simulated by using the state equations to model the plate. The state estimate and state are then approximated by simultaneously integrating the corresponding equations. The integration of the state equation can be performed using any sufficiently accurate ODE routine which is efficient for the system under consideration; for example, a variable order, variable stepsize method was used to solve the stiff system which arose when simulating plate dynamics. If simulations demonstrating the levels of control that can be obtained under "optimal" conditions are desired, the estimator equation can be integrated using the same high-order routine. On the other hand, simulations representing the attenuation levels that can be expected under "implementation" conditions can be obtained by incorporating values of y_{ob} and g calculated at discrete times and approximating $y_c(t_j)$ using the techniques employed when implementing the method. Simulation results using both techniques can be found in [14].

4.1. Approximation of the estimator and tracking equations. In order to obtain tracking values and state estimates to be used when calculating controlling voltages, the solution to the tracking and state estimator equations must be numerically approximated. If the goal is solely to perform simulations, this can be easily accomplished using the same ODE solver used to integrate the state equation (indeed, the state and estimator equations can be combined into a single system and integrated simultaneously). This is not practical when experimentally implementing the method, however, and one must typically perform the work subject to the following criteria. The method must be sufficiently efficient so as to facilitate real-time implementation and sufficiently accurate so as to resolve system dynamics. The systems are quite often stiff, which implies that either a-stability or α-stability is important. Finally, the difficulties in storing past data make it prohibitive to use many popular multistep methods.

For the experiments performed with the circular plate, the sample rate was sufficiently fast (and hence Δt was sufficiently small) that a modified backward Euler

ALGORITHM 4.1. *Discrete-time control of initial displacement and velocity.*

Offline	(i)	**Construct matrices** $A, B, C, Q, R, \tilde{Q}, \tilde{R}$
	(ii)	**Solve Riccati equations (3.3) for** Π **and** P
	(iii)	**Construct** $K = R^{-1}B^T\Pi$
		$F = PC^T\tilde{R}^{-1}$
		$A_c = A - BK - FC$
	(iv)	**Choose appropriate** Δt **(determined by sample rate)**
	(v)	**Construct** $\mathcal{A}_c = (I - \Delta t A_c)^{-1}$
		$\mathcal{F}_c = (I - \Delta t A_c)^{-1} F$
Online	(i)	**Collect acceleration data** $a(t_j)$
	(ii)	**Integrate to obtain** $y_{ob}(t_j) = v(t_j)$
	(iii)	**Time step the discrete estimator system**
		$y_{c_{j+1}} = \mathcal{A}_c y_{c_j} + \mathcal{F}_c y_{ob}(t_j)$
	(iv)	**Calculate the voltage** $u(t_j) = -K y_{c_j}$

or trapezoidal method produced adequate results. We illustrate here such a modified backward Euler method.

4.1.1. Initial displacement and velocity. Considering first the compensator/estimator system with no primary exogenous force, we find that the modified Euler approximate to the solution at time t_{j+1} is given by

$$\begin{aligned} y_{c_{j+1}} &= (I - \Delta t A_c)^{-1} y_{c_j} + (I - \Delta t A_c)^{-1} F y_{ob}(t_j) \\ &= \mathcal{A}_c y_{c_j} + \mathcal{F}_c y_{ob}(t_j). \end{aligned}$$

The method is modified in the sense that current observation values $y_{ob}(t_j)$ are used as input since futures values at t_{j+1} are unknown. The time step Δt is dictated by the sample rate. We point out that the matrix $\mathcal{A}_c = (I - \Delta t A_c)^{-1}$ and vector $\mathcal{F}_c = (I - \Delta t A_c)^{-1} F$ can be computed offline and then loaded as datafiles for the online computations. Hence the implicit nature of the method, which is necessary to ensure stability, does not slow the implementation. The discrete-time implementation of the method is summarized in Algorithm 4.1. The definitions of the component matrices can be found in section 3, and Algorithm 4.1 can be compared with the corresponding continuous-time Algorithm 3.1 given in that section.

4.1.2. Periodic exogenous force. The application of these control techniques to systems with both transient and steady-state behavior involves the approximation of both the tracking and the state estimator equations before a control input is calculated. While a variety of techniques and strategies can be used to obtain approximate values of $r(t_j)$, which are then used when computing $y(t_j)$, these calculations must ultimately be performed in real time when implementing the method. In the experiments involving the circular plate, the exogenous force was measured for several periods and the solutions to the tracking equation were approximated and stored over a time period commensurate with the driving frequency. These stored tracking values were then used as a filter when approximating the estimated state during the remainder of the experiment.

Illustrating with the backward Euler discretization, the approximate to the tracking solution was determined from the difference equation

$$\begin{aligned} r_{j+1} &= (I - \Delta t A_{tr})^{-1} r_j + (I - \Delta t A_{tr})^{-1} \hat{g}(t_j) \\ &= \mathcal{A}_{tr} r_j + \mathcal{A}_{tr} \hat{g}(t_j) \end{aligned}$$

ALGORITHM 4.2. *Discrete-time control of periodic excitation.*

Offline	(i)	**Construct matrices** $A, B, C, Q, R, \tilde{Q}, \tilde{R}$
	(ii)	**Solve Riccati equations (3.3) for** Π **and** P
	(iii)	**Construct** $K = R^{-1}B^T\Pi$ $F = PC^T\tilde{R}^{-1}$ $A_c = A - BK - FC$ $A_{tr} = -[A - BK]^T$
	(iv)	**Choose appropriate** Δt **(determined by sample rate)**
	(v)	**Construct** $\mathcal{A}_c = (I - \Delta t A_c)^{-1}$ $\mathcal{A}_r = (I - \Delta t A_c)^{-1} A_r$ $\mathcal{A}_{tr} = (I - \Delta t A_{tr})^{-1}$ $\mathcal{F}_c = (I - \Delta t A_c)^{-1} F$
Online	(i)	**Collect acceleration data** $a(t_j)$
	(ii)	**Integrate to obtain** $y_{ob}(t_j) = v(t_j)$
	(iii)	**Approximate and store tracking values** $r_{j+1} = \mathcal{A}_{tr} r_j + \mathcal{A}_{tr}\hat{g}(t_j)$ $r(\tau) = 0$
	(iii)	**Time step the discrete estimator system** $y_{c_{j+1}} = \mathcal{A}_c y_{c_j} + \mathcal{F}_c y_{ob}(t_j) + \mathcal{A}_r r_j + \mathcal{A}_c g(t_j)$ $y_c(k\tau) = 0$
	(iv)	**Calculate the voltage** $u(t_j) = -Ky_{c_j} + R^{-1}B^T r_j$

subject to the final condition $r(\tau) = 0$. (When implementing the method, one can simply search for a "suitable" zero crossing to start the approximation.)

This approximation was continued throughout several periods of the driving force, with r_j being stored in a circular buffer. This buffer was then treated as a filter when estimating the state using the difference equations

$$\begin{aligned} y_{c_{j+1}} &= (I - \Delta t A_c)^{-1} y_{c_j} + (I - \Delta t A_c)^{-1} F y_{ob}(t_j) + (I - \Delta t A_c)^{-1} A_r r_j \\ &\quad + (I - \Delta t A_c)^{-1} g(t_j) \\ &= \mathcal{A}_c y_{c_j} + \mathcal{F}_c y_{ob}(t_j) + \mathcal{A}_r r_j + \mathcal{A}_c g(t_j). \end{aligned}$$

As indicated in Algorithm 4.2, the time-intensive calculations involving matrix construction, solution of the Riccati equations, and matrix inversion were performed offline prior to the experiments, and the matrices $\mathcal{A}_{tr}, \mathcal{A}_c, \mathcal{A}_r$ and vector $\mathcal{F}_c$ were simply loaded as datafiles. This, combined with the solution of the tracking filter before state estimation, yielded an algorithm which was sufficiently fast for implementation. Current efforts are aimed toward the simultaneous approximation of the tracking and state estimator equations during implementation.

4.1.3. Higher-order approximations. In the previous discussion, a modified backward Euler method was used to discretize the state estimator and tracking equations. As indicated, by numerical simulations reported in [14] and experimental results in the next section, for small Δt, this provides sufficient accuracy to calculate an effective feedback voltage. If more accuracy is needed, a trapezoid rule or hybrid method of the nature discussed on page 225 of [34] can be used. These provide increased accuracy without adding complexity during implementation since the components $\mathcal{A}_{tr}, \mathcal{A}_c, \mathcal{A}_r$, and $\mathcal{F}_c$ can still be computed offline.

TABLE 4.2
System and control matrices used in the circular plate experiments.

Component	Size	Comments
$A = \begin{bmatrix} 0 & I \\ -(M^{\mathcal{N}})^{-1} K_D^{\mathcal{N}} & -(M^{\mathcal{N}})^{-1} K_{c_D}^{\mathcal{N}} \end{bmatrix}$	32×32	The elements composing the 16×16 matrices $M^{\mathcal{N}}, K_D^{\mathcal{N}}$, and $K_{c_D}^{\mathcal{N}}$ are summarized in (2.5)
$B = \begin{bmatrix} 0 \\ (M^{\mathcal{N}})^{-1} \tilde{B}^{\mathcal{N}} \end{bmatrix}$	32×1	See (2.5) for the description of $\tilde{B}^{\mathcal{N}}$
$g(t) = \begin{bmatrix} 0 \\ \hat{g}^{\mathcal{N}}(t) \end{bmatrix}$	32×1	The elements of $\hat{g}^{\mathcal{N}}(t)$ are detailed in (2.5)
$C = \left[\begin{array}{c\|c} 0, \cdots, 0 & B_1^{\mathcal{N}}(r_1, \theta_1), \cdots, B_{\mathcal{N}}^{\mathcal{N}}(r_1, \theta_1) \\ \vdots & \vdots \\ 0, \cdots, 0 & B_1^{\mathcal{N}}(r_p, \theta_p), \cdots, B_{\mathcal{N}}^{\mathcal{N}}(r_p, \theta_p) \end{array}\right]$	$p \times 32$	In the experiments, acceleration data were integrated to obtain velocity values which are the second-state values in the second-order formulation. Since one accelerometer was used, $p = 1$. See Example 2 of section 4.3 for a discussion regarding the duality between control and observation.

4.2. Example 1—matrix construction. The control discussion thus far has been general in the sense that it holds for general systems of the form

$$\begin{aligned} \dot{y}(t) &= Ay(t) + Bu(t), \\ y_{ob}(t) &= Cy(t), \end{aligned} \quad \text{or} \quad \begin{aligned} \dot{y}(t) &= Ay(t) + Bu(t) + g(t), \\ y_{ob}(t) &= Cy(t) \end{aligned}$$

as long as the pair (A, B) is stabilizable and (A, C) is detectable. Moreover, the cost functional matrices Q, R and observation matrices $\tilde{Q}, \tilde{R}$ have been treated as general design criteria to be specified according to the application under consideration. In this example, we illustrate explicitly the matrices and filters used when implementing these control techniques for a vibrating circular plate.

We first note that $16 (= N)$ modified cubic splines (see (2.4)) were sufficient for resolving the plate dynamics in the frequency range under consideration. Due to the axisymmetric excitation and response of the plate, the Fourier limit $M = 0$ was used in all calculations. Hence a total of $\mathcal{N} = 16$ basis functions were used, which led to 32 coefficients in the vector y.

The formulation and sizes of all components in the control system for the circular plate are summarized in Tables 4.2 and 4.3. The component matrices and vectors are then employed in Algorithm 4.1 or 4.2 to create the implementation matrices and filters which were ultimately used in the experiments.

4.3. Example 2—duality between control and observation. The duality between control and observation can be noted by considering the form of the control matrix B and observation matrix C. Illustrating with the case in which s patch pairs are used for control, the $2\mathcal{N} \times s$ control matrix has the form

$$B = \begin{bmatrix} 0 \\ (M^{\mathcal{N}})^{-1} \tilde{B}^{\mathcal{N}} \end{bmatrix},$$

TABLE 4.3

Control and observation matrices used in the circular plate experiments. The experimental results in section 5.1 *demonstrate control of transient dynamics, while the periodic case is illustrated in section* 5.2.

Component	Size	Comments
$Q = \begin{bmatrix} d_1 I^{\mathcal{N}} & 0 \\ 0 & d_2 I^{\mathcal{N}} \end{bmatrix} \begin{bmatrix} K_D^{\mathcal{N}} & 0 \\ 0 & M^{\mathcal{N}} \end{bmatrix}$	32×32	A weighted mass matrix was used for the penalty term Q. As discussed in [6], this provides a means of weighting the kinetic and potential energy of the plate. Section 5.1: $d_1 = d_2 = 1$. Section 5.2: $d_1 = d_2 = 5$.
$R = \begin{bmatrix} R_{11} & & \\ & \ddots & \\ & & R_{ss} \end{bmatrix}$	$s \times s$	In the plate experiments, one controlling patch was used, so $s = 1$. Section 5.1: $R_{11} = 10^{-7}$. Section 5.2: $R_{11} = 10^{-10}$.
$\tilde{Q} = \begin{bmatrix} c_1 I^{\mathcal{N}} & 0 \\ 0 & c_2 I^{\mathcal{N}} \end{bmatrix}$	32×32	For the plate experiments, $\tilde{Q}$ was treated solely as a design parameter as compared with the choice in [4], where physical arguments were used to construct the matrix. The identity weights were taken to be $c_1 = c_2 = 1$ in the experiments.
$\tilde{R} = \begin{bmatrix} \tilde{R}_{11} & & \\ & \ddots & \\ & & \tilde{R}_{pp} \end{bmatrix}$	$p \times p$	In the experiments, $p = 1$ since one accelerometer was used for data collection. The weight was taken to be $\tilde{R}_{11} = 1$.

where the $\mathcal{N} \times s$ matrix $\tilde{B}^{\mathcal{N}}$ has elements

$$\begin{aligned} \left[\tilde{B}^{\mathcal{N}}\right]_{\ell,j} &= \int_{\Gamma_0} \mathcal{K}_j \overline{\nabla^2 B_\ell^{\mathcal{N}}} \chi_j(r,\theta)\, d\gamma \\ &= \int_{jth\ patch} \mathcal{K}_j \overline{\nabla^2 B_\ell^{\mathcal{N}}}\, d\gamma \end{aligned}$$

(here $\chi_j(r,\theta)$ denotes the characteristic function over the jth patch).

When data from accelerometers located at the points $(r_1,\theta_1),\ldots,(r_p,\theta_p)$ are integrated to obtain velocity values, the $p \times 2\mathcal{N}$ observation matrix is given by

$$C = [0 \quad 1] \begin{bmatrix} \tilde{C} & 0 \\ 0 & \tilde{C} \end{bmatrix},$$

where, using the $2-D$ Dirac delta notation δ,

$$\begin{aligned} \left[\tilde{C}\right]_{j,\ell} &= \int_{\Gamma_0} B_\ell^{\mathcal{N}} \delta(r - r_j, \theta - \theta_j)\, d\gamma \\ &= B_\ell^{\mathcal{N}}(r_j, \theta_j). \end{aligned}$$

With C thus defined, it can immediately be noted that

$$\begin{aligned} y_{ob}(t) &= Cy(t) \\ &= \sum_{k=1}^{\mathcal{N}} \dot{w}_k(t) B_k^{\mathcal{N}}(r_j, \theta_j) \end{aligned}$$

denotes the physical value of the velocity at the point (r_j, θ_j) given by the state equations at time t. This is an approximation (to within modeling and processing error) of the actual plate plate velocity $v(t)$ which is measured in experiments.

Similarly, multiplication of the state estimator coefficients y_c by C produces an estimate of the velocity which is then compared with the measured plate velocity when integrating the state estimator equation

$$\dot{y}_c(t) = \left[A - BR^{-1}B^T\Pi\right] y_c(t) + PC^T\tilde{R}^{-1}\left[v(t) - Cy_c(t)\right].$$

It should be noted that as the state estimates approach the measured plate values, the estimator equation approaches the state equation

$$\dot{y}(t) = Ay(t) - BR^{-1}B^T\Pi y_c(t),$$

which is used to model the plate dynamics.

4.4. Integration of experimental data. For the experiments involving the control of circular plate vibrations, data consisted of acceleration measurements obtained from one or more accelerometers on the plate. It was then necessary to approximately integrate these data to obtain velocity values so as to have a state variable for control calculations. An issue which turns out to be crucial when approximately integrating experimental data concerns the robustness of the integrator with respect to inexact initial conditions and DC gains or biases (added constants or offsets) in the data. The inexact initial conditions can be due to unknown system contributions, static shocks during system connections, et cetera. While careful calibration can alleviate some of the uncertainty in initial conditions, it cannot fully eliminate the problem. The problem of gains or biases due to small DC voltages in the system can also be minimized but never fully eliminated. Hence an integrator which is minimally affected by uncertain initial conditions and DC offsets in the data is crucial for success when approximately integrating data.

Here we consider two techniques for approximately integrating acceleration data to obtain velocity values in accordance with the relation

$$\dot{v}(t) = a(t).$$

Essentially, the idea is to replace the integration by either the first-order differential equation

$$\dot{v} + \Omega v = \frac{1}{RC} a \tag{4.1}$$

or the second-order equation

$$\ddot{v} + \Omega\dot{v} + \Omega^2 v = \frac{1}{RC}\dot{a} \tag{4.2}$$

(see [29]). The design parameters Ω and RC are frequency and time constants, respectively, which are chosen so that $RC = 1$ and $\omega > 6\Omega$, where ω is the smallest

observed frequency. For solution, (4.2) is written as the first-order system

$$\begin{aligned} \begin{bmatrix} \dot{v} \\ \dot{e} \end{bmatrix} &= \begin{bmatrix} -\Omega & \Omega \\ -\Omega & 0 \end{bmatrix} \begin{bmatrix} v \\ e \end{bmatrix} + \begin{bmatrix} \frac{a(t)}{RC} \\ 0 \end{bmatrix} \\ &\Rightarrow \dot{z} = Az + f, \end{aligned} \tag{4.3}$$

where $\dot{e} = -\Omega v$. The integration of (4.3) is subject to the initial conditions

$$\begin{bmatrix} v(0) \\ e(0) \end{bmatrix} = \begin{bmatrix} v_0 \\ e_0 \end{bmatrix}. \tag{4.4}$$

As will be detailed in the subsequent discussion, the first-order integrator (4.1) is robust with respect to inexact initial conditions but propagates DC offsets in the data. In the second-order integrator (4.2), or system (4.3), both disturbance in initial conditions and DC offsets in acceleration data are exponentially attenuated. Moreover, the frequency response of this approximator is very close to that in the original signal for $\omega > 6\Omega$. Hence this latter method provides an accurate and robust means of approximately integrating experimental data.

4.4.1. First-order approximate integrator. Here we examine the properties of the first-order approximate integrator (4.1). We consider first the case in which the exact initial condition v_0 is known and the acceleration $a(t)$ is free from DC gains or biases (added constants). If we let $V(s) = \mathcal{L}\{v(t)\}$ and $A(s) = \mathcal{L}\{a(t)\}$, then Laplace transformation of the system (4.1) yields

$$V(s) = \frac{1}{s+\Omega} v_0 + \frac{1}{s+\Omega} \cdot \frac{1}{RC} A(s).$$

Hence the approximate integrator (4.1) is a single pole filter. Inverse transformation then yields

$$v(t) = e^{-\Omega t} v_0 + \frac{1}{RC} \int_0^t e^{-\Omega(t-s)} a(s) ds \tag{4.5}$$

as the solution to (4.1).

Similarly, if $\tilde{v}_0$ denotes a perturbed initial condition and a DC gain $\tilde{g}_a$ is present in the data, then the solution is given by

$$\tilde{v}(t) = e^{-\Omega t} \tilde{v}_0 + \frac{1}{RC} \int_0^t e^{-\Omega(t-s)} \left[a(s) + \tilde{g}_a \right] ds.$$

It follows immediately that

$$\tilde{v}(t) = v(t) + e^{-\Omega t} \left(\tilde{v}_0 - v_0 - \frac{\tilde{g}_a}{RC\Omega} \right) + \frac{\tilde{g}_a}{RC\Omega}.$$

It is first noted that the perturbations in initial conditions exponentially decay with the rate of decay influenced by the magnitude of the parameter Ω. DC gains of the order $\tilde{g}_a/(RC\Omega)$ remain, however, thus leading to difficulties when such biases are present in the data. Both properties are numerically illustrated through examples in [14].

The manner through which the solution (4.5) approximates the solution to the original relation $\dot{v}(t) = a(t)$ can be illustrated with a simple example. Consider $a(t) = 120\pi\cos(120\pi t)$. The solution to (4.1) for this acceleration is

$$v(t) = \frac{1}{RC}\cdot\frac{120\pi}{\Omega^2+(120\pi)^2}\left[\Omega\cos(120\pi t)+120\pi\sin(120\pi t)\right]-\frac{1}{RC}\cdot\frac{120\pi\Omega}{\Omega^2+(120\pi)^2}e^{-\Omega t},$$

which reduces to the solution of the original relation with $\Omega = 0$ and $RC = 1$. For $\Omega = 16\pi$, the solution v still provides an adequate approximation to the original, whereas it is a very poor approximation with $\Omega = 120\pi$. This phenomenon is illustrated in [14].

4.4.2. Second-order approximate integrator. The second-order approximate integrator (4.2), or equivalent system (4.3), eliminates the difficulties associated with both inexact initial conditions and DC gains or biases in the data. The elimination of constants in the data can heuristically be attributed to the differentiation of the acceleration data. This can be made rigorous by analytically solving the problem.

We again let $V(s) = \mathcal{L}\{v(t)\}$ and $A(s) = \mathcal{L}\{a(t)\}$ and let v_0, v_1, and a_0 denote initial conditions. Transformation of (4.2) yields

$$\left[s^2V(s)-sv_0-v_1\right]+\Omega\left[sV(s)-v_0\right]+\Omega^2V(s)=\frac{1}{RC}\left[sA(s)-a_0\right],$$

from which it follows that

$$V(s)=\frac{s+\Omega}{s^2+\Omega s+\Omega^2}v_0+\frac{(v_1-a_0/RC)}{s^2+\Omega s+\Omega^2}+\frac{s}{s^2+\Omega s+\Omega^2}\cdot\frac{1}{RC}A(s).$$

Inverse transformation then yields the solution

$$\begin{aligned}v(t) = e^{-\Omega t/2}&\left[\cos\left(\sqrt{3}\Omega t/2\right)v_0+\frac{1}{\sqrt{3}}\sin\left(\sqrt{3}\Omega t/2\right)v_0\right.\\&\left.+\frac{2}{\sqrt{3}\Omega}\sin\left(\sqrt{3}\Omega t/2\right)\left(v_1-\frac{a_0}{RC}\right)\right]\\+\frac{1}{RC}\int_0^t e^{-\Omega(t-s)/2}&\left[\cos\left(\sqrt{3}\Omega(t-s)/2\right)-\frac{1}{\sqrt{3}}\sin\left(\sqrt{3}\Omega(t-s)/2\right)\right]a(s)ds.\end{aligned}$$

As in the discussion of (4.1), we then consider the corresponding solution with perturbed initial conditions $\tilde{v}_0, \tilde{v}_1$, and $\tilde{a}_0$ and DC gain $\tilde{g}_a$. In this case, the perturbed solution $\tilde{v}(t)$ is given by

$$\begin{aligned}\tilde{v}(t) = v(t)&+e^{-\Omega t/2}\left[\cos\left(\sqrt{3}\Omega t/2\right)+\frac{1}{\sqrt{3}}\sin\left(\sqrt{3}\Omega t/2\right)\right]\left[\tilde{v}_0-v_0\right]\\&+e^{-\Omega t/2}\cdot\frac{2}{\sqrt{3}\Omega}\sin\left(\sqrt{3}\Omega t/2\right)\left[\left(\tilde{v}_1-\frac{\tilde{a}_0}{RC}\right)-\left(v_1-\frac{a_0}{RC}\right)\right]\\&+e^{-\Omega t/2}\cdot\frac{2\tilde{g}_a}{\sqrt{3}RC\Omega}.\end{aligned}$$

Here both the perturbations in initial conditions and the added constants in the data exponentially decay with the rate of decay dependent on the magnitude of Ω. This is illustrated in examples given in [14]. We reiterate that while increased values of Ω lead to more rapid decay of perturbations and biases, the solution to the differential equation less accurately approximates the true velocity. This, in combination with the goal of accurately preserving signal frequencies, leads to the condition $\Omega < \omega/6$.

4.4.3. Numerical approximation. Because of the potential for problems involving DC gains with the first-order formulation (4.1), we concentrate primarily on the second-order filter (4.2) and the corresponding system (4.3). We note that similar scalar techniques can be used to approximate the solution to (4.1). In considering numerical techniques for integrating (4.3), emphasis was placed on using a technique which could easily be implemented in real time. The two methods considered here are Euler's method and a backward Euler's method. The two are summarized below.

Euler's method.

$$z(t_{k+1}) = [I + \Delta t A]\, z(t_k) + \Delta t f(t_k)$$
$$\Rightarrow \begin{bmatrix} v \\ e \end{bmatrix}(t_{k+1}) = \begin{bmatrix} 1 - \Delta t\Omega & \Delta t\Omega \\ -\Delta t\Omega & 1 \end{bmatrix} \begin{bmatrix} v \\ e \end{bmatrix}(t_k) + \begin{bmatrix} \Delta t \dfrac{a(t_k)}{RC} \\ 0 \end{bmatrix}.$$

Backward Euler's method.

$$z(t_k) = [I - \Delta t A]^{-1} z(t_{k-1}) + [I - \Delta t A]^{-1} \Delta t f(t_{k-1}),$$

where

$$[I - \Delta t A]^{-1} = \begin{bmatrix} \dfrac{1}{1 + \Delta t\Omega + (\Delta t\Omega)^2} & \dfrac{\Delta t\Omega}{1 + \Delta t\Omega + (\Delta t\Omega)^2} \\ \dfrac{-\Delta t\Omega}{1 + \Delta t\Omega + (\Delta t\Omega)^2} & \dfrac{1 + \Delta t}{1 + \Delta t\Omega + (\Delta t\Omega)^2} \end{bmatrix},$$

$$\Delta t\, [I - \Delta t A]^{-1} f(t_{k-1}) = \begin{bmatrix} \dfrac{(\Delta t/RC) a(t_{k-1})}{1 + \Delta t\Omega + (\Delta t\Omega)^2} \\ \dfrac{(-(\Delta t)^2\Omega/RC) a(t_{k-1})}{1 + \Delta t\Omega + (\Delta t\Omega)^2} \end{bmatrix}.$$

The advantage of the backward Euler's method over Euler's method is its stability properties, with slightly more involved matrices being the disadvantage. Numerical examples demonstrating both methods with a variety of exogenous forces can be found in [14].

5. Experimental results. Experimental results demonstrating both the transient and steady-state capabilities of the control methodology are presented in this section. The circular plate used in these experiments had a radius of 9 inches (.2276m) and a thickness of .05 inch (.00127m). A pair of piezoceramic patches having radius .75 inch (.01905m) and thickness .007 inch (.0001778m) was bonded to the center of the plate. In both the transient and the steady-state experiments, only one patch was used for control. In the steady-state experiments, the opposite patch was used to drive the plate while it was allowed to remain uncharged in the transient case. The plate was mounted in a wooden frame by a circular aluminum collar which provided boundary conditions which were sufficiently close to clamped (zero displacement and slope).

The first step in the process was the estimation of physical parameters through fit-to-data techniques. As detailed in [2, 15], transient plate vibrations were excited though an impact hammer strike or the input of a voltage spike to the patches, and acceleration data were measured. The parameter values summarized in Table 5.1 were obtained through a least squares minimization of the difference between the model response and the measured data. These values were then employed when constructing the component matrices A, B, and Q (see Table 4.2) used during the experimental implementation of the controller.

TABLE 5.1
Physical parameters used in the experiments.

		Physical parameters
$\rho \cdot h$ (kg/m^2)	Plate	3.170
	Plate + Pzt	3.216
D $(N \cdot m)$	Plate	11.151
	Plate + Pzt	11.506
c_D $(N \cdot m \cdot sec)$	Plate	$1.443 - 4$
	Plate + Pzt	$2.031 - 4$
ν	Plate	.326
	Plate + Pzt	.325
$\mu\ (sec \cdot N/m^3)$		17.021
$\mathcal{K}^B\ (N/V)$ — Controlling Patch		.016
$\mathcal{K}^B\ (N/V)$ — Driving Patch		.017

5.1. Transient control. To investigate the capabilities of the method for controlling transient vibrations, decaying plate responses generated by an impact hammer strike were considered. In each case, the strike was directed to the center of the plate and hence the plate response was axisymmetric. In the first set of experiments, data were collected from an accelerometer located at the plate center on the side opposite from the hammer impact; thus $P_1 = (r_1, \theta_1) = (0, 0)$ in the construction of the observation matrix C described in Table 4.2. The experiments were then repeated with the accelerometer placed at the off-center point $P_2 = (r_1, \theta_1) = (2'', 0)$ to illustrate that collocation between the sensor and actuator is unnecessary in this control method (see Figure 5.1 for sensor, actuator, and impact locations). The results obtained with the off-center accelerometer are presented here, and the reader is referred to [14] for a discussion of the transient control results obtained with observations from the centered accelerometer.

Since no exogenous force was applied to the plate, the state estimator and control law summarized in Algorithm 4.1 were used to compute the controlling voltage to the patch. The component matrices as well as the cost functional and observation parameters used in these experiments are summarized in Tables 4.2 and 4.3.

Data acquisition and processing were performed with a PC-based Texas Instruments TMS 320-C30 digital signal processing (DSP) board. A schematic of the amplifiers, filters, DSP configuration and PC algorithm is given Figure 5.2. In the experiments, the accelerometer voltage was initially boosted by a factor of 10 and then reduced by 5 dB before reaching the DSP board. The controlling voltage output from the DSP was also boosted by an amplifier before input to the patches. This was necessary since the maximum voltage output by the DSP is 2.5 V whereas $60-70$ V were needed at the patch. The reader is also referred to Figure 4.1 for an illustration of the experimental process and to [14] for details regarding the implementation process. We point out that the ratio $\frac{2048}{2.5V}$, illustrated in the A/D conversion, results when the voltage range -2.5 to 2.5 is discretized into 4096 possible digital values. The reciprocal process occurs when digital values are converted to analog voltages in the D/A converter.

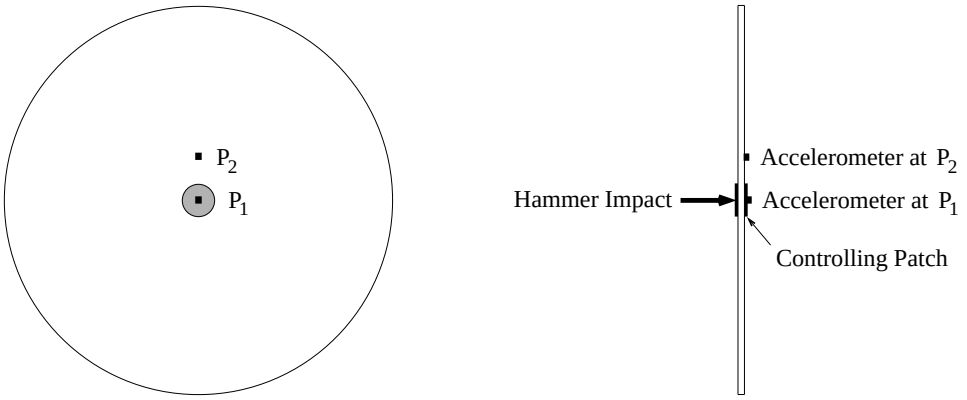


FIG. 5.1. *Patch, accelerometer, and impact locations for the experiments involving control of transient vibrations.*

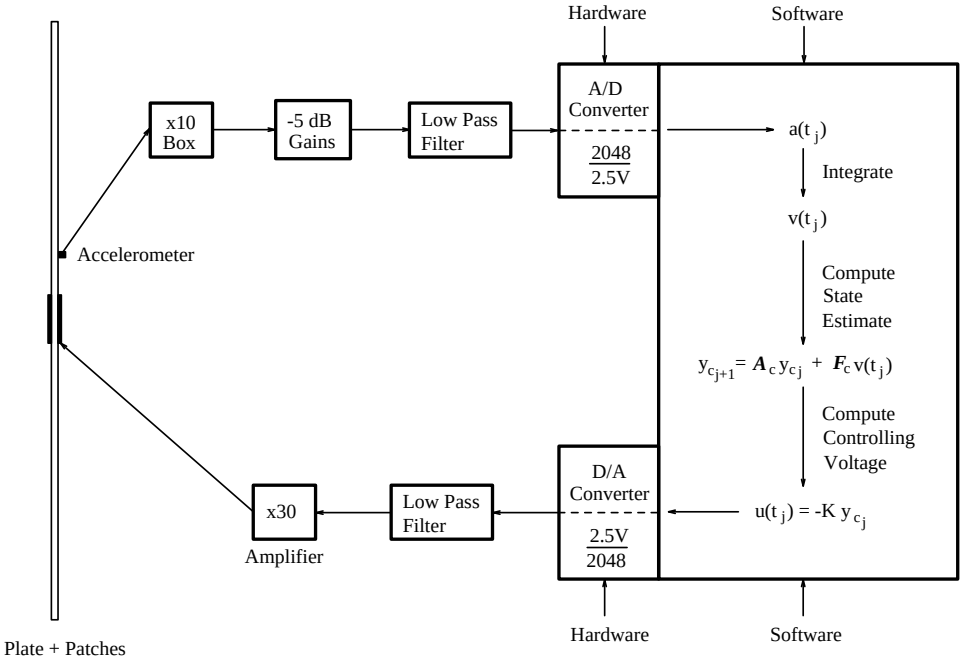


FIG. 5.2. *Amplifiers, DSP configuration, and PC Algorithm* 4.1 *for controlling a plate excited by an initial impact. Component matrices are defined in Figures* 4.2 *and* 4.3.

The control code was written in assembler in order to attain sufficiently fast sample rates for resolving transient frequencies excited by the hammer impact. While the code ran at rates greater than 7 KHz, a sample rate of 3.5 KHz was used in the experiments. This proved to be sufficient for resolving the three axisymmetric modes (with frequencies of 60 Hz, 227 Hz, and 512 Hz) excited in the experiments.

Representative plots of the plate velocity (integrated from the data) at the off-center point P_2 in the uncontrolled and controlled cases are given in Figure 5.3, with reduction levels at times $t = 0.5, 1.0, 1.5$ sec summarized in Table 5.2. The percentage

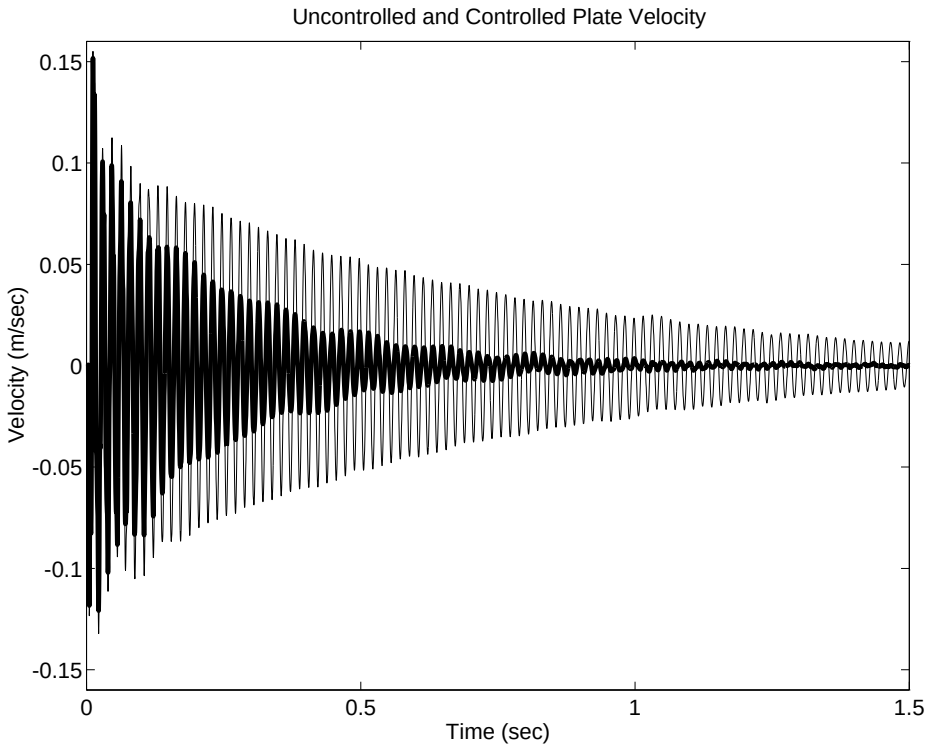


FIG. 5.3. *Uncontrolled and controlled plate vibrations at* $(2'', 0)$ *in response to an impact hammer hit:* —— *(uncontrolled),* ▬▬ *(controlled).*

TABLE 5.2
The percent reductions in acceleration and velocity levels at the point $P_2 = (2'', 0)$ *when feedback control is implemented.*

Time	Acceleration	Velocity
.5 sec	69.5	68.2
1 sec	88.9	84.7
1.5 sec	93.8	97.8

reductions were calculated by determining the ratio between the maximum values of the controlled and uncontrolled trajectories through one period containing the time point of interest. As illustrated by the results in Table 5.2 and Figure 5.3, the velocity level in the controlled case has been reduced by 50% before .5 sec and is essentially fully attenuated by 1.5 sec. We reiterate that these results were obtained with data obtained from an accelerometer at $(2'', 0)$ and a centered actuating patch, thus illustrating that collocation is unnecessary for this control method.

The voltage $u(t_j) = -Ky_{c_j}$ was recorded in each experiment and that voltage yielding the control results reported here is plotted in Figure 5.4. It is noted that the voltage has a maximum magnitude of 70 V. In practice, it has been observed that the patches can be used for extended periods at the frequencies of interest without damage or degradation of performance if the voltage levels are maintained below 8–10 rms V/mil [38]. Hence control voltage levels required for control of the transient vibrations is well within the tolerance of the 7-mil patches used in the experiments.

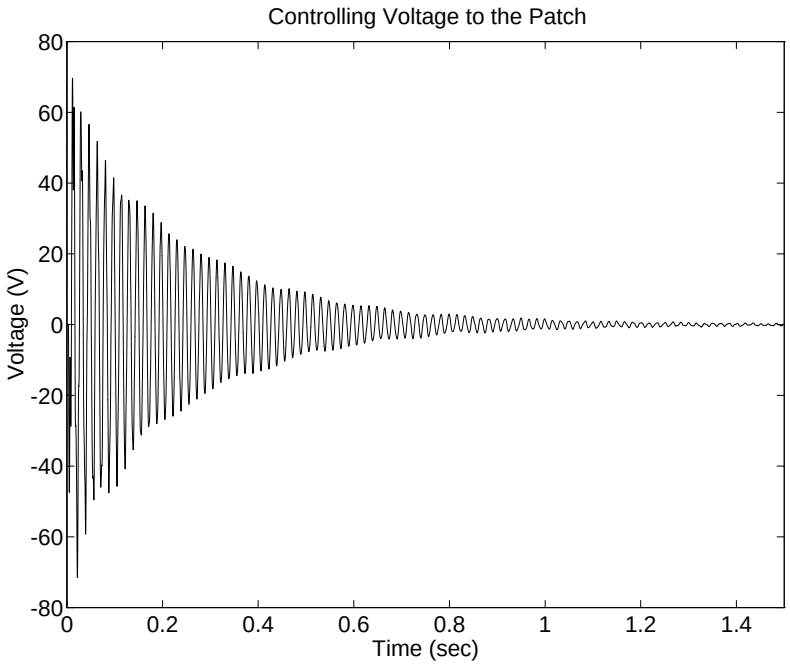


FIG. 5.4. *The controlling voltage for acceleration data observed at* $(2'', 0)$.

Finally, the force delivered by the hammer impact in the uncontrolled and controlled cases is plotted in Figure 5.5 so as to provide a means of testing the equity of excitation levels in the uncontrolled and controlled experiments (the initial velocities in the two cases can also be compared to determine whether the same level of energy is being delivered in each experiment). As indicated by the results in this latter figure, the force delivered in the two cases is nearly identical (a slight double hit was always present when an impact hammer was used to excite the axisymmetric modes).

5.2. Control of a periodic exogenous force. A second problem under consideration concerns the control of plate vibrations when the plate is driven by a periodic exogenous force. To demonstrate the control capabilities in this case, a periodic driving voltage was supplied to one centered patch on the plate, and the patch on the opposite side of the plate was used as the control actuator. Experimental tests indicated that a 350-Hz driving voltage produced a strong plate response and all tests were conducted with the exogenous voltage at that frequency.

Two sources were used to generate this exogenous signal: namely, an external oscillator and the PC running the control algorithm. As reported under Case 1 below, a purely steady-state response could be considered with the oscillator-generated exogenous force since the plate was driven to steady state before the control program was initiated. Both a transient and steady-state response were noted in the PC-generated signal since the input of the exogenous force to the plate began at the same time that the control algorithm was started. This latter means of excitation is considered in Case 2.

Since the system was driven by a periodic exogenous force, the discrete-time Algorithm 4.2 (corresponding to continuous-time Algorithm 3.2) was used to calculate the controlling voltage to the actuating patch. Again, component matrices as well as control and observation parameters are summarized in Tables 4.2 and 4.3.

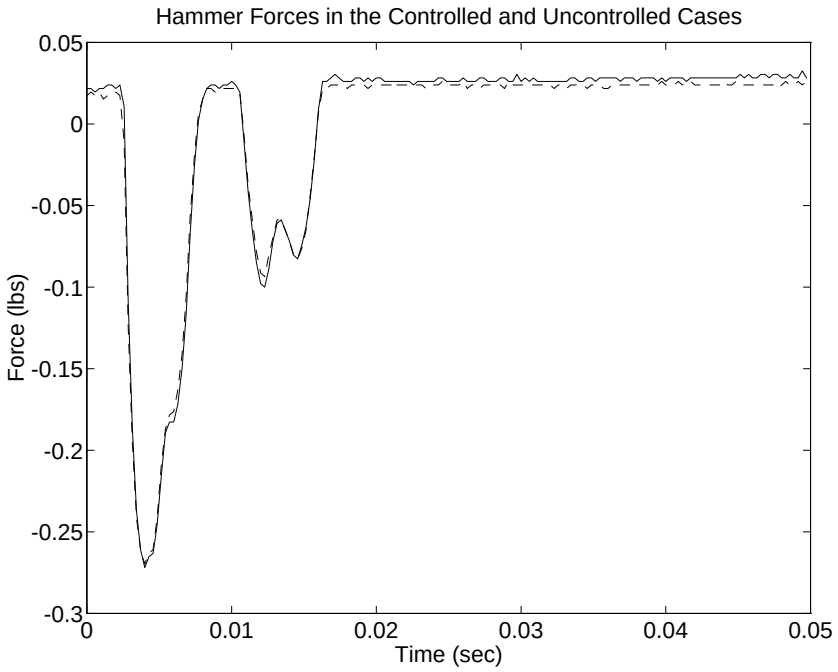


FIG. 5.5. *The force delivered by the impact hammer in the uncontrolled and controlled cases:* ——– *(controlled case),* – – – *(uncontrolled case).*

For the off-center results reported here, velocity observation values were obtained by integrating data from an off-center accelerometer located at the point $P_2 = (2'', 0)$ as depicted in Figure 5.1. The second input to the algorithm consisted of measurements $g(t_j)$ of the force driving the plate. We point out that when implementing this control method, both phase and magnitude information for the driving force were required, as compared with many other methods (e.g., feedforward) which require only phase information.

In these experiments, the tracking components r_j were calculated first and stored in a circular buffer. These values were then used when calculating the state estimates y_j and voltages $u(t_j)$. While the implementation in this manner facilitated running the algorithm with sample rates on the order of 7 KHz (again, the algorithm was coded in assembler), it limits the robustness of the method with respect to changes and variations in the driving force. Current efforts are directed toward simultaneous solution of the tracking and estimator difference equations.

A crucial issue when implementing the method concerns the handling of delays and phase shifts produced by the filters, A/D and D/A conversions, computation of the control voltages, et cetera. While the amount of delay and phase shift is frequency dependent, experiments indicated that at 350 Hz, 30°–40° phase shifts were introduced by the hardware. This was sufficient to destabilize the controller if left uncompensated. In the experiments, we compensated by first conducting an offline, numerical "identification" to determine the amount of added delay necessary for stabilizing the controller in the presence of phase shifts of the order introduced by the experimental hardware. A summary of these results can be found in [14]. This is analogous to the online tuning or phase locking which is necessary for ensuring stability in other control methods. The numerical tests, summarized in [14],

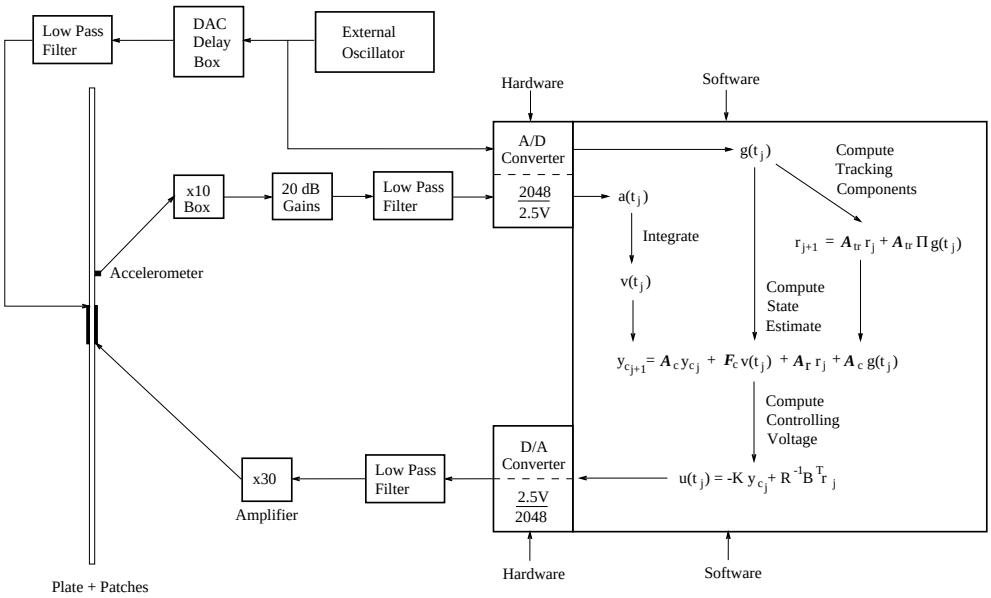


FIG. 5.6. *Amplifiers, DSP configuration, and PC Algorithm* 4.2 *for controlling a plate driven by a periodic exogenous force. Component matrices are defined in Tables* 4.2 *and* 4.3.

indicated that the introduction of a 216° delay in accelerometer or exogenous force data would stabilize the systems, and this was implemented in the experiments by using a DAC delay box. The compensation for phase delays in this manner provides merely a first step toward optimal implementation of the method, and one aspect of current research efforts is directed toward online compensation in the algorithm.

Data acquisition and processing was again performed with a PC-based Texas Instruments TMS 320-C30 board. A schematic of the setup is given in Figure 5.6.

Case 1. Oscillator-generated driving signal. For the results described here, the signal to the driving patch was generated by an external oscillator. The plate was allowed to reach steady state, and then the control program was initiated. Acceleration levels measured by the accelerometer located at $P_2 = (2'', 0)$ and integrated velocity values for the uncontrolled and controlled cases are plotted in Figure 5.7. As noted from the controlled trajectories in that figure, it takes the algorithm approximately 0.06 sec to calculate and store a sufficient number of tracking components r_j. During that time interval, no voltage is fed to the actuating patch. Once the tracking calculations are completed, state estimation begins and the controlling voltage is computed and fed back into the system. The vibration levels decay for approximately 0.3 sec and then are maintained at levels that are approximately 15% of those for the uncontrolled case for the remainder of the time interval. This corresponds to a $20\log(a_{con}/a_{uncon}) \approx -16.5$ dB reduction in acceleration levels.

While the magnitude of the controlling voltage is dependent upon the amplitude of the driving signal, magnitudes less than 40 V (28.3 V rms) were required to attain the levels reported here. At 350 Hz, this was well within the range (56 V–70 V rms in this case) that was considered to be safe for the patch being used (see section 5.1 for further discussion regarding the voltage levels at which the patches can be driven without damage or degradation in performance).

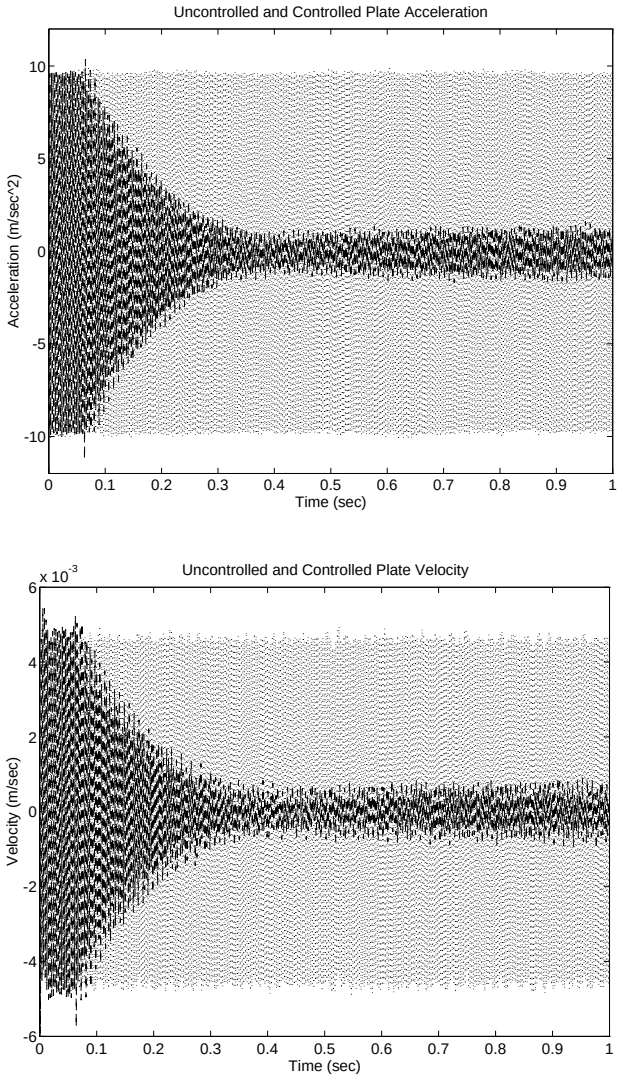


FIG. 5.7. *Uncontrolled and controlled plate acceleration and velocity at* $(2, 0)$ *for system driven by periodic exogenous force:* ——— *(uncontrolled),* ▬▬ *(controlled).*

The results in this experiment demonstrate the effectiveness of the algorithm for controlling a system that has reached steady state. Hence one is not required to start the control with a system at rest.

Case 2. PC-generated driving signal. A second mechanism for generating the driving signal is with the PC that is used to process data and run the control algorithm. Acceleration and velocity plots of the uncontrolled and controlled plate vibrations excited in this manner are given in Figure 5.8. It can be seen that in this case, the plate starts from rest and is still being driven through a transient stage when the tracking calculations are completed and control begins. At that point, the controlled trajectories are reduced to the levels noted in the purely steady-state case, whereas the uncontrolled trajectories are driven to steady state. Here, an 82% (15 dB) reduction in levels is noted at time $T = 1$ sec. This was obtained with a controlling

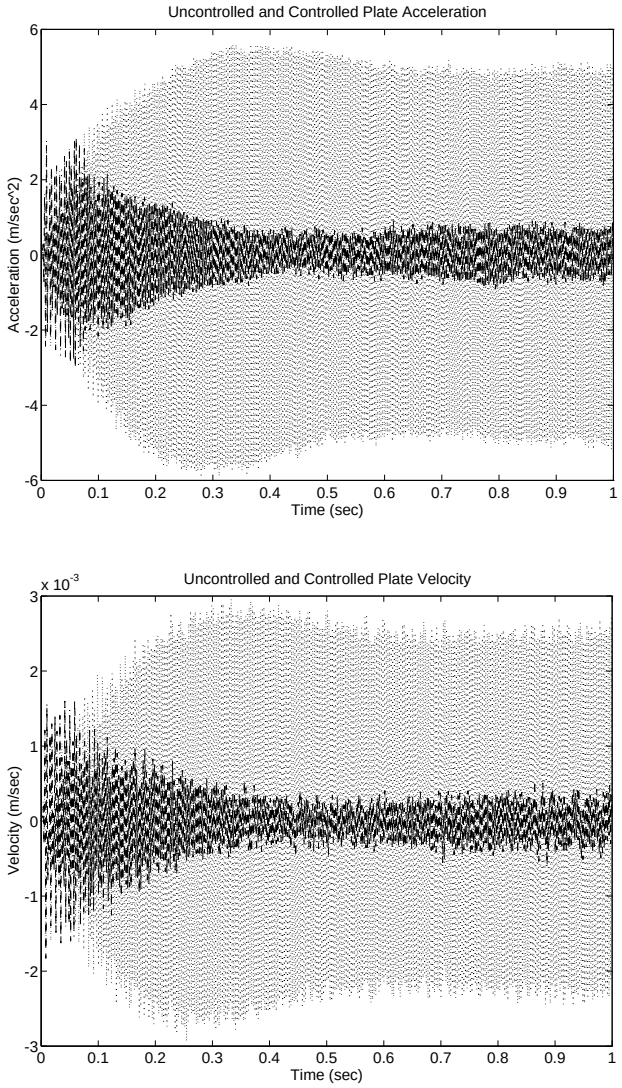


FIG. 5.8. *Uncontrolled and controlled plate acceleration and velocity at* $(2, 0)$ *for system driven by periodic exogenous force:* —— *(uncontrolled),* ▬▬ *(controlled).*

voltage of magnitude $12\,\mathrm{V}_{max}$ $(8.5\,\mathrm{V}_{rms})$. These results demonstrate the effectiveness of the control algorithm for a system undergoing transient oscillations before reaching steady state in response to a periodic driving force.

6. Conclusions. In this work, the experimental implementation of a PDE-based controller was considered. While the motivating application involves the control of vibration levels for a circular plate through the excitation of surface-mounted piezoceramic patches, the general techniques described here will extend to a variety of applications.

For such control techniques, the first step is the derivation of a PDE model which accurately describes the dynamics of the system under consideration. Physical parameters in these models must typically be estimated through fit-to-data techniques be-

fore control applications can be considered. Following a brief discussion regarding the strong and weak forms of a thin plate model with a discontinuous control input term (due to the piecewise constant nature of the piezoceramic patches), continuous-time LQG and H^∞ methods for systems with no exogenous force or a periodic exogenous force were discussed. The discrete-time approximations necessary for implementing the methods with digital measurements were also presented. A crucial step when implementing the discrete-time controllers involves the approximate integration of data (e.g., accelerometer data integrated to obtain velocity state values), and first- and second-order filters for accomplishing this were discussed. Without such filters, DC biases, which are always present in the data, would render the integrated values useless.

Experimental results demonstrating the control of transient and steady-state vibrations were then presented. One advantage of the PDE-based controllers over standard frequency response input/output techniques is the capability for direct control of transient responses, and this was demonstrated in the first set of examples. A centered hammer impact was used to excite the plate, and integrated data from an off-center accelerometer were used to reconstruct the state. The results demonstrate that attenuation levels on the order of 95% reduction can be attained by 1.5 sec using the PDE-based controller.

The second example demonstrates the control of transient and steady-state responses when the plate was driven by a periodic exogenous voltage to a secondary piezoceramic patch. These results demonstrate that, after accounting for hardware delays, attenuation levels on the order of 85% were attained when control was implemented. While implementation techniques are still being refined, these results demonstrate the effectiveness of the PDE-based controller for this system and indicate the potential of these control techniques for reducing transient and steady-state dynamics in other structural and structural acoustic systems.

Acknowledgments. The authors extend their sincere thanks to Yun Wang, Brooks Air Force Base, for substantial collaboration throughout the duration of this work.

REFERENCES

[1] T. Bailey and J.E. Hubbard, Jr., *Distributed piezoelectric-polymer active vibration control of a cantilever beam*, J. Guidance, 8 (1985), pp. 605–611.

[2] H.T. Banks, D.E. Brown, V. Metcalf, R.J. Silcox, R.C. Smith, and Y. Wang, *A PDE-based methodology for modeling, parameter estimation and feedback control in structural and structural acoustic systems*, in Proc. North American Conference on Smart Structures and Materials, Orlando, FL, 1994, pp. 311–320.

[3] H.T. Banks and J. Burns, *Hereditary control problems: Numerical methods based on averaging approximations*, SIAM J. Control Optim., 16 (1978), pp. 169–208.

[4] H.T. Banks, M.A. Demetriou, and R.C. Smith, *An H^∞/MinMax periodic control in a 2-D structural acoustic model with piezoceramic actuators*, IEEE Trans. Automat. Control, 41 (1996), pp. 943–959.

[5] H.T. Banks, M.A. Demetriou, and R.C. Smith, *Robustness studies for H^∞ feedback control in a structural acoustic model with periodic excitation*, Internat. J. Robust and Nonlinear Control, 6 (1996), pp. 453–478.

[6] H.T. Banks, W. Fang, R.J. Silcox, and R.C. Smith, *Approximation methods for control of acoustic/structure models with piezoceramic actuators*, J. Intelligent Material Systems Structures, 4 (1993), pp. 98–116.

[7] H.T. Banks and K. Ito, *Approximation in LQR Problems for Infinite Dimensional Systems with Unbounded Input Operators*, Technical Report, CRSC-TR94-22, Center for Research in Scientific Computation, North Carolina State University, Raleigh, NC, 1994; summary in J. Math. Systems Estim. Control, 7 (1997), pp. 119–122.

[8] H.T. Banks, K. Ito, and B.B. King, *Theoretical and computational aspects of feedback in structural systems with piezoceramic controllers*, in Computation and Control III, K.L. Bowers and J. Lund, eds., Birkhäuser Boston, Cambridge, MA, 1993, pp. 1–27.

[9] H.T. Banks, K. Ito, and Y. Wang, *Computational methods for identification and feedback control in structures with piezoceramic actuators and sensors*, in Proc. Conference on Recent Advances in Adaptive and Sensory Materials and Their Applications, VPI and SU, Blacksburg, VA, 1992, pp. 111–119.

[10] H.T. Banks and K. Kunisch, *The linear regulator problem for parabolic systems*, SIAM J. Control Optim., 22 (1984), pp. 684–698.

[11] H.T. Banks, I.G. Rosen, and K. Ito, *A spline-based technique for computing riccati operators and feedback controls in regulator problems for delay systems*, SIAM J. Sci. Statist. Comput., 5 (1984), pp. 830–855.

[12] H.T. Banks and R.C. Smith, *The Modeling and Approximation of a Structural Acoustics Problem in a Hard-Walled Cylindrical Domain*, Technical Report, CRSC-TR94-26, Center for Research in Scientific Computation, North Carolina State University, Raleigh, NC, 1994.

[13] H.T. Banks and R.C. Smith, *Noise control in a 3-D structural acoustic system: Numerical simulations*, Proc. Second International Conference on Intelligent Materials, Williamsburg, VA, 1994, pp. 128–139.

[14] H.T. Banks and R.C. Smith, *Implementation Issues Regarding PDE-Based Controllers—Control of Transient and Periodic Plate Vibrations*, CRSC Technical Report, CRSC-TR95-16, Center for Research in Scientific Computation, North Carolina State University, Raleigh, NC, 1995.

[15] H.T. Banks, R.C. Smith, D.E. Brown, V.L. Metcalf, and R.J. Silcox, *The estimation of material and patch parameters in a PDE-based circular plate model*, J. Sound Vibration, 199 (1996), pp. 777–799.

[16] H.T. Banks, R.C. Smith, and Y. Wang, *The modeling of piezoceramic patch interactions with shells, plates and beams*, Quart. Appl. Math., 53 (1995), pp. 353–387.

[17] H.T. Banks, R.C. Smith, and Y. Wang, *Vibration suppression with approximate finite dimensional compensators for distributed systems: Computational methods and experimental results*, in Proc. Second International Conference on Intelligent Materials, Williamsburg, VA, 1994, pp. 140–154.

[18] H.T. Banks, R.C. Smith, and Y. Wang, *Modeling and parameter estimation for an imperfectly clamped plate*, in Computation and Control IV, K.L. Bowers and J. Lund, eds., Birkhäuser Boston, Cambridge, MA, 1995, pp. 23–42.

[19] H.T. Banks, R.C. Smith and Y. Wang, *Smart Material Structures: Modeling, Estimation and Control*, Recherches en Mathématiques Appliquées, Masson/John Wiley & Sons, Paris/New York, 1996.

[20] T. Başar and P. Bernhard, *H^∞-Optimal Control and Related Minimax Design Problems*, Birkhäuser Boston, Cambridge, MA, 1991.

[21] D.S. Bernstein and D.C. Hyland, *The optimal projection equations for finite dimensional fixed order dynamic compensator of infinite dimensional systems*, SIAM J. Control Optim., 24 (1986), pp. 122–151.

[22] H.T. Burns, K. Ito, and G. Propst, *On nonconvergence of adjoint semigroups for control systems with delay*, SIAM J. Control Optim., 26 (1988), pp. 1442–1454.

[23] G. Da Prato, *Synthesis of optimal control for an infinite dimensional periodic problem*, SIAM J. Control Optim., 25 (1987), pp. 706–714.

[24] J. D'Cruz, *The active control of panel vibrations with piezoelectric actuators*, in Proc. Conference on Recent Advances in Adaptive and Sensory Materials and Their Applications, VPI and SU, Blacksburg, VA, 1992, pp. 665–674.

[25] C.R. Fuller, G.P. Gibbs, and R.J. Silcox, *Simultaneous active control flexural and extension waves in beams*, J. Intelligent Material Systems Structures, 1 (1990), pp. 235–247.

[26] C.R. Fuller, C.H. Hansen, and S.D. Snyder, *Active control of structurally radiated noise using piezoceramic actuators*, in Proc. Inter-Noise 89, Newport Beach, CA, 1989, pp. 509–512.

[27] C.R. Fuller, S.D. Snyder, C.H. Hansen, and R.J. Silcox, *Active control of interior noise in model aircraft fuselages using piezoceramic actuators*, in AIAA 13th Aeroacoustics Conference, Tallahassee, FL, 1990, paper 90-3922.

[28] J.S. Gibson and A. Adamian, *Approximation theory for linear-quadratic-Gaussian optimal control of flexible structures*, SIAM J. Control Optim., 29 (1991), pp. 1–37.

[29] W.L. Hallauer and S.E. Lamberson, *Experimental active vibration damping of a plane truss using hybrid actuation*, in Proc. 30th AIAA/ASME/ASCE/AHS/ASC Structural Dynamics and Materials Conference, paper 89-1169-CP, 1989, pp. 80–90.

[30] K. Ito, *Finite-dimensional compensators for infinite-dimensional systems via Galerkin-type approximation*, SIAM J. Control Optim., 28 (1990), pp. 1251–1269.
[31] B. van Keulen, *H_∞-Control for Distributed Parameter Systems: A State-Space Approach*, Birkhäuser Boston, Cambridge, MA, 1993.
[32] S. Koshigoe and J.W. Murdock, *A unified analysis of both active and passive damping for a plate with piezoelectric transducers*, J. Acoust. Soc. Amer., 93 (1993), pp. 346–355.
[33] H. Kwakernaak and R. Sivan, *Linear Optimal Control System*, John Wiley, New York, 1972.
[34] J.D. Lambert, *Numerical Methods for Ordinary Differential Systems: The Initial Value Problem,* John Wiley, New York, 1991.
[35] I. Lasiecka, *Galerkin approximations of infinite-dimensional compensators for flexible structures with unbounded control action*, Acta Appl. Math., 28 (1992), pp. 101–133.
[36] I. Lasiecka, *Finite element approximations of compensator design for analytic generators with fully unbounded controls/observations*, SIAM J. Control Optim., 33 (1995), pp. 67–88.
[37] E.H. Mansfield, *The Bending and Stretching of Plates*, International Series of Monographs on Aeronautics and Astronautics 6, Macmillan, New York, 1964.
[38] V.L. Metcalf, personal communication, U.S. Army Research Laboratory, NASA Langley Research Center, Hampton, VA, 1992.
[39] R.C. Smith, *A Galerkin method for linear PDE systems in circular geometries with structural acoustic applications*, SIAM J. Sci. Comput., 18 (1997), pp. 371–402.

SIAM J. CONTROL OPTIM.
Vol. 35, No. 4, pp. 1297–1327, July 1997

010

PONTRYAGIN'S PRINCIPLE FOR STATE-CONSTRAINED BOUNDARY CONTROL PROBLEMS OF SEMILINEAR PARABOLIC EQUATIONS*

EDUARDO CASAS†

Abstract. This paper deals with state-constrained optimal control problems governed by semilinear parabolic equations. We establish a minimum principle of Pontryagin's type. To deal with the state constraints, we introduce a penalty problem by using Ekeland's principle. The key tool for the proof is the use of a special kind of spike perturbations distributed in the domain where the controls are defined. Conditions for normality of optimality conditions are given.

Key words. Pontryagin principle, boundary control, semilinear parabolic equations, optimality conditions, state constraints

AMS subject classifications. 49K20, 35K20

PII. S0363012995283637

1. Introduction.

In the last years, some proofs of minimum principles of Pontryagin's type have appeared. For long time, the optimality conditions for control problems governed by partial differential equations (PDEs) have been given in an integral form, assuming the convexity of the control set and the differentiability with respect to the control and state of all functions involved in the problem. This makes a big difference with the control theory for problems governed by ordinary differential equations (ODEs), where a Pontryagin principle is derived without the previous assumptions. In my opinion, the reason for this difference is the difficulty of extending the methods used for ODEs to infinite-dimensional systems. In particular, the classical spike perturbations of the controls localized around a point do not work properly for PDEs because they lead to some equations with Dirac measures as data, which produce noncontinuous solutions. This makes it difficult to treat the state constraints, especially the pointwise state constraints.

A new type of spike perturbation was developed by a group of mathematicians from Fudan University; see Li [25], Li and Yao [26], and Li and Yong [27]. They used these perturbations to study control problems of evolution equations. The spike perturbations were defined by using the representation of the state given by the corresponding semigroup. This idea was also followed by Fattorini [17], [18]; Fattorini and Frankowska [19]; and Fattorini and Murphy [20], [21]. Later Yong [33] and Casas and Yong [14] built a similar kind of spike perturbations for elliptic equations by using the representation of the solution with the aid of the Green function. Afterwards, Casas suggested a new construction of the set where the perturbations were localized; see Casas [11] and Bei Hu and Yong [22]. This construction was independent of the equation. For a different viewpoint explaining the true nature of this new type of spike perturbations, the reader is referred to Casas [12], where the boundary control of a quasi-linear elliptic equation was considered.

Bonnans and Casas [5], [6] followed a different approach to derive Pontryagin's principle that did not use this type of spike perturbations. However, it was necessary

*Received by the editors March 24, 1995; accepted for publication (in revised form) May 13, 1996. This research was partially supported by Dirección General de Investigación Científica y Técnica (Madrid).

http://www.siam.org/journals/sicon/35-4/28363.html

†Departamento de Matemática Aplicada y Ciencias de la Computación, E.T.S.I. Industriales y de Telecomunicación, Universidad de Cantabria, 39071 Santander, Spain (casas@etsiso.macc.unican.es).

to assume a stability condition of the optimal cost functional with respect to small perturbation of the feasible state set.

In this paper, we consider a boundary control problem governed by a parabolic semilinear equation. General state constraints are included in the formulation of the problem. The idea developed in [12] is used here. To deal with the state constraints we penalize them. The lack of convexity of the control set and the noncontinuity with respect to the control of the functions involved in the control problem make it difficult to formulate a penalty problem having a solution converging to the optimal control of the original problem, however. Ekeland's variational principle is the key to obtaining the suitable penalization.

Pontryagin's principle is often established in a nonqualified form, which implies that the cost functional does not appear in the conditions for optimality. In the absence of equality state constraints, we give a condition that leads to a qualified optimality system. This condition was introduced by Bonnans [4] and Bonnans and Casas [6]. It consists of assuming a certain kind of Lipschitz dependence of the optimal cost functional with respect to small perturbations of the state constraint. It is proved that this condition is satisfied "almost everywhere (a.e.)." We will distinguish strong and weak Pontryagin principles, depending on whether the optimality system is qualified or not. To prove the strong principle we make an exact penalization of the state contraints.

One of the difficulties found in the optimality system is the adjoint state equation. This equation can have measures as data in the domain, on the boundary, and as a final condition. There are not many papers written about parabolic equations involving measures. For these equations the reader is referred to Barbu and Precupanu [1], Lasiecka [24], Tröltzsch [32], and Boccardo and Gallouët [3], the last one dealing with quasi-linear equations. Here we use the transposition method to derive a general result of existence and "uniqueness" of solution. Since we do not assume continuity of the coefficients of the state equation, we need to be precise in which sense the solution is unique; see Serrin [30] for a nonuniqueness result in $W_0^{1,p}(\Omega)$ $(p < 2)$ of an elliptic problem well posed in $H^1(\Omega)$.

The paper is organized as follows. In the next section, the control problem is formulated. The state constraints are presented in an abstract framework. We show through some examples how the usual state constraints are included in the abstract formulation. The weak and strong Pontryagin principles are formulated in sections 3 and 4, respectively. In section 5, the state equation is studied and the spike perturbations are defined. The linear parabolic equations involving measures are analyzed in section 6. All the mentioned papers dealing with control of evolution equations, except [22], followed the semigroup approach to analyze the state and adjoint state equations. Here we will follow the variational approach, which allows us to obtain some pointwise information of the solutions of the PDEs. This information is very important for studying the control problems with pointwise state constraints. Finally, the proofs of weak and strong principles are given in section 7.

2. Setting of the control problem. Let $\Omega \subset \mathbb{R}^n$, $n \geq 1$, be an open and bounded set, with Lipschitz boundary Γ. Given $0 < T < +\infty$, we set $\Omega_T = \Omega \times (0,T)$ and $\Sigma_T = \Gamma \times (0,T)$. Let $(\mathcal{K}, d)$ be a metric space and let us consider a function $f : \Sigma_T \times \mathbb{R} \times \mathcal{K} \longrightarrow \mathbb{R}$ of class C^1 with respect to the second variable and satisfying the following assumptions:

$$\frac{\partial f}{\partial y}(x,t,y,u) \leq 0 \quad \forall (x,t,y,u) \in \Sigma_T \times \mathbb{R} \times \mathcal{K}; \tag{2.1}$$

$$\text{(2.2)}\qquad \begin{cases} \forall M>0\ \exists C_M>0 \text{ such that } \forall (x,t,u)\in \Sigma_T\times\mathcal{K} \text{ and } |y|\le M, \\ |f(x,t,0,u)|+\left|\dfrac{\partial f}{\partial y}(x,t,y,u)\right|\le C_M. \end{cases}$$

The state equation is as follows:

$$\text{(2.3)}\qquad \begin{cases} \dfrac{\partial y}{\partial t}(x,t)+Ay(x,t)+a_0(x,t,y(x,t))=0 & \text{in } \Omega_T, \\ \partial_{\nu_A}y(x,t)=f(x,t,y(x,t),u(x,t)) & \text{on } \Sigma_T, \\ y(x,0)=y_0(x) & \text{in } \Omega, \end{cases}$$

where $y_0\in C(\bar\Omega)$, A is the linear operator

$$\text{(2.4)}\qquad \begin{aligned} Ay= & -\sum_{j=1}^n \partial_{x_j}\left\{\sum_{i=1}^n [a_{ij}(x,t)\partial_{x_i}y(x,t)]+b_j(x,t)y(x,t)\right\} \\ & +\sum_{j=1}^n d_j(x,t)\partial_{x_j}y(x,t)+c(x,t)y(x,t), \end{aligned}$$

and

$$\text{(2.5)}\qquad \partial_{\nu_A}y(x,t)=\sum_{j=1}^n\left\{\sum_{i=1}^n [a_{ij}(x,t)\partial_{x_i}y(x,t)]+b_j(x,t)y(x,t)\right\}\nu_j(x),$$

$\nu(x)$ being the outward unit normal vector to Γ at the point x; see Casas [9] or Casas and Fernández [13] for an interpretation of this Neumann condition in a trace sense. Function $a_0:\Omega_T\times\mathbb{R}\longrightarrow\mathbb{R}$ is a Carathéodory function of class C^1 with respect to the second variable and satisfies the following assumptions:

$$\text{(2.6)}\qquad \begin{cases} \exists \psi_0\in L^{\hat p}([0,T],L^{\hat q}(\Omega)) \text{ and } C_1>0 \text{ such that} \\ a_0(x,t,y)y\ge \psi_0(x,t)-C_1y^2 \quad \forall (x,t,y)\in\Omega_T\times\mathbb{R}; \end{cases}$$

$$\text{(2.7)}\qquad \begin{cases} a_0(\cdot,\cdot,0)\in L^{\hat p}([0,T],L^{\hat q}(\Omega)) \text{ and } \forall M>0\ \exists C_M>0 \text{ such that} \\ \left|\dfrac{\partial a_0}{\partial y}(x,t,y)\right|\le C_M\ \forall (x,t)\in\Omega_T, |y|\le M; \end{cases}$$

where $\hat q,\hat p\in[1,+\infty]$ and $1/\hat p+n/2\hat q<1$.

As usual, we assume the following hypotheses on A:

$$\text{(2.8)}\qquad \begin{cases} a_{ij},b_j,d_j,c\in L^\infty(\Omega_T)\quad \forall i,j=1,\dots,n; \\ \displaystyle\sum_{i.j=1}^n a_{ij}(x,t)\xi_i\xi_j\ge\Lambda|\xi|^2\quad \forall\xi\in\mathbb{R}^n \text{ a.e. } (x,t)\in\Omega_T, \text{ with } \Lambda>0. \end{cases}$$

Once given the state equation, we introduce the cost functional

$$J(u)=\int_{\Omega_T}L(x,t,y_u(x,t))dxdt+\int_{\Sigma_T}l(x,t,y_u(x,t),u(x,t))d\sigma(x)dt,$$

where y_u is the solution of (2.3) associated with u; σ denotes the usual $(n-1)$-dimensional measure on Γ induced by the parametrization (remember that Γ is a Lipschitz manifold); and $L : \Omega_T \times \mathbb{R} \longrightarrow \mathbb{R}$ and $l : \Sigma_T \times \mathbb{R} \times \mathcal{K} \longrightarrow \mathbb{R}$ are of class C^1 with respect to the second variable, L being measurable with respect to the first one, satisfying

$$\begin{cases} \forall M > 0\ \exists \psi_{dM} \in L^1(\Omega_T) \text{ such that } \forall (x,t) \in \Omega_T, |y| \le M, \\ |L(x,t,0)| + \left| \dfrac{\partial L}{\partial y}(x,t,y) \right| \le \psi_{dM}(x,t) \end{cases} \tag{2.9}$$

and

$$\begin{cases} \forall M > 0\ \exists \psi_{bM} \in L^1(\Sigma_T) \text{ such that } \forall (x,t,u) \in \Sigma_T \times \mathcal{K}, |y| \le M, \\ |l(x,t,0,u)| + \left| \dfrac{\partial l}{\partial y}(x,t,y,u) \right| \le \psi_{bM}(x,t). \end{cases} \tag{2.10}$$

The space of controls $\mathcal{U}$ is formed by the measurable functions $u : \Sigma_T \longrightarrow \mathcal{K}$ such that the mapping

$$(x,t) \in \Sigma_T \longrightarrow (f(x,t,y,u(x,t)), l(x,t,y,u(x,t))) \in \mathbb{R}^2$$

is measurable for every $y \in \mathbb{R}$. In section 5 we will prove that there exists a unique solution of (2.3) in the space $Y = C(\bar{\Omega}_T) \cap L^2([0,T], H^1(\Omega))$ for every $u \in \mathcal{U}$, so that functional $J : \mathcal{U} \longrightarrow \mathbb{R}$ is well defined.

Finally we introduce the state constraints. Let Z be a separable Banach space and $Q \subset Z$ a closed convex subset with nonempty interior. Given two mappings of class C^1, $G : Y \longrightarrow Z$ and $F : C(\bar{\Omega}_T) \longrightarrow \mathbb{R}^s$, $s \ge 1$, we formulate the optimal control problem as follows:

$$\text{Minimize } \{J(u) : u \in \mathcal{U}, G(y_u) \in Q, F(y_u) = 0\}. \tag{P}$$

Let us show how the usual examples of state constraints can be handled with this formulation.

Example 2.1. Given a continuous function $g : \bar{\Omega}_T \times \mathbb{R} \longrightarrow \mathbb{R}$ of class C^1 in respect to the second variable, the constraint $g(x,t,y_u(x,t)) \le \delta$ for all $(x,t) \in \bar{\Omega}_T$, with $\delta > 0$ being a given number, can be written in the above framework by putting $Z = C(\bar{\Omega}_T)$, $G : Y \longrightarrow C(\bar{\Omega}_T)$, defined by $G(y) = g(\cdot, y(\cdot))$, and

$$Q = \{z \in C(\bar{\Omega}_T) : z(x,t) \le \delta \quad \forall (x,t) \in \bar{\Omega}_T\}.$$

Example 2.2. Let $\{(x_j,t_j)\}_{j=1}^s \subset \bar{\Omega}_T$; then we can include the equality constraints $y_u(x_j,t_j) = \delta_j$, $1 \le j \le s$, in the above formulation. Indeed, it is enough to define the functions $F_j : C(\bar{\Omega}_T) \longrightarrow \mathbb{R}$ given by $F_j(y) = y(x_j) - \delta_j$ and to take $F = (F_1, \dots, F_s)^T$. Then F is of class C^1.

Example 2.3. Let $g : \Omega \times [0,T] \times \mathbb{R} \longrightarrow \mathbb{R}$ be a function measurable with respect to the first variable, continuous with respect to the second, of class C^1 with respect to the third, and such that $\partial g/\partial y$ is also continuous in the last two variables. Moreover, it is assumed that for every $M > 0$ there exists a function $\psi_M \in L^1(\Omega)$ such that

$$|g(x,t,0)| + \left| \frac{\partial g}{\partial y}(x,t,y) \right| \le \psi_M(x) \quad \text{a.e. } x \in \Omega \ \ \forall t \in [0,T] \text{ and } |y| \le M.$$

Then the constraint

$$\int_\Omega g(x,t,y_u(x,t))dx \le \delta \quad \forall t \in [0,T]$$

is included in the above formulation by taking $Z = C[0,T]$,

$$Q = \{z \in C[0,T] : z(t) \le \delta \quad \forall t \in [0,T]\},$$

and $G : Y \longrightarrow C[0,T]$ given by

$$G(y) = \int_\Omega g(x,\cdot,y(x,\cdot))dx.$$

Example 2.4. The constraint

$$\int_{\Omega_T} |y_u(x,t)|dxdt \le \delta$$

is considered by taking $Z = L^1(\Omega_T)$, $G : Y \longrightarrow L^1(\Omega)$, with $G(y) = y$, and Q the closed ball in $L^1(\Omega)$ of center at 0 and radius δ.

Example 2.5. For every $1 \le j \le k$ let $g_j : \Omega_T \times \mathbb{R} \longrightarrow \mathbb{R}$ be a measurable function of class C^1 with respect to the second variable such that for each $M > 0$ there exists a function $\eta_M^j \in L^1(\Omega_T)$ satisfying

$$|g_j(x,t,0)| + \left|\frac{\partial g_j}{\partial y}(x,t,y)\right| \le \eta_M^j(x,t) \quad \text{a.e. } (x,t) \in \Omega_T \ \ \forall |y| \le M.$$

Then the constraints

$$\int_\Omega g_j(x,t,y_u(x,t))dxdt \le \delta_j, \quad 1 \le j \le k,$$

are included in the formulation of (P) by choosing $G = (G_1,\dots,G_k)^T$, with

$$G_j(y) = \int_\Omega g_j(x,t,y(x,t))dxdt,$$

$Z = \mathbb{R}^k$, and $Q = (-\infty,\delta_1] \times \cdots \times (-\infty,\delta_k]$.

Example 2.6. The equality constraints

$$\int_\Omega f_j(x,t,y_u(x,t))dx = \delta_j, \quad 1 \le j \le l,$$

can also be included in problem (P) in the obvious way by assuming the same hypotheses as in Example 2.5.

Example 2.7. Integral constraints on the gradient of the state can be considered within our formulation of problem (P):

$$G(y_u) = \int_0^T \int_\Omega |\nabla_x y_u(x,t)|^2 dxdt \le \delta.$$

In this case we can take $Z = \mathbb{R}$ and $Q = (-\infty,\delta]$.

3. The weak Pontryagin principle. Before formulating the weak Pontryagin principle, we introduce some notation. Given $\alpha \geq 0$, we define the Hamiltonian $H_\alpha : \Sigma_T \times \mathbb{R} \times \mathcal{K} \times \mathbb{R} \longrightarrow \mathbb{R}$ as follows:

$$H_\alpha(x,t,y,u,\varphi) = \alpha l(x,t,y,u) + \varphi f(x,t,y,u).$$

Now we can establish Pontryagin's principle.

THEOREM 3.1. *If $\bar{u} \in \mathcal{U}$ is a solution of (P), then there exist $\bar{\alpha} \geq 0$, $\bar{y} \in C(\bar{\Omega}_T) \cap L^2([0,T], H^1(\Omega))$, and $\bar{\varphi} \in L^r([0,T], W^{1,p}(\Omega))$ for all $p, r \in [1,2)$ with $(2/r)+(n/p) > n+1$, $\bar{\mu} \in Z'$ and $\bar{\lambda} \in R^s$ such that*

$$\bar{\alpha} + \|\bar{\mu}\|_{Z'} + |\bar{\lambda}| > 0; \tag{3.1}$$

$$\begin{cases} \dfrac{\partial \bar{y}}{\partial t} + A\bar{y} + a_0(x,t,\bar{y}(x,t)) = 0 \quad \text{in } \Omega_T, \\ \partial_{\nu_A} \bar{y}(x,t) = f(x,t,\bar{y}(x,t),\bar{u}(x,t)) \text{ on } \Sigma_T, \\ \bar{y}(0) = y_0 \quad \text{in } \Omega; \end{cases} \tag{3.2}$$

$$\begin{cases} -\dfrac{\partial \bar{\varphi}}{\partial t} + A^*\bar{\varphi} + \dfrac{\partial a_0}{\partial y}(x,t,\bar{y})\bar{\varphi} = \bar{\alpha}\dfrac{\partial L}{\partial y}(x,t,\bar{y}) \\ \qquad\qquad + [DG(\bar{y})^*\bar{\mu}]|_{\Omega_T} + [DF(\bar{y})^*\bar{\lambda}]|_{\Omega_T} \quad \text{in } \Omega_T, \\ \partial_{\nu_{A^*}}\bar{\varphi} = \dfrac{\partial f}{\partial y}(x,t,\bar{y},\bar{u})\bar{\varphi} + \bar{\alpha}\dfrac{\partial l}{\partial y}(x,t,\bar{y},\bar{u}) \\ \qquad\qquad + [DG(\bar{y})^*\bar{\mu}]|_{\Sigma_T} + [DF(\bar{y})^*\bar{\lambda}]|_{\Sigma_T} \text{ on } \Sigma_T, \\ \bar{\varphi}(T) = [DG(\bar{y})^*\bar{\mu}]|_{\bar{\Omega}\times\{T\}} + [DF(\bar{y})^*\bar{\lambda}]|_{\bar{\Omega}\times\{T\}} \text{ in } \bar{\Omega}; \end{cases} \tag{3.3}$$

$$\langle \bar{\mu}, z - G(\bar{y}) \rangle \leq 0 \quad \forall z \in Q; \tag{3.4}$$

$$\begin{aligned} &\int_{\Sigma_T} H_{\bar{\alpha}}(x,t,\bar{y}(x,t),\bar{u}(x,t),\bar{\varphi}(x,t))d\sigma(x)dt \\ &= \min_{u\in\mathcal{U}} \int_{\Sigma_T} H_{\bar{\alpha}}(x,t,\bar{y}(x,t),u(x,t),\bar{\varphi}(x,t))d\sigma(x)dt; \end{aligned} \tag{3.5}$$

where A^* denotes the formal adjoint operator of A. Moreover, if one of the following assumptions is satisfied,

(A1) *Functions f and l are continuous with respect to the third variable on $(\mathcal{K}, d)$ and this space is separable;*

(A2) *There exists a set $\Sigma_T^0 \subset \Sigma_T$, with $m_{\Sigma_T}(\Sigma_T^0) = m_{\Sigma_T}(\Sigma_T)$, such that the function*

$$(x,t) \in \Sigma_T \longrightarrow (f(x,t,y,u), l(x,t,y,u)) \in \mathbb{R}^2$$

is continuous in Σ_T^0 for every $(y,u) \in \mathbb{R} \times \mathcal{K}$,
then the following pointwise relation holds:

$$\begin{aligned} &H_{\bar{\alpha}}(x,t,\bar{y}(x,t),\bar{u}(x,t),\bar{\varphi}(x,t)) \\ &= \min_{u\in\mathcal{K}} H_{\bar{\alpha}}(x,t,\bar{y}(x,t),u,\bar{\varphi}(x,t)) \quad \text{a.e.}[\sigma]\ x \in \Gamma \text{ and a.e. } t \in [0,T]. \end{aligned} \tag{3.6}$$

Remark 3.2. In the previous theorem, $[DG(\bar{y})]^*\bar{\mu}$ and $[DF(\bar{y})]^*\bar{\lambda}$ are elements of

$$Y' = C(\bar{\Omega}_T)' + L^2([0,T], H^1(\Omega))' = M(\bar{\Omega}_T) + L^2([0,T], H^1(\Omega)'),$$

where $M(\bar{\Omega}_T)$ is the space of the real and regular Borel measures in $\bar{\Omega}_T$. Let us assume that $[DG(\bar{y})]^*\bar{\mu} = \bar{\phi} + \bar{\nu}$, with $\bar{\phi} \in L^2([0,T], H^1(\Omega)')$ and $\bar{\nu} \in M(\bar{\Omega}_T)$, then we can write

$$[DG(\bar{y})]^*\bar{\mu}|_{\Omega_T} = \bar{\phi} + \bar{\nu}|_{\Omega_T}, \quad [DG(\bar{y})]^*\bar{\mu}|_{\Sigma_T} = \bar{\nu}|_{\Sigma_T}, \text{ and } [DG(\bar{y})]^*\bar{\mu}|_{\bar{\Omega}\times\{T\}} = \bar{\nu}|_{\bar{\Omega}\times\{T\}}.$$

Analogous considerations can be made for $[DF(\bar{y})]^*\bar{\lambda}$.

Let us apply the above principle to the examples given in section 2.

Example 3.3. In Example 2.1, $Z = C(\bar{\Omega}_T)$; therefore, the Lagrange multiplier $\bar{\mu}$ whose existence is established in Theorem 3.1 is a measure in $\bar{\Omega}_T$. In this case the transversality condition (3.4) is written as follows:

$$\int_{\bar{\Omega}_T} (z(x,t) - g(x,t,\bar{y}(x,t)))d\bar{\mu}(x,t) \le 0 \quad \forall z \in C(\bar{\Omega}_T) \text{ with } z(x,t) \le \delta.$$

From this relation we can deduce that $\bar{\mu}$ is a positive measure concentrated in the set of points $(x,t) \in \bar{\Omega}_T$, where $g(x,t,\bar{y}(x,t)) = \delta$. In particular, it could be a Dirac measure or a combination of Dirac measures; see Casas [7].

The adjoint state equation (3.2) now becomes

$$\begin{cases} -\dfrac{\partial\bar{\varphi}}{\partial t} + A^*\bar{\varphi} + \dfrac{\partial a_0}{\partial y}(x,t,\bar{y})\bar{\varphi} = \bar{\alpha}\dfrac{\partial L}{\partial y}(x,t,\bar{y}) + \dfrac{\partial g}{\partial y}(x,t,\bar{y})\bar{\mu}|_{\Omega_T} & \text{in } \Omega_T, \\ \partial_{\nu_{A^*}}\bar{\varphi} = \dfrac{\partial f}{\partial y}(x,t,\bar{y},\bar{u})\bar{\varphi} + \bar{\alpha}\dfrac{\partial l}{\partial y}(x,t,\bar{y},\bar{u}) + \dfrac{\partial g}{\partial y}(x,t,\bar{y})\bar{\mu}|_{\Sigma_T} & \text{on } \Sigma_T, \\ \bar{\varphi}(T) = \dfrac{\partial g}{\partial y}(x,T,\bar{y}(x,T))\bar{\mu}|_{\bar{\Omega}\times\{T\}} & \text{in } \bar{\Omega}. \end{cases}$$

Since $\partial g/\partial y$ is a continuous function in $\bar{\Omega}_T$, then the product $(\partial g/\partial y)\bar{\mu}$ is well defined and can be identified again with a measure.

Example 3.4. In Example 2.2

$$[DF(\bar{y})]^*\bar{\lambda} = \sum_{j=1}^{l} \bar{\lambda}_j \delta_{(x_j,t_j)}.$$

If the points (x_j,t_j) are all of them included in Ω_T, then the adjoint state equation is

$$\begin{cases} -\dfrac{\partial\bar{\varphi}}{\partial t} + A^*\bar{\varphi} + \dfrac{\partial a_0}{\partial y}(x,\bar{y}(x))\bar{\varphi} = \bar{\alpha}\dfrac{\partial L}{\partial y}(x,\bar{y}(x)) + \displaystyle\sum_{j=1}^{l} \bar{\lambda}_j\delta_{(x_j,t_j)} & \text{in } \Omega_T, \\ \partial_{\nu_{A^*}}\bar{\varphi} = \dfrac{\partial f}{\partial y}(x,t,\bar{y},\bar{u})\bar{\varphi} + \bar{\alpha}\dfrac{\partial l}{\partial y}(x,t,\bar{y},\bar{u}) & \text{on } \Sigma_T, \\ \bar{\varphi}(T) = 0 & \text{in } \Omega. \end{cases}$$

If some points x_j are in Γ, then the corresponding term $\bar{\lambda}_j\delta_{(x_j,t_j)}$ should appear on the Neumann condition. Analogously, if $t_j = T$ for some index j, then $\bar{\lambda}\delta_{(x_j,T)}$ should be included in the final condition.

Example 3.5. In Example 2.3, the Lagrange multiplier $\bar\mu$ is a positive Borel measure in $[0,T]$ concentrated in the set of points t where the state constraint is active and

$$DG(\bar y)^*\bar\mu = \frac{\partial g}{\partial y}(x,t,\bar y(x,t))\bar\mu(t).$$

Then we have the following equation for $\bar\varphi$:

$$\begin{cases} -\dfrac{\partial\bar\varphi}{\partial t} + A^*\bar\varphi + \dfrac{\partial a_0}{\partial y}(x,t,\bar y)\bar\varphi = \bar\alpha\dfrac{\partial L}{\partial y}(x,t,\bar y) + \dfrac{\partial g}{\partial y}(x,t,\bar y)\bar\mu|_{(0,T)} \ \text{ in } \Omega_T, \\ \partial_{\nu_{A^*}}\bar\varphi = \dfrac{\partial f}{\partial y}(x,t,\bar y,\bar u)\bar\varphi + \bar\alpha\dfrac{\partial l}{\partial y}(x,t,\bar y,\bar u) \text{ on } \Sigma_T, \\ \bar\varphi(T) = \dfrac{\partial g}{\partial y}(x,T,\bar y(x,T))\bar\mu(\{T\}) \text{ in } \Omega. \end{cases}$$

So, in particular, we have that $\bar\varphi(T) = 0$ if the state constraint is not active in T. This type of state constraints has been studied by many authors; see Barbu and Precupanu [1], Lasiecka [24], and Tröltzsch [32]. All of them consider the semigroup theory approach to deal with the state and adjoint state equations. They prove some regularity of the adjoint state $\bar\varphi$; see section 6.

Example 3.6. In Example 2.4, the Lagrange multiplier $\bar\mu$ is an element of $Z' = L^\infty(\Omega_T)$; therefore, (3.2) reduces in this case to

$$\begin{cases} -\dfrac{\partial\bar\varphi}{\partial t} + A^*\bar\varphi + \dfrac{\partial a_0}{\partial y}(x,t,\bar y)\bar\varphi = \bar\alpha\dfrac{\partial L}{\partial y}(x,t,\bar y) + \bar\mu \ \text{ in } \Omega_T, \\ \partial_{\nu_{A^*}}\bar\varphi = \dfrac{\partial f}{\partial y}(x,t,\bar y,\bar u)\bar\varphi + \bar\alpha\dfrac{\partial l}{\partial y}(x,t,\bar y,\bar u) \text{ on } \Sigma_T, \\ \bar\varphi(T) = 0 \text{ in } \Omega. \end{cases}$$

In this case, assuming more regularity for the functions ψ_{dM} and ψ_{bM} given in (2.8)–(2.9), we can obtain additional regularity for $\bar\varphi$. For instance, if we take function $\psi_{bM} \in L^{\hat p}([0,T], L^{\hat q}(\Omega))$, then $\bar\varphi \in Y$. $H^{2,1}(\Omega)$-regularity is also obtained provided that Γ is of class C^2 and the coefficients a_{ij} of A are Lipschitz in the variable x.

Example 3.7. The Lagrange multipliers in Example 2.5 are positive real numbers $\{\bar\mu_j\}_{j=1}^k$. The positivity is a consequence of the transversality condition (3.3). The adjoint state equation can be written as follows:

$$\begin{cases} -\dfrac{\partial\bar\varphi}{\partial t} + A^*\bar\varphi + \dfrac{\partial a_0}{\partial y}(x,t,\bar y)\bar\varphi = \bar\alpha\dfrac{\partial L}{\partial y}(x,t,\bar y) + \displaystyle\sum_{j=1}^k \bar\mu_j\frac{\partial g_j}{\partial y}(x,t,\bar y) \ \text{ in } \Omega_T, \\ \partial_{\nu_{A^*}}\bar\varphi = \dfrac{\partial f}{\partial y}(x,t,\bar y,\bar u)\bar\varphi + \bar\alpha\dfrac{\partial l}{\partial y}(x,t,\bar y,\bar u) \text{ on } \Sigma_T, \\ \bar\varphi(T) = 0 \text{ in } \Omega. \end{cases}$$

By increasing the regularity of functions η_j, we can improve the regularity of $\bar\varphi$ such as it was described in Example 3.6.

For the equality constraints considered in Example 2.6 the adjoint state equation is similar to the above one. The only difference is that the Lagrange multipliers can be negative.

Example 3.8. In Example 2.7, the Lagrange multiplier $\bar{\mu}$ is a nonnegative real number, $\bar{\varphi} \in Y$, and the adjoint state equation is

$$\begin{cases} -\dfrac{\partial \bar{\varphi}}{\partial t} + A^*\bar{\varphi} + \dfrac{\partial a_0}{\partial y}(x,t,\bar{y})\bar{\varphi} = \bar{\alpha}\dfrac{\partial L}{\partial y}(x,t,\bar{y}) + 2\bar{\mu}\nabla^*\nabla_x y(x,t) \ \text{ in } \Omega_T, \\ \partial_{\nu_{A^*}}\bar{\varphi} = \dfrac{\partial f}{\partial y}(x,t,\bar{y},\bar{u})\bar{\varphi} + \bar{\alpha}\dfrac{\partial l}{\partial y}(x,t,\bar{y},\bar{u}) \text{ on } \Sigma_T, \\ \bar{\varphi}(T) = 0 \text{ in } \Omega, \end{cases}$$

where $\nabla^*\nabla_x y \in L^2([0,T], H^1(\Omega)')$ is given by

$$\langle \nabla^*\nabla_x y, z\rangle = \int_{\Omega_T} \nabla_x y(x,t)\nabla_x z(x,t)dxdt.$$

The restriction of $\nabla^*\nabla_x y$ to $L^2([0,T], H_0^1(\Omega))$ is equal to $-\Delta_x y$.

4. The strong Pontryagin principle. In this section we will prove that, in the absence of equality constraints, Theorem 3.1 holds with $\bar{\alpha} = 1$ for "almost all" control problems. We will be precise about this term later. The key to achieving this result is the introduction of a stability assumption of the optimal cost functional with respect to small perturbations of the set of feasible controls. This stability allows us to accomplish an exact penalization of the state constraints. First of all let us formulate the following control problem:

$$(\mathrm{P}_\delta) \begin{cases} \text{Minimize } J(u), \\ u \in \mathcal{U},\ G(y_u) \in Q_\delta \end{cases}$$

with the same notation and assumptions of section 2 and setting $Q_\delta = Q + \bar{B}_\delta(0)$ for every $\delta > 0$.

DEFINITION 4.1. *We say that* (P_δ) *is strongly stable if there exist* $\epsilon > 0$ *and* $C > 0$ *such that*

$$\inf(\mathrm{P}_\delta) - \inf(\mathrm{P}_{\delta'}) \le C(\delta' - \delta) \quad \forall \delta' \in [\delta, \delta + \epsilon]. \tag{4.1}$$

This concept was first introduced in relation with optimal control problems by Bonnans [4]; see also Bonnans and Casas [6]. A weaker stability concept was used by Casas [8] to analyze the convergence of the numerical discretizations of optimal control problems. The following proposition states that almost all problems (P_δ) are strongly stable.

PROPOSITION 4.2. *Let* $\delta_0 \ge 0$ *be the smallest number such that* (P_δ) *has feasible controls for every* $\delta > \delta_0$. *Then* (P_δ) *is strongly stable for all* $\delta > \delta_0$ *except at most a zero Lebesgue measure set.*

Proof. It is enough to consider the function $h : (\delta_0, +\infty) \longrightarrow \mathbb{R}$ defined by

$$h(\delta) = \inf(\mathrm{P}_\delta)$$

and remark that it is a nonincreasing monotone function and, consequently, differentiable at every point of $(\delta_0, +\infty)$ except at a zero measure set. Now it is obvious to check that (P_δ) is strongly stable at every point where h is differentiable. ∎

Now we state the strong Pontryagin principle.

THEOREM 4.3. *If* (P_δ) *is strongly stable and* $\bar{u}$ *is a solution of this problem, then Theorem* 3.1 *remains to be true with* $\bar{\alpha} = 1$.

The proof of this theorem is postponed until section 7.

5. Analysis of the state equation. In this section we will see that (2.3) is well posed in $Y = C(\bar{\Omega}_T) \cap L^2([0,T], H^1(\Omega))$ for every control $u \in \mathcal{U}$. Also we will study the variations of the state with respect to some pointwise perturbations of the control which are the crucial point in the proof of Pontryagin's principle. In $\mathcal{U}$ we consider Ekeland's distance

$$d_E(u,v) = m_{\Sigma_T}\left(\{(x,t) \in \Sigma_T : u(x,t) \neq v(x,t)\}\right), \tag{5.1}$$

where m_{Σ_T} is the measure on Σ_T obtained as the product of σ and the Lebesgue measure in the interval $(0,T)$. It is easy to check that $(\mathcal{U}, d_E)$ is a complete metric space. Indeed the proof given by Ekeland [16] can be repeated in this framework.

THEOREM 5.1. *Under assumptions* (2.1)–(2.8), *problem* (2.3) *has a unique solution in* $Y = C(\bar{\Omega}_T) \cap L^2([0,T], H^1(\Omega))$ *for every control* $u \in \mathcal{U}$. *Moreover, there exists a constant* $M > 0$ *such that*

$$\|y_u\|_\infty + \|y_u\|_{L^2([0,T],H^1(\Omega))} \leq M \quad \forall u \in \mathcal{U}. \tag{5.2}$$

Finally, if $\{u_k\}_{k=1}^\infty \subset \mathcal{U}$ *is a sequence converging to* u *in* $\mathcal{U}$, *i.e.* $d_E(u_k, u) \to 0$, *then* $\{y_{u_k}\}_{k=1}^\infty$ *converges to* y_u *strongly in* Y.

Proof. The uniqueness of the solution in Y can be proved by using the Gronwall inequality in the standard way along with the monotonicity of the nonlinear terms. Let us prove the existence.

If a_0 and f are bounded functions, then the existence and uniqueness of a solution in $L^\infty([0,T], L^2(\Omega)) \cap L^2([0,T], H^1(\Omega))$ is a consequence of the monotonicity of f imposed in (2.1) and the condition on a_0 given in (2.6); see Lions [29] or Ladyzhenskaya, Solonnikov, and Ural'tseva [23] for a proof based in Galerkin's approximation of the problem. If f is not bounded, we can consider the usual truncation of the function

$$f_m(x,t,y,u) = \begin{cases} f(x,t,y,u) & \text{if } |y| \leq m, \\ f(x,t,m,u) & \text{if } y > m, \\ f(x,t,m,u) & \text{if } y < -m. \end{cases}$$

Thus hypothesis (2.2) implies the boundedness of f_m.

An analogous modification can be made on a_0. Then we deduce the existence and uniqueness of a solution $y_m \in L^\infty([0,T], L^2(\Omega)) \cap L^2([0,T], H^1(\Omega))$ for problem (2.3) with a_0 and f replaced by a_{0m} and f_m, respectively. Now thanks to the assumptions (2.1)–(2.8), we can apply the procedure of Ladyzhenskaya, Solonnikov, and Ural'tseva [23] to deduce the existence of a constant $M > 0$ independent of m and $u \in \mathcal{U}$ such that (5.2) holds for y_u replaced by y_m. This implies that

$$a_m(x,t,y_m(x,t)) = a(x,t,y_m(x,t)) \ \ \forall m \geq M$$

and

$$f_m(x,t,y_m(x,t),u(x,t)) = f(x,t,y_m(x,t),u(x,t)) \ \ \forall m \geq M.$$

Consequently, the uniqueness of a solution of (2.3) lets us obtain the identity $y_m = y_u$ and the inequality (5.2).

In order to prove the continuity of y_u, we first suppose that $y_0 \in C^\theta(\bar{\Omega}_T)$ for some constant $\theta \in (0,1]$. Then, by applying the results of di Benedetto [2], we deduce

that $y_u \in C^{\beta,\beta/2}(\bar{\Omega}_T)$ for some $\beta \in (0,\theta]$. When y_0 is not a Hölder function, we can take a sequence $\{y_{0k}\}_{k=1}^{\infty} \subset C^{\theta}(\bar{\Omega}_T)$ converging uniformly to y_0 in $\bar{\Omega}_T$. Then the corresponding solutions of (2.3), denoted by y_k, are Hölder functions. Now, by applying the methods of [23] is easy to deduce the convergence $y_k \to y_u$ in $L^{\infty}(\Omega_T)$, which proves the continuity of y_u.

Finally, the convergence $y_{u_k} \to y_u$ in $L^2([0,T],H^1(\Omega))$ when $d_E(u_k,u) \to 0$ is easily derived. The uniform convergence is obtained again by using the arguments of [23]. □

The rest of the section is devoted to the proof of the following theorem

THEOREM 5.2. *Let $u,v \in \mathcal{U}$. Given $\rho \in (0,1)$, there exist m_{Σ_T}-measurable sets $E_\rho \subset \Sigma_T$, with $m_{\Sigma_T}(E_\rho) = \rho m_{\Sigma_T}(\Sigma_T)$, such that if we define*

$$u_\rho(x,t) = \begin{cases} u(x,t) & \text{if } (x,t) \in \Sigma_T \setminus E_\rho, \\ v(x,t) & \text{if } (x,t) \in E_\rho, \end{cases}$$

and if we denote by y_ρ and y the states corresponding to u_ρ and u, respectively, then the following equalities hold:

$$y_\rho = y + \rho z + r_\rho, \quad \lim_{\rho \to 0} \frac{1}{\rho} \|r_\rho\|_Y = 0, \tag{5.3}$$

and

$$J(u_\rho) = J(u) + \rho z^0 + r_\rho^0, \quad \lim_{\rho \to 0} \frac{1}{\rho} r_\rho^0 = 0, \tag{5.4}$$

where $z \in Y$ satisfies

$$\begin{cases} \dfrac{\partial z}{\partial t} + Az + \dfrac{\partial a_0}{\partial y}(x,t,y(x,t))z = 0 \quad \text{in } \Omega_T, \\ \partial_{\nu_A} z = \dfrac{\partial f}{\partial y}(x,t,y(x,t),u(x,t))z \\ +f(x,t,y(x,t),v(x,t)) - f(x,t,y(x,t),u(x,t)) \text{ on } \Sigma_T, \\ z(x,0) = 0 \text{ in } \Omega \end{cases} \tag{5.5}$$

and

$$\begin{aligned} z^0 = &\int_{\Omega_T} \frac{\partial L}{\partial y}(x,t,y(x,t))z(x,t)dxdt + \int_{\Sigma_T} \frac{\partial l}{\partial y}(x,t,y(x,t),u(x,t))z(x,t)d\sigma(x)dt \\ &+ \int_{\Sigma_T} [l(x,t,y(x,t),v(x,t)) - l(x,t,y(x,t),u(x,t))]d\sigma(x)dt. \end{aligned} \tag{5.6}$$

The first step is the proof of the following result

PROPOSITION 5.3. *For every $0 < \rho < 1$ there exists a sequence of m_{Σ_T}-measurable sets $\{E_k\}_{k=1}^{\infty}$ satisfying*

(1) *$E_k = E_\Gamma^k \times J^k$, with $E_k \subset \Gamma$ and $J^k \subset (0,T)$, $\sigma(E_\Gamma^k) = \sqrt{\rho}\sigma(\Gamma)$, and $|J^k| = \sqrt{\rho}T$.*
(2) *$(1/\sqrt{\rho})\chi_{E_\Gamma^k} \to 1$ *weakly in $L^{\infty}(\Gamma)$; $(1/\sqrt{\rho})\chi_{J^k} \to 1$ *weakly in $L^{\infty}(0,T)$; and $(1/\rho)\chi_{E_k} \to 1$ *weakly in $L^{\infty}(\Sigma_T)$.*

Proof. We divide the proof into several steps.

Step 1. *The sets* E_Γ^k.

Let us construct the sets E_Γ^k. Since Ω is bounded and Γ is a Lipschitz manifold, we can obtain a finite collection of σ-measurable sets $\{\Gamma_r\}_{r=1}^d$ and functions $\{a_r\}_{r=1}^d$ satisfying

(i) $\bigcup_{r=1}^d \Gamma_r = \Gamma$, $\overset{o}{\Gamma}_i \bigcap \overset{o}{\Gamma}_j = \emptyset$ if $i \neq j$ and $\sigma(\Gamma) = \sum_{r=1}^d \sigma(\overset{o}{\Gamma}_r)$.

(ii) The functions $a_r : (-\Lambda_\Gamma, +\Lambda_\Gamma)^{n-1} \longrightarrow \mathbb{R}$ are Lipschitz, and for some coordinate system $(x_r', x_{r,n}) = (x_{r,1}, \dots, x_{r,n})$ in $\mathbb{R}^n$ we have that

$$\overset{o}{\Gamma}_r = \{(x_r', a_r(x_r')) : x_r' \in (-\Lambda_\Gamma, +\Lambda_\Gamma)^{n-1}\}$$

and for every set $E = \{(x_r', a_r(x_r')) : x_r' \in F\}$, with $F \subset (-\Lambda_\Gamma, +\Lambda_\Gamma)^{n-1}$ Lebesgue measurable, the following identity holds:

$$\sigma(E) = \int_F \sqrt{1 + \sum_{i=1}^{n-1} \left| \frac{\partial a_r}{\partial x_{r,i}}(x_r') \right|^2}\, dx_r'.$$

For every $k \in \mathbb{N}$ we decompose the interval $[-\Lambda_\Gamma, +\Lambda_\Gamma]$ into k closed subintervals of length $2\Lambda_\Gamma/k$ and disjoint interiors. Now we make all possible Cartesian products of these subintervals and obtain a family of cubes $\{Q_{k,i}\}_{i=1}^{k^{n-1}}$ of equal Lebesgue measure, covering $[-\Lambda_\Gamma, +\Lambda_\Gamma]^{n-1}$ and with disjoint interiors. For every $r = 1, \dots, d$ and every cube we take a measurable set $F_{k,j}^r \subset \overset{o}{Q}_{k,j}$ such that

$$\int_{F_{k,j}^r} \sqrt{1 + \sum_{i=1}^{n-1} \left| \frac{\partial a_r}{\partial x_{r,i}}(x_r') \right|^2}\, dx_r' = \sqrt{\rho} \int_{Q_{k,j}} \sqrt{1 + \sum_{i=1}^{n-1} \left| \frac{\partial a_r}{\partial x_{r,i}}(x_r') \right|^2}\, dx_r'.$$

Let us see that such an $F_{k,j}^r$ exists. For every $t \in [0,1]$ we define $Q_{k,j}(t)$ as the cube with the same center as $Q_{k,j}$ and the length of each side being equal to t times the length of the sides of $Q_{k,j}$. So $Q_{k,j}(1) = Q_{k,j}$ and $Q_{k,j}(0)$ is reduced to one point: the center of $Q_{k,j}$. Let us consider the function $g : [0,1] \longrightarrow \mathbb{R}$ defined by

$$g(t) = \int_{Q_{k,j}(t)} \sqrt{1 + \sum_{i=1}^{n-1} \left| \frac{\partial a_r}{\partial x_{r,i}}(x_r') \right|^2}\, dx_r'.$$

Then it is obvious that g is continuous and

$$0 = g(0) < \sqrt{\rho} \int_{Q_{k,j}} \sqrt{1 + \sum_{i=1}^{n-1} \left| \frac{\partial a_r}{\partial x_{r,i}}(x_r') \right|^2}\, dx_r' < g(1).$$

Therefore there exists $0 < t_0 < 1$ such that

$$g(t_0) = \sqrt{\rho} \int_{Q_{k,j}} \sqrt{1 + \sum_{i=1}^{n-1} \left| \frac{\partial a_r}{\partial x_{r,i}}(x_r') \right|^2}\, dx_r'.$$

Thus we can choose $F_{k,j}^r = Q_{k,j}(t_0)$.

Now we set

$$F_k^r = \bigcup_{i=1}^{k^{n-1}} F_{k,i}^r, \quad E_k^r = \{(x_r', a_r(x_r')) : x_r' \in F_k^r\} \subset \overset{o}{\Gamma}_r, \quad E_\Gamma^k = \bigcup_{r=1}^d E_k^r.$$

Then

$$\sigma(E_\Gamma^k)=\sum_{r=1}^d\sigma(E_k^r)=\sum_{r=1}^d\int_{F_k^r}\sqrt{1+\sum_{i=1}^{n-1}\left|\frac{\partial a_r}{\partial x_{r,i}}(x_r')\right|^2}\,dx_r'$$

$$=\sqrt{\rho}\sum_{r=1}^d\int_{[-\Lambda_\Gamma,+\Lambda_\Gamma]^{n-1}}\sqrt{1+\sum_{i=1}^{n-1}\left|\frac{\partial a_r}{\partial x_{r,i}}(x_r')\right|^2}\,dx_r'=\sqrt{\rho}\sum_{r=1}^d\sigma(\overset{o}{\Gamma}_r)=\sqrt{\rho}\sigma(\Gamma).$$

We are going to prove that

$$\frac{1}{\sqrt{\rho}}\lim_{k\to\infty}\sigma(A\cap E_\Gamma^k)=\sigma(A)\quad\forall A\subset\Gamma\ \sigma\text{ measurable.}\tag{5.7}$$

Once this is proved, the convergence $(1/\sqrt{\rho})\chi_{E_\Gamma^k}\to 1$ *weakly in $L^\infty(\Gamma)$ follows from the density of the simple functions in $L^1(\Gamma)$.

First, let us assume that $A\subset\overset{o}{\Gamma}_r$ is an open set. Let us take the open set $B\subset(-\Lambda_\Gamma,+\Lambda_\Gamma)^{n-1}$ such that $A=\{(x_r',a_r(x_r')):x_r'\in B\}$. Then, from Lemma 5.4 proved below, we deduce

$$\sigma(A)=\int_B\sqrt{1+\sum_{i=1}^{n-1}\left|\frac{\partial a_r}{\partial x_{r,i}}(x_r')\right|^2}\,dx_r'$$

$$=\frac{1}{\sqrt{\rho}}\lim_{k\to\infty}\int_{B\cap F_k^r}\sqrt{1+\sum_{i=1}^{n-1}\left|\frac{\partial a_r}{\partial x_{r,i}}(x_r')\right|^2}\,dx_r'$$

$$=\frac{1}{\sqrt{\rho}}\lim_{k\to\infty}\sigma(A\cap E_k^r)=\frac{1}{\sqrt{\rho}}\lim_{k\to\infty}\sigma(A\cap E_\Gamma^k).$$

If $A\subset\Gamma$ is an open set, then

$$\sigma(A)=\sum_{r=1}^d\sigma(A\cap\overset{0}{\Gamma}_r)=\sum_{r=1}^d\frac{1}{\sqrt{\rho}}\lim_{k\to\infty}\sigma(A\cap\overset{0}{\Gamma}_r\cap E_\Gamma^k)=\frac{1}{\sqrt{\rho}}\lim_{k\to\infty}\sigma(A\cap E_\Gamma^k).$$

Thus (5.7) holds for every open subset of Γ. Let us take a closed set $K\subset\Gamma$,

$$\sigma(K)=\sigma(\Gamma)-\sigma(\Gamma\setminus K)=\sigma(\Gamma)-\frac{1}{\sqrt{\rho}}\lim_{k\to\infty}\sigma([\Gamma\setminus K]\cap E_\Gamma^k)$$

$$=\sigma(\Gamma)-\frac{1}{\sqrt{\rho}}\lim_{k\to\infty}\left\{\sigma(E_\Gamma^k)-\sigma(K\cap E_\Gamma^k)\right\}=\frac{1}{\sqrt{\rho}}\lim_{k\to\infty}\sigma(K\cap E_\Gamma^k).$$

Finally, let $A\subset\Gamma$ be a σ-measurable set. Given $\epsilon>0$ arbitrary, we can take $K\subset\Gamma$ closed and $V\subset\Gamma$ open such that $K\subset A\subset V$ and

$$\sigma(A)-\epsilon\le\sigma(K)\le\sigma(V)\le\sigma(A)+\epsilon.$$

Then

$$\sigma(A)-\epsilon\le\sigma(K)\le\frac{1}{\sqrt{\rho}}\lim_{k\to\infty}\sigma(K\cap E_\Gamma^k)\le\frac{1}{\sqrt{\rho}}\liminf_{k\to\infty}\sigma(A\cap E_\Gamma^k)$$
$$\le\frac{1}{\sqrt{\rho}}\limsup_{k\to\infty}\sigma(A\cap E_\Gamma^k)\le\frac{1}{\sqrt{\rho}}\lim_{k\to\infty}\sigma(V\cap E_\Gamma^k)=\sigma(V)\le\sigma(A)+\epsilon,$$

which concludes the proof of (5.7).

Step 2. *The sets* J^k.

To construct the sets J^k, we decompose the interval $[0,T]$ into k closed intervals I_j^k of length T/k and disjoint interiors. For each $j=1,\dots,k$ we take a subinterval $J_j^k\subset\overset{0}{I}{}_j^k$ of length $\sqrt{\rho}T/k$ and the same center as I_j^k. Finally, we define J^k as the union of the intervals $\{J_j^k\}_{j=1}^k$. Then $|J^k|=\sqrt{\rho}T$ and the convergence $(1/\sqrt{\rho})\chi_{J^k}\to 1$ *weakly in $L^\infty(0,T)$ can be proved following the same ideas as in the previous step.

Step 3. *The sets* E_k.

Taking $E_k=E_\Gamma^k\times J^k$, it remains to prove the convergence $(1/\rho)\chi_{E_k}\to 1$ *weakly in $L^\infty(\Sigma_T)$. Given $f\in L^1(\Gamma)$ and $h\in L^1(0,T)$, we get from Steps 1 and 2 that

$$\lim_{k\to\infty}\int_{\Sigma_T}\frac{1}{\rho}\chi_{E_k}(x,t)f(x)h(t)dm_{\Sigma_T}(x,t)$$
$$=\left(\lim_{k\to\infty}\int_\Gamma\frac{1}{\sqrt{\rho}}\chi_{E_\Gamma^k}(x)f(x)d\sigma(x)\right)\left(\lim_{k\to\infty}\int_0^T\frac{1}{\sqrt{\rho}}\chi_{J^k}(t)h(t)dt\right)$$
$$=\int_{\Sigma_T}f(x)h(t)dm_{\Sigma_T}(x,t).$$

Since the functions $f(x)h(t)$, with $f\in L^1(\Gamma)$ and $h\in L^1(0,T)$, expand a subspace dense in $L^1(\Sigma_T)$, we conclude the proof. □

LEMMA 5.4. *With the notations of the above proof, the following identity holds for all open sets* $B\subset(-\Lambda_\Gamma,+\Lambda_\Gamma)^{n-1}$*:*

$$\int_B\sqrt{1+\sum_{i=1}^{n-1}\left|\frac{\partial a_r}{\partial x_{r,i}}(x_r')\right|^2}\,dx_r'=\frac{1}{\sqrt{\rho}}\lim_{k\to\infty}\int_{B\cap F_k^r}\sqrt{1+\sum_{i=1}^{n-1}\left|\frac{\partial a_r}{\partial x_{r,i}}(x_r')\right|^2}\,dx_r'\tag{5.8}$$

for every $r=1,\dots,d$.

Proof. Let us take a sequence $\{C_k\}_{k=1}^\infty$ of closed cubes with sides parallel to the axes and $\overset{o}{C}{}_k\cap\overset{o}{C}{}_i=\emptyset$ if $i\ne k$, so that $B=\bigcup_{k=1}^\infty C_k$; see Stein [31, pp. 167–170].

Fixed r, for each cube C_l, it is obvious that

$$\int_{C_l}\sqrt{1+\sum_{i=1}^{n-1}\left|\frac{\partial a_r}{\partial x_{r,i}}(x_r')\right|^2}\,dx_r'=\lim_{k\to\infty}\sum_{Q_{k,j}\subset C_l}\int_{Q_{k,j}}\sqrt{1+\sum_{i=1}^{n-1}\left|\frac{\partial a_r}{\partial x_{r,i}}(x_r')\right|^2}\,dx_r'$$
$$=\frac{1}{\sqrt{\rho}}\lim_{k\to\infty}\sum_{Q_{k,j}\subset C_l}\int_{F_{k,j}^r}\sqrt{1+\sum_{i=1}^{n-1}\left|\frac{\partial a_r}{\partial x_{r,i}}(x_r')\right|^2}\,dx_r'$$
$$=\frac{1}{\sqrt{\rho}}\lim_{k\to\infty}\int_{C_l\cap F_k^r}\sqrt{1+\sum_{i=1}^{n-1}\left|\frac{\partial a_r}{\partial x_{r,i}}(x_r')\right|^2}\,dx_r'.$$

Now, given $\epsilon > 0$ there exists $k_\epsilon \in \mathbb{N}$ such that

$$\left| \int_B \sqrt{1 + \sum_{i=1}^{n-1} \left| \frac{\partial a_r}{\partial x_{r,i}}(x'_r) \right|^2} dx'_r - \sum_{l=1}^{k_\epsilon} \int_{C_l} \sqrt{1 + \sum_{i=1}^{n-1} \left| \frac{\partial a_r}{\partial x_{r,i}}(x'_r) \right|^2} dx'_r \right| < \epsilon.$$

From here it follows

$$\int_B \sqrt{1 + \sum_{i=1}^{n-1} \left| \frac{\partial a_r}{\partial x_{r,i}}(x'_r) \right|^2} dx'_r - \epsilon$$

$$\leq \sum_{l=1}^{k_\epsilon} \int_{C_l} \sqrt{1 + \sum_{i=1}^{n-1} \left| \frac{\partial a_r}{\partial x_{r,i}}(x'_r) \right|^2} dx'_r$$

$$= \lim_{k\to\infty} \frac{1}{\sqrt{\rho}} \sum_{l=1}^{k_\epsilon} \int_{C_l \cap F_k^r} \sqrt{1 + \sum_{i=1}^{n-1} \left| \frac{\partial a_r}{\partial x_{r,i}}(x'_r) \right|^2} dx'_r$$

$$\leq \liminf_{k\to\infty} \frac{1}{\sqrt{\rho}} \int_{B \cap F_k^r} \sqrt{1 + \sum_{i=1}^{n-1} \left| \frac{\partial a_r}{\partial x_{r,i}}(x'_r) \right|^2} dx'_r \leq$$

$$\limsup_{k\to\infty} \frac{1}{\sqrt{\rho}} \int_{B \cap F_k^r} \sqrt{1 + \sum_{i=1}^{n-1} \left| \frac{\partial a_r}{\partial x_{r,i}}(x'_r) \right|^2} dx'_r$$

$$\leq \limsup_{k\to\infty} \frac{1}{\sqrt{\rho}} \sum_{l=1}^{k_\epsilon} \int_{C_l \cap F_k^r} \sqrt{1 + \sum_{i=1}^{n-1} \left| \frac{\partial a_r}{\partial x_{r,i}}(x'_r) \right|^2} dx'_r + \frac{\epsilon}{\sqrt{\rho}}$$

$$= \sum_{l=1}^{k_\epsilon} \int_{C_l} \sqrt{1 + \sum_{i=1}^{n-1} \left| \frac{\partial a_r}{\partial x_{r,i}}(x'_r) \right|^2} dx'_r + \frac{\epsilon}{\sqrt{\rho}}$$

$$\leq \int_B \sqrt{1 + \sum_{i=1}^{n-1} \left| \frac{\partial a_r}{\partial x_{r,i}}(x'_r) \right|^2} dx'_r + \left(1 + \frac{1}{\sqrt{\rho}}\right) \epsilon.$$

Since $\epsilon > 0$ is arbitrary, the previous relations conclude the proof. □

Finally, we are ready to prove Theorem 5.2.

Proof of Theorem 5.2. Let $\rho \in (0,1)$ be fixed. Applying Proposition 5.3, we deduce the existence of measurable sets $\{E_k\}_{k=1}^\infty$ such that $m_{\Sigma_T}(E_k) = \rho m_{\Sigma_T}(\Sigma_T)$ and $(1/\rho)\chi_{E_k} \to 1$ *weakly in $L^\infty(\Sigma_T)$. For every $k \in \mathbb{N}$, we set

$$u_k(x,t) = \begin{cases} u(x,t) & \text{if } (x,t) \in \Sigma_T \setminus E_k, \\ v(x,t) & \text{if } (x,t) \in E_k, \end{cases}$$

and we denote by y_k and y the states corresponding to u_k and u, respectively. Now, subtracting the equations satisfied by y_k and y, and putting $z_k = (y_k - y)/\rho$ we obtain

$$\begin{cases} \dfrac{\partial z_k}{\partial t} + Az_k + c_k(x,t)z_k = 0 \quad \text{in } \Omega_T, \\ \partial_{\nu_A} z_k = b_k(x,t)z_k + h(x,t)\dfrac{1}{\rho}\chi_{E_k} \text{ on } \Sigma_T, \\ z(x,0) = 0 \text{ in } \Omega, \end{cases} \tag{5.9}$$

where

$$c_k(x,t) = \int_0^1 \frac{\partial a_0}{\partial y}(x,t,y(x,t) + \tau[y_k(x,t) - y(x,t)])d\tau,$$

$$b_k(x,t) = \int_0^1 \frac{\partial f}{\partial y}(x,t,y(x,t) + \tau[y_k(x,t) - y(x,t)], u_k(x,t))d\tau,$$

and

$$h(x,t) = f(x,t,y(x,t),v(x,t)) - f(x,t,y(x,t),u(x,t)).$$

By subtracting (5.9) and (5.5) and writing $\zeta_k = z_k - z$, we deduce

$$\begin{cases} \dfrac{\partial \zeta_k}{\partial t} + A\zeta_k + c_k(x,t)\zeta_k = \left[\dfrac{\partial a_0}{\partial y}(x,t,y(x,t)) - c_k(x,t)\right] z \quad \text{in } \Omega_T, \\ \partial_{\nu_A}\zeta_k = b_k(x,t)\zeta_k + \left[b_k(x,t) - \dfrac{\partial f}{\partial y}(x,t,y(x,t),u(x,t))\right] z \\ + h(x,t)\left(\dfrac{1}{\rho}\chi_{E_k} - 1\right) \text{ on } \Sigma_T, \\ \zeta_k(x,0) = 0 \text{ in } \Omega. \end{cases} \tag{5.10}$$

Now we decompose $\zeta_k = \zeta_k^1 + \zeta_k^2$, with

$$\begin{cases} \dfrac{\partial \zeta_k^1}{\partial t} + A\zeta_k^1 + c_k(x,t)\zeta_k^1 = \left[\dfrac{\partial a_0}{\partial y}(x,t,y(x,t)) - c_k(x,t)\right] z \quad \text{in } \Omega_T, \\ \partial_{\nu_A}\zeta_k^1 = b_k(x,t)\zeta_k^1 + \left[b_k(x,t) - \dfrac{\partial f}{\partial y}(x,t,y(x,t),u(x,t))\right] z \text{ on } \Sigma_T, \\ \zeta_k^1(x,0) = 0 \text{ in } \Omega \end{cases} \tag{5.11}$$

and

$$\begin{cases} \dfrac{\partial \zeta_k^2}{\partial t} + A\zeta_k^2 + c_k(x,t)\zeta_k^2 = 0 \quad \text{in } \Omega_T, \\ \partial_{\nu_A}\zeta_k^2 = b_k(x,t)\zeta_k^2 + h(x,t)\left(\dfrac{1}{\rho}\chi_{E_k} - 1\right) \text{ on } \Sigma_T, \\ \zeta_k^2(x,0) = 0 \text{ in } \Omega. \end{cases} \tag{5.12}$$

Taking into account (5.2) and (2.1)–(2.8), multiplying equation (5.12) by the function $\exp(-\omega t)\zeta_k^2$, with $\omega > 0$ large enough, and integrating by parts, we deduce

$$C\left(\|\zeta_k^2\|^2_{L^2(\Omega_T)} + \|\zeta_k^2\|^2_{L^2([0,T],H^1(\Omega))}\right)$$

$$\leq \frac{\exp(-\omega T)}{2}\|\zeta_k^2(T)\|^2_{L^2(\Omega)} + \frac{\omega}{2}\int_0^T \exp(-\omega t)\int_\Omega |\zeta_k^2(x,t)|^2 dxdt$$

$$+\int_0^T \exp(-\omega t)\langle A\zeta_k^2, \zeta_k^2\rangle dt + \int_0^T \exp(-\omega t)\int_\Omega c_k(x,t)|\zeta_k^2(x,t)|^2 dxdt$$

$$= \int_0^T\int_\Gamma \exp(-\omega t) b_k(x,t)|\zeta_k^2(x,t)|^2 d\sigma(x)dt$$

$$+\int_0^T\int_\Gamma \exp(-\omega t)h(x,t)\left(\frac{1}{\rho}\chi_{E_k}(x,t)-1\right)\zeta_k^2(x,t)d\sigma(x)dt$$

$$\leq \int_0^T\int_\Gamma \exp(-\omega t)h(x,t)\left(\frac{1}{\rho}\chi_{E_k}(x,t)-1\right)\zeta_k^2(x,t)d\sigma(x)dt. \tag{5.13}$$

From here it follows that

$$\|\zeta_k^2\|^2_{L^2(\Omega_T)} \leq C'\left\|h\left(\frac{1}{\rho}\chi_{E_k}-1\right)\right\|_{C^{\beta,\beta/2}(\bar\Omega_T)'}\|\zeta_k^2\|_{C^{\beta,\beta/2}(\bar\Omega_T)} \tag{5.14}$$

for some $\beta \in (0,1]$. The Hölder regularity of ζ_k^2 follows from the assumptions (2.1)–(2.8) and the results of di Benedetto [2].

On the other hand, for $\theta \in (0,\beta)$, the inclusions

$$C^{\beta,\beta/2}(\bar\Omega_T) \subset C^{\theta,\theta/2}(\bar\Omega_T) \subset L^2(\Omega_T)$$

are compact. Then we can apply the Lions lemma [28] to obtain

$$\|\zeta_k^2\|_{C^{\theta,\theta/2}(\bar\Omega_T)} \leq \epsilon\|\zeta_k^2\|_{C^{\beta,\beta/2}(\bar\Omega_T)} + C_\epsilon\|\zeta_k^2\|_{L^2(\Omega_T)}. \tag{5.15}$$

Since y, y_k, and h are uniformly bounded, the Hölder estimate of ζ_k^2 can be chosen depending only on ρ:

$$\|\zeta_k^2\|_{C^{\beta,\beta/2}(\bar\Omega_T)} \leq C_\rho \quad \forall k \in \mathbb{N}. \tag{5.16}$$

Taking $\epsilon = \rho/(2[1+C_\rho])$ in (5.15) and using (5.14) and (5.16), it follows

$$\|\zeta_k^2\|_{C^{\theta,\theta/2}(\bar\Omega_T)} \leq \frac{\rho}{2} + C_\epsilon\left\{C'\left\|h\left(\frac{1}{\rho}\chi_{E_k}-1\right)\right\|_{C^{\beta,\beta/2}(\bar\Omega_T)'}C_\rho\right\}^{1/2}$$

$$= \frac{\rho}{2} + C'_\rho\left\|h\left(\frac{1}{\rho}\chi_{E_k}-1\right)\right\|^{1/2}_{C^{\beta,\beta/2}(\bar\Omega_T)'}. \tag{5.17}$$

Then, for ρ fixed, the convergence $(1/\rho)\chi_{E_k} \to 1$ *weakly in $L^\infty(\Sigma_T)$, the boundedness of h, and the compactness of the inclusion $L^\infty(\Sigma_T) \subset C^{\beta,\beta/2}(\bar\Omega_T)'$ implies strong convergence $(1/\rho)h\chi_{E_k} \to h$ in $C^{\beta,\beta/2}(\bar\Omega_T)'$. Therefore we can take $k_\rho \in \mathbb{N}$ large enough in such a way that

$$\left|\int_{\Sigma_T} h_0(x,t)\left(\frac{1}{\rho}\chi_{E_k}-1\right)d\sigma(x)dt\right| + \left\|h(x,t)\left(\frac{1}{\rho}\chi_{E_k}-1\right)\right\|_{C^{\beta,\beta/2}(\bar\Omega_T)'}$$

$$< \frac{\rho^2}{4(1+C'_\rho)^2} \quad \forall k \ge k_\rho, \tag{5.18}$$

where

$$h_0(x,t) = l(x,t,y(x,t),v(x,t)) - l(x,t,y(x,t),u(x,t)).$$

Let us set $E_\rho = E_{k_\rho}$, $u_\rho = u_{k_\rho}$, and the analogous changes for y_ρ, ζ_ρ, ζ^i_ρ, $i = 1, 2$. It is obvious that $d_E(u_\rho, u) \to 0$ when $\rho \to 0$. Hence Theorem 5.1 implies that $y_\rho \to y$ in Y. This convergence along with the estimates of di Benedetto [2] allow us to deduce from (5.11) the strong convergence $\zeta^1_\rho \to 0$ in Y when $\rho \to 0$. Combining this with (5.13), (5.17), and (5.18), it is easy to derive the strong convergence $\zeta_\rho \to 0$ in Y, which proves (5.3).

To conclude the proof it is enough to note that

$$\frac{J(u_\rho)-J(u)}{\rho} - z^0$$

$$= \int_{\Omega_T}\left\{\frac{L(x,t,y_\rho(x,t))-L(x,t,y(x,t))}{\rho} - \frac{\partial L}{\partial y}(x,t,y(x,t))z(x,t)\right\}dxdt$$

$$\int_{\Sigma_T}\left\{\frac{l(x,t,y_\rho(x,t),u_\rho(x,t))-l(x,t,y(x,t),u_\rho(x,t))}{\rho}\right.$$

$$\left.-\frac{\partial l}{\partial y}(x,t,y(x,t),u(x,t))z(x,t)\right\}d\sigma(x)dt$$

$$+\int_{\Sigma_T} h_0(x,t)\left(\frac{1}{\rho}\chi_{E_\rho}(x,t)-1\right)d\sigma(x)dt$$

and to take into account the convergences previously established and (5.18). □

6. Linear parabolic equations involving measure data. Let μ be a regular Borel measure in $\bar\Omega_T$. We can write $\mu = \mu_{\Omega_T} + \mu_{\Sigma_T} + \mu_T + \mu_0$, where $\mu_{\Omega_T} = \mu|_{\Omega_T}$, $\mu_{\Sigma_T} = \mu|_{\Sigma_T}$, $\mu_T = \mu|_{\bar\Omega\times\{T\}}$, and $\mu_0 = \mu|_{\bar\Omega\times\{0\}}$. The aim of this section is the study of the following problem:

$$\begin{cases} -\dfrac{\partial\varphi}{\partial t} + A^*\varphi = \mu_{\Omega_T} \text{ in } \Omega_T, \\ \partial_{\nu_{A^*}}\varphi = \mu_{\Sigma_T} \text{ on } \Sigma_T, \\ \varphi(T) = \mu_T \text{ in } \bar\Omega. \end{cases} \tag{6.1}$$

The reader is referred to Boccardo and Gallouët [3] for the study of a quasi-linear parabolic equation with a measure in Ω_T as a datum. Here we improve the results of [3] by exploiting the linearity of the equation.

Let us denote

$$Y_0 = \{y \in Y : y(x,0) = 0 \ \forall x \in \Omega\}.$$

DEFINITION 6.1. *Given $p, r \in [1,2)$, with $(2/r) + (n/p) > n+1$, we will say that a function $\varphi \in L^r([0,T], W^{1,p}(\Omega))$ is a solution of (6.1) if for every $y \in Y_0 \cap C^1(\bar{\Omega}_T)$*

$$\int_{\Omega_T} \left\{ \frac{\partial y}{\partial t}\varphi + \sum_{j=1}^{n} \left[\sum_{i=1}^{n} a_{ij} \partial_{x_i} y \partial_{x_j} \varphi + b_j y \partial_{x_j} \varphi + d_j \partial_{x_j} y \varphi \right] + cy\varphi \right\} dxdt$$

$$= \int_{\bar{\Omega}_T} y d\mu(x,t) = \int_{\Omega_T} y d\mu_{\Omega_T}(x,t) + \int_{\Sigma_T} y d\mu_{\Sigma_T}(x,t) + \int_{\bar{\Omega}} y(x,T) d\mu_T(x). \tag{6.2}$$

Let us note that (6.2) implies that $-(\partial\varphi/\partial t) + A^*\varphi = \mu_{\Omega_T}$ in the distribution sense in Ω_T. Let us take $\vec{w} = (w_1, \ldots, w_{n+1})$, with

$$w_i = \sum_{j=1}^{n} a_{ij} \partial_{x_j} \varphi + d_i \varphi, \ 1 \le i \le n, \text{ and } \ w_{n+1} = \varphi.$$

Then $\vec{w} \in L^q(\Omega_T)^{n+1}$, $q = \min\{r,p\} < (n+1)/n$, and

$$\operatorname{div}_{(x,t)} \vec{w} = \frac{\partial \varphi}{\partial t} + \sum_{i=1}^{n} \partial_{x_i} \left[\sum_{j=1}^{n} a_{ij} \partial_{x_j} \varphi + d_i \varphi \right] = \frac{\partial \varphi}{\partial t} - A^*\varphi + \sum_{i=1}^{n} b_i \partial_{x_i} \varphi + c\varphi$$

$$= -\mu_{\Omega_T} + \sum_{i=1}^{n} b_i \partial_{x_i} \varphi + c\varphi \in M(\Omega_T). \tag{6.3}$$

Thus we have $\vec{w} \in V^q(\Omega_T)$,

$$V^q(\Omega_T) = \{\vec{w} \in L^q(\Omega_T)^{n+1} : \operatorname{div}_{(x,t)} \vec{w} \in M(\Omega_T)\}.$$

This space, endowed with the graph norm, is a Banach space. We have the following result.

THEOREM 6.2 (see Casas [10]). *Given $q \in (1, (n+1)/n)$, there exists a unique continuous linear mapping $\gamma_{\nu_T} : V^q(\Omega_T) \longrightarrow W^{-1/q,q}(\partial\Omega_T)$ satisfying*

$$\gamma_{\nu_T}(\vec{w}) = \vec{w} \cdot \vec{\nu}_T \quad \forall \vec{w} \in C^1(\bar{\Omega}_T) \tag{6.4}$$

and

$$\int_{\Omega_T} \vec{w} \cdot \nabla_{(x,t)} \phi dxdt + \langle \operatorname{div}_{(x,t)} \vec{w}, \phi \rangle_{M(\Omega_T), C_b(\Omega_T)}$$

$$= \langle \gamma_{\nu_T}(\vec{w}), \gamma(\phi) \rangle_{W^{-1/q,q}(\partial\Omega_T), W^{1/q,q'}(\partial\Omega_T)} \quad \forall \phi \in W^{1,q'}(\Omega_T), \tag{6.5}$$

where $C_b(\Omega_T)$ is the space of bounded and continuous functions in Ω_T and $\vec{\nu}_T(x,t)$ is the outward unit normal vector to $\partial\Omega_T$ at the point (x,t).

By applying this theorem to the function $\vec{w}$ defined above and using (6.2) and (6.3), we have for all $y \in Y_0 \cap C^1(\bar{\Omega}_T)$

$$\langle \gamma_{\nu_T}(\vec{w}), \gamma(y) \rangle_{W^{-1/q,q}(\partial\Omega_T), W^{1/q,q'}(\partial\Omega_T)}$$

$$= \int_{\Omega_T} \vec{w} \cdot \nabla_{(x,t)} y dx dt + \langle \mathrm{div}_{(x,t)} \vec{w}, y \rangle_{M(\Omega_T), C_b(\Omega_T)}$$

$$= \int_{\Omega_T} \left\{ \frac{\partial y}{\partial t}\varphi + \sum_{i=1}^n \left[\sum_{j=1}^n a_{ij} \partial_{x_i} y \partial_{x_j} \varphi + b_i y \partial_{x_i} \varphi + d_i \partial_{x_i} y \varphi \right] + cy\varphi \right\} dx dt$$

$$- \int_{\Omega_T} y d\mu_{\Omega_T} = \int_0^T \int_\Gamma y d\mu_{\Sigma_T}(x,t) + \int_\Omega y(x,T) d\mu_T(x).$$

From the identity

$$\langle \gamma_{\nu_T}(\vec{w}), \gamma(y) \rangle_{W^{-1/q,q}(\partial\Omega_T), W^{1/q,q'}(\partial\Omega_T)} = \int_0^T \int_\Gamma y d\mu_{\Sigma_T}(x,t) + \int_\Omega y(x,T) d\mu_T(x)$$

and taking into account that

$$\vec{\nu}_T(x,t) = \begin{pmatrix} \vec{\nu}(x) \\ 0 \end{pmatrix} \quad \forall (x,t) \in \Sigma_T \quad \text{and} \quad \vec{\nu}_T(x,T) = \begin{pmatrix} \vec{0} \\ 1 \end{pmatrix} \quad \forall x \in \Omega,$$

we can identify

$$\partial_{\nu_{A^*}} \varphi = \gamma_{\nu_T}(\vec{\omega})|_{\Sigma_T} = \mu_{\Sigma_T} \quad \text{and} \quad \varphi(x,T) = \gamma_{\nu_T}(\vec{\omega})|_{\bar{\Omega}\times\{T\}} = \mu_T.$$

Now we have the following result of existence and uniqueness of solution for problem (6.1).

THEOREM 6.3. *There exists a unique function $\varphi \in L^r([0,T], W^{1,p}(\Omega))$ $\forall r,p \in [1,2)$ with $(2/r) + (n/p) > n+1$ such that it is a solution of* (6.1) *and*

$$\int_{\Omega_T} \left(\frac{\partial y}{\partial t} + Ay \right) \varphi dx dt + \int_{\Sigma_T} \partial_{\nu_A} y \varphi d\sigma(x) dt = \int_{\bar{\Omega}_T} y d\mu(x,t) \quad \forall y \in Y_0^\infty, \tag{6.6}$$

with

$$Y_0^\infty = \left\{ y \in Y_0 : \frac{\partial y}{\partial t} + Ay \in L^\infty(\Omega_T) \quad \textit{and} \quad \partial_{\nu_A} y \in L^\infty(\Sigma_T) \right\}.$$

Moreover, there exists a constant $C_{r,p} > 0$ independent of μ such that

$$\|\varphi\|_{L^r([0,T],W^{1,p}(\Omega))} \le C_{r,p} \|\mu\|_{M(\bar{\Omega}_T)}. \tag{6.7}$$

Proof. Let $\{f_k\}_k \subset C(\bar{\Omega}_T)$, $\{g_k\}_k \subset C(\Gamma \times [0,T])$ and $\{h_k\}_k \subset C(\bar{\Omega})$ such that $f_k \to \mu_{\Omega_T}$, $g_k \to \mu_{\Sigma_T}$, and $h_k \to \mu_T$ *weakly in $M(\Omega_T)$, $M(\Sigma_T)$, and $M(\bar{\Omega})$, respectively. Moreover, we can assume that

$$\|f_k\|_{L^1(\Omega_T)} \le \|\mu_{\Omega_T}\|_{M(\Omega_T)}, \quad \|g_k\|_{L^1(\Sigma_T)} \le \|\mu_{\Sigma_T}\|_{M(\Sigma_T)}, \text{ and } \|h_k\|_{L^1(\Omega)} \le \|\mu_T\|_{M(\bar{\Omega})}.$$

Let us take $\varphi_k \in Y$ such that

$$\begin{cases} -\dfrac{\partial \varphi_k}{\partial t} + A^*\varphi_k = f_k \text{ in } \Omega_T, \\ \partial_{\nu_{A^*}} \varphi_k = g_k \text{ on } \Sigma_T, \\ \varphi_k(T) = h_k \text{ in } \Omega. \end{cases} \tag{6.8}$$

Now for every $\psi = (\psi_0, \psi_1, \ldots, \psi_n) \in \mathcal{D}(\Omega_T)^{n+1}$, we denote by y_ψ the solution in Y of

$$\begin{cases} \dfrac{\partial y}{\partial t} + Ay = \psi_0 - \displaystyle\sum_{j=1}^{n} \partial_{x_j} \psi_j \text{ in } \Omega_T, \\ \partial_{\nu_A} y = 0 \text{ on } \Sigma_T, \\ y(0) = 0 \text{ in } \Omega. \end{cases} \tag{6.9}$$

Then

$$\int_{\Omega_T} \left(\psi_0 \varphi_k + \sum_{j=1}^{n} \psi_j \partial_{x_j} \varphi_k \right) dxdt = \int_{\Omega_T} \left(\frac{\partial y_\psi}{\partial t} + A y_\psi \right) \varphi_k dxdt$$

$$= \int_{\Omega_T} \left(-\frac{\partial \varphi_k}{\partial t} + A^*\varphi_k \right) y_\psi dxdt + \int_{\Sigma_T} \partial_{\nu_{A^*}} \varphi_k y_\psi d\sigma(x)dt + \int_{\Omega} \varphi_k(T) y_\psi(T) dx. \tag{6.10}$$

Using (6.8) and the properties of f_k, g_k, and h_k, we deduce from (6.8)

$$\int_{\Omega_T} \left(\psi_0 \varphi_k + \sum_{j=1}^{n} \psi_j \partial_{x_j} \varphi_k \right) dxdt$$

$$\leq \|\mu\|_{M(\bar{\Omega}_T)} \|y_\psi\|_{C(\bar{\Omega}_T)} \leq C_{r,p} \|\mu\|_{M(\bar{\Omega}_T)} \sum_{j=0}^{n} \|\psi_j\|_{L^{r'}([0,T], L^{p'}(\Omega))}, \tag{6.11}$$

the last inequality being a consequence of the estimates for the solution of (6.9); see di Benedetto [2] and Ladyzhenskaya, Solonnikov, and Ural'tseva [23]. From the density of the space $\{\psi_0 - \sum_{j=1}^{n} \partial_{x_j} \psi_j : \psi \in \mathcal{D}(\Omega_T)^{n+1}\}$ in $L^{r'}([0,T], W^{1,p}(\Omega)')$ and estimate (6.11) follows the boundedness of $\{\varphi_k\}_k$ in the space $L^r([0,T], W^{1,p}(\Omega))$. Moreover, by taking a subsequence if necessary, we can assume that $\varphi_k \to \varphi$ weakly in $L^r([0,T], W^{1,p}(\Omega))$ and (6.7) is satisfied.

Let us prove that φ does not depend on r and p. Indeed, passing to the limit in (6.10) and remembering that $y_\psi(0) = 0$, we get

$$\int_{\Omega_T} \left(\psi_0 \varphi + \sum_{j=1}^{n} \psi_j \partial_{x_j} \varphi \right) dxdt = \int_{\bar{\Omega}_T} y_\psi d\mu \quad \forall \psi \in \mathcal{D}(\Omega_T)^{n+1}. \tag{6.12}$$

It is obvious that there is at most one function φ in $L^1([0,T], W^{1,1}(\Omega))$ satisfying (6.12), which proves that φ is independent of r and p.

Given $y \in Y_0 \cap C^1(\bar{\Omega}_T)$, multiplying (6.8) by y and integrating by parts, it follows that

$$\int_{\Omega_T} \left\{ \frac{\partial y}{\partial t}\varphi_k + \sum_{j=1}^{n} \left[\sum_{i=1}^{n} a_{ij}\partial_{x_i} y \partial_{x_j}\varphi_k + b_j y \partial_{x_j}\varphi_k + d_j \partial_{x_j} y \varphi_k \right] + cy\varphi_k \right\} dxdt$$
$$= \int_{\Omega_T} f_k y dxdt + \int_{\Sigma_T} g_k y d\sigma(x)dt + \int_{\Omega} h_k y(T)dx.$$

Now passing to the limit we deduce (6.2) and consequently φ is a solution of (6.1).

Let us prove (6.6). Given $y \in Y_0^\infty$, multiplying (6.8) by y and integrating by parts, we deduce

$$\int_{\Omega_T} f_k y dxdt + \int_{\Sigma_T} g_k y d\sigma(x)dt + \int_{\Omega} h_k y(T)dx$$
$$= \int_{\Omega_T} \left(\frac{\partial y}{\partial t} + Ay \right) \varphi_k dxdt + \int_{\Sigma_T} \partial_{\nu_A} y \varphi_k d\sigma(x)dt.$$

Now (6.6) is obtained by passing to the limit.

Finally, the uniqueness of φ follows from (6.6). Indeed, the regularity results for the Neumann problem associated with the operator $(\partial/\partial t) + A$ (see [2] or [23]) prove the surjectivity of the mapping

$$y \in Y_0^\infty \longrightarrow \left(\frac{\partial y}{\partial t} + Ay, \partial_{\nu_A} y \right) \in L^\infty(\Omega_T) \times L^\infty(\Sigma_T).$$

This along with (6.6) implies that the zero function of $L^r([0,T], W^{1,p}(\Omega))$ is the only one satisfying

$$\int_{\Omega_T} \left(\frac{\partial y}{\partial t} + Ay \right) \varphi dxdt + \int_{\Sigma_T} \partial_{\nu_A} y \varphi d\sigma(x)dt = 0 \quad \forall y \in Y_0^\infty.$$

This shows the uniqueness of φ. □

An interesting case arises when $\mu = g\omega$, with $g \in C([0,T], L^2(\Omega))$ and $\omega \in M[0,T]$

$$\int_{\bar{\Omega}_T} z d\mu = \int_0^T \left(\int_\Omega z(x,t)g(x,t)dx \right) d\omega(t) \quad \forall z \in C([0,T], L^2(\Omega));$$

see Example 3.5. In this particular case we have the following result.

THEOREM 6.4. *With the above notation, there exists a unique function φ in the space $L^2([0,T], H^1(\Omega)) \cap L^\infty([0,T], L^2(\Omega))$ solution of the problem*

$$\begin{cases} -\dfrac{\partial \varphi}{\partial t} + A^*\varphi = g\omega \text{ in } \Omega_T, \\ \partial_{\nu_{A^*}} \varphi = 0 \text{ on } \Sigma_T, \\ \varphi(T) = g(T)\omega(\{T\}) \text{ in } \Omega. \end{cases} \tag{6.13}$$

Proof. Uniqueness can be obtained in the standard way. For the proof of the existence we take a sequence $\{\omega_k\}_k \subset C[0,T]$ converging *weakly to ω in $M[0,T]$ and satisfying

$$\|\omega_k\|_{L^1(\Omega_T)} \le \|\omega\|_{M[0,T]}.$$

Let us take $\varphi_k \in Y$ such that

$$\begin{cases} -\dfrac{\partial \varphi_k}{\partial t} + A^*\varphi_k = g\omega_k \text{ in } \Omega_T, \\ \partial_{\nu_{A^*}}\varphi_k = 0 \text{ on } \Sigma_T, \\ \varphi_k(T) = g(T)\omega(\{T\}) \text{ in } \Omega. \end{cases} \tag{6.14}$$

Given $f \in \mathcal{D}(\Omega_T)$, let us denote by y_f the solution in Y of the problem

$$\begin{cases} \dfrac{\partial y}{\partial t} + Ay = f \text{ in } \Omega_T, \\ \partial_{\nu_A} y = 0 \text{ on } \Sigma_T, \\ y(0) = 0 \text{ in } \Omega. \end{cases} \tag{6.15}$$

Then

$$\int_{\Omega_T} f\varphi_k dxdt = \int_{\Omega_T} \left(\frac{\partial y}{\partial t} + Ay\right)\varphi_k dxdt = \int_{\Omega_T} g\omega_k y dxdt + \int_{\Omega} \omega(\{T\})g(T)y(T)dx$$

$$\leq \|g\|_{C([0,T],L^2(\Omega))}\|\omega\|_{M[0,T]}\|y\|_{C([0,T],L^2(\Omega))}. \tag{6.16}$$

From (6.15) it follows by using the classical arguments that

$$\|y\|_{C([0,T],L^2(\Omega))} \leq C_1\|f\|_{L^1([0,T],L^2(\Omega))} \text{ and } \|y\|_{C([0,T],L^2(\Omega))} \leq C_2\|f\|_{L^2([0,T],H^1(\Omega)')}.$$

From the first inequality and (6.16) we deduce the boundedness of the sequence $\{\varphi_k\}_k$ in the space $L^\infty([0,T],L^2(\Omega))$. The second inequality leads to the boundedness of the same sequence in $L^2([0,T],H^1(\Omega))$. The rest of the proof is easy. □

As mentioned in section 3, problems of type (6.13) have been studied by Barbu and Precupanu [1], Lasiecka [24], and Tröltzsch [32].

In the case of a measure $\mu = g\omega$, with $g \in L^1[0,T]$ and $\omega \in \mathcal{M}(\bar{\Omega})$, we deduce from Theorem 6.3 and the inclusion $W^{1,p}(\Omega) \subset \mathcal{M}(\bar{\Omega}) \subset W^{1,p'}(\Omega)'$the existence of a solution $\varphi \in L^1([0,T],W^{1,p}(\Omega))$ for all $p \in [1, n/(n-1))$ and such that $\partial\varphi/\partial t \in L^1([0,T],W^{1,p'}(\Omega)')$. Hence we deduce that $\varphi \in C([0,T],W^{1,p'}(\Omega)')$ after a modification on a set of zero measure.

7. Proof of Pontryagin principle. In this section we prove Theorems 3.1 and 4.3. A crucial point in the proofs is the use of Ekeland's variational principle that we state now.

LEMMA 7.1 (see Ekeland [16]). *Let (E,d) be a complete metric space and $F : E \longrightarrow \mathbb{R} \cup \{+\infty\}$ a lower semicontinuous function, and let $e_\epsilon \in E$ satisfy*

$$F(e_\epsilon) \leq \inf_{e\in E} F(e) + \epsilon.$$

Then there exists an element $\bar{e}_\epsilon \in E$ such that

$$F(\bar{e}_\epsilon) \leq F(e_\epsilon), \quad d(\bar{e}_\epsilon, e_\epsilon) \leq \sqrt{\epsilon},$$

and

$$F(\bar{e}_\epsilon) \leq F(e) + \sqrt{\epsilon}d(e,\bar{e}_\epsilon) \quad \forall e \in E.$$

Proof of Theorem 3.1. Since Z is separable, we can take in Z a norm $\|\cdot\|_Z$ such that Z' endowed with the dual norm $\|\cdot\|_{Z'}$ is strictly convex. Then the function

$$d_Q : (Z, \|\cdot\|_Z) \longrightarrow \mathbb{R},$$

$$d_Q(z) = \inf_{y\in Q} \|y - z\|_Z$$

is convex, Lipschitz and Gâteaux differentiable at every point $z \notin Q$, with $\partial d_Q(z) = \{\nabla d_Q(z)\}$, where the Clarke's generalized gradient and the subdifferential in the sense of the convex analysis coincide for this function. Therefore, given $\xi \in \partial d_Q(y)$, we have that

$$\langle \xi, z - y\rangle + d_Q(y) \le d_Q(z) \quad \forall z \in Z. \tag{7.1}$$

Moreover, $\|\nabla d_Q(z)\|_{Z'} = 1$ for every $z \notin Q$; see Clarke [15] and Casas and Yong [14].

Let us take $J_\epsilon : \mathcal{U} \longrightarrow \mathbb{R}$ defined by

$$J_\epsilon(u) = \left\{[(J(u) - J(\bar u) + \epsilon)^+]^2 + d_Q(G(y_u))^2 + |F(y_u)|^2\right\}^{1/2}.$$

It is obvious that $J_\epsilon(u) > 0$ for every $u \in \mathcal{U}$ and $J_\epsilon(\bar u) = \epsilon$. On the other hand, thanks to Theorem 5.1 we have that J_ϵ is continuous in $(\mathcal{U}, d_E)$, with d_E defined by (5.1). Therefore we can apply Ekeland's variational principle and deduce the existence of $u^\epsilon \in \mathcal{U}$ such that

$$d_E(u^\epsilon, \bar u) \le \sqrt{\epsilon} \quad \text{and} \quad 0 < J_\epsilon(u^\epsilon) \le J_\epsilon(u) + \sqrt{\epsilon} d_E(u^\epsilon, u) \ \ \forall u \in \mathcal{U}. \tag{7.2}$$

Given $v \in \mathcal{U}$ arbitrary, let us take E_ρ and u^ϵ_ρ as in Theorem 5.2,

$$u^\epsilon_\rho(x) = \begin{cases} u^\epsilon(x) & \text{if } x \in \Sigma_T \setminus E_\rho, \\ v(x) & \text{if } x \in E_\rho. \end{cases}$$

Then with the help of (5.3) and (5.4) we get

$$\begin{aligned} -\sqrt{\epsilon} m_{\Sigma_T}(\Sigma) \le \frac{J_\epsilon(u^\epsilon_\rho) - J_\epsilon(u^\epsilon)}{\rho} &= \frac{[(J(u^\epsilon_\rho) - J(\bar u) + \epsilon)^+]^2 - [(J(u^\epsilon) - J(\bar u) + \epsilon)^+]^2}{\rho[J_\epsilon(u^\epsilon_\rho) + J_\epsilon(u^\epsilon)]} \\ &+ \frac{d_Q(G(y^\epsilon_\rho))^2 - d_Q(G(y^\epsilon))^2 + |F(y^\epsilon_\rho)|^2 - |F(y^\epsilon)|^2}{\rho[J_\epsilon(u^\epsilon_\rho) + J_\epsilon(u^\epsilon)]} \\ \xrightarrow{\rho\to 0} \left\{(J(u^\epsilon) - J(\bar u) + \epsilon)^+ z^{0,\epsilon} + \langle \xi^\epsilon, DG(y^\epsilon)z^\epsilon\rangle + \langle F(y^\epsilon), DF(y^\epsilon)z^\epsilon\rangle\right\} &/ J_\epsilon(u^\epsilon) \\ = \alpha_\epsilon z^{0,\epsilon} + \langle [DG(y^\epsilon)]^* \mu^\epsilon, z^\epsilon\rangle + \langle [DF(y^\epsilon)]^* \lambda^\epsilon, z^\epsilon\rangle, & \end{aligned} \tag{7.3}$$

where y^ϵ and y^ϵ_ρ are the states associated with u^ϵ and u^ϵ_ρ, respectively, and $z^\epsilon \in Y$ satisfies

$$\begin{cases} \dfrac{\partial z^\epsilon}{\partial t} + Az^\epsilon + \dfrac{\partial a_0}{\partial y}(x, t, y^\epsilon(x))z^\epsilon = 0 \quad \text{in } \Omega_T, \\ \partial_{\nu_A} z^\epsilon = \dfrac{\partial f}{\partial y}(x, t, y^\epsilon(x,t), u^\epsilon(x,t))z^\epsilon \\ + f(x, t, y^\epsilon(x,t), v(x,t)) - f(x, t, y^\epsilon(x,t), u^\epsilon(x,t)) \text{ on } \Sigma_T, \\ z^\epsilon(x, 0) = 0 \text{ in } \Omega, \end{cases} \tag{7.4}$$

$$z^{0,\epsilon} = \int_{\Omega_T} \frac{\partial L}{\partial y}(x,t,y^\epsilon(x,t))z^\epsilon(x,t)dxdt + \int_{\Sigma_T} \frac{\partial l}{\partial y}(x,t,y^\epsilon(x,t),u^\epsilon(x,t))z^\epsilon(x,t)d\sigma(x)dt$$

$$+ \int_{\Sigma_T} [l(x,t,y^\epsilon(x,t),v(x,t)) - l(x,t,y^\epsilon(x,t),u(x,t))]d\sigma(x)dt, \tag{7.5}$$

$$\alpha_\epsilon = \frac{(J(u^\epsilon) - J(\bar{u}) + \epsilon)^+}{J_\epsilon(u^\epsilon)}, \quad \mu^\epsilon = \frac{\xi^\epsilon}{J_\epsilon(u^\epsilon)}, \quad \lambda^\epsilon = \frac{F(y^\epsilon)}{J_\epsilon(u^\epsilon)}, \tag{7.6}$$

$$\xi^\epsilon = \begin{cases} d_Q(G(y^\epsilon))\nabla d_Q G(y^\epsilon)) & \text{if } G(y^\epsilon) \notin Q, \\ 0 & \text{otherwise.} \end{cases} \tag{7.7}$$

By using Theorem 6.3, we can take a function $\varphi^\epsilon \in L^r([0,T],W^{1,p}(\Omega))$ $\forall r,p \in [1,2)$ with $(2/r)+(n/p) > n+1$ such that

$$\begin{cases} -\dfrac{\partial \varphi^\epsilon}{\partial t} + A^*\varphi^\epsilon + \dfrac{\partial a_0}{\partial y}(x,t,y^\epsilon)\varphi^\epsilon = \alpha_\epsilon \dfrac{\partial L}{\partial y}(x,t,y^\epsilon) \\ \qquad\qquad +[DG(y^\epsilon)^*\mu_\epsilon]|_{\Omega_T} + [DF(y^\epsilon)^*\lambda_\epsilon]|_{\Omega_T} \quad \text{in } \Omega_T, \\ \partial_{\nu_{A^*}}\varphi^\epsilon = \dfrac{\partial f}{\partial y}(x,t,y^\epsilon,u^\epsilon)\varphi^\epsilon + \alpha_\epsilon \dfrac{\partial l}{\partial y}(x,t,y^\epsilon,u^\epsilon) \\ \qquad\qquad +[DG(y^\epsilon)^*\mu_\epsilon]|_{\Sigma_T} + [DF(y^\epsilon)^*\lambda_\epsilon]|_{\Sigma_T} \text{ on } \Sigma_T, \\ \varphi^\epsilon(T) = [DG(y^\epsilon)^*\mu_\epsilon]|_{\Omega\times\{T\}} + [DF(y^\epsilon)^*\lambda_\epsilon]|_{\Omega\times\{T\}} \text{ in } \Omega. \end{cases} \tag{7.8}$$

Thanks to the assumptions (2.2) and (2.7), we have that $z^\epsilon \in Y_0^\infty$. Then we can apply (6.6) with $y = z^\epsilon$ and deduce from (7.3)–(7.5) and the definition of H_α given in section 3 the inequality

$$\int_{\Sigma_T} H_{\alpha_\epsilon}(x,t,y^\epsilon(x,t),u^\epsilon(x,t),\varphi^\epsilon(x,t))d\sigma(x)dt$$

$$\leq \int_{\Sigma_T} H_{\alpha_\epsilon}(x,t,y^\epsilon(x,t),v(x,t),\varphi^\epsilon(x,t))d\sigma(x)dt + \sqrt{\epsilon}m_{\Sigma_T}(\Sigma_T) \quad \forall v \in \mathcal{U}. \tag{7.9}$$

Now we pass to the limit when $\epsilon \to 0$. To do this, let us remark that

$$\alpha_\epsilon^2 + \|\mu^\epsilon\|_{Z'}^2 + |\lambda^\epsilon|^2 = 1. \tag{7.10}$$

Then we take subsequences, denoted in the same way, satisfying

$$\begin{cases} \alpha_\epsilon \to \bar{\alpha} \ \text{ in } \mathbb{R}, \quad \lambda^\epsilon \to \bar{\lambda} \ \text{ in } \mathbb{R}^n, \\ \mu^\epsilon \to \bar{\mu} \ \text{ in the } {}^*\text{weak topology of } Z'. \end{cases} \tag{7.11}$$

On the other hand, the convergence $y^\epsilon \to \bar{y}$ in Y follows from Theorem 5.1. The boundedness of $\{\varphi^\epsilon\}$ in $L^r([0,T],W^{1,p}(\Omega))$ follows from (6.7) and (7.10). Then, using (7.11), it is easy to pass to the limit in (7.8) and (7.9) and to deduce (3.3) and (3.5). Now remembering the definition of μ^ϵ and ξ^ϵ and (7.1), we deduce

$$\langle \mu^\epsilon, z - G(y^\epsilon)\rangle \leq 0 \quad \forall z \in Q. \tag{7.12}$$

Passing to the limit in this expression we obtain (3.4). Let us prove (3.1). To do this, let us suppose that $\bar{\alpha} = |\bar{\lambda}| = 0$; then from (7.10) it follows $\|\mu^\epsilon\|_{Z'} \to 1$ as $\epsilon \to 0$. Let us take $z_0 \in \overset{o}{Q}$ and $\rho > 0$ such that $\bar{B}_\rho(z_0) \subset \overset{o}{Q}$. Then (7.12) implies that

$$\langle \mu^\epsilon, z + z_0 - G(y^\epsilon)\rangle \leq 0 \quad \forall z \in \bar{B}_\rho(0).$$

Hence

$$\rho\|\mu^\epsilon\|_{Z'} = \sup_{z\in\bar B_\rho(0)} \langle\mu^\epsilon, z\rangle \le \langle\mu^\epsilon, G(y^\epsilon) - z_0\rangle.$$

Passing to the limit

$$0 < \rho \le \lim_{\epsilon\to 0}\langle\mu^\epsilon, G(y^\epsilon) - z_0\rangle = \langle\bar\mu, G(\bar y) - z_0\rangle,$$

which proves that $\bar\mu \neq 0$.

It remains to prove (3.6); see Bonnans and Casas [5] or Casas [11] for the study of analogous situations. To do this we consider the coordinate system $\{(\Gamma_r, a_r)\}_{r=1}^d$ of Γ introduced in the proof of Proposition 5.3. Given a point $x_0 \in \overset{o}{\Gamma}_r$ for some $1 \le r \le d$ we denote for each $\epsilon > 0$ small enough

$$\Gamma_\epsilon(x_0) = \{x = (x'_r, a_r(x'_r)) : x'_r \in B_\epsilon(x'_{0r}) \subset (0,1)^{n-1}\},$$

where $B_\epsilon(x'_{0r})$ is the ball in $\mathbb{R}^{n-1}$ centered at x'_{0r} and having radius ϵ. Now given $0 < t_0 < T$, we set

$$\Sigma_T^\epsilon(x_0, t_0) = \Gamma_\epsilon(x_0) \times (t_0 - \epsilon, t_0 + \epsilon).$$

The following lemma is used in this proof.

LEMMA 7.2. *Given $f \in L^1(\Sigma_T)$, there exists a m_{Σ_T}-measurable set $S \subset \bigcup_{r=1}^d \overset{o}{\Gamma}_r \times(0,T)$, with $m_{\Sigma_T}(S) = m_{\Sigma_T}(\Sigma_T)$, such that for every $(x_0, t_0) \in S$ we have*

$$\lim_{\epsilon\to 0} \frac{1}{m_{\Sigma_T}(\Sigma_T^\epsilon(x_0,t_0))} \int_{\Sigma_T^\epsilon(x_0,t_0)} |f(x,t) - f(x_0,t_0)| dm_{\Sigma_T}(x,t) = 0. \tag{7.13}$$

Proof. Let us denote for all $(x'_r, t) \in (0,1)^{n-1} \times (0,T)$

$$\omega_r(x'_r) = \sqrt{1 + \sum_{i=1}^{n-1} \left|\frac{\partial a_r}{\partial x_i}(x'_r)\right|^2} \quad \text{and} \quad f_r(x'_r, t) = \omega_j(x'_r) f(x'_r, a_r(x'_r), t).$$

Since ω_r and f_r are Lebesgue integrable functions in $(0,1)^{n-1}$ and $(0,1)^{n-1}\times(0,T)$, respectively, we know that the set of Lebesgue points of these functions U_r and V_r, respectively, have measure equal to 1 and T, respectively. Let us define

$$S_r = \{(x,t) \in V_r = (x'_r, a_r(x'_r), t) : x'_r \in U_r\} \quad \text{and} \quad S = \bigcap_{r=1}^d S_r.$$

Then $m_{\Sigma_T}(S) = m_{\Sigma_T}(\Sigma_T)$ and $S_r \subset \overset{o}{\Gamma}_r \times(0,T)$, $1 \le r \le d$.

Let us take $(x_0, t_0) = (x'_{0r}, a(x'_{0r}), t_0) \in S_r$. Then x'_{0j} and (x_0, t_0) are Lebesgue points of ω_r and f_r; consequently,

$$\lim_{\epsilon\to 0} \frac{1}{m_{\Sigma_T}(\Sigma_T^\epsilon(x_0,t_0))} \int_{\Sigma_T^\epsilon(x_0,t_0)} |f(x,t) - f(x_0,t_0)| dm_{\Sigma_T}(x,t)$$

$$= \lim_{\epsilon\to 0} \left(\frac{1}{2\epsilon|B_\epsilon(x'_{0r})|} \int_{t_0-\epsilon}^{t_0+\epsilon} \int_{B_\epsilon(x'_{0r})} |f_r(x'_r,t) - f_r(x'_{0r},t)| dx'_r dt \right),$$

$$\lim_{\epsilon\to 0}\left(\frac{1}{|B_\epsilon(x'_{0r})|}\int_{B_\epsilon(x'_{0r})}\omega_r(x'_r)dx'_r\right)^{-1}$$

$$= f_r(x'_{0r},t_0)/\omega_r(x'_{0r}) = f(x_0,t_0),$$

where $|B_\epsilon(x'_{0r})|$ denotes the $(n-1)$-measure of $B_\epsilon(x'_{0r})$. □

The set points of S will be called the Lebesgue points of f. This set depends on the system of coordinates $\{(\Gamma_r,a_r)\}_{r=1}^d$, but this dependence only affects a set of σ-measure equal to zero.

We return to the proof of (3.6). Assume first that (A1) holds. Let us take a numerable dense subset $\{v_r\}_{j=1}^\infty$ of $\mathcal{K}$. Let F and $\{F_r\}_{j=1}^\infty$ be measurable subsets of Ω, with $m_{\Sigma_T}(F) = m_{\Sigma_T}(\Sigma_T) = m_{\Sigma_T}(F_r)$ for every j, such that the Lebesgue point sets of functions $(x,t)\in\Sigma_T \longrightarrow H_{\bar\alpha}(x,t,\bar y(x,t),\bar u(x,t),\bar\varphi(x,t))$ and $(x,t)\in\Omega\longrightarrow H_{\bar\alpha}(x,t,\bar y(x,t),v_j,\bar\varphi(x,t))$ are F and F_j, respectively. Let us set $F_0 = F\cap[\cap_{j=1}^\infty F_j]$. Then we have $m_{\Sigma_T}(F_0) = m_{\Sigma_T}(\Sigma_T)$. Now given $(x_0,t_0)\in F_0$ arbitrary, for every $\epsilon>0$ small enough and $j\ge 1$ we define the admissible controls

$$u_j^\epsilon(x,t) = \begin{cases} \bar u(x,t) & \text{if } (x,t)\notin\Sigma_T^\epsilon(x_0,t_0),\\ v_j & \text{otherwise.}\end{cases}$$

Then from (3.5) we deduce

$$\frac{1}{m_{\Sigma_T}(\Sigma_T^\epsilon(x_0,t_0))}\int_{\Sigma_T^\epsilon(x_0,t_0)} H_{\bar\alpha}(x,t,\bar y(x,t),\bar u(x,t),\bar\varphi(x,t))d\sigma(x)dt$$

$$\le \frac{1}{m_{\Sigma_T}(\Sigma_T^\epsilon(x_0,t_0))}\int_{\Sigma_T^\epsilon(x_0,t_0)} H_{\bar\alpha}(x,t,\bar y(x,t),v_j,\bar\varphi(x,t))d\sigma(x)dt,\quad 1\le j.$$

Passing to the limit where $\epsilon\to 0$, with the help of Lemma 7.2 we get

$$H_{\bar\alpha}(x_0,t_0,\bar y(x_0,t_0),\bar u(x_0,t_0),\bar\varphi(x_0,t_0)) \le H_{\bar\alpha}(x_0,t_0,\bar y(x_0,t_0),v_j,\bar\varphi(x_0,t_0))$$

for every $(x_0,t_0)\in F_0$ and $j\ge 1$. Taking into account that function

$$v\longrightarrow H_{\bar\alpha}(x_0,t_0,\bar y(x_0,t_0),v,\bar\varphi(x_0,t_0))$$

is continuous and that $\{v_j\}_{j=1}^\infty$ is dense in $\mathcal{K}$, (3.6) follows from the above inequality.

Now let us suppose that assumption (A2) holds. Let $F_{\bar\varphi}$ be a measurable subset of Σ_T such that for every $(x_0,t_0)\in F_{\bar\varphi}$

$$\lim_{\epsilon\to 0}\frac{1}{m_{\Sigma_T}(\Sigma_T^\epsilon(x_0,t_0))}\int_{\Sigma_T^\epsilon(x_0,t_0)}|\bar\varphi(x,t)-\bar\varphi(x_0,t_0)|d\sigma(x)dt = 0. \tag{7.14}$$

Let $F_0 = F_{\bar\varphi}\cap\Sigma_T^0\cap F$, where F is taken as above. Thus we have that $m_{\Sigma_T}(F_0) = m_{\Sigma_T}(\Sigma_T)$, and taking spike perturbations as before we deduce

$$\frac{1}{m_{\Sigma_T}(\Sigma_T^\epsilon(x_0,t_0))}\int_{\Sigma_T^\epsilon(x_0,t_0)} H_{\bar\alpha}(x,t,\bar y(x,t),\bar u(x,t),\bar\varphi(x,t))d\sigma(x)dt$$

$$\le \frac{1}{m_{\Sigma_T}(\Sigma_T^\epsilon(x_0,t_0))}\int_{\Sigma_T^\epsilon(x_0,t_0)} H_{\bar\alpha}(x,t,\bar y(x,t),v,\bar\varphi(x,t))d\sigma(x)dt$$

for every $(x_0,t_0)\in F_0$ and $v\in\mathcal{K}$. Since $(x_0,t_0)\in F$, we can pass to the limit on the left-hand side of the inequality. Let us study the right-hand side:

$$\begin{aligned}
&\frac{1}{m_{\Sigma_T}(\Sigma_T^\epsilon(x_0,t_0))}\int_{\Sigma_T^\epsilon(x_0,t_0)} H_{\bar\alpha}(x,t,\bar y(x,t),v,\bar\varphi(x,t))d\sigma(x)dt\\
&\quad=\frac{1}{m_{\Sigma_T}(\Sigma_T^\epsilon(x_0,t_0))}\int_{\Sigma_T^\epsilon(x_0,t_0)}\bar\alpha l(x,t,\bar y(x,t),v)d\sigma(x)dt\\
&\quad+\frac{1}{m_{\Sigma_T}(\Sigma_T^\epsilon(x_0,t_0))}\int_{\Sigma_T^\epsilon(x_0,t_0)} f(x,t,\bar y(x,t),v)d\sigma(x)dt\bar\varphi(x_0,t_0)\\
&+\frac{1}{m_{\Sigma_T}(\Sigma_T^\epsilon(x_0,t_0))}\int_{\Sigma_T^\epsilon(x_0,t_0)}[\bar\varphi(x,t)-\bar\varphi(x_0,t_0)]f(x,t,\bar y(x,t),v)d\sigma(x)dt.
\end{aligned}$$

The first two terms converge to $H_{\bar\alpha}(x_0,t_0,\bar y(x_0,t_0),v,\bar\varphi(x_0,t_0))$ because of the continuity of the integrands in $(x_0,t_0)\in\Sigma_T^0$. Let us prove that the last term goes to zero.

$$\begin{aligned}
&\left|\frac{1}{m_{\Sigma_T}(\Sigma_T^\epsilon(x_0,t_0))}\int_{\Sigma_T^\epsilon(x_0,t_0)}[\bar\varphi(x,t)-\bar\varphi(x_0,t_0)]f(x,t,\bar y(x,t),v)d\sigma(x)dt\right|\\
&\quad\le C\frac{1}{m_{\Sigma_T}(\Sigma_T^\epsilon(x_0,t_0))}\int_{\Sigma_T^\epsilon(x_0,t_0)}|\bar\varphi(x,t)-\bar\varphi(x_0,t_0)|d\sigma(x)dt\longrightarrow 0,
\end{aligned}$$

thanks to (7.14) and the fact that $(x,t)\to f(x,t,\bar y(x,t),v)$ is bounded in Σ_T because of the assumption (2.2) and the boundedness of $\bar y$. ∎

Now we will prove Theorem 4.3. The key to achieving this result is to carry out an exact penalization of the state constraint. To do this, we will use the distance function d_{Q_δ} associated with the set Q_δ and defined in the same way as in the proof of Theorem 4.3.

PROPOSITION 7.3. *If (P_δ) is strongly stable and $\bar u$ is a solution of this problem, then there exists $q_0>0$ such that $\bar u$ is also a solution of*

$$\inf_{u\in\mathcal{U}} J_q(u)=J(u)+qd_{Q_\delta}(G(y_u)) \tag{7.15}$$

for every $q\ge q_0$.

Proof. Let us suppose that it is false. Then there exists a sequence $\{q_k\}_{k=1}^\infty$ of real numbers, with $q_k\to+\infty$ and elements $\{u_k\}_{k=1}^\infty\subset\mathcal{U}$ such that

$$J(u_k)+q_kd_{Q_\delta}(G(y_k))<J(\bar u)\quad\forall k\ge 1,$$

where y_k is the state corresponding to u_k. From here we obtain that

$$d_{Q_\delta}(G(y_k))<\frac{J(\bar u)-J(u_k)}{q_k}\longrightarrow 0\quad\text{when } k\to+\infty$$

and $G(y_k)\notin Q_\delta$. Let $\delta_k>\delta$ be the smallest number such that $G(y_k)\in Q_{\delta_k}$. Since $\delta_k\to\delta$, we can use (4.1) to deduce

$$\begin{aligned}
C(\delta_k-\delta)&\ge\inf(\mathrm{P}_\delta)-\inf(\mathrm{P}_{\delta_k})\ge J(\bar u)-J(u_k)\\
&>q_kd_{Q_\delta}(G(y_k))=q_k(\delta_k-\delta)\quad\forall k\ge k_\epsilon,
\end{aligned}$$

which is not possible. ∎

Since J_q is not Gâteaux differentiable on Q_δ, we are going to modify slightly this functional to attain the differentiability necessary for the proof.

PROPOSITION 7.4. *Let us take $q \geq q_0$ and for every $\epsilon > 0$ let us consider the problem*

$$(P_{\delta,\epsilon}) \quad \inf_{u\in\mathcal{U}} J_{q,\epsilon}(u) = J(u) + q\left\{d_{Q_\delta}(G(y_u))^2 + \epsilon^2\right\}^{1/2}.$$

Then $\inf(P_{\delta,\epsilon}) \to \inf(P_\delta)$ *when* $\epsilon \to 0$.

Proof. It is an immediate consequence of the inequality

$$J_q(u) \leq J_{q,\epsilon}(u) \leq J_q(u) + q\epsilon \quad \forall u \in \mathcal{U}. \qquad \square$$

Finally we are ready to prove the strong Pontryagin principle.

Proof of Theorem 4.3. Propositions 7.3 and 7.4 imply that $\bar{u}$ is a σ_ϵ^2–solution of $(\mathrm{P}_{\delta,\epsilon})$, with $\sigma_\epsilon \to 0$ when $\epsilon \to 0$; i.e.

$$J_{q,\epsilon}(\bar{u}) \leq \inf(\mathrm{P}_{\delta,\epsilon}) + \sigma_\epsilon^2.$$

Then we can apply again Ekeland's principle and deduce the existence of an element $u^\epsilon \in \mathcal{U}$ such that

$$d(u^\epsilon, \bar{u}) \leq \sigma_\epsilon, \quad J_{q,\epsilon}(u^\epsilon) \leq J_{q,\epsilon}(\bar{u}),$$

and

$$J_{q,\epsilon}(u^\epsilon) \leq J_{q,\epsilon}(u) + \sigma_\epsilon d_E(u^\epsilon, u) \quad \forall u \in \mathcal{U}.$$

Now we argue as in the proof of Theorem 3.1 and replace (7.3) by

$$-\sigma_\epsilon m_{\Sigma_T}(\Sigma_T) \leq \lim_{\rho\to 0} \frac{J_{q,\epsilon}(u_\rho^\epsilon) - J_{q,\epsilon}(u^\epsilon)}{\rho} = z^{0,\epsilon} + \langle \mu^\epsilon, DG(y^\epsilon)z^\epsilon\rangle,$$

where $\mu^\epsilon \in Z'$ is given by

$$\mu^\epsilon = \begin{cases} \dfrac{q d_{Q_\delta}(G(y^\epsilon))}{\{d_{Q_\delta}(G(y^\epsilon))^2 + \epsilon^2\}^{1/2}} \nabla d_{Q_\delta}(G(y^\epsilon)) & \text{if } G(y^\epsilon) \notin Q_\delta, \\ 0 & \text{otherwise.} \end{cases}$$

Therefore we have $\|\mu^\epsilon\|_{Z'} \leq q$ for every $\epsilon > 0$. Now we can take a subsequence that converges weakly* to an element $\bar{\mu} \in Z'$. The rest is as in the proof of Theorem 3.1, taking $\alpha_\epsilon = 1$. $\square$

REFERENCES

[1] V. BARBU AND T. PRECUPANU, *Convexity and Optimization in Banach Spaces*, Editura Academiei, Sijthoff & Noordhoff, Bucharest, 1978.

[2] E. DI BENEDETTO, *On the local behaviour of solutions of degenerate parabolic equations with measurable coefficients*, Ann. Scuola Norm. Sup. Pisa Cl. Sci. (4), 13 (1986), pp. 487–535.

[3] L. BOCCARDO AND T. GALLOUËT, *Non–linear elliptic and parabolic equations involving measure data*, J. Funct. Anal., 87 (1989), pp. 149–169.

[4] J. BONNANS, *Pontryagin's principle for the optimal control of semilinear elliptic systems with state constraints*, in 30th IEEE Conference on Control and Decision, Brighton, England, 1991, pp. 1976–1979.

[5] J. Bonnans and E. Casas, *Un principe de Pontryagine pour le contrôle des systèmes elliptiques*, J. Differential Equations, 90 (1991), pp. 288–303.
[6] J. Bonnans and E. Casas, *An extension of Pontryagin's principle for state-constrained optimal control of semilinear elliptic equations and variational inequalities*, SIAM J. Control Optim., 33 (1995), pp. 274–298.
[7] E. Casas, *Control of an elliptic problem with pointwise state constraints*, SIAM J. Control Optim., 24 (1986), pp. 1309–1318.
[8] E. Casas, *Finite element approximations for some state-constrained optimal control problems*, in Mathematics of the Analysis and Design in Process Control, P. Borne, S. Tzafestas, and N. Radhy, eds., North Holland, Amsterdam, 1992, pp. 293–301.
[9] E. Casas, *Introducción a las Ecuaciones en Derivadas Parciales*, University of Cantabria, Santander, 1992.
[10] E. Casas, *Boundary control of semilinear elliptic equations with pointwise state constraints*, SIAM J. Control Optim., 31 (1993), pp. 993–1006.
[11] E. Casas, *Pontryagin's principle for optimal control problems governed by semilinear elliptic equations*, in International Conference on Control and Estimation of Distributed Parameter Systems: Nonlinear Phenomena, F. Kappel and K. Kunisch, eds., Birkhäuser, Basel, 1994, pp. 97–114.
[12] E. Casas, *Boundary control problems of quasilinear elliptic equations: A Pontryagin's principle*, Appl. Math. Optim., 33 (1996), pp. 265–291.
[13] E. Casas and L. Fernández, *A Green's formula for quasilinear parabolic operators*, in Equadiff 91. International Conference on Differential Equations, C. Perelló, C. Simó, and J. Solá-Morales, eds., World Scientific Publishing, Singapore, 1993, pp. 363–367.
[14] E. Casas and J. Yong, *Maximum principle for state-constrained optimal control problems governed by quasilinear elliptic equations*, Differential Integral Equations, 8 (1995), pp. 1–18.
[15] F. Clarke, *Optimization and Nonsmooth Analysis*, John Wiley, Toronto, 1983.
[16] I. Ekeland, *Nonconvex minimization problems*, Bull. Amer. Math. Soc., 1 (1979), pp. 76–91.
[17] H. Fattorini, *Optimal control problems for distributed parameter systems governed by semilinear parabolic equations in L^1 and L^∞ spaces*, in Optimal Control of Partial Differential Equations, Lecture Notes in Control and Information Sciences 149, K. Hoffmann and W. Krabs, eds., Springer-Verlag, Berlin, 1991, pp. 60–80.
[18] H. Fattorini, *Optimal control problems for distributed parameter systems in Banach spaces*, Appl. Math. Optim., 4 (1993), pp. 225–257.
[19] H. Fattorini and H. Frankowska, *Infinite dimensional control problems with state constraints*, in Proceedings of IFIP-IIASA Conference on Modelling and Inverse Problems of Control for Distributed Parameter Systems, Lecture Notes in Control and Information Sciences 154, Springer-Verlag, Berlin, 1991, pp. 52–62.
[20] H. Fattorini and T. Murphy, *Optimal controls problems for nonlinear parabolic boundary control systems: the Dirichlet boundary condition*, Differential Integral Equations, 6 (1994), pp. 1367–1388.
[21] H. Fattorini and T. Murphy, *Optimal controls problems for nonlinear parabolic boundary control systems*, SIAM J. Control Optim., 32 (1994), pp. 1577–1596.
[22] B. Hu and J. Yong, *Pontryagin Maximum Principle for Semilinear and Quasilinear Parabolic Equations with Pointwise State Constraints*, Tech. Report 1141, Institute for Mathematics and Its Applications, University of Minnesota, IMA Preprint Series, June 1993.
[23] O. Ladyzhenskaya, V. Solonnikov, and N. Ural'tseva, *Linear and Quasilinear Equations of Parabolic Type*, AMS, Providence, RI, 1968.
[24] I. Lasiecka, *State constrained control problems for parabolic systems: Regularity of optimal solutions*, Appl. Math. Optim., 6 (1980), pp. 1–29.
[25] X. Li, *Vector-valued measure and the necessary conditions for the optimal control problems of linear systems*, in Proc. IFAC 3rd Symposium on Control of Distributed Parameter Systems, Toulouse, France, 1982.
[26] X. Li and Y. Yao, *Maximum principle of distributed parameter systems with time lags*, in Distributed Parameter Systems, Lecture Notes in Control and Information Sciences 75, Springer–Verlag, New York, 1985, pp. 410–427.
[27] X. Li and J. Yong, *Necessary conditions of optimal control for distributed parameter systems*, SIAM J. Control Optim., 29 (1991), pp. 895–908.
[28] J. Lions, *Problèmes aux limites non homogènes IV*, Ann. Scuola Norm. Sup. Pisa, 15 (1961), pp. 311–236.
[29] J. Lions, *Quelques Méthodes de Résolution des Problèmes aux Limites non Linéaires*, Dunod, Paris, 1969.

[30] J. Serrin, *Pathological solutions of elliptic differential equations*, Ann. Scuola Norm. Sup. Pisa, 18 (1964), pp. 385–387.

[31] E. Stein, *Singular Integrals and Differentiability Properties of Functions*, Princeton University Press, Princeton, NJ, 1970.

[32] F. Tröltzsch, *On some parabolic boundary control problems with constraints on the control and functional-constraints on the state*, Z. Anal. Anwendungen, 1 (1982), pp. 1–13.

[33] J. Yong, *Pontryagin maximum principle for semilinear second order elliptic partial differential equations and variational inequalities with state constraints*, Differential Integral Equations, 5 (1992), pp. 1307–1334.

SIAM J. CONTROL OPTIM.
Vol. 35, No. 4, pp. 1328–1365, July 1997

011

AN APPROXIMATION ALGORITHM FOR NONHOLONOMIC SYSTEMS*

WENSHENG LIU†

Abstract. In [*SIAM J. Control Optim.*, 37 (1997), to appear], [*Limiting process of control-affine systems with Hölder continuous inputs*, submitted], we have studied the limiting behavior of trajectories of control affine systems $\Sigma : \dot{x} = \sum_{k=1}^{m} u_k f_k(x)$ generated by a sequence $\{u^j\} \subseteq L^1([0,T],\mathbb{R}^m)$, where the f_k are smooth vector fields on a smooth manifold M. We have shown that under very general conditions the trajectories of Σ generated by the u^j converge to trajectories of an *extended system* of Σ of the form $\Sigma_{ext} : \dot{x} = \sum_{k=1}^{r} v_k f_k(x)$, where $f_k, k = 1,\dots,m$, are the same as in Σ and $f_{m+1},\dots,f_r$ are Lie brackets of $f_1,\dots,f_m$. In this paper, we will apply these convergence results to solve the *inverse* problem; i.e., given any trajectory γ of an extended system Σ_{ext}, find trajectories of Σ that converge to γ uniformly. This is done by means of a universal construction that only involves the knowledge of the $v_k, k = 1,\dots,r$, and the structure of the Lie brackets in Σ_{ext} but does not depend on the manifold M and the vector fields $f_1,\dots,f_m$. These results can be applied to approximately track an arbitrary smooth path in M for controllable systems Σ, which in particular gives an alternative approach to the motion planning problem for nonholonomic systems.

Key words. control affine systems, extended inputs, free associative algebras, free Lie algebras, Chen–Fliess series, nonholonomic motion planning

AMS subject classifications. 34E50, 34E10, 93B05, 93B27, 93C15

PII. S0363012993260501

1. Introduction.

The purpose of this paper is to study the relation between trajectories of control affine systems

$$\dot{x} = \sum_{k=1}^{m} u_k(t) f_k(x), \tag{1}$$

where $f_1,\dots,f_m$ are smooth vector fields on a smooth manifold M, and those of a "Lie bracket extension" of (1), which will be called an *extended system*, given by

$$\dot{x} = \sum_{k=1}^{r} v_k(t) f_k(x), \tag{2}$$

where the first m vector fields $f_1,\dots,f_m$ are the same as in (1) and the new admissible directions of motion $f_{m+1},\dots,f_r$ are Lie brackets of the f_k, $k \in \{1,\dots,m\}$.

Using averaging techniques and developing proper algebraic formalisms, in [19] and [20] we have studied the limiting behavior of trajectories of (1). We have shown that under very general conditions trajectories of (1) generated by a sequence $\{u^j\} \subseteq L^1([0,T],\mathbb{R}^m)$ converge to trajectories of (2). In this paper we will apply these convergence results to solve the *inverse* problem, i.e., to find a sequence of trajectories of (1) approximating a given trajectory of (2).

*Received by the editors December 21, 1993; accepted for publication (in revised form) May 21, 1996. This research was supported in part by National Science Foundation grant DMS92-02554.
http://www.siam.org/journals/sicon/35-4/26050.html
†Department of Mathematics, Rutgers University, New Brunswick, NJ 08903 (wliu@math.rutgers.edu).

It was proved by Haynes and Hermes in [6] that if system (1) satisfies the *Lie algebra rank condition*[1] (LARC), every trajectory of (2) can be uniformly approximated by trajectories of (1). However, their proof does not easily lead to an explicit constructive procedure for producing such a sequence. Here we present a different approach. Based on the convergence results described in [19], we give an algorithm that, for each extended system (2), produces a sequence $\{u^j\}$ of inputs that explicitly generates trajectories for (1) converging to those of (2). This is done in a universal way in the sense that the construction of the u^j only involves the knowledge of the $v_k, k = 1, \ldots, r$ and the structure of the Lie brackets in (2) but does not depend on the manifold M and the vector fields $f_1, \ldots, f_m$.

The motivation for our interest in an explicit construction of trajectories of (1) that converge to a prescribed trajectory of (2) is the *motion planning problem* (MPP) for nonholonomic systems.[2] The extended system (2) clearly has more trajectories than (1). So, if we want to construct paths that satisfy some extra conditions (e.g., the solution of MPP), it is easier to do it for (2) than for (1). In the extreme case when system (1) satisfies the LARC, every smooth curve $\gamma : [0, T] \to M$ is a trajectory of a suitably chosen extended system (2). (To see this, take any compact subset K of M that contains γ in its interior. Then, by the LARC, in principle we can take r large enough such that the set of admissible directions $\{f_1, \ldots, f_r\}$ at each point $x \in K$ is simply the set of all possible directions. Then by the span condition, $\dot{\gamma}(t)$ can be written as a linear combination of $f_1(\gamma(t)), \ldots, f_r(\gamma(t))$.) If we could somehow produce for any given trajectory γ of (2) a sequence $\{\gamma^j\}$ of trajectories of (1) that converges to γ uniformly as $j \to \infty$, then we would have solved (for systems satisfying the LARC) the problem of approximating any given smooth curve by admissible trajectories. On the other hand, the LARC is a natural condition for these kinds of problems. For example, if system (1) is analytic and M is connected, the LARC is equivalent to the notion of *complete controllability*, which is the property that any two points in M can be joined by a trajectory of (1). Suppose that, in addition, M is such that it is easy to find a smooth curve in M that joins any two given points p, q. (This happens, for instance, if M is Euclidean space $\mathbb{R}^n$—in which case we can always use a straight-line segment—or if $M = \mathbb{R}^n - C$, where the "obstacle" C is a closed set, provided that we can somehow solve the problem of "path finding with obstacle avoidance," i.e., that we can find for any two points p, q a path that goes from p to q and avoids C.) In that case we will have solved the problem of producing an admissible trajectory that approximately steers p to q. This leads to an alternative approach to the MPP for nonholonomic systems studied by many authors, e.g., Brockett and Dai [2]; Fernandes, Gurvits, and Li [4]; Gurvits and Li [5]; Lafferriere and Sussmann [8]; Jacobs and Canny [7]; Laumond [9]; Murray and Sastry [12], [11]; Sastry and Li [14]; and Sussmann and Liu [17].

Our strategy for constructing the approximation sequences relies on the algebraic formalisms and the convergence results given in [19]. The key point for the algebraic formulations in [19] is the reformulation of the problem of convergence of trajectories in terms of convergence of inputs. This is done by introducing the concept of extended inputs. More precisely, let $X_1, \ldots, X_m$ be noncommuting *in-*

[1]System (1) is said to satisfy the Lie algebra rank condition if for every $x \in M$ we have $\Lambda(x) = T_xM$, where Λ is the Lie algebra of vector fields generated by the f_k, $\Lambda(x) = \{X(x) : X \in \Lambda\}$, and T_xM is the tangent space of M at x.

[2]The MPP for system (1) is the problem of finding an input $u \in L^1([0, T], \mathbb{R}^m)$ that steers p to q for any prescribed points p and q in the state space.

determinates. We define an *extended input* to be an integrable function on $[0,T]$ whose values are linear combinations of $X_1,\dots,X_m$ and various Lie brackets such as $[X_1,X_2]$, $[X_1,\,[X_1,X_2]]$, etc. For example, an expression (for $m=2$) like $v = v_1X_1 + v_2X_2 + v_3[X_1,X_2] + v_4[X_2,[X_1,X_2]]$ with integrable coefficients v_k on $[0,T]$ is an extended input. Clearly an ordinary input $u=(u_1,\dots,u_m)$ can be regarded as an extended input by identifying it with $u = u_1X_1 + \cdots + u_mX_m$, i.e., a linear combination of $X_1,\dots,X_m$ only without higher-order Lie brackets. The point is that extended inputs can be plugged into any system (1) exactly the same as ordinary inputs can, giving rise to ordinary differential equations. (This requires only that the vector fields $f_1,\dots,f_m$ be smooth enough so that the various Lie brackets exist.) For example, the extended input above gives rise to the following differential equation: $\dot{x} = v_1(t)f_1(x) + v_2(t)f_2(x) + v_3(t)[f_1,f_2](x) + v_4(t)[f_2,[f_1,f_2]](x)$. Therefore, we can talk about trajectories of extended inputs once the vector fields $f_1,\dots,f_m$ are known. We say that a sequence $\{u^j\}$ of ordinary inputs converges to an extended input u^∞ if for every initial condition $x(0)=p$, every choice of sufficiently smooth vector fields $f_1,\dots,f_m$, the trajectories x^j generated by the u^j converge uniformly to the trajectory x^∞ generated by u^∞. In [19] we have presented various sufficient conditions for a sequence of ordinary inputs to converge to an extended input. In this paper we show how to use those conditions to solve the inverse problem of producing, for a given extended input v, a sequence $\{u^j\}$ of ordinary inputs that converges to v.

The basic idea for solving the inverse problem is to use *highly oscillatory sequences*. (For short, HOSs. The precise definition of HOSs will be given later.) Applying the convergence results in [19], we can explicitly compute the limiting extended inputs of certain HOSs. An HOS used in our approximation algorithm involves several arbitrary functions η_ω, and the limiting extended input turns out to be given by an explicit formula in terms of these η_ω. We then show that this formula implies that any prescribed extended input can be achieved by a suitable choice of the η_ω.

An outline of this paper is as follows. Section 2 introduces some algebraic machinery that is needed throughout this paper. We start by reviewing the algebraic formalism developed in [19] and making it more appropriate for our needs here. We then state a convergence theorem proved in [19] which will be used to establish the approximation algorithm. Beginning in section 3, we proceed to solve the problem of finding a sequence of ordinary inputs that converges to a prescribed extended input. We first examine the limiting behavior of some special sequences of ordinary inputs. This leads naturally to the definition of some functions associated with brackets in a *P. Hall basis* of a free Lie algebra. These functions have nice algebraic properties and are naturally related to the *Chen–Fliess* product expansions of formal trajectories of some HOSs. The approximation algorithm is described in section 5, and its proof is given in sections 6 and 7. In section 8, we first present some examples that show how the algorithm is applied and then discuss briefly the feedback control laws and the case in which systems have a drift term.

2. Algebraic preliminaries and a convergence result. In this section we present briefly some algebraic preliminaries and state a convergence result that is needed later. For a more detailed treatment of the materials presented in this section, we refer to [19]. We follow the notations and definitions in [19].

2.1. Algebraic preliminaries. As in [19], let $\mathbf{X} = \{X_1,\dots,X_m\}$ be a finite sequence of objects that will be called *indeterminates*. We let $A(\mathbf{X})$ denote the free associative algebra generated over $\mathbb{R}$ by $\mathbf{X}$. For any multi-index $I=(i_1,\dots,i_k)$ with $i_1,\dots,i_k \in \{1,\dots,m\}$, we let $X_I = X_{i_1}\cdots X_{i_k}$. (There is a special multi-index $I=\emptyset$.

It is understood that $X_\emptyset = 1$.) Then $A(\mathbf{X})$ is the set of all sums $\sum_I a_I X_I$, where the coefficients a_I are real numbers, the summation runs over all possible multi-indices I, and all but finitely many a_I vanish. Therefore, the *monomials* X_I form a basis of $A(\mathbf{X})$, and every element of $A(\mathbf{X})$ is a finite linear combination of the X_I.

We also consider the algebra $\hat{A}(\mathbf{X})$ of all formal power series in $\mathbf{X}$. The elements of $\hat{A}(\mathbf{X})$ are the formal sums $\sum_I a_I X_I$, where I ranges over all multi-indices. This is the sum as above except that the a_I are no longer required to vanish for all but finitely many I. In both $A(\mathbf{X})$ and $\hat{A}(\mathbf{X})$, addition is done componentwise, and multiplication is carried out using the formula $X_I X_J = X_{IJ}$, where IJ is the concatenation of I and J, namely the multi-index obtained by writing, in order, first the components of I and then those of J.

For any integer $r \geq 0$, we use $A^r(\mathbf{X})$ to denote the *free nilpotent associative algebra of step* $r+1$ in the indeterminates $\mathbf{X}$. Therefore, $A^r(\mathbf{X})$ is defined like $\hat{A}(\mathbf{X})$, except that now all the monomials X_I with $|I| > r$ are set equal to zero. (Here $|I|$ is the length of I, i.e., $|I| = k$ if $I = (i_1, \ldots, i_k)$.) Clearly $A^r(\mathbf{X})$ can be thought of as the quotient of $A(\mathbf{X})$ or $\hat{A}(\mathbf{X})$ modulo the two side ideal of all sums of monomials of degree strictly larger than r. (The degree of a monomial X_I is $|I|$.) The canonical projection $\mathbf{Tr}(r)$ from $\hat{A}(\mathbf{X})$ to $A^r(\mathbf{X})$ is the operator that assigns to each series $S \in \hat{A}(\mathbf{X})$ the finite series $\mathbf{Tr}(r)(S)$ obtained from S by deleting all the terms of degree $> r$. (The symbol **Tr** comes from the word "truncation." It is used to indicate that the map $\mathbf{Tr}(r) : \hat{A}(\mathbf{X}) \to A^r(\mathbf{X})$ is in essence a truncation map, i.e., for any $S \in \hat{A}(\mathbf{X})$, $\mathbf{Tr}(r)(S)$ is the truncation of S "up to order r.") The kernel of $\mathbf{Tr}(r)$ is denoted by $\hat{A}_r(\mathbf{X})$. In particular, $\hat{A}_0(\mathbf{X})$ is the set of all formal power series $\sum_I a_I X_I$ for which $a_\emptyset = 0$. The exponential map is a well-defined bijection $\exp : \hat{A}_0(\mathbf{X}) \to 1 + \hat{A}_0(\mathbf{X})$, whose inverse is a map from $1+\hat{A}_0(\mathbf{X})$ to $\hat{A}_0(\mathbf{X})$ denoted by "log." (Here $1+\hat{A}_0(\mathbf{X})$ is the subset of $\hat{A}(\mathbf{X})$ that contains all the elements S such that $S-1 \in \hat{A}_0(\mathbf{X})$.) If $S \in \hat{A}_0(\mathbf{X})$, then $\exp(S)$ and $\log(1+S)$ are given by the usual series $\exp(S) = \sum_{n=0}^{\infty} \frac{S^n}{n!}$, $\log(1+S) = \sum_{n=1}^{\infty} \frac{(-1)^{(n-1)} S^n}{n}$. The algebras $A(\mathbf{X}), \hat{A}(\mathbf{X}), A^r(\mathbf{X})$ are Lie algebras in the usual way. We let $L(\mathbf{X})$ denote the Lie subalgebra of $A(\mathbf{X})$ generated by the indeterminates $X_1, \ldots, X_m$. An element S of $A(\mathbf{X})$ will be said to be a *Lie element* if $S \in L(\mathbf{X})$. It is clear that an $S \in A(\mathbf{X})$ is a Lie element iff all the *homogeneous* components of S are Lie elements. (An $S \in \hat{A}(\mathbf{X})$ is homogeneous if it is a linear combination of monomials with equal degree.)

For $\mathbf{X} = \{X_1, \ldots, X_m\}$, we let $\mathcal{FB}r(\mathbf{X})$ denote the set of *formal brackets*. The elements of $\mathcal{FB}r(\mathbf{X})$ are purely formal expressions in the indeterminates X_i, the left and the right brackets, and the commas. Precisely, cf. [16], let $\mathcal{A}$ be the alphabet that consists of the X_i, the left and the right brackets, and the commas. Then $\mathcal{FB}r(\mathbf{X})$ is the smallest set Ω of words in $\mathcal{A}$ that contains $X_1, \ldots, X_m$ and has the property that whenever α, β are two words that belong to Ω, then the word "$[\alpha, \beta]$" also belongs to Ω. Every formal bracket $B \in \mathcal{FB}r(\mathbf{X})$ has a well-defined *degree* $\delta(B)$, which is an integer ≥ 1. (The degree $\delta(B)$ of $B \in \mathcal{FB}r(\mathbf{X})$ is inductively defined as follows: $\delta(X_i) = 1, i = 1, \ldots, m$, and $\delta([\alpha, \beta]) = \delta(\alpha) + \delta(\beta)$.) If degree $\delta(B) > 1$, then B can be written in a unique way as $[B_1, B_2]$ with $B_1, B_2 \in \mathcal{FB}r(\mathbf{X})$. The formal brackets B_1, B_2 are called the *left* and the *right* factors of B, respectively. There is a natural mapping μ which associates with each $B \in \mathcal{FB}r(\mathbf{X})$ an element of $L(\mathbf{X})$. The elements of $L(\mathbf{X})$ of the form $\mu(B), B \in \mathcal{FB}r(\mathbf{X})$, are called *Lie monomials*. They will also be referred to as Lie brackets in $L(\mathbf{X})$ in the indeterminates $X_1, \ldots, X_m$. The Lie algebra $L(\mathbf{X})$ is spanned by all the Lie brackets of $X_1, \ldots, X_m$. Naturally Lie brackets in the indeterminates $X_1, \ldots, X_m$ are not linearly independent. There are

several systematic procedures for singling out a basis of $L(\mathbf{X})$. We will come back to this later.

Remark 2.1. In the following we will use $\mathcal{B}r(\mathbf{X})$ to denote the set of Lie monomials in $L(\mathbf{X})$, i.e., the set $\{\mu(B) \ : \ B \in \mathcal{FB}r(\mathbf{X})\}$ in $L(\mathbf{X})$. For simplicity, we will drop the symbol μ and just use B to denote a Lie bracket in $\mathcal{B}r(\mathbf{X})$. Each bracket $B \in \mathcal{B}r(\mathbf{X})$ has a well-defined *degree* $\delta(B)$. For later use, we let $\delta_i(B), i = 1, \ldots, m$, be the *degree* of B in X_i, so $\delta_1(B) + \cdots + \delta_m(B) = \delta(B)$. For example if $B = [X_1, [X_1, X_2]]$, then $\delta_1(B) = 2, \delta_2(B) = 1$, and $\delta(B) = 3$.

We define $\hat{L}(\mathbf{X})$ to be the set of all those elements of $\hat{A}(\mathbf{X})$ whose components are Lie elements. The elements of $\hat{L}(\mathbf{X})$ are called a *Lie series* in $X_1, \ldots, X_m$. Let $\hat{G}(\mathbf{X}) = \{\exp(Z), Z \in \hat{L}(\mathbf{X})\}$, the set of exponentials of the elements of $\hat{L}(\mathbf{X})$. The Campbell–Hausdorff formula implies (cf., e.g., [1]) that $\hat{G}(\mathbf{X})$ is a group under the operation of multiplication in $\hat{A}(\mathbf{X})$. The elements of $\hat{G}(\mathbf{X})$ are called an *exponential Lie series*. If we let $L^r(\mathbf{X})$ be the Lie subalgebra of $A^r(\mathbf{X})$ generated by the $\mathbf{X}$ and define $G^r(\mathbf{X})$ to be the subset of $A^r(\mathbf{X})$ consisting of all the exponentials of elements of $L^r(\mathbf{X})$, then $L^r(\mathbf{X})$ is a finite-dimensional Lie algebra and $G^r(\mathbf{X})$ is its corresponding simply connected Lie group.

Let $v(t) = \sum_I v_I(t) X_I$ be a function on $[0, T]$ with values in $\hat{A}(\mathbf{X})$. We say that v is *integrable* if all the functions v_I are in $L^1[0, T]$. A *polynomial input* is an integrable function $\mathbf{v}$ on an interval $[0, T]$ with values in $\hat{A}_0(\mathbf{X})$. An *extended input* is a polynomial input which is $\hat{L}(\mathbf{X})$ valued. An ordinary input $u = (u_1, \ldots, u_m) \in L^1([0, T], \mathbb{R}^m)$ can be regarded as a polynomial input by identifying it with the function $\mathbf{u} = u_1 X_1 + \cdots + u_m X_m$. It is an extended input in fact by the above definition. In most cases we will make no difference between $u = (u_1, \ldots, u_m)$ and $\mathbf{u} = u_1 X_1 + \cdots + u_m X_m$ and call both ordinary inputs.

To each polynomial input $\mathbf{v}$, the *Chen–Fliess series* $S_{\mathbf{v}}$ determined by $\mathbf{v}$ is the absolutely continuous $\hat{A}(\mathbf{X})$-valued function on $[0, T]$ that satisfies

$$\dot{S} = S\mathbf{v}, \quad S(0) = 1, \quad S(t) \in \hat{A}(\mathbf{X}). \tag{3}$$

So if $S_{\mathbf{v}}$ is the solution of (3), then it is clear that

$$S_{\mathbf{v}} = 1 + \sum_{k=1}^{\infty} \int_0^t \int_0^{t_1} \cdots \int_0^{t_{k-1}} \mathbf{v}(t_k)\mathbf{v}(t_{k-1}) \cdots \mathbf{v}(t_1)\, dt_k \cdots dt_1.$$

The Chen–Fliess series $S_{\mathbf{v}}$ will be called the *formal trajectory* of $\mathbf{v}$.

By definition, the function $t \to S_{\mathbf{v}}(t)$ is an absolutely continuous $\hat{A}(\mathbf{X})$-valued function. It is in fact $1 + \hat{A}_0(\mathbf{X})$-valued. Conversely, define a *formal trajectory* to be an absolutely continuous $1 + \hat{A}_0(\mathbf{X})$-valued function S on $[0, T]$. Then every formal trajectory S is the formal trajectory of a polynomial input $\mathbf{v}$ given by $\mathbf{v} = S^{-1}\dot{S}$. Therefore the map $\mathbf{v} \to S_{\mathbf{v}}$ is a one-to-one correspondence between the set of polynomial inputs and that of formal trajectories whose inverse is given by $\mathbf{v} = S_{\mathbf{v}}^{-1}\dot{S}_{\mathbf{v}}$.

The polynomial input $\mathbf{v}$ can be computed from $S_{\mathbf{v}}$ using standard algebraic tools. If $S_{\mathbf{v}} = 1 + \sum_{|I|>0} H_I X_I$, then we know that

$$S_{\mathbf{v}}^{-1} = 1 + \sum_{|I|>0} \left(\sum_{k=1}^{\infty} (-1)^k \sum_{J_1 \ldots J_k = I} H_{J_1} \cdots H_{J_k} \right) X_I.$$

We have

$$S_{\mathbf{v}}^{-1}\dot{S}_{\mathbf{v}} = \sum_{|I|>0}\left(\dot{H}_I + \sum_{k=1}^{\infty}(-1)^k \sum_{J_1\dots J_k J_{k+1}=I} H_{J_1}\cdots H_{J_k}\dot{H}_{J_{k+1}}\right) X_I,$$

where the inner summation above runs over all ways of expressing the multi-index I as a concatenation $J_1\cdots J_kJ_{k+1}$ of $k+1$ indices. So, if we let $\mathbf{v} = S_{\mathbf{v}}^{-1}\dot{S}_{\mathbf{v}} = \sum_{|I|>0} v_I X_I$, then the v_I are given by

$$v_I = \dot{H}_I + \sum_{k=1}^{\infty}(-1)^k \sum_{J_1\cdots J_k J_{k+1}=I} H_{J_1}H_{J_2}\cdots H_{J_k}\dot{H}_{J_{k+1}}. \tag{4}$$

Remark 2.2. It follows from [15] that $S_{\mathbf{v}}$ is $\hat{G}(\mathbf{X})$-valued if $\mathbf{v}$ is an extended input. The converse of this is also true; i.e., if S is a $\hat{G}(\mathbf{X})$-valued formal trajectory, then $\mathbf{v} = S^{-1}\dot{S}$ is an extended input. To see this, it suffices to prove that in this case the function $S^{-1}\dot{S}$ is $\hat{L}(\mathbf{X})$ valued. Notice that $S^{-1}(t)\dot{S}(t) = \lim_{h\to 0}\frac{1}{h}S^{-1}(t)(S(t+h) - S(t)) = \lim_{h\to 0}\frac{1}{h}(S^{-1}(t)S(t+h) - 1)$. Using the Campbell–Hausdorff formula we conclude that $S^{-1}(t)S(t+h) = e^{\Lambda(t,h)}$, where $\Lambda(t,h)$ is a Lie series that goes to zero as $h\to 0$. So $S^{-1}(t)\dot{S}(t) = \lim_{h\to 0}\frac{1}{h}(\Lambda(t,h) + \frac{1}{2}\Lambda(t,h)^2 + \cdots) = \lim_{h\to 0}\frac{\Lambda(t,h)}{h}$, which implies that $S^{-1}\dot{S}$ is Lie series valued.

Let π be the linear map of $A(\mathbf{X})$ onto $L(\mathbf{X})$ defined by $\pi(X_I) = \frac{1}{|I|}[X_I]$ for $|I| > 0$, where if $I = (i_1,\cdots,i_k)$, then $[X_I] \stackrel{\text{def}}{=} [X_{i_1},[X_{i_2},\dots,[X_{i_{k-1}},X_{i_k}]\cdots]]$. It is well known, cf., e.g., [1], that the restriction of π to $L(\mathbf{X})$ is the identity map; i.e., π is a projector of $A(\mathbf{X})$ onto $L(\mathbf{X})$. Let $\hat{\pi}$ be the linear projection map from $\hat{A}(\mathbf{X})$ to $\hat{L}(\mathbf{X})$ that extends π. From Remark 2.2 we know that if $S = 1 + \sum_{|I|>0} H_I X_I$ is a $\hat{G}(\mathbf{X})$-valued formal trajectory, $\mathbf{v} = S^{-1}\dot{S}$ is an extended input. In that case we have $\hat{\pi}(\mathbf{v}) = \mathbf{v}$, so

$$\mathbf{v} = \sum_{|I|>0}\frac{1}{|I|}v_I[X_I], \tag{5}$$

where the v_I are given by (4).

In the particular case where $\mathbf{v}$ is an ordinary input $\mathbf{u} = u_1X_1 + \cdots + u_mX_m$, the Chen–Fliess series $S_{\mathbf{u}}$ is given by the formula $S_{\mathbf{u}}(t) = 1 + \sum_{|I|>0} U_I(t)X_I$, where if $I = (i_1,\dots,i_k)$, then U_I is the iterated integral defined by

$$U_I(t) \stackrel{\text{def}}{=} \int_0^t u_{i_k}(t_k)\int_0^{t_k} u_{i_{k-1}}(t_{k-1})\cdots\int_0^{t_2} u_{i_1}(t_1)\,dt_1\cdots dt_k. \tag{6}$$

Let $\mathbf{u}$ be an ordinary input and $\mathbf{v}$ be a polynomial input. We define a *generalized difference* of $\mathbf{u}$ and $\mathbf{v}$ to be an absolutely continuous $\hat{A}_0(\mathbf{X})$-valued function W on $[0,T]$ that satisfies

$$\dot{W} = -\mathbf{u}W + \mathbf{v} - \mathbf{u}, \quad W(t)\in\hat{A}_0(\mathbf{X}). \tag{7}$$

Clearly a solution W of (7) is uniquely determined by its initial value $W(0)$.

Let W be a solution of (7) with initial condition $W(0) = W_0 \in \hat{A}_0(\mathbf{X})$. Then W satisfies the integral equation

$$W(t) = W_0 - \int_0^t \mathbf{u}(s)W(s)ds + \int_0^t (\mathbf{v}(s) - \mathbf{u}(s))\,ds. \tag{8}$$

If $\mathbf{u} = u_1X_1+\cdots+u_mX_m$ and $\mathbf{v} = \sum_{|I|>0} v_IX_I$, letting $(\mathbf{u} \overset{\text{g.d.}}{-} \mathbf{v})(t) = \sum_{|I|>0} \widetilde{UV}_I(t)X_I$ be a generalized difference of $\mathbf{u}$ and $\mathbf{v}$ with initial value $(\mathbf{u} \overset{\text{g.d.}}{-} \mathbf{v})(0) = \sum_{|I|>0} \hat{W}_IX_I$, then from (8) we have the recursive formulas

$$\widetilde{UV}_i(t) = \hat{W}_i + \int_0^t v_i(s)ds - \int_0^t u_i(s)ds, \tag{9}$$

$$\widetilde{UV}_{i_1,\dots,i_k}(t) = \hat{W}_{i_1,\dots,i_k} + \int_0^t v_{i_1,\dots,i_k}(s) - \int_0^t u_{i_1}(s)\widetilde{UV}_{i_2,\dots,i_k}(s)ds. \tag{10}$$

If we let $\widetilde{uv}_I = \dot{\widetilde{UV}}_I$, from the above we see that the $\widetilde{uv}_I$ satisfy $\widetilde{uv}_i(t) = v_i(t) - u_i(t)$ and $\widetilde{uv}_{i_1,\dots,i_k}(t) = v_{i_1,\dots,i_k}(t) - u_{i_1}(t)\widetilde{UV}_{i_2,\dots,i_k}(t)$.

The concepts of polynomial inputs, extended inputs, formal trajectories, generalized differences, etc., have truncated analogues. Let us say that a polynomial input $\mathbf{v}$ has *order* $\le r$ if the values of $\mathbf{v}$ are linear combinations of monomials of degree $\le r$. The smallest such r is called the *order* of $\mathbf{v}$. We say that $\mathbf{v}$ is of *finite order* if it has order r for some integer $r > 0$.

If $\mathbf{v}$ is a polynomial input of order $\le r$, we can regard $\mathbf{v}$ as an $A_0^r(\mathbf{X})$-valued rather than $\hat{A}_0(\mathbf{X})$-valued function, and we define the *rth-order truncated formal trajectory* determined by $\mathbf{v}$ to be the solution of the initial value problem $\dot{S} = S\mathbf{v}$, $S(0) = 1$, $S(t) \in A^r(\mathbf{X})$ on $[0,T]$ (here $A_0^r(\mathbf{X}) = A^r(\mathbf{X}) \cap A_0(\mathbf{X})$). We will use $S_{\mathbf{v}}^r$ to denote the rth-order truncated formal trajectory determined by $\mathbf{v}$ in $A^r(\mathbf{X})$.

If $\mathbf{u}$ is an ordinary input and $\mathbf{v}$ is a polynomial input of order $\le r$, we can also define an *rth-order truncated generalized difference* of $\mathbf{u}$ and $\mathbf{v}$, which is an absolutely continuous $A_0^r(\mathbf{X})$-valued function on $[0,T]$ that satisfies

$$\dot{W} = -\mathbf{u}W + \mathbf{v} - \mathbf{u}, \quad W(t) \in A_0^r(\mathbf{X}). \tag{11}$$

We will use the notation $(\mathbf{u} \overset{\text{g.d.(r)}}{-} \mathbf{v})$ to denote an rth-order truncated generalized difference of $\mathbf{u}$ and $\mathbf{v}$. It is uniquely determined by $(\mathbf{u} \overset{\text{g.d.(r)}}{-} \mathbf{v})(0)$. Let $\mathbf{u} = u_1X_1 + \cdots + u_mX_m$ and $\mathbf{v} = \sum_{0<|I|\le r} v_IX_I$. Let $(\mathbf{u} \overset{\text{g.d.(r)}}{-} \mathbf{v})(t) = \sum_{0<|I|\le r} \widetilde{UV}_I(t)X_I$ be an rth-order truncated generalized difference of $\mathbf{u}$ and $\mathbf{v}$. Then the $\widetilde{UV}_I, 0 < |I| \le r$, can still be calculated by (9) and (10) once the initial values $\widetilde{UV}_I(0)$ are known.

For more general results and some of the generalizations, we refer to [20], [21].

2.2. A convergence theorem. Now we state a convergence theorem from [19] that is needed later.

Let $\mathbf{v} = \sum_{0<|I|\le r} v_IX_I$ be an extended input of order $\le r$. Then we know that $\mathbf{v} = \sum_{0<|I|\le r} \frac{v_I}{|I|}[X_I]$. As was said in the introduction, $\mathbf{v}$ can be plugged into any system (1) if the vector fields $f_1,\dots,f_m$ are of class C^{r-1}. The result is a time-varying ordinary differential equation $\dot{x} = \sum_{0<|I|\le r} \frac{v_I(t)}{|I|}[f_I](x)$, where we write $[f_I] \overset{\text{def}}{=} [f_{i_1},[f_{i_2},[\cdots,[f_{i_{k-1}},f_{i_k}]\cdots]]]$ for $I = (i_1,\dots,i_k)$.

Let $\mathbf{v} = \sum_{0<|I|\le r} v_IX_I$ be an extended input of order $\le r$. A sequence $\{\mathbf{u}^j\}$ of ordinary inputs *EI(r)-converges* to $\mathbf{v}$ if the following condition is satisfied:

For every integer $n > 0$, every point $p \in \mathbb{R}^n$, any sequence $\{p^j\} \subseteq \mathbb{R}^n$ that converges to p, and any vector fields $f_1,\dots,f_m$ of class C^{r-1} on $\mathbb{R}^n$, if the initial

value problem

$$\dot{x} = \sum_{0<|I|\le r} \frac{v_I(t)}{|I|}[f_I](x),\ x(0) = p,$$

has a unique solution x^∞ which is defined on $[0,T]$ and if x^j is a maximal solution of the initial value problem

$$\dot{x} = \sum_{k=1}^{m} u_k^j(t) f_k(x),\ x(0) = p^j,$$

then the x^j are defined on $[0,T]$ for j large enough and converge uniformly to x^∞ on $[0,T]$ as $j \to \infty$.

With these preliminaries, we have the following convergence theorem from [19].

THEOREM 2.1. *Let r be a positive integer. Let $\{\mathbf{u}^j = (u_1^j, \dots, u_m^j)\} \subseteq L^1([0,T], \mathbb{R}^m)$ be a sequence of ordinary inputs. Let $\mathbf{v} = \sum_{0<|I|\le r} v_I X_I$ be a polynomial input of order $\le r$. Assume that there exist a sequence $\{\mathbf{v}^j = \sum_{0<|I|\le r} v_I^j X_I\}$ of polynomial inputs of order $\le r$ and rth-order truncated generalized differences $(\mathbf{u}^j \overset{\text{g.d.(r)}}{-} \mathbf{v}^j) = \sum_{0<|I|\le r} \widetilde{UV}_I^j X_I$ of $\mathbf{u}^j$ and $\mathbf{v}^j$ such that*

c1(r). *the indefinite integrals $\int_0^t v_I^j(s)ds$ converge to $\int_0^t v_I(s)ds$ uniformly on $[0,T]$ as $j \to \infty$ for all $0 < |I| \le r$,*

c2(r). *the $\widetilde{UV}_I^j$ converge to 0 uniformly as $j \to \infty$ for $0 < |I| \le r$,*

c3(r). *the L^1 norms of the $\widetilde{uv}_I^j = \dot{\widetilde{UV}}_I^j$ for $|I| = r$ and of the v_I^j for $0 < |I| \le r$ are uniformly bounded.*

Then

C1. $\mathbf{v}$ *is an extended input of order $\le r$;*

C2. *the $\mathbf{u}^j$ EI(r)-converge to $\mathbf{v}$.*

Remark 2.3. It is shown in [19] that if a sequence $\{\mathbf{u}^j\}$ of ordinary inputs EI(r)-converges to an extended input $\mathbf{v}$ of order r, then the $S^r_{\mathbf{u}^j}$ converge to $S^r_{\mathbf{v}}$ uniformly; i.e., if we write $S^r_{\mathbf{u}^j} = 1 + \sum_{0<|I|\le r} U_I^j X_I$ and $S^r_{\mathbf{v}} = 1 + \sum_{0<|I|\le r} H_I V_I$, then the U_I^j converge to H_I uniformly for all $0 < |I| \le r$.

Remark 2.4. In practice, we just know the sequence $\{\mathbf{u}^j\}$. We do not know a priori if it is convergent or what its limit $\mathbf{v}$ is even if it converges. We have to find sequences $\{\mathbf{v}^j\}$ and $(\mathbf{u}^j \overset{\text{g.d.(r)}}{-} \mathbf{v}^j)$ so that c1(r), c2(r), and c3(r) are satisfied for some r. From (9) and (10) we see that

$$\widetilde{UV}_i^j(t) = \widetilde{UV}_i^j(0) + \int_0^t v_i^j(s)ds - \int_0^t u_i^j(s)ds,$$

$$\widetilde{UV}_{i_1,\dots,i_k}^j(t) = \widetilde{UV}_{i_1,\dots,i_k}^j(0) + \int_0^t v_{i_1,\dots,i_k}^j(s) - \int_0^t u_{i_1}^j(s)\widetilde{UV}_{i_2,\dots,i_k}^j(s)ds.$$

If we let $V_I^j(t) = \widetilde{UV}_I^j(0) + \int_0^t v_I^j(s)ds$, then we can rewrite the above as

$$\widetilde{UV}_i^j(t) = V_i^j(t) - \int_0^t u_i^j(s)ds, \tag{12}$$

$$\widetilde{UV}_{i_1,\dots,i_k}^j(t) = V_{i_1,\dots,i_k}^j(t) - \int_0^t u_{i_1}^j(s)\widetilde{UV}_{i_2,\dots,i_k}^j(s)ds. \tag{13}$$

If we can find absolutely continuous functions V_I^j for $0 < |I| \le r$ such that

c1′(r). the V_I^j converge to absolutely continuous functions V_I uniformly for all $0 < |I| \le r$,

c2′(r). the $\widetilde{UV}_I^j$ determined by (12) and (13) converge to 0 uniformly for $0 < |I| \le r$,

c3′(r). the L^1 norms of the $\dot{\widetilde{UV}}_I^j$ for $|I| = r$ and of the $\dot{V}_I^j$ for $0 < |I| \le r$ are uniformly bounded,

then the conclusions C1 and C2 in Theorem 2.1 hold, and the EI(r)-limit of the $\mathbf{u}^j$ is equal to $\mathbf{v} = \sum_{0<|I|\le r} \dot{V}_I(t) X_I$.

To see an explicit example of how Theorem 2.1 is used, see Example 3.1 in the next section.

3. Approximate tracking: A simple example. Theorem 2.1 can be applied to the problem of approximating a given trajectory of an extended system. Our final goal is, for any given extended input $\mathbf{v}$ of finite order, find a sequence $\{\mathbf{u}^j\}$ of ordinary inputs that converges to $\mathbf{v}$. Before we describe how this can be done, we present some simple examples to show the basic ideas of how Theorem 2.1 can be applied to compute the limit of sequences of some highly oscillatory ordinary inputs.

Example 3.1. Let

$$u_1^j(t) = \eta_1(t) + j^{\frac{2}{3}} \cos \omega_1 jt, \qquad u_2^j(t) = \eta_2(t) + j^{\frac{2}{3}} \eta_3(t) \cos \omega_2 jt,$$

where η_i are functions of class C^1 on $[0, T]$, ω_1, ω_2 are nonzero numbers such that $2\omega_1 + \omega_2 = 0$. Using Theorem 2.1 we can show that $\{\mathbf{u}^j\}$ is convergent to some $\mathbf{v}$, and we can find $\mathbf{v}$ explicitly.

As the first step, we let

$$U_1^j(t) = \int_0^t u_1^j(s)ds = \int_0^t \eta_1(s)\, ds + \frac{j^{-\frac{1}{3}}}{\omega_1} \sin \omega_1 jt,$$
$$U_2^j(t) = \int_0^t u_2^j(s)ds = \int_0^t \eta_2(s)\, ds + \frac{j^{-\frac{1}{3}}}{\omega_2} \eta_3(t) \sin \omega_2 jt - \frac{j^{-\frac{1}{3}}}{\omega_2} \int_0^t \eta_3'(s) \sin \omega_2 js\, ds.$$

Letting

$$V_1^j(t) = \int_0^t \eta_1(s)\, ds, \qquad V_2^j(t) = \int_0^t \eta_2(s)\, ds - \frac{j^{-\frac{1}{3}}}{\omega_2} \int_0^t \eta_3'(s) \sin \omega_2 js\, ds,$$

we get

$$\widetilde{UV}_1^j(t) = -\frac{j^{-\frac{1}{3}}}{\omega_1} \sin \omega_1 jt, \qquad \widetilde{UV}_2^j(t) = -\frac{j^{-\frac{1}{3}}}{\omega_2} \eta_3(t) \sin \omega_2 jt.$$

Now it is easily computed that

$$\int_0^t u_1^j(s) \widetilde{UV}_1^j(s)\, ds = \int_0^t \eta_1(s) \widetilde{UV}_1^j(s) ds + \frac{j^{-\frac{2}{3}}}{4\omega_1^2} \left[\cos 2\omega_1 jt - 1\right],$$
$$\int_0^t u_1^j(s) \widetilde{UV}_2^j(s)\, ds = \int_0^t \eta_1(s) \widetilde{UV}_2^j(s) ds + \frac{j^{-\frac{2}{3}} \eta_3(t)}{2\omega_2} \left[\frac{\cos(\omega_1 + \omega_2)jt}{\omega_1 + \omega_2} + \frac{\cos(\omega_2 - \omega_1)jt}{\omega_2 - \omega_1}\right]$$

$$- \frac{j^{-\frac{2}{3}} \eta_3(0)}{2\omega_2} \left[\frac{1}{\omega_1 + \omega_2} + \frac{1}{\omega_2 - \omega_1}\right] - \frac{j^{-\frac{2}{3}}}{2\omega_2} \int_0^t \eta_3'(s) \left[\frac{\cos(\omega_1 + \omega_2)js}{\omega_1 + \omega_2} + \frac{\cos(\omega_2 - \omega_1)js}{\omega_2 - \omega_1}\right] ds,$$

$$\int_0^t u_2^j(s)\widetilde{UV}_1^j(s)\,ds = \int_0^t \eta_2(s)\widetilde{UV}_1^j(s)ds + \frac{j^{-\frac{2}{3}}\eta_3(t)}{2\omega_1}\left[\frac{\cos(\omega_1+\omega_2)jt}{\omega_1+\omega_2} + \frac{\cos(\omega_1-\omega_2)jt}{\omega_1-\omega_2}\right]$$

$$-\frac{j^{-\frac{2}{3}}\eta_3(0)}{2\omega_1}\left[\frac{1}{\omega_1+\omega_2} + \frac{1}{\omega_1-\omega_2}\right] - \frac{j^{-\frac{2}{3}}}{2\omega_1}\int_0^t \eta_3'(s)\left[\frac{\cos(\omega_1+\omega_2)js}{\omega_1+\omega_2} + \frac{\cos(\omega_1-\omega_2)js}{\omega_1-\omega_2}\right] ds,$$

$$\int_0^t u_2^j(s)\widetilde{UV}_2^j(s)\,ds = \int_0^t \eta_2(s)\widetilde{UV}_2^j(s)ds + \frac{j^{-\frac{2}{3}}}{4\omega_2^2}\left[\eta_3^2(t)\cos 2\omega_2 jt - \eta_3^2(0)\right]$$

$$-\frac{j^{-\frac{2}{3}}}{2\omega_2^2}\int_0^t \eta_3(s)\eta_3'(s)\cos 2\omega_2 js\,ds.$$

So if we let

$$\begin{aligned}
V_{1,1}^j(t) &= \int_0^t \eta_1(s)\widetilde{UV}_1^j(s)ds - \frac{j^{-\frac{2}{3}}}{4\omega_1^2},\\
V_{1,2}^j(t) &= \int_0^t \eta_1(s)\widetilde{UV}_2^j(s)ds - \frac{j^{-\frac{2}{3}}\eta_3(0)}{2\omega_2}\left[\frac{1}{\omega_1+\omega_2} + \frac{1}{\omega_2-\omega_1}\right]\\
&\quad -\frac{j^{-\frac{2}{3}}}{2\omega_2}\int_0^t \eta_3'(s)\left[\frac{\cos(\omega_1+\omega_2)js}{\omega_1+\omega_2} + \frac{\cos(\omega_2-\omega_1)js}{\omega_2-\omega_1}\right] ds,\\
V_{2,1}^j(t) &= \int_0^t \eta_2(s)\widetilde{UV}_1^j(s)ds - \frac{j^{-\frac{2}{3}}\eta_3(0)}{2\omega_1}\left[\frac{1}{\omega_1+\omega_2} + \frac{1}{\omega_1-\omega_2}\right]\\
&\quad -\frac{j^{-\frac{2}{3}}}{2\omega_1}\int_0^t \eta_3'(s)\left[\frac{\cos(\omega_1+\omega_2)js}{\omega_1+\omega_2} + \frac{\cos(\omega_1-\omega_2)js}{\omega_1-\omega_2}\right] ds,\\
V_{2,2}^j(t) &= \int_0^t \eta_2(s)\widetilde{UV}_2^j(s)ds - \frac{j^{-\frac{2}{3}}\eta_3^2(0)}{4\omega_2^2} - \frac{j^{-\frac{2}{3}}}{2\omega_2^2}\int_0^t \eta_3(s)\eta_3'(s)\cos 2\omega_2 js\,ds,
\end{aligned}$$

then we get

$$\begin{aligned}
\widetilde{UV}_{1,1}^j(t) &= -\frac{j^{-\frac{2}{3}}}{4\omega_1^2}\cos 2\omega_1 jt,\\
\widetilde{UV}_{1,2}^j(t) &= -\frac{j^{-\frac{2}{3}}\eta_3(t)}{2\omega_2}\left[\frac{\cos(\omega_1+\omega_2)jt}{\omega_1+\omega_2} + \frac{\cos(\omega_2-\omega_1)jt}{\omega_2-\omega_1}\right],\\
\widetilde{UV}_{2,1}^j(t) &= -\frac{j^{-\frac{2}{3}}\eta_3(t)}{2\omega_1}\left[\frac{\cos(\omega_1+\omega_2)jt}{\omega_1+\omega_2} + \frac{\cos(\omega_1-\omega_2)jt}{\omega_1-\omega_2}\right],\\
\widetilde{UV}_{2,2}^j(t) &= -\frac{j^{-\frac{2}{3}}\eta_3^2(t)}{4\omega_2^2}\cos 2\omega_2 jt.
\end{aligned}$$

There are eight indices of degree 3, namely, $(1,1,1)$, $(1,1,2)$, $(1,2,1)$, $(1,2,2)$, $(2,1,1)$, $(2,1,2)$, $(2,2,1)$, $(2,2,2)$. We simply let $V_{i_1,i_2,i_3}^j(t) = \int_0^t u_{i_1}^j(s)\widetilde{UV}_{i_2,i_3}^j(s)ds$, so that $\widetilde{UV}_{i_1,i_2,i_3}^j(t) \equiv 0$. The V_{i_1,i_2,i_3}^j can be computed explicitly. We just compute $V_{1,1,2}^j$.

By definition we have

$$\begin{aligned} V^j_{1,1,2}(t) &= \int_0^t \eta_1(s)\widetilde{UV}^j_{1,2}(s) \\ &\quad - \int_0^t \eta_3(s)\left[\frac{\cos\omega_1 js\cos(\omega_1+\omega_2)js}{2\omega_2(\omega_1+\omega_2)} + \frac{\cos\omega_1 js\cos(\omega_2-\omega_1)js}{2\omega_2(\omega_2-\omega_1)}\right]ds \\ &= -\int_0^t \frac{\eta_3(s)}{4\omega_2(\omega_1+\omega_2)}ds + o(1). \end{aligned}$$

In the above $o(1)$ denotes the terms that converge to 0 uniformly as $j\to\infty$. Also we have used the fact $2\omega_1+\omega_2=0$ to get the first term, since

$$\cos\omega_1 js\cos(\omega_1+\omega_2)js = \frac{1}{2}[\cos(2\omega_1+\omega_2)js + \cos(\omega_1-\omega_1-\omega_2)js] = \frac{1}{2}[1+\cos\omega_2 js].$$

Similarly, one can compute all the other $V^j_{i_1,i_2,i_3}$.

From the definition of the V^j_I we can get that

$$\lim_{j\to\infty} V^j_1(t) = \int_0^t \eta_1(s)\,ds, \quad \lim_{j\to\infty} V^j_2(t) = \int_0^t \eta_2(s)\,ds,$$

$$\lim_{j\to\infty} V^j_{1,1,2}(t) = -\int_0^t \frac{\eta_3(s)}{4\omega_2(\omega_1+\omega_2)}ds,$$

$$\lim_{j\to\infty} V^j_{1,2,1}(t) = -\int_0^t \frac{\eta_3(s)}{4\omega_1(\omega_1+\omega_2)}ds,$$

$$\lim_{j\to\infty} V^j_{2,1,1}(t) = -\int_0^t \frac{\eta_3(s)}{8\omega_1^2}ds,$$

and all the other V^j_I converge to 0 uniformly. It is also clear that if we let $v^j_I = \dot V^j_I$, then the $\|v^j_I\|_{L^1[0,T]}$ are uniformly bounded. Therefore if we let

$$\mathbf{v} = \eta_1(t)X_1 + \eta_2(t)X_2 - \frac{\eta_3(t)}{4\omega_2(\omega_1+\omega_2)}X_1X_1X_2 - \frac{\eta_3(t)}{4\omega_1(\omega_1+\omega_2)}X_1X_2X_1 - \frac{\eta_3(t)}{8\omega_1^2}X_2X_1X_1,$$

using the fact that $\omega_2=-2\omega_1$, and $[X_1,[X_1,X_2]] = X_1X_1X_2 - 2X_1X_2X_1 + X_2X_1X_1$, we get that $\mathbf{v} = \eta_1(t)X_1 + \eta_2(t)X_2 - \frac{\eta_3(t)}{8\omega_1^2}[X_1,[X_1,X_2]]$. From Theorem 2.1 we can conclude that the $\mathbf{u}^j$ EI(3)-converge to $\mathbf{v}$.

4. Some definitions and a special sequence of ordinary inputs. In this section we do some preliminary work for the general approximate tracking algorithm. We first introduce a number of new definitions and review some known facts about the free Lie algebra $L(\mathbf{X})$ generated by the indeterminates $\{X_1,\dots,X_m\}$ and about the *Chen–Fliess* product expansion of formal trajectories. Then we analyze one special sequence $\{u^j = (j^{\frac{m-1}{m}}e^{ij\omega_1 t},\dots,j^{\frac{m-1}{m}}e^{ij\omega_m t})\}$ of ordinary inputs in detail and show that $\{u^j\}$ is convergent to some extended input of finite order (here the ω_k are constants, $i=\sqrt{-1}$). The sequences of this form will be generalized later, and they play a crucial rule in our approximate tracking algorithm described in the next section.

4.1. P. Hall basis and Chen–Fliess product expansions of formal trajectories. Suppose that $\mathbf{v}$ is an rth-order extended input with coefficients of class C^1. Our goal is to write an explicit formula for a sequence of ordinary inputs that converges to $\mathbf{v}$. The basic idea for solving this problem is to use highly oscillatory inputs as was done in the example of section 3. The HOSs used in our general approximation algorithm involve a finite set of frequencies ω chosen so as to satisfy some special resonance conditions. We proceed by trying to handle each bracket in $\mathbf{v}$ separately. Let $L(\mathbf{X})_{\nu_1,\dots,\nu_m}$ denote the linear span of all the brackets $B \in \mathcal{B}r(\mathbf{X})$ such that $\delta_k(B) = \nu_k, k = 1,\dots,m$. Then $L(\mathbf{X})$ is the direct sum of the spaces $L(\mathbf{X})_{\nu_1,\dots,\nu_m}$. In particular an extended input $\mathbf{v}$ of finite order has a unique decomposition into a sum $\sum_{\nu_1,\dots,\nu_m} \mathbf{v}_{\nu_1,\dots,\nu_m}$, where each $\mathbf{v}_{\nu_1,\dots,\nu_m}$ is $L(\mathbf{X})_{\nu_1,\dots,\nu_m}$ valued. We try to produce, for each $\mathbf{v}_{\nu_1,\dots,\nu_m}$, an HOS $\{u^j_{\nu_1,\dots,\nu_m}\}$ that converges to $\mathbf{v}_{\nu_1,\dots,\nu_m}$. We then let $u^j = \sum_{\nu_1,\dots,\nu_m} u^j_{\nu_1,\dots,\nu_m}$ and hope that this will work. It turns out that even though the "input-to-trajectory" map is highly nonlinear a kind of "high frequency superposition principle" holds, and the u^j converge to $\mathbf{v}$, *provided that the frequencies associated to the various components* $\mathbf{v}_{\nu_1,\dots,\nu_m}$ *are independent* in a sense that will be made precise below.

If B is a Lie bracket in $L(\mathbf{X})$, $\delta_k(B) = \nu_k, k = 1,\dots,m$, then we define the *multiplicity* of B to be the dimension of $L(\mathbf{X})_{\nu_1,\dots,\nu_m}$.

In order to carry out the above program, one has to take into account the fact that it is not obvious how to decompose an extended input into parts in a canonical way (although there is a unique decomposition of $\mathbf{v}$ into a sum $\sum_{\nu_1,\dots,\nu_m} \mathbf{v}_{\nu_1,\dots,\nu_m}$). The general expression for an extended input obtained in (5) is not suitable because the brackets $[X_I]$ are not independent. What is needed is to write in an explicit way a *basis* of $L(\mathbf{X})$. One way of doing that is by using a *P. Hall basis*. (Cf. [1]. For an explanation of the reason why this is the right kind of basis for our problem, cf. [15].)

We recall that a *P. Hall set* $\mathcal{B}$ of formal brackets is a subset of $\mathcal{FB}r(\mathbf{X})$, endowed with a total ordering $\preceq$, that satisfies the following:

PH1. If $B, B' \in \mathcal{B}$ and $\delta(B) < \delta(B')$, then $B \preceq B'$;

PH2. Every X_k is in $\mathcal{B}$;

PH3. If B is a formal bracket and $\delta(B) > 1$, so that B can be written in a unique way as $[B_1, B_2]$, then $B \in \mathcal{B}$ iff (i) $B_1 \in \mathcal{B}$, (ii) $B_2 \in \mathcal{B}$, (iii) $B_1 \preceq B_2$, and (iv) either $\delta(B_2) = 1$ or $B_2 = [B_3, B_4]$ and $B_3 \preceq B_1$.

We will impose the additional requirement (which is not usually included in the definition of a P. Hall set) that

PH4. $X_{k_1} \preceq X_{k_2}$ iff $k_1 < k_2$.

The canonical map μ from $\mathcal{FB}r(\mathbf{X})$ to $L(\mathbf{X})$ is not one to one as shown by the example $B_1 = [X_1, [X_1, X_2]], B_2 = [[X_2, X_1], X_1]$, which are different as formal brackets, but $B_1 = B_2$ in $L(\mathbf{X})$. If we restrict μ to $\mathcal{B}$, then μ is one to one, and it turns out that $\mu(\mathcal{B})$ is a basis of $L(\mathbf{X})$, cf., e.g., [1]. Later on, we will make no difference between $\mathcal{B}$ and $\mu(\mathcal{B})$ and use $(\mathcal{B}, \preceq)$ to denote a basis of $L(\mathbf{X})$. A basis $(\mathcal{B}, \preceq)$ coming from this way is called a *P. Hall basis* of $L(\mathbf{X})$.

We remark for future use that every $B \in \mathcal{B}$ such that $\delta(B) > 1$ can be written in a unique way as $\mathrm{ad}^{\kappa}_{B_1}(B_2)$, where $B_1 \preceq B_2$ and either $\delta(B_2) = 1$ or the left factor B_3 of B_2 satisfies $B_3 \preceq B_1$. (Here we are using the standard notation ad_B to denote the operator $Z \to [B, Z]$.) From now on we fix a choice of P. Hall basis $(\mathcal{B}, \preceq)$ of $L(\mathbf{X})$ for which PH4 holds. We let $\mathcal{B}_n$ be the set of members of $\mathcal{B}$ that are of degree n.

Next we define the *Chen–Fliess product coefficients* associated with functions $u \in L^1([0,T],\mathbb{R}^m)$ and brackets $B \in \mathcal{B}$. We follow the notations in [15].

Associate with each bracket $B \in \mathcal{B}$ and each $u = (u_1,\dots,u_m) \in L^1([0,T],\mathbb{R}^m)$, two functions $c_B(u), C_B(u)$, defined on the interval $[0,T]$. The functions $c_B(u)$ will be in $L^1([0,T],\mathbb{R})$, and then $C_B(u)$ will be given by

$$C_B(u)(t) = \int_0^t c_B(s)\,ds. \tag{14}$$

The $c_B(u), C_B(u)$ are defined recursively as follows. For $B = X_k$, we let $c_B(u)(t) = u_k(t), 0 \le t \le T$, and then define $C_B(u)(t)$ by (14). Assume that $c_B(u), C_B(u)$ have been defined for all $B \in \mathcal{B}$ of degree $\le n$, and let $B \in \mathcal{B}_{n+1}$. Then B can be written in a unique way as $\mathrm{ad}^{\kappa}_{B_1}(B_2)$, where either $\delta(B_2) = 1$ or the left factor B_3 of B_2 satisfies $B_3 \preceq B_1$. We then define $c_B(u) = \frac{1}{\kappa!}(C_{B_1}(u))^{\kappa} c_{B_2}(u)$ and again define $C_B(u)$ by (14). This completes the recursive definition of the functions $c_B(u), C_B(u)$.

For each $u = (u_1,\dots,u_m) \in L^1([0,T],\mathbb{R}^m)$, let $S_{\mathbf{u}}$ be the formal trajectory determined by $\mathbf{u} = u_1X_1 + \cdots + u_mX_m$. We will simply say that $S_{\mathbf{u}}$ is the formal trajectory determined by u. It is proved in [15] that with the $C_B(u)$ defined as above, the formal trajectory $S_{\mathbf{u}}$ determined by u can be written as $S_{\mathbf{u}}(t) = \overleftarrow{\prod}_{B\in\mathcal{B}} \exp(C_B(u)(t)\,B)$, where "$\leftarrow$" means that the product is taken from left to right according to the total ordering $\preceq$ in $\mathcal{B}$. The above formula is the Chen–Fliess product expansion of $S_{\mathbf{u}}$, cf. [15].

4.2. A special sequence of ordinary inputs. For any finite set of real numbers F, let us write $|F|$ to denote the number of elements of F. Call a finite set $F \subseteq \mathbb{R} - \{0\}$ *canceling* if the sum of all the members of F is equal to 0. Call F *properly noncanceling* (PNC) if no proper subset of F is canceling. Call F *minimally canceling* (MC) if the only linear combinations $\sum_{\omega\in F} a_\omega\omega$ that are equal to zero and have integer coefficients such that $\sum_{\omega\in F}|a_\omega| \le |F|$ are those for which the a_ω are all equal (in which case, of course, they all have to be equal to 0, 1 or -1). It is clear that every MC set is PNC.

LEMMA 4.1. *Let $\{\omega_1,\dots,\omega_m\} \subseteq \mathbb{R} - \{0\}$ be MC. Let $\{u^j = (u_1^j,\dots,u_m^j)\}$, with $u_k^j(t) = j^{\frac{m-1}{m}} e^{ij\omega_k t}$, be a sequence of inputs. Then the u^j EI(m)-converge to*

$$\begin{aligned}\mathbf{u}^\infty &= \sum_{\ell_1\ne\ell_2\ne\cdots\ne\ell_m} \frac{1}{i^{m-1}\omega_{\ell_1}(\omega_{\ell_1}+\omega_{\ell_2})\cdots(\omega_{\ell_1}+\cdots+\omega_{\ell_{m-1}})} X_{\ell_1}\cdots X_{\ell_m}\\ &= \sum_{\ell_1\ne\ell_2\ne\cdots\ne\ell_m} \frac{1}{i^{m-1}m\,\omega_{\ell_1}(\omega_{\ell_1}+\omega_{\ell_2})\cdots(\omega_{\ell_1}+\cdots+\omega_{\ell_{m-1}})} [X_{\ell_1,\dots,\ell_m}].\end{aligned}$$

The proof follows from a direct verification that the conditions of Theorem 2.1 are satisfied with u^j and $\mathbf{u}^\infty$. We show how this is done by an example.

Example 4.1. Let $m = 3$, $\{\omega_1,\omega_2,\omega_3\} \subseteq \mathbb{R} - \{0\}$ be MC, and $u^j = (u_1^j, u_2^j, u_3^j)$ with $u_k^j(t) = j^{\frac{2}{3}} e^{ij\omega_k t}$. As in Example 3.1, we let $U_\ell^j(t) = \int_0^t u_\ell^j(s)\,ds = \frac{j^{-\frac{1}{3}}}{i\omega_\ell}[e^{ij\omega_\ell t} - 1]$. So if we let $V_\ell^j(t) = -\frac{j^{-\frac{1}{3}}}{i\omega_\ell}$, we get $\widetilde{UV}_\ell^j(t) = -\frac{j^{-\frac{1}{3}}}{i\omega_\ell} e^{ij\omega_\ell t}$. Then $\int_0^t u_{\ell_1}^j(s)\widetilde{UV}_{\ell_2}^j(s)ds = -\frac{j^{-\frac{2}{3}}}{i^2\omega_{\ell_2}(\omega_{\ell_1}+\omega_{\ell_2})}[e^{ij(\omega_{\ell_1}+\omega_{\ell_2})t} - 1]$. Letting $V_{\ell_1,\ell_2}^j(t) = \frac{j^{-\frac{2}{3}}}{i^2\omega_{\ell_2}(\omega_{\ell_1}+\omega_{\ell_2})}$, we have $\widetilde{UV}_{\ell_1,\ell_2}^j(t)$ $= \frac{j^{-\frac{2}{3}}}{i^2\omega_{\ell_2}(\omega_{\ell_1}+\omega_{\ell_2})} e^{ij(\omega_{\ell_1}+\omega_{\ell_2})t}$. Finally we simply let

$$V^j_{\ell_1,\ell_2,\ell_3}(t) = \int_0^t u^j_{\ell_1}(s)\widetilde{UV}^j_{\ell_2,\ell_3}(s)\,ds = \frac{1}{i^2\omega_{\ell_3}(\omega_{\ell_3}+\omega_{\ell_2})}\int_0^t e^{ij(\omega_{\ell_1}+\omega_{\ell_2}+\omega_{\ell_3})s}\,ds$$
$$= \begin{cases} \frac{t}{i^2\omega_{\ell_3}(\omega_{\ell_3}+\omega_{\ell_2})} & \text{if } \ell_1\neq\ell_2\neq\ell_3,\\ \frac{1}{i^3 j\omega_{\ell_3}(\omega_{\ell_3}+\omega_{\ell_2})(\omega_{\ell_3}+\omega_{\ell_2}+\omega_{\ell_1})}\left[e^{ij(\omega_{\ell_1}+\omega_{\ell_2}+\omega_{\ell_3})t}-1\right] & \text{otherwise.}\end{cases}$$

Then we have $\widetilde{UV}^j_{\ell_1,\ell_2,\ell_3}(t)\equiv 0$. Clearly the conditions of Theorem 2.1 are satisfied and the u^j EI(3)-converge to $\mathbf{u}^\infty = \sum_{\ell_1\neq\ell_2\neq\ell_3}\frac{1}{i^2\omega_{\ell_3}(\omega_{\ell_3}+\omega_{\ell_2})}X_{\ell_1}X_{\ell_2}X_{\ell_3}$. Since $\omega_1+\omega_2+\omega_3=0$, we have $\mathbf{u}^\infty = \sum_{\ell_1\neq\ell_2\neq\ell_3}\frac{1}{i^2\omega_{\ell_1}(\omega_{\ell_1}+\omega_{\ell_2})}X_{\ell_1}X_{\ell_2}X_{\ell_3}$, which is as claimed in Lemma 4.1.

Now come back to Lemma 4.1. Since $\mathcal{B}$ is a basis of $L(\mathbf{X})$, we can write each $[X_{\ell_1,\dots,\ell_m}]$ as a linear combination of brackets $B\in\mathcal{B}_m$. Then we get $\mathbf{u}^\infty = \sum_{B\in E^m} i^{1-m}g_B(\omega_1,\dots,\omega_m)B$, where E^m is the subset of $\mathcal{B}_m$ that contains all the brackets B such that $\delta_k(B)=1, k=1,\dots,m$, and $g_B(\omega_1,\dots,\omega_m)$ denote the coefficients of $B\in E^m$. Let $S^m_{\mathbf{u}^j}$ be the mth-order truncated formal trajectory determined by $\mathbf{u}^j$. By the Chen–Fliess product expansion, we get that $S^m_{\mathbf{u}^j}(t) = \mathbf{Tr}(m)(\prod^{\leftarrow}_{B\in\mathcal{B}}\exp(C_B(u^j)(t)\,B))$ in $A^m(\mathbf{X})$. It is clear also that in $A^m(\mathbf{X})$, $S^m_{\mathbf{u}^\infty}(t) = 1+\int_0^t \mathbf{u}^\infty(s)\,ds = 1+\sum_{B\in E^m} i^{1-m}g_B(\omega_1,\dots,\omega_m)t\,B$. Since $S^m_{\mathbf{u}^j}\to S^m_{\mathbf{u}^\infty}$ uniformly, cf. Remark 2.3, we see that $C_B(u^j)\to 0$ if $B\in\cup_{k=1}^m\mathcal{B}_k - E^m$ and $C_B(u^j)(t)\to i^{1-m}g_B(\omega_1,\dots,\omega_m)t$ if $B\in E^m$. So $g_B(\omega_1,\dots,\omega_m) = \frac{i^{m-1}}{t}\lim_{j\to\infty}C_B(u^j)(t)$ for $B\in E^m$.

Next we generalize this to an arbitrary bracket $B\in\mathcal{B}$. Let $\bar B\in\mathcal{B}$ be a bracket of degree $\delta(\bar B)=n>1$. Let $\{\omega_1,\dots,\omega_n\}$ be MC. We define a sequence $\{u^j\}$ of ordinary inputs associated to $\bar B$ by the following

$$u^j_k(t) = j^{\frac{n-1}{n}}\sum_{\ell=\delta_1(\bar B)+\cdots+\delta_{k-1}(\bar B)+1}^{\delta_1(\bar B)+\cdots+\delta_k(\bar B)} e^{ij\omega_\ell t},\quad k=1,\dots,m. \tag{15}$$

(Note that if $\delta_k(\bar B)=0$ we just let $u^j_k(t)=0$.) Then it is easy to see that the u^j EI(n)-converge to some extended input $\mathbf{u}^\infty$ of order n and $\mathbf{u}^\infty$ can be written as $\mathbf{u}^\infty = \sum_{B\in E(\bar B)} i^{1-n}g_B(\omega_1,\dots,\omega_n)\,B$ for some constants $g_B(\omega_1,\dots,\omega_n)$ determined by $(\omega_1,\dots,\omega_n)$ and B, where $E(\bar B)$ denotes the set of the brackets $B\in\mathcal{B}_n$ which are *equivalent* to $\bar B$ in the sense that $B\in E(\bar B)$ iff $\delta_k(B)=\delta_k(\bar B)$ for $k=1,\dots,m$. It follows from the Chen–Fliess product expansion that the $g_B(\omega_1,\dots,\omega_n)$ are equal to $g_B(\omega_1,\dots,\omega_n) = \frac{i^{n-1}}{t}\lim_{j\to\infty}C_B(u^j)(t)$, which in fact do not depend on t. In particular this is true for $\bar B\in E(\bar B)$.

Remark 4.1. An alternative way of getting the general g_B is as follows. Let $\bar B\in\mathcal{B}$ be a bracket of degree $\delta(\bar B)=n$. Take another set $\mathbf{Y}=\{Y_1,\dots,Y_n\}$ of indeterminates. Fix a choice of a P. Hall basis $\mathcal{B}^{\mathbf{Y}}$ of $L(\mathbf{Y})$. Let $\{\omega_1,\dots,\omega_n\}$ be MC. Consider the sequence $\bar u^j = (\bar u^j_1\ \dots,\bar u^j_n)$ with $\bar u^j_k(t) = j^{\frac{n-1}{n}}\exp(ij\omega_k t)$, $k=1,\dots,n$. Let $\bar{\mathbf{u}}^j = \bar u^j_1 Y_1+\cdots+\bar u^j_n Y_n$. From Lemma 4.1 we know that the $\bar{\mathbf{u}}^j$ EI(n)-converge to

$$\bar{\mathbf{u}}^\infty_{\mathbf{Y}} = \sum_{B^{\mathbf{Y}}\in E^{\mathbf{Y}}} i^{1-n}g_{B^{\mathbf{Y}}}(\omega_1,\dots,\omega_n)\,B^{\mathbf{Y}}, \tag{16}$$

where $E^{\mathbf{Y}}$ denotes the set of brackets $B^{\mathbf{Y}}\in\mathcal{B}^{\mathbf{Y}}$ such that $\delta_k(B^{\mathbf{Y}})=1, k=1,\dots,n$, and

$$g_{B^{\mathbf{Y}}}(\omega_1,\dots,\omega_n) = \frac{i^{n-1}}{t}\lim_{j\to\infty}C_{B^{\mathbf{Y}}}(\bar u^j)(t).$$

Now let $\theta_{\bar{B}}$ be the algebra homomorphism from $\hat{A}(\mathbf{Y})$ to $\hat{A}(\mathbf{X})$ defined by

$$\theta_{\bar{B}}(Y_k) = X_\ell, \text{ for } k = \delta_1(\bar{B}) + \cdots + \delta_{\ell-1}(\bar{B}) + 1, \ldots, \delta_1(\bar{B}) + \cdots + \delta_\ell(\bar{B}).$$

Letting $\theta_{\bar{B}}$ act on (16) we get an $L(\mathbf{X})$-valued function $\bar{\mathbf{u}}^\infty$. This $\bar{\mathbf{u}}^\infty$ is clearly equal to $\mathbf{u}^\infty$, which is the limit of the u^j associated to $\bar{B}$ in (15). If we write the brackets in $\bar{\mathbf{u}}^\infty$ into linear combinations of brackets in $\mathcal{B}$ we see that the $g_B(\omega_1, \ldots, \omega_n), B \in E(\bar{B})$, are linear combinations of the $g_{B^{\mathbf{Y}}}(\omega_1, \ldots, \omega_n)$ for $B^{\mathbf{Y}} \in E^{\mathbf{Y}}$. This observation will be used later.

So we have associated with each bracket $B \in \mathcal{B}$, $\delta(B) > 1$, each MC set $\{\omega_1, \ldots, \omega_{\delta(B)}\}$, a number $g_B(\omega_1, \ldots, \omega_{\delta(B)})$. We will think of these $g_B(\omega_1, \ldots, \omega_{\delta(B)})$ thus defined as functions in $\delta(B)$ variables $(\omega_1, \ldots, \omega_{\delta(B)})$. (For $B = X_k$, we define $g_{X_k}(\omega) = 1$.) Although they are only defined for $(\omega_1, \ldots, \omega_{\delta(B)})$ with $\{\omega_1, \ldots, \omega_{\delta(B)}\}$ being MC, as rational functions, they are completely determined.

4.3. An alternative definition. The definition of $g_B(\omega_1, \ldots, \omega_{\delta(B)})$ given in the previous section as limits of the Chen–Fliess product coefficients of some sequences of ordinary inputs is good for analyzing their structures. Next we give an alternative definition of g_B which is good for explicit calculations. First we need some definitions.

Any set M equipped with a map $M \times M \to M$ denoted by $(a, b) \to ab$ is called a *magma*.

Example 4.2. For any two functions $f, g \in L^1[0, T]$, we let $(f, g) \to f\#g$ be defined by $(f\#g)(t) = \int_0^t f(s)dsg(t)$. Then $(L^1[0, T], \#)$ is a magma.

Example 4.3. Let $AC[0, T]$ be the set of real-valued absolutely continuous functions on $[0, T]$. We define a product $* : AC[0, T] \times AC[0, T] \to AC[0, T]$ by $(U * V)(t) = \int_0^t U(s)\dot{V}(s)ds$. Then $(AC[0, T], *)$ is a magma.

Example 4.4. Let RF_n be the set of all real-valued rational functions on $\mathbb{R}^n$. Let $\mathrm{RF} = \cup_{n=1}^\infty \mathrm{RF}_n$ with the product $\tilde{\#} : \mathrm{RF}_p \times \mathrm{RF}_q \to \mathrm{RF}_{p+q}$ given by

$$(f\tilde{\#}g)(x_1, \ldots, x_p, y_1, \ldots, y_q) = \frac{f(x_1, \ldots, x_p)}{x_1 + \cdots + x_p} g(y_1, \ldots, y_q).$$

Then $(\mathrm{RF}, \tilde{\#})$ is a magma.

Example 4.5. Let $AC([0, T], \mathbb{R}^n)$ be the set of absolutely continuous $\mathbb{R}^n$-valued functions on $[0, T]$. Let $\mathcal{FL}_n$ be the set of all maps from $AC([0, T], \mathbb{R}^n)$ to $AC[0, T]$. Let $\mathcal{FL} = \cup_{n=1}^\infty \mathcal{FL}_n$. For any $h_1 \in \mathcal{FL}_p, h_2 \in \mathcal{FL}_q$, we define the product $(h_1, h_2) \to h_1.h_2 \in \mathcal{FL}_{p+q}$ by

$$(h_1.h_2)(U_1, \ldots, U_p, V_1, \ldots, V_q)(t) = (h_1(U_1, \ldots, U_p) * h_2(V_1, \ldots, V_q))(t).$$

Then with this product $\mathcal{FL}$ is a magma.

Equipped with the bracket product, i.e., $\mathcal{FB}r(\mathbf{X}) \times \mathcal{FB}r(\mathbf{X}) \ni (B_1, B_2) \to [B_1, B_2] \in \mathcal{FB}r(\mathbf{X})$, $(\mathcal{FB}r(\mathbf{X}), [\cdot, \cdot])$ is a magma. This magma is *isomorphic* to the free magma generated by $\mathbf{X}$ (for the definition of a free magma generated by $\mathbf{X}$, cf. [1]) and has the following property: let N be any magma; then every mapping of $\mathbf{X}$ into N can be uniquely extended to a *magma homomorphism* of $\mathcal{FB}r(\mathbf{X})$ into N. (A magma homomorphism $\nu : M \to N$ is a map from $M \to N$ that satisfies $\nu(ab) = \nu(a)\nu(b)$. It is an isomorphism if it is also $1 - 1$ and onto.)

Now let $u = (u_1, \ldots, u_m)$ be a function in $L^1([0, T], \mathbb{R}^m)$. We define $\phi(u) : \mathbf{X} \to (L^1[0, T], \#)$ by

$$X_k \to \phi_{X_k}(u)(t) = u_k(t),\ k = 1, \ldots, m.$$

Then we have a unique extension of $\phi(u)$ (still denoted by $\phi(u)$) from $\mathcal{FB}r(\mathbf{X}) \to (L^1[0,T],\#)$ satisfying $\phi_{[B_1,B_2]}(u)(t) = (\phi_{B_1}(u)\#\phi_{B_2}(u))(t)$.

Define a map Θ from $\mathbf{X}$ to $\mathcal{FL}_1 \subset \mathcal{FL}$ by $\Theta_{X_k}(U)(t) = U(t)$ for $U \in AC[0,T]$ and extend Θ to a unique map (denoted by Θ again) $\Theta : \mathcal{FB}r(\mathbf{X}) \to \mathcal{FL}$. Let $B = [B_1, B_2]$. Then Θ satisfies

$$\Theta_B(U_1,\dots,U_{\delta(B)})(t) = \left(\Theta_{B_1}(U_1,\dots,U_{\delta(B_1)}) * \Theta_{B_2}(U_{\delta(B_1)+1},\dots,U_{\delta(B)})\right)(t).$$

For each fixed $B \in \mathcal{FB}r(\mathbf{X}), \delta(B) = n$, it is clear that $\Theta_B : AC([0,T],\mathbb{R}^n) \to AC[0,T]$ is multiple linear in $(U_1,\dots,U_n)$; i.e., if $U_k = \sum_{\ell\in\Omega(k)} U_k^\ell(t),\ k = 1,\dots,n$, where $\Omega(k)$ are some finite index sets and $U_k^\ell \in AC[0,T]$ for $\ell \in \Omega(k)$, then

$$\Theta_B(U_1,\dots,U_n)(t) = \sum_{(\ell_1,\dots,\ell_n)\in\Omega} \Theta_B(U_1^{\ell_1},\dots,U_n^{\ell_n})(t),$$

where $\Omega = \Omega(1)\times\Omega(2)\times\cdots\times\Omega(n)$.

Let $\Psi : \mathbf{X} \to \mathbb{RF}_1 \subset \mathbb{RF}$ be defined as follows: $\Psi_{X_k}(\omega) = 1$ for $k = 1,\dots,m$. Then extend this to a homomorphism $\Psi : \mathcal{FB}r(\mathbf{X}) \to \mathbb{RF}$. So if $B = [B_1, B_2]$, Ψ_B is a rational function in $\delta(B)$ variables satisfying

$$\begin{aligned}\Psi_B(\omega_1,\dots,\omega_{\delta(B)}) &= \Psi_{B_1}(\omega_1,\dots,\omega_{\delta(B_1)})\tilde{\#}\Psi_{B_2}(\omega_{\delta(B_1)+1},\dots,\omega_{\delta(B)}) \\ &= \frac{\Psi_{B_1}(\omega_1,\dots,\omega_{\delta(B_1)})\Psi_{B_2}(\omega_{\delta(B_1)+1},\dots,\omega_{\delta(B)})}{\omega_1+\cdots+\omega_{\delta(B_1)}}.\end{aligned}$$

Since the map $\mu : \mathcal{FB}r(\mathbf{X}) \to L(\mathbf{X})$ is one-to-one restricted to any P. Hall set in $\mathcal{FB}r(\mathbf{X})$, the maps ϕ, Θ, Ψ are well defined on any P. Hall basis of $L(\mathbf{X})$.

Let $\mathcal{B}$ be a fixed P. Hall basis of $L(\mathbf{X})$. Let $u = (u_1,\dots,u_m)$ be an ordinary input. From the definition of the $c_B(u)$ and $\phi_B(u)$ we see that $c_B(u)(t) = æ_B\phi_B(u)(t)$ for some constant $æ_B$. Clearly $æ_{X_k} = 1$. If $\delta(B) = n > 1$, write B into the unique decomposition of $ad^\kappa_{B_1}(B_2)$, where $B_1, B_2 \in \mathcal{B}$ and either $\delta(B_2) = 1$ or the left factor of B_2 is not equal to B_1. We see that $æ_B = \frac{1}{\kappa!}æ^\kappa_{B_1}æ_{B_2}$.

For any $B \in \mathcal{B}$, we define a rational function $\hat{g}_B$, depending on $\delta(B)$ variables, by

$$\hat{g}_B(\omega_1,\dots,\omega_{\delta(B)}) = æ_B\Psi_B(\omega_1,\dots,\omega_{\delta(B)}). \tag{17}$$

It is clear that $\hat{g}_{X_k}(\omega) = 1$. Let $B \in \mathcal{B}_n, n > 1$. Write the unique decomposition $B = ad^\kappa_{B_1}(B_2)$, where $B_1, B_2 \in \mathcal{B}$ and either $\delta(B_2) = 1$ or the left factor of B_2 is not equal to B_1. Let n_1, n_2 be the degrees of B_1, B_2, so that $\kappa n_1 + n_2 = n$. Then formula (17) implies

$$\begin{aligned}&\hat{g}_B(\omega_1,\dots,\omega_n) \\ &\quad= \frac{1}{\kappa!}\prod_{q=1}^{\kappa}\left(\frac{\hat{g}_{B_1}(\omega_{(q-1)n_1+1},\omega_{(q-1)n_1+2},\dots,\omega_{qn_1})}{\omega_{(q-1)n_1+1}+\omega_{(q-1)n_1+2}+\ldots+\omega_{qn_1}}\right)\hat{g}_{B_2}(\omega_{\kappa n_1+1},\omega_{\kappa n_1+2},\dots,\omega_n).\end{aligned} \tag{18}$$

Our purpose is to define g_B in terms of $\hat{g}_B$. It is clear that in order to get g_B from $\hat{g}_B$, we need to take a symmetrization procedure. To each $B \in \mathcal{B}$, associate the sequence Σ_B of indeterminates obtained by just deleting all the brackets and commas, so that, for instance, if $B = [X_1,[X_1,X_2]]$, we associate the sequence $\Sigma_B = X_1X_1X_2$. We then define a map $\theta_B : \{1,\dots,\delta(B)\} \to \{1,\dots,m\}$ by letting $\theta_B(s) = k$ if the

indeterminate in the sth place of Σ_B is X_k. Next let P_B be the set of all permutations of the set $\{1,2,\dots,\delta(B)\}$ that map the set $\{s:\theta_B(s)=k\}$ for each $k\in\{1,\dots,m\}$ to the interval $\mathcal{I}_{B,k}=\{\delta_1(B)+\cdots+\delta_{k-1}(B)+1,\dots,\delta_1(B)+\cdots+\delta_k(B)\}$. We then define $g_B(\omega_1,\dots,\omega_{\delta(B)})$ by

$$g_B(\omega_1,\dots,\omega_{\delta(B)})=\sum_{\pi\in P_B}\hat{g}_B(\omega_{\pi(1)},\dots,\omega_{\pi(\delta(B))}). \tag{19}$$

Next we prove that the functions g_B thus defined from the symmetrization of the $\hat{g}_B$ are the same as those defined to be the limits of the Chen–Fliess product coefficients $C_B(u^j)$ of some sequences $\{u^j\}$ defined in section 4.2. First we need an auxiliary result. Let $\hat{\omega}=(\omega_1,\dots,\omega_n)$ be an $n$tuple of real numbers. Then we say that $\hat{\omega}$ is *essentially noncanceling* (ENC) if $\sum_{\ell\in\Omega}\omega_\ell\neq 0$ for any proper subset Ω of $\{1,\dots,n\}$. (Note that in the definition of ENC, the entries ω_k of $(\omega_1,\dots,\omega_n)$ are allowed to be equal.) Clearly if $\Omega=\{\omega_1,\dots,\omega_n\}$ is MC, then for any $\tilde{\omega}_i\in\Omega$, $(\tilde{\omega}_1,\dots,\tilde{\omega}_n)\in\Omega\times\cdots\times\Omega$ is ENC.

LEMMA 4.2. *Let $B\in\mathcal{B}$ be a bracket of degree n. Let $\hat{\omega}=(\omega_1,\dots,\omega_n)$ be ENC. Let $\{U^j=(U_1^j,\dots,U_n^j)\}$ be a sequence of $\mathbb{R}^n$-valued functions on $[0,T]$ defined by $U_k^j(t)=j^{\frac{n-1}{n}}\int_0^t e^{ij\omega_k s}\,ds$, $k=1,\dots,n$. Then for $j=1,2,\dots$, and $t\in[0,T]$, we have*

$$i^{n-1}\Theta_B(U_1^j,\dots,U_n^j)(t)=\int_0^t\Psi_B(\omega_1,\dots,\omega_n)e^{ij(\omega_1+\cdots+\omega_n)s}\,ds+R_B^j(t), \tag{20}$$

where the R_B^j are finite linear combinations of integrals of the form $\int_0^t h_\xi e^{ij\sum_{\ell\in\Omega(\xi)}\omega_\ell s}\,ds$, where h_ξ are some constants and $\Omega(\xi)$ are some proper subsets of $\{1,2,\dots,n\}$.

Proof. We use induction on n to prove the lemma. Obviously (20) is true if $B=X_k$. Assume that (20) is true for all $B\in\mathcal{B}$ of degree $\le n-1$. Let $B\in\mathcal{B}_n$, $n>1$. Write $B=[B_1,B_2]$ with $B_1,B_2\in\mathcal{B}$, $B_1\preceq B_2$. Assume that $\delta(B_1)=n_1,\delta(B_2)=n_2$, so $n_1+n_2=n$. Let $(\omega_1,\dots,\omega_n)$ be ENC. Then both $(\omega_1,\dots,\omega_{n_1})$ and $(\omega_{n_1+1},\dots,\omega_n)$ are ENC. Let $\bar{U}_k^j(t)=j^{\frac{n_1-1}{n_1}}\int_0^t e^{ij\omega_k s}ds$ for $k=1,\dots,n_1$, $\bar{U}_k^j(t)=j^{\frac{n_2-1}{n_2}}\int_0^t e^{ij\omega_k s}ds$ for $k=n_1+1,\dots,n$. So $U_k^j(t)=j^{\frac{n-1}{n}-\frac{n_1-1}{n_1}}\bar{U}_k^j(t)$ for $k=1,\dots,n_1$, $U_k^j(t)=j^{\frac{n-1}{n}-\frac{n_2-1}{n_2}}\bar{U}_k^j(t)$ for $k=n_1+1,\dots,n$. By induction we know that

$$i^{n_1-1}\Theta_{B_1}(\bar{U}_1^j,\dots,\bar{U}_{n_1}^j)(t)=\int_0^t\Psi_{B_1}(\omega_1,\dots,\omega_{n_1})e^{ij(\omega_1+\cdots+\omega_{n_1})s}\,ds+R_{B_1}^j(t),$$

$$i^{n_2-1}\Theta_{B_2}(\bar{U}_{n_1+1}^j,\dots,\bar{U}_n^j)(t)=\int_0^t\Psi_{B_2}(\omega_{n_1+1},\dots,\omega_n)e^{ij(\omega_{n_1+1}+\cdots+\omega_n)s}\,ds+R_{B_2}^j(t),$$

with $R_{B_1}^j,R_{B_2}^j$ satisfying the induction assumptions. We have

$$\begin{aligned}
i^{n-1}\Theta_B(U_1^j,\dots,U_n^j)(t)&=ij((i^{n_1-1}\Theta_{B_1}(\bar{U}_1^j,\dots,\bar{U}_{n_1}^j))*(i^{n_2-1}\Theta_{B_2}(\bar{U}_{n_1+1}^j,\dots,\bar{U}_n^j)))(t)\\
&=ij\int_0^t\left\{\left(\int_0^s\Psi_{B_1}(\omega_1,\dots,\omega_{n_1})e^{ij(\omega_1+\cdots+\omega_{n_1})\tau}\,d\tau+R_{B_1}^j(s)\right)\right.\\
&\qquad\left.\times\,\Psi_{B_2}(\omega_{n_1+1},\dots,\omega_n)e^{ij(\omega_{n_1+1}+\cdots+\omega_n)s}+\dot{R}_{B_2}^j(s)\right\}ds\\
&=\int_0^t\Psi_B(\omega_1,\dots,\omega_n)e^{ij(\omega_1+\cdots+\omega_n)s}\,ds+R_B^j(t),
\end{aligned}$$

where

$$R_B^j(t) = ij\int_0^t \Big\{ R_{B_1}^j(s)\Psi_{B_2}(\omega_{n_1+1},\dots,\omega_n)e^{ij(\omega_{n_1+1}+\cdots+\omega_n)s} + \dot{R}_{B_2}^j(s)\int_0^s \Psi_{B_1}(\omega_1,\dots,\omega_{n_1})e^{ij(\omega_1+\cdots+\omega_{n_1})\tau}\,d\tau + R_{B_1}^j(s)\dot{R}_{B_2}^j(s)\Big\}\,ds - \int_0^t \frac{\Psi_{B_1}(\omega_1,\dots,\omega_{n_1})\Psi_{B_2}(\omega_{n_1+1},\dots,\omega_n)e^{ij(\omega_{n_1+1}+\cdots+\omega_n)s}}{\omega_1+\cdots+\omega_{n_1}}\,ds.$$

It is clear that R_B^j satisfies the induction assumption. This finishes the proof of the lemma.

Now we are ready to show that the two definitions of the g_B give rise to the same functions. Let $B \in \mathcal{B}_n$ with $\Sigma_B = X_{\ell_1}\cdots X_{\ell_n}$. Let $u \in L^1([0,T],\mathbb{R}^m)$ be a function and $U(t) = \int_0^t u(s)ds$. Using induction, we can easily show that $\int_0^t \phi_B(u)(s)ds = \Theta_B(U_{\ell_1},\dots,U_{\ell_n})(t)$. Let $\{\omega_1,\dots,\omega_n\}$ be MC. Recall that $g_B(\omega_1,\dots,\omega_n)$ is the limit of $\frac{i^{n-1}}{t}C_B(u^j)(t) = \frac{i^{n-1}\text{æ}_B}{t}\int_0^t \phi_B(u^j)(s)ds$, where the u^j are defined by

$$u_k^j(t) = j^{\frac{n-1}{n}} \sum_{\ell=\delta_1(B)+\cdots+\delta_{k-1}(B)+1}^{\delta_1(B)+\cdots+\delta_k(B)} e^{ij\omega_\ell t}. \tag{21}$$

Let $U^j(t) = \int_0^t u^j(s)ds$. We have

$$\frac{i^{n-1}\text{æ}_B}{t}\int_0^t \phi_B(u^j)(s)ds = \frac{i^{n-1}\text{æ}_B}{t}\Theta_B(U_{\ell_1}^j,\dots,U_{\ell_n}^j)(t) = \frac{i^{n-1}\text{æ}_B}{t}\sum_{\pi\in Q_B}\Theta_B(\bar{U}_{\pi(1)}^j,\dots,\bar{U}_{\pi(n)}^j)(t),$$

where (1) $Q_B = \mathcal{I}_{B,\ell_1}\times\mathcal{I}_{B,\ell_2}\times\cdots\times\mathcal{I}_{B,\ell_n}$, $\pi = (\pi(1),\dots,\pi(n)) \in Q_B$; (2) $\bar{U}_{\pi(l)}^j(t) = j^{\frac{n-1}{n}}\int_0^t e^{ij\omega_{\pi(l)}s}ds$. Clearly $P_B \subset Q_B$. Now from the requirement that $\{\omega_1,\dots,\omega_n\}$ be MC, using Lemma 4.2 and the definition of $\hat{g}_B$, we have

$$i^{n-1}\text{æ}_B\Theta_B(\bar{U}_{\pi(1)}^j,\dots,\bar{U}_{\pi(n)}^j)(t) = \int_0^t \hat{g}_B(\omega_{\pi(1)},\dots,\omega_{\pi(n)})e^{ij(\omega_{\pi(1)}+\cdots+\omega_{\pi(n)})s}\,ds + \text{æ}_B R_B^j(t),$$

where the R_B^j converge to 0 as $j\to\infty$. So if $\omega_{\pi(1)}+\cdots+\omega_{\pi(n)} \neq 0$, then $\Theta_B(\bar{U}_{\pi(1)}^j,\dots,\bar{U}_{\pi(n)}^j)(t) \to 0$. By the minimally canceling property of the $\{\omega_1,\dots,\omega_n\}$, $\omega_{\pi(1)}+\cdots+\omega_{\pi(n)} = 0$ iff π is a permutation of $(1,\dots,n)$, i.e., $\pi \in P_B$. In that case we get

$$\lim_{j\to\infty} i^{n-1}\text{æ}_B\Theta_B(\bar{U}_{\pi(1)}^j,\dots,\bar{U}_{\pi(n)}^j)(t) = \hat{g}_B(\omega_{\pi(1)},\dots,\omega_{\pi(n)})t.$$

So we conclude that

$$\lim_{j\to\infty}\frac{i^{n-1}\text{æ}_B}{t}\Phi_B(U^j)(t) = \lim_{j\to\infty}\frac{i^{n-1}\text{æ}_B}{t}\Theta_B(U_{\ell_1}^j,\dots,U_{\ell_m}^j)(t) = \sum_{\pi\in P_B}\hat{g}_B(\omega_{\pi(1)},\dots,\omega_{\pi(n)}).$$

This shows that the functions g_B defined to be the limits of the Chen–Fliess product coefficients of some control sequences $\{u^j\}$ in (21) coincide with the functions obtained from the symmetrization of the $\hat{g}_B$ on the set $\{(\omega_1, \ldots, \omega_n) \mid \{\omega_1, \ldots, \omega_n\} \text{ is MC}\}$. So they coincide as rational functions.

5. Approximate tracking: The general case. We now describe the algorithm of approximating an arbitrary rth-order extended input by ordinary inputs. We will limit ourselves to the presentation of the algorithm in this section. The proof that it actually works will be given in the next two sections.

Let's fix a choice of P. Hall basis $(\mathcal{B}, \preceq)$ of $L(\mathbf{X})$ for which PH4 holds. We express our extended input $\mathbf{v}$ as a sum $\mathbf{v}(t) = \sum_{n=1}^{r} \sum_{B \in \mathcal{B}_n} v_B(t) B$, where the v_B are functions of class C^1 on $[0, T]$.

We take our approximating sequence to be of the form

$$u_k^j(t) = \eta_{k,0}(t) + j^{\frac{1}{2}} \sum_{\omega \in \Omega(2,k)} \eta_{\omega,k}(t) \mathrm{e}^{ij\omega t} + \sum_{n=3}^{r} j^{\frac{n-1}{n}} \sum_{\omega \in \Omega(n,k)} \eta_{\omega}(t) \mathrm{e}^{ij\omega t}. \tag{22}$$

Here

(1) the $\Omega(n,k)$ for $(n,k) \in \{2, \ldots, r\} \times \{1, \ldots, m\}$ are pairwise disjoint finite subsets of $\mathbb{R} - \{0\}$.

(2) each $\Omega(n,k)$ is symmetric; i.e., $\omega \in \Omega(n,k)$ implies $-\omega \in \Omega(n,k)$.

(3) the $\eta_{k,0}$, $\eta_{\omega,k}$, η_{ω} are complex-valued functions of class C^1 on $[0,T]$.

(4) the $\eta_{k,0}$ are in fact real valued.

(5) the identity $\eta_{-\omega} = \overline{\eta_{\omega}}$ holds for each $\omega \in \cup_{n=3}^{r} \cup_{k=1}^{m} \Omega(n,k)$, and the functions $\eta_{\omega,k}$ satisfy $\eta_{-\omega,k} = \overline{\eta_{\omega,k}}$ for $\omega \in \Omega(2,k)$.

(Here $\bar{z}$ denotes the complex conjugate of z.) Notice that the various terms of (22) are in principle complex, but conditions (4) and (5) imply that the u_k^j are real.

We now describe in detail how the sets $\Omega(n,k)$ are chosen. Let us call F *symmetrically minimally canceling* (SMC) if it is symmetric (i.e., $\omega \in F$ implies $-\omega \in F$) and contains an MC subset of cardinality $\frac{1}{2}|F|$. (In that case, it is easy to see that F has exactly two such subsets F_1, F_2, and these subsets are disjoint and satisfy $\omega \in F_1$ iff $-\omega \in F_2$.)

At this point we have to take care of an additional technical issue. The "nonlinear superposition principle" that was referred to in section 4 only makes it possible to separate brackets in $B \in \mathcal{B}$ *as long as not all their degrees* $\delta_k(B)$ *are equal.* (So, for instance, the brackets $[X_2, [X_1, [X_1, [X_1, X_2]]]]$ and $[[X_1, X_2], [X_1, [X_1, X_2]]]$, both of which are in $\mathcal{B}$, cannot be separated.) This requires that we group brackets in $\mathcal{B}$ into classes, separate the classes rather than the individual brackets, and handle the problem of fitting the coefficients of the individual brackets by a different procedure.

Divide each $\mathcal{B}_n$ into equivalence classes by declaring two members B_1, B_2 of $\mathcal{B}_n$ to be equivalent if $\delta_k(B_1) = \delta_k(B_2)$ for $k = 1, \ldots, m$. Let $\mathcal{E}_n$ be the set of equivalence classes of $\mathcal{B}_n$. For each $E \in \mathcal{E}_n$, we can define $\delta_k(E)$, $\delta(E)$ to be $\delta_k(B)$, $\delta(B)$ for any $B \in E$. Clearly the cardinality $|E|$ of E is equal to the multiplicity of the brackets $B \in E$.

For $n \geq 2$, $E \in \mathcal{E}_n$, $k \in \{1, \ldots, m\}$, choose $|E|$ sets $\Omega_{E,\rho,k}$, $\rho = 1, \ldots, |E|$ of $\mathbb{R} - \{0\}$ such that

(6) each $\Omega_{E,\rho,k}$, for $E \in \cup_{n=2}^{r} \mathcal{E}_n$, $\rho \in \{1, \ldots, |E|\}$, $k \in \{1, \ldots, m\}$, is a symmetric subset of $\mathbb{R} - \{0\}$ of cardinality $|\Omega_{E,\rho,k}| = 2\delta_k(E)$.

About the sets $\Omega_{E,\rho,k}$ for $n \geq 3$, assume

(7) for each $E \in \cup_{n=3}^{r} \mathcal{E}_n$, $\rho \in \{1, \dots, |E|\}$, the sets $\Omega_{E,\rho,k}$, $k \in \{1, \dots, m\}$, are pairwise disjoint.

For $n = 2$ we make a different assumption. We first observe that every $E \in \mathcal{E}_2$ consists of exactly one bracket which is of the form $[X_{k_1}, X_{k_2}]$ with $k_1 < k_2$. (And, conversely, every such bracket gives rise to an $E \in \mathcal{E}_2$.) We take $\Omega_{E,\rho,k_1} = \Omega_{E,\rho,k_2}$ and let $\Omega_{E,\rho,k} = \emptyset$ if $k \notin \{k_1, k_2\}$. So

(8) for each $E \in \mathcal{E}_2$, $E = [X_{k_1}, X_{k_2}]$, we have $\Omega_{E,1,k_1} = \Omega_{E,1,k_2}$.

We then let $\Omega_{E,\rho} = \bigcup_{k=1}^{m} \Omega_{E,\rho,k}$, so that $|\Omega_{E,\rho}| = 2\delta(E)$ if $\delta(E) > 2$, and $|\Omega_{E,\rho}| = 2$ if $\delta(E) = 2$. We assume

(9) each $\Omega_{E,\rho}$, for $E \in \cup_{n=3}^{r} \mathcal{E}_n$, $\rho \in \{1, \dots, |E|\}$, is SMC.

Next define $\Omega(n,k)$ by $\Omega(n,k) = \bigcup_{E \in \mathcal{E}_n} \bigcup_{\rho=1}^{|E|} \Omega_{E,\rho,k}$. Since the sets $\Omega_{E,\rho}$ are SMC, each $\Omega_{E,\rho}$ has exactly two MC subsets, and these subsets are disjoint and have cardinality $\frac{1}{2}|\Omega_{E,\rho}|$. Let $Q_{E,\rho}$ be the set whose two elements are these two MC subsets of $\Omega_{E,\rho}$.

There are two further conditions that have to be imposed on the sets $\Omega_{E,\rho}$. Let us call a finite collection $\{S_\lambda\}_{\lambda \in \Lambda}$ of finite subsets of $\mathbb{R}$ *independent* with respect to r if the sets S_λ are pairwise disjoint and, whenever a linear combination $\sum_{\lambda \in \Lambda} \sum_{s \in S_\lambda} a_s s$ vanishes, and the a_s are integers such that $\sum_{\lambda \in \Lambda} \sum_{s \in S_\lambda} |a_s| \le r$, it follows that each of the sums $\sum_{s \in S_\lambda} a_s s$ vanishes. (In other words, there should not be any integral relations among the members of $\cup_\lambda S_\lambda$ other than those that come from the S_λ themselves.)

We then require

(10) that the $\Omega_{E,\rho}$, as (E,ρ) ranges over all pairs such that $E \in \cup_{n=2}^{r} \mathcal{E}_n$, $\rho \in \{1, \dots, |E|\}$, are independent with respect to r.

In the special case when $\delta(E) = 2$, each E is of the form $\{[X_{k_1}, X_{k_2}]\}$, where $k_1 < k_2$. In that case, the set $\Omega_{E,1} = \Omega_{E,1,k_1} = \Omega_{E,1,k_2}$ consists of a frequency $\omega_{k_1,k_2} > 0$ together with its negative.

We then have the following.

THEOREM 5.1. *Let m, r be positive integers. Let $\{u^j\}_{j=1}^{\infty}$ be the sequence of input functions defined by* (22)*, where the functions $\eta_{0,k}$, $\eta_{\omega,k}$, η_ω and the sets $\Omega(n,k)$, $\Omega_{E,\rho}$, $\Omega_{E,\rho,k}$ satisfy conditions* (1)–(10)*. Then the sequence $\{u^j\}$ EI(r)-converges to the extended input*

$$\begin{aligned} \mathbf{u}^\infty(t) = &\sum_{k=1}^{m} \eta_{0,k}(t) X_k + \sum_{1 \le k_1 < k_2 \le m} \frac{1}{i\omega_{k_1,k_2}} \xi_{k_1,k_2}(t) [X_{k_1}, X_{k_2}] \\ &+ \sum_{n=3}^{r} \sum_{E \in \mathcal{E}_n} \sum_{B \in E} \sum_{\rho=1}^{|E|} \left(\sum_{F \in Q_{E,\rho}} i^{1-\delta(E)} \xi_{B,\rho}^{F} \prod_{\omega \in F} \eta_\omega(t) \right) B, \end{aligned} \tag{23}$$

where $\xi_{k_1,k_2}(t) = \eta_{\omega_{k_1,k_2},k_1} \eta_{-\omega_{k_1,k_2},k_2} - \eta_{-\omega_{k_1,k_2},k_1} \eta_{\omega_{k_1,k_2},k_2}$.

It is now easy to see how to choose the η_ω so that the limiting extended input $\mathbf{u}^\infty$ has a desired value. Fix an E and a ρ.

CH1. Assume first that $\delta(E)$ is odd and $\delta(E) > 1$. Then the numbers $\xi_{B,\rho}^{F}$ for the two members of $Q_{E,\rho}$ are equal. Call their common value $\xi_{B,\rho}$. Also, $i^{1-\delta(E)} = \pm 1$. Let $\hat{\xi}_{B,\rho} = \xi_{B,\rho}$. Pick all the η_ω for $\omega \in \Omega_{E,\rho}$ to be equal to 1, except for one pair $\{\omega, -\omega\}$ of members of $\Omega_{E,\rho}$, for which we pick $\eta_\omega = \eta_{-\omega} = \frac{i^{1-\delta(E)}}{2} \zeta_{E,\rho}$, where $\zeta_{E,\rho} \in C^1([0,T], \mathbb{R})$.

CH2. Assume that $\delta(E)$ is even and $\delta(E) > 2$. Then the numbers $\xi_{B,\rho}^{F}$, for the two members of $Q_{E,\rho}$, are negatives of each other. Pick one of the two members of

$Q_{E,\rho}$ and call it $F_{E,\rho}$. Also, $i^{1-\delta(E)} = \pm i$. Let $\hat{\xi}_{B,\rho} = \xi_{B,\rho}^{F_{E,\rho}}$. Pick all the η_ω for $\omega \in \Omega_{E,\rho}$ to be equal to 1, except for one pair $\{\omega, -\omega\}$ of members of $\Omega_{E,\rho}$ such that $\omega \in F_{E,\rho}$. Then pick $\eta_\omega = -\frac{i^{1-\delta(E)}}{2}\zeta_{E,\rho}$, where $\zeta_{E,\rho} \in C^1([0,T],\mathbb{R})$, and let $\eta_{-\omega} = -\eta_\omega = \overline{\eta_\omega}$.

CH3. Assume that $\delta(E) = 2$. Let $E = \{[X_{k_1}, X_{k_2}]\}$, $k_1 < k_2$. We pick $\eta_{\omega_{k_1,k_2},k_2} = \eta_{-\omega_{k_1,k_2},k_2} = 1$, and $\eta_{\omega_{k_1,k_2},k_1} = -\eta_{-\omega_{k_1,k_2},k_1} = \frac{i\omega_{k_1,k_2}}{2}\zeta_{k_1,k_2}$, where $\zeta_{k_1,k_2} \in C^1([0,T],\mathbb{R})$. If $k_1 \neq k \neq k_2$, we let $\eta_{\omega_{k_1,k_2},k} = \eta_{-\omega_{k_1,k_2},k} = 0$.

CH4. Choose $\eta_{k,0} = \zeta_{0,k} \in C^1([0,T],\mathbb{R})$.

With these choices, we have

$$\mathbf{u}^\infty(t) = \sum_{k=1}^{m} \zeta_{k,0}(t)X_k + \sum_{1\le k_1<k_2\le m} \zeta_{k_1,k_2}(t)[X_{k_1},X_{k_2}] + \sum_{n=3}^{r}\sum_{E\in\mathcal{E}_n}\sum_{B\in E}\sum_{\rho=1}^{|E|}(\hat{\xi}_{B,\rho}\zeta_{E,\rho}(t))B. \tag{24}$$

In order to get $\mathbf{u}^\infty$ to be equal to $\mathbf{v}$, we need to choose

$$\zeta_{0,k} = v_{X_k}, \qquad \zeta_{k_1,k_2} = v_{[X_{k_1},X_{k_2}]} \tag{25}$$

and to let the $\zeta_{E,\rho}$ be solutions of

$$\sum_{\rho=1}^{|E|} \hat{\xi}_{B,\rho}\zeta_{E,\rho}(t) = v_B(t) \text{ for each } E \in \mathcal{E}_n,\ n > 2. \tag{26}$$

The possibility of solving (26) is guaranteed if the frequency sets $\Omega_{E,\rho,k}$ are chosen so that

(11) the square matrix $\{\hat{\xi}_{B,\rho}\}_{B\in E,\,1\le\rho\le|E|}$ is invertible for each $E \in \cup_{n=3}^{r}\mathcal{E}_n$. It turns out that

(A) it is always possible to choose frequencies so that conditions (1), (2), (6), (7), (8), (9), (10), (11) hold;

(B) if the frequencies are chosen so that conditions (1), (2), (6), (7), (8), (9), (10), (11) hold, and if the η's are chosen according to CH1, CH2, CH3, CH4 (so that, in particular, (3), (4), (5) hold), where the ζ's satisfy (25) and (26), then the u^j EI(r)-converge to $\mathbf{v}$.

This completes the description of the algorithm. Its justification requires, of course, that we prove Theorem 5.1 and statement A. The proofs are given in the next two sections.

6. Proof of Theorem 5.1**.** Now we prove Theorem 5.1. In order to justify the algorithm we still need to prove that it is always possible to choose the frequency sets so that condition (11) holds in addition to the other requirements. This will be done in the next section.

Proof of Theorem 5.1. We will apply Theorem 2.1. For this we need to find a sequence $\{\mathbf{v}^j = \sum_{0<|I|\le r} v_I^j X_I\}$ of polynomial inputs of order $\le r$ and rth-order truncated generalized differences $(\mathbf{u}^j \overset{\text{g.d.(r)}}{-} \mathbf{v}^j) = \sum_{0<|I|\le r} \widetilde{UV}_I^j X_I$ of $\mathbf{u}^j$ and $\mathbf{v}^j$ such that $\{\mathbf{v}^j\}$, $\mathbf{u}^\infty$, and $\{(\mathbf{u}^j \overset{\text{g.d.(r)}}{-} \mathbf{v}^j)\}$ satisfy the conditions of Theorem 2.1.

We define absolutely continuous functions V_I^j and $\widetilde{UV}_I^j$ for $0 < |I| \le r$ such that (12) and (13) hold and conditions c1′(r), c2′(r), c3′(r) are satisfied; cf. Remark 2.4. For this, we use induction to define V_I^j and $\widetilde{UV}_I^j$ such that

(A) the V_I^j have the form $V_I^j = V_I + R_I^j$, where
(i)

$$V_\ell(t) = \int_0^t \eta_{\ell,0}(s)\,ds, \tag{27}$$

$$V_{\ell_1,\ell_2}(t) = -\sum_{\substack{(\omega_1,\omega_2)\in\Omega(2,\ell_1)\times\Omega(2,\ell_2)\\ \omega_1+\omega_2=0}} \int_0^t \frac{\eta_{\omega_1,\ell_1}(s)\eta_{\omega_2,\ell_2}(s)}{i\omega_2}\,ds, \tag{28}$$

$$V_I(t) = (-1)^{k-1}\sum_{\hat\omega\in\Omega(I)}\int_0^t \frac{\eta_{\hat\omega}(s)}{i^{k-1}\omega_k(\omega_k+\omega_{k-1})\cdots(\omega_k+\cdots+\omega_2)}\,ds \tag{29}$$

for $I = (\ell_1,\dots,\ell_k), 3 \le k \le r$, where $\hat\omega = (\omega_1,\dots,\omega_k)$, $\eta_{\hat\omega}(t) = \eta_{\omega_1}(t)\cdots\eta_{\omega_k}(t)$, and $\Omega(I)$ is the subset of $\Omega(k,\ell_1)\times\cdots\times\Omega(k,\ell_k)$ such that $\hat\omega = (\omega_1,\dots,\omega_k)$ belongs to $\Omega(I)$ iff the set $\{\omega_1,\dots,\omega_k\} \in Q_{E,\rho}$ for some $E \in \mathcal{E}_k, \rho \in \{1,\dots,|E|\}$;

(ii) the R_I^j converge to 0 uniformly as $j \to \infty$ and the $\|\dot R_I^j\|_{L^1}$ are uniformly bounded;

(B) for $I = (\ell_1,\dots,\ell_k)$, the $\widetilde{UV}_I^j$ can be written as

$$\widetilde{UV}_I^j(t) = (-1)^k \sum_{\hat n\in\Omega_1(k)} j^{-\alpha_{\hat n}} \sum_{\hat\omega\in\Omega(\hat n,I)} \frac{\eta_{\hat\omega}(t)e^{ij\sum\hat\omega t}}{i^k\omega_k(\omega_k+\omega_{k-1})\cdots(\omega_k+\cdots+\omega_1)}$$

$$+(-1)^k \sum_{\hat n\in\Omega_2(k)} j^{-\alpha_{\hat n}} \sum_{\hat\omega\in\Omega(\hat n,I)} \frac{\eta_{\hat\omega}(t)e^{ij\sum\hat\omega t}}{i^k\omega_k(\omega_k+\omega_{k-1})\cdots(\omega_k+\cdots+\omega_1)}. \tag{30}$$

Here
(1) we write $\hat n = (n_1,\dots,n_k)$, $\alpha_{\hat n} = \frac{1}{n_1}+\cdots+\frac{1}{n_k}$, $\sum\hat\omega = \omega_1+\cdots+\omega_k$;
(2) $\Omega_1(k)$ is a subset of

$$\overbrace{\{2,\dots,r\}\times\{2,\dots,r\}\times\cdots\times\{2,\dots,r\}}^{k}$$

and $\Omega_2(k)$ is a subset of

$$\overbrace{\{3,\dots,r\}\times\{3,\dots,r\}\times\cdots\times\{3,\dots,r\}}^{k}$$

such that
(a) $\hat n = (n_1,\dots,n_k) \in \Omega_2(k)$ iff $\alpha_{\hat n} < 1$;
(b) $\hat n = (n_1,\dots,n_k) \in \Omega_1(k)$ iff one of $n_1,\dots,n_k$, say n_τ, is equal to 2 and the others are between 3 and r; or $\alpha_{\hat n} < 1$;
(3) for each $\hat n = (n_1,\dots,n_k)$, $\Omega(\hat n, I)$ is a subset of $\Omega(n_1,\ell_1)\times\cdots\times\Omega(n_k,\ell_k)$ such that $\hat\omega = (\omega_1,\dots,\omega_k) \in \Omega(\hat n, I)$ iff $\omega_k(\omega_k+\omega_{k-1})\cdots(\omega_k+\cdots+\omega_1) \ne 0$;
(4) if $\hat n \in \Omega_2(k)$, for each $\hat\omega = (\omega_1,\dots,\omega_k) \in \Omega(\hat n, I)$ we write $\eta_{\hat\omega}(t) = \eta_{\omega_1}(t)\cdots\eta_{\omega_k}(t)$. If $\hat n = (n_1,\dots,n_k) \in \Omega_1(k)$ with $n_\tau = 2$, for each $\hat\omega = (\omega_1,\dots,\omega_k) \in \Omega(\hat n, I)$, we write $\eta_{\hat\omega}(t) = \eta_{\omega_1}(t)\cdots\eta_{\omega_\tau,\ell_\tau}(t)\cdots\eta_{\omega_k}(t)$.

Using integration by parts we get

$$\begin{aligned}
U_\ell^j(t) &= \int_0^t \eta_{\ell,0}(s)\,ds + j^{\frac{1}{2}} \sum_{\omega\in\Omega(2,\ell)} \int_0^t \eta_{\omega,\ell}(s) e^{ij\omega s}\,ds \\
&\quad + \int_0^t \sum_{n=3}^{r} j^{\frac{n-1}{n}} \sum_{\omega\in\Omega(n,\ell)} \eta_\omega(s) e^{ij\omega s}\,ds \\
&= \int_0^t \eta_{\ell,0}(s)\,ds + j^{-\frac{1}{2}} \sum_{\omega\in\Omega(2,\ell)} \left\{ \frac{\eta_{\omega,\ell}(t)e^{ij\omega t}}{i\omega} - \frac{\eta_{\omega,\ell}(0)}{i\omega} - \int_0^t \frac{\eta'_{\omega,\ell}(s)e^{ij\omega s}}{i\omega}\,ds \right\} \\
&\quad + \sum_{n=3}^{r} j^{-\frac{1}{n}} \sum_{\omega\in\Omega(n,\ell)} \left\{ \frac{\eta_\omega(t)e^{ij\omega t}}{i\omega} - \frac{\eta_\omega(0)}{i\omega} - \int_0^t \frac{\eta'_\omega(s)e^{ij\omega s}}{i\omega}\,ds \right\}.
\end{aligned}$$

We then let

$$\begin{aligned}
V_\ell(t) &= \int_0^t \eta_{\ell,0}(s)\,ds, \\
R_\ell^j(t) &= -j^{-\frac{1}{2}} \sum_{\omega\in\Omega(2,\ell)} \left\{ \frac{\eta_{\omega,\ell}(0)}{i\omega} + \int_0^t \frac{\eta'_{\omega,\ell}(s)}{i\omega} e^{ij\omega s}\,ds \right\} \\
&\quad - \sum_{n=3}^{r} j^{-\frac{1}{n}} \sum_{\omega\in\Omega(n,\ell)} \left\{ \frac{\eta_\omega(0)}{i\omega} + \int_0^t \frac{\eta'_\omega(s)}{i\omega} e^{ij\omega s}\,ds \right\}, \\
V_\ell^j(t) &= V_\ell(t) + R_\ell^j(t)
\end{aligned}$$

and define

$$\widetilde{UV}_\ell^j(t) = V_\ell^j(t) - U_\ell^j(t) = -j^{-\frac{1}{2}} \sum_{\omega\in\Omega(2,\ell)} \frac{\eta_{\omega,\ell}(t)}{i\omega} e^{ij\omega t} - \sum_{n=3}^{r} j^{-\frac{1}{n}} \sum_{\omega\in\Omega(n,\ell)} \frac{\eta_\omega(t)}{i\omega} e^{ij\omega t}.$$

Now in order to show the idea of how the general V_I^j and $\widetilde{UV}_I^j$ are defined, let us proceed one more step and write down V_{ℓ_1,ℓ_2}^j and $\widetilde{UV}_{\ell_1,\ell_2}^j$ explicitly. If we multiply $\widetilde{UV}_{\ell_2}^j(t)$ by $u_{\ell_1}^j(t)$ and integrate we get

$$\int_0^t u_{\ell_1}^j(s)\widetilde{UV}_{\ell_2}^j(s)\,ds = A_{\ell_1,\ell_2}^j + B_{\ell_1,\ell_2}^j + C_{\ell_1,\ell_2}^j,$$

where

$$\begin{aligned}
A_{\ell_1,\ell_2}^j &= -j^{-\frac{1}{2}} \sum_{\omega\in\Omega(2,\ell_2)} \int_0^t \frac{\eta_{\ell_1,0}(s)\eta_{\omega,\ell_2}(s)}{i\omega} e^{ij\omega s}\,ds \\
&\quad - \sum_{n=3}^{r} j^{-\frac{1}{n}} \sum_{\omega\in\Omega(n,\ell_2)} \int_0^t \frac{\eta_{\ell_1,0}(s)\eta_\omega(s)}{i\omega} e^{ij\omega s}\,ds, \\
B_{\ell_1,\ell_2}^j &= - \sum_{(\omega_1,\omega_2)\in\Omega(2,\ell_1)\times\Omega(2,\ell_2)} \int_0^t \frac{\eta_{\omega_1,\ell_1}(s)\eta_{\omega_2,\ell_2}(s)}{i\omega_2} e^{ij(\omega_1+\omega_2)s}\,ds, \\
C_{\ell_1,\ell_2}^j &= - \sum_{n=3}^{r} j^{\frac{1}{2}-\frac{1}{n}} \sum_{(\omega_1,\omega_2)\in\Omega(2,\ell_1)\times\Omega(n,\ell_2)} \int_0^t \frac{\eta_{\omega_1,\ell_1}(s)\eta_{\omega_2}(s)}{i\omega_2} e^{ij(\omega_1+\omega_2)s}\,ds
\end{aligned}$$

$$-\sum_{n=3}^{r} j^{\frac{1}{2}-\frac{1}{n}} \sum_{(\omega_1,\omega_2)\in\Omega(n,\ell_1)\times\Omega(2,\ell_2)} \int_0^t \frac{\eta_{\omega_1}(s)\eta_{\omega_2,\ell_2}(s)}{i\omega_2} e^{ij(\omega_1+\omega_2)s}\,ds$$

$$-\sum_{n_1,n_2=3}^{r} j^{1-\frac{1}{n_1}-\frac{1}{n_2}} \sum_{(\omega_1,\omega_2)\in\Omega(n_1,\ell_1)\times\Omega(n_2,\ell_2)} \int_0^t \frac{\eta_{\omega_1}(s)\eta_{\omega_2}(s)}{i\omega_2} e^{ij(\omega_1+\omega_2)s}\,ds.$$

Clearly, the $A^j_{\ell_1,\ell_2}$ converge to 0 uniformly and have uniformly bounded derivatives in j. In $B^j_{\ell_1,\ell_2}$, the terms with $\omega_1+\omega_2\neq 0$ go to 0 uniformly and have uniformly bounded derivatives. These terms will be moved to $R^j_{\ell_1,\ell_2}$. The terms in $B^j_{\ell_1,\ell_2}$ that correspond to $\omega_1+\omega_2=0$ give rise to V_{ℓ_1,ℓ_2}. Precisely, we let

$$V_{\ell_1,\ell_2}(t) = -\sum_{\substack{(\omega_1,\omega_2)\in\Omega(2,\ell_1)\times\Omega(2,\ell_2)\\ \omega_1+\omega_2=0}} \int_0^t \frac{\eta_{\omega_1,\ell_1}(s)\eta_{\omega_2,\ell_2}(s)}{i\omega_2}\,ds.$$

In $C^j_{\ell_1,\ell_2}$ for those terms with $\omega_1+\omega_2\neq 0$, we can use integration by parts to bring a factor j^{-1}. We then move the terms coming from the integration by parts that converge to 0 uniformly and have uniformly bounded derivatives in L^1 to $R^j_{\ell_1,\ell_2}$. For example, for the terms

$$j^{\frac{1}{2}-\frac{1}{n}} \int_0^t \frac{\eta_{\omega_1,\ell_1}(s)\eta_{\omega_2}(s)}{i\omega_2} e^{ij(\omega_1+\omega_2)s}\,ds$$

in the first summation in the right-hand side of $C^j_{\ell_1,\ell_2}$, we know that $\omega_1+\omega_2\neq 0$. Via integration by parts we get

$$j^{\frac{1}{2}-\frac{1}{n}} \int_0^t \frac{\eta_{\omega_1,\ell_1}(s)\eta_{\omega_2}(s)}{i\omega_2} e^{ij(\omega_1+\omega_2)s}\,ds = j^{-\frac{1}{2}-\frac{1}{n}} \frac{\eta_{\omega_1,\ell_1}(t)\eta_{\omega_2}(t)}{i^2\omega_2(\omega_1+\omega_2)} e^{ij(\omega_1+\omega_2)t} + D^j_{\omega_1,\omega_2},$$

where

$$D^j_{\omega_1,\omega_2} = -j^{-\frac{1}{2}-\frac{1}{n}} \frac{\eta_{\omega_1,\ell_1}(0)\eta_{\omega_2}(0)}{i^2\omega_2(\omega_1+\omega_2)} - j^{-\frac{1}{2}-\frac{1}{n}} \int_0^t \frac{(\eta_{\omega_1,\ell_1}(s)\eta_{\omega_2}(s))'}{i^2\omega_2(\omega_1+\omega_2)} e^{ij(\omega_1+\omega_2)s}\,ds.$$

The $D^j_{\omega_1,\omega_2}$ go to zero uniformly and have uniformly bounded derivatives. We then move them to $R^j_{\ell_1,\ell_2}$. Similarly, we can take care of the terms in the second summation and the terms with $\omega_1+\omega_2\neq 0$ in the third summation in the right-hand side of $C^j_{\ell_1,\ell_2}$. The terms in $C^j_{\ell_1,\ell_2}$ that correspond to $\omega_1+\omega_2=0$ add up to 0. Precisely,

$$\sum_{n_1,n_2=3}^{r} j^{1-\frac{1}{n_1}-\frac{1}{n_2}} \sum_{\substack{(\omega_1,\omega_2)\in\Omega(n_1,\ell_1)\times\Omega(n_2,\ell_2)\\ \omega_1+\omega_2=0}} \int_0^t \frac{\eta_{\omega_1}(s)\eta_{\omega_2}(s)}{i\omega_2}\,ds = 0.$$

This is because, by the symmetry of $\Omega(n,k)$, if $\frac{\eta_{\omega_1}\eta_{-\omega_1}}{i\omega_1} = \frac{|\eta_{\omega_1}|^2}{i\omega_1}$ is in the summation, then $\frac{\eta_{-\omega_1}\eta_{\omega_1}}{-i\omega_1} = -\frac{|\eta_{\omega_1}|^2}{i\omega_1}$ is also in the summation. So they add up to 0. We then let

$$R^j_{\ell_1,\ell_2}(t) = A^j_{\ell_1,\ell_2} - \sum_{\substack{(\omega_1,\omega_2)\in\Omega(2,\ell_1)\times\Omega(2,\ell_2)\\ \omega_1+\omega_2\neq 0}} \int_0^t \frac{\eta_{\omega_1,\ell_1}(s)\eta_{\omega_2,\ell_2}(s)}{i\omega_2} e^{ij(\omega_1+\omega_2)s}\,ds$$

$$+\sum_{n=3}^{r} j^{-\frac{1}{2}-\frac{1}{n}} \sum_{(\omega_1,\omega_2)\in\Omega(2,\ell_1)\times\Omega(n,\ell_2)} \left\{ \frac{\eta_{\omega_1,\ell_1}(0)\eta_{\omega_2}(0)}{i^2\omega_2(\omega_2+\omega_1)} + \int_0^t \frac{(\eta_{\omega_1,\ell_1}(s)\eta_{\omega_2}(s))'}{i^2\omega_2(\omega_2+\omega_1)} e^{ij(\omega_1+\omega_2)s} ds \right\}$$

$$+\sum_{n=3}^{r} j^{-\frac{1}{2}-\frac{1}{n}} \sum_{(\omega_1,\omega_2)\in\Omega(n,\ell_1)\times\Omega(2,\ell_2)} \left\{ \frac{\eta_{\omega_1}(0)\eta_{\omega_2,\ell_2}(0)}{i^2\omega_2(\omega_2+\omega_1)} + \int_0^t \frac{(\eta_{\omega_1}(s)\eta_{\omega_2,\ell_2}(s))'}{i^2\omega_2(\omega_2+\omega_1)} e^{ij(\omega_1+\omega_2)s} ds \right\}$$

$$+ \sum_{n_1,n_2=3}^{r} j^{-\frac{1}{n_1}-\frac{1}{n_2}} \sum_{\substack{(\omega_1,\omega_2)\in\Omega(n_1,\ell_1)\times\Omega(n_2,\ell_2)\\ \omega_1+\omega_2\neq 0}} \left\{ \frac{\eta_{\omega_1}(0)\eta_{\omega_2}(0)}{i^2\omega_2(\omega_2+\omega_1)} + \int_0^t \frac{(\eta_{\omega_1}(s)\eta_{\omega_2}(s))'}{i^2\omega_2(\omega_2+\omega_1)} e^{ij(\omega_1+\omega_2)s}\, ds \right\},$$

$V^j_{\ell_1,\ell_2}(t) = V_{\ell_1,\ell_2}(t) + R^j_{\ell_1,\ell_2}(t)$.

So we define

$$\begin{aligned}
\widetilde{UV}^j_{\ell_1,\ell_2}(t) &= V^j_{\ell_1,\ell_2}(t) - \int_0^t u^j_{\ell_1}(s)\widetilde{UV}^j_{\ell_2}(s)\, ds \\
&= \sum_{n=3}^{r} j^{-\frac{1}{2}-\frac{1}{n}} \sum_{(\omega_1,\omega_2)\in\Omega(2,\ell_1)\times\Omega(n,\ell_2)} \frac{\eta_{\omega_1,\ell_1}(t)\eta_{\omega_2}(t)}{i^2\omega_2(\omega_2+\omega_1)} e^{ij(\omega_1+\omega_2)t} \\
&\quad + \sum_{n=3}^{r} j^{-\frac{1}{2}-\frac{1}{n}} \sum_{(\omega_1,\omega_2)\in\Omega(n,\ell_1)\times\Omega(2,\ell_2)} \frac{\eta_{\omega_1}(t)\eta_{\omega_2,\ell_2}(t)}{i^2\omega_2(\omega_2+\omega_1)} e^{ij(\omega_1+\omega_2)t} \\
&\quad + \sum_{n_1,n_2=3}^{r} j^{-\frac{1}{n_1}-\frac{1}{n_2}} \sum_{\substack{(\omega_1,\omega_2)\in\Omega(n_1,\ell_1)\times\Omega(n_2,\ell_2)\\ \omega_1+\omega_2\neq 0}} \frac{\eta_{\omega_1}(t)\eta_{\omega_2}(t)}{i^2\omega_2(\omega_2+\omega_1)} e^{ij(\omega_1+\omega_2)t}.
\end{aligned}$$

Clearly the $\widetilde{UV}^j_{\ell_1,\ell_2}(t)$ are of the form (30).

Assume that we have defined $V^j_I, \widetilde{UV}^j_I$ such that $V^j_I, \widetilde{UV}^j_I$ are as in (29) and (30) for $|I|$ up to $|I| \le k-1, 3 \le k \le r$. Let $I = (\ell_1, \dots, \ell_k)$. For simplicity we introduce the following notations:

$$\begin{aligned}
\bar{I} &= (\ell_2, \dots, \ell_k), \\
h(\hat{\omega}) &= \omega_k(\omega_k+\omega_{k-1})\cdots(\omega_k+\cdots+\omega_1) \quad \text{if} \quad \hat{\omega} = (\omega_1, \dots, \omega_k).
\end{aligned}$$

Multiplying $u^j_{\ell_1}$ by $\widetilde{UV}^j_{\bar{I}}$ and integrating we get

$$\int_0^t u^j_{\ell_1}(s)\widetilde{UV}^j_{\bar{I}}(s)\, ds = \int_0^t \eta_{\ell_1,0}(s)\widetilde{UV}^j_{\bar{I}}(s)\, ds + A^j_I + B^j_I + C^j_I + D^j_I.$$

Here

$$A_I^j = (-1)^{k-1} \sum_{\hat{n}\in\Omega_1(k-1)} j^{\frac{1}{2}-\alpha_{\hat{n}}} \sum_{(\omega_1,\hat{\omega})\in\Omega(2,\ell_1)\times\Omega(\hat{n},\bar{I})} \int_0^t \frac{\eta_{\hat{\omega}}(s)\eta_{\omega_1,\ell_1}(s)e^{ij(\omega_1+\sum\hat{\omega})s}}{i^{k-1}h(\hat{\omega})} ds,$$

$$B_I^j = (-1)^{k-1} \sum_{\hat{n}\in\Omega_2(k-1)} j^{\frac{1}{2}-\alpha_{\hat{n}}} \sum_{(\omega_1,\hat{\omega})\in\Omega(2,\ell_1)\times\Omega(\hat{n},\bar{I})} \int_0^t \frac{\eta_{\hat{\omega}}(s)\eta_{\omega_1,\ell_1}(s)e^{ij(\omega_1+\sum\hat{\omega})s}}{i^{k-1}h(\hat{\omega})} ds,$$

$$C_I^j = (-1)^{k-1} \sum_{n_1=3}^{r} \sum_{\hat{n}\in\Omega_1(k-1)} j^{1-\frac{1}{n_1}-\alpha_{\hat{n}}} \sum_{(\omega_1,\hat{\omega})\in\Omega(n_1,\ell_1)\times\Omega(\hat{n},\bar{I})} \int_0^t \frac{\eta_{\hat{\omega}}(s)\eta_{\omega_1}(s)e^{ij(\omega_1+\sum\hat{\omega})s}}{i^{k-1}h(\hat{\omega})} ds,$$

$$D_I^j = (-1)^{k-1} \sum_{n_1=3}^{r} \sum_{\hat{n}\in\Omega_2(k-1)} j^{1-\frac{1}{n_1}-\alpha_{\hat{n}}} \sum_{(\omega_1,\hat{\omega})\in\Omega(n_1,\ell_1)\times\Omega(\hat{n},\bar{I})} \int_0^t \frac{\eta_{\omega_1}(s)\eta_{\hat{\omega}}(s)e^{ij(\omega_1+\sum\hat{\omega})s}}{i^{k-1}h(\hat{\omega})} ds.$$

Notice that for the terms in A_I^j, we know that $\frac{1}{2}-\alpha_{\hat{n}} < 0$. So the A_I^j converge to 0 and have uniformly bounded derivatives. For all the terms in B_I^j, by the independence of the sets $\Omega_{E,\rho}$, we know that $\omega_1+\sum\hat{\omega} \neq 0$. So in B_I^j, if $\frac{1}{2}-\alpha_{\hat{n}} \leq 0$, the corresponding terms converge to 0 uniformly and have uniformly bounded derivatives already. For the terms with $\frac{1}{2}-\alpha_{\hat{n}}$ positive, we apply integration by parts to get a j^{-1}. We then move those terms coming from the integration by parts that converge to 0 and have uniformly bounded derivatives to R_I^j. This is done in exactly the same way as we take care of the terms in the first two summations in the right hand of C_{ℓ_1,ℓ_2}^j. The C_I^j are taken care of in exactly the same way as for B_I^j. Finally we take care of the terms in D_I^j by the following:

(i) We move all the terms in D_I^j whose j powers are negative or equal 0 but $\omega_1+\hat{\omega} \neq 0$ to R_I^j.

(ii) For the terms with j powers positive and $\omega_1+\hat{\omega} \neq 0$, we apply integration by parts and then move all the terms coming from the integration by parts that go to 0 and have uniformly bounded derivatives to R_I^j.

This takes care of the terms in D_I^j with $\omega_1+\sum\hat{\omega} \neq 0$ and those with j powers negative. For the terms in D_I^j with $\omega_1+\sum\hat{\omega}=0$ and j powers ≥ 0, there are two possibilities:

(iii) $\omega_1+\sum\hat{\omega}=0$ and the j powers $1-\frac{1}{n_1}-\alpha_{\hat{n}}=0$. These terms give rise to the V_I; i.e., we let

$$V_I(t) = (-1)^{k-1} \sum_{n_1=3}^{r} \sum_{\hat{n}\in\Omega_2(k-1)} \sum_{\substack{(\omega_1,\hat{\omega})\in\Omega(n_1,\ell_1)\times\Omega(\hat{n},\bar{I}) \\ \omega_1+\sum\hat{\omega}=0 \\ 1-\frac{1}{n_1}-\alpha_{\hat{n}}=0}} \int_0^t \frac{\eta_{\omega_1}(s)\eta_{\hat{\omega}}(s)}{i^{k-1}h(\hat{\omega})} ds\,;$$

(iv) $\omega_1+\sum\hat{\omega}=0$ and the j powers $1-\frac{1}{n_1}-\alpha_{\hat{n}}>0$.

In order to take care of the terms in (iv), we first notice the following fact. Let $(\omega_1,\ldots,\omega_k)$ be a k tuple of numbers with $\{\omega_1,\ldots,\omega_k\} \subset \cup_{\ell=1}^{m}\cup_{n=2}^{r}\Omega(n,\ell)$. Because of the linear independence of the $\Omega_{E,\rho}$ and the minimal cancelation requirement of each $F\in Q_{E,\rho}$, $\omega_1+\cdots+\omega_k=0$ is possible only in the following cases:

(a) k is even and each ω_ℓ is canceled out by its negative $-\omega_\ell$. This case will be referred to as *pure cancelation by pairs*;

(b) $\omega_1 + \cdots + \omega_k = 0$ because the set $\{\omega_1, \ldots, \omega_k\}$ is equal to some $F \in Q_{E,\rho}$ for some E and ρ. (Note that each $F \in Q_{E,\rho}$ is canceling.) This case will be referred to as *pure cancelation by F*;

(c) *mixed cancelation*; i.e., some of the ω_ℓ are canceled out by $-\omega_\ell$, some others are canceled out because the set of them is equal to some $F \in Q_{E,\rho}$.

Let $\bar{D}_I^j$ denote the sum of the terms in D_I^j for which (iv) happens; i.e., let

$$\bar{D}_I^j = (-1)^{k-1} \sum_{n_1=3}^{r} \sum_{\hat{n} \in \Omega_2(k-1)} j^{1-\frac{1}{n_1}-\alpha_{\hat{n}}} \sum_{\substack{(\omega_1,\hat{\omega}) \in \Omega(n_1,\ell_1)\times\Omega(\hat{n},\bar{I}) \\ \omega_1+\sum\hat{\omega}=0 \\ 1-\frac{1}{n_1}-\alpha_{\hat{n}}>0}} \int_0^t \frac{\eta_{\omega_1}(s)\eta_{\hat{\omega}}(s)}{i^{k-1}h(\hat{\omega})}\, ds.$$

Now the two conditions $1-\frac{1}{n_1}-\alpha_{\hat{n}} > 0$, $\omega_1+\sum\hat{\omega} = 0$ imply that the $(\omega_1, \hat{\omega})$ is canceled out purely by pairs. (Notice that if the $(\omega_1, \hat{\omega}) \in \Omega(n_1, \ell_1) \times \Omega(\hat{n}, \bar{I})$ is canceled by an $F \in Q_{E,\rho}$, then $1 - \frac{1}{n_1} - \alpha_{\hat{n}} \le 0$. Similarly, since we require that the $\alpha_{\hat{n}}$ be less than 1, the mixed cancelations cannot occur too.) In particular $|I| = k$ has to be even. But now, by the symmetry of the $\Omega(\hat{n}, \bar{I})$ and $\Omega(n_1, \ell_1)$, we know that if $\int_0^t \frac{\eta_{\omega_1}(s)\eta_{\hat{\omega}}(s)}{i^{k-1}h(\hat{\omega})}\, ds$ is in $\bar{D}_I^j$, then, since $\eta_{-\omega_1}\eta_{-\hat{\omega}} = \eta_{\omega_1}\eta_{\hat{\omega}}, h(-\hat{\omega}) = (-1)^{k-1}h(\hat{\omega}) = -h(\hat{\omega})$, the integral

$$\int_0^t \frac{\eta_{-\omega_1}\eta_{-\hat{\omega}}(s)}{i^{k-1}h(-\hat{\omega})}\, ds = -\int_0^t \frac{\eta_{\omega_1}(s)\eta_{\hat{\omega}}(s)}{i^{k-1}h(\hat{\omega})}\, ds$$

is also in $\bar{D}_I^j$. So they add up to 0; i.e. the contribution of each term and that of its negative cancel. So we get that $\bar{D}_I^j = 0$. Summarizing the above we define

$$\begin{aligned}
R_I^j(t) &= \int_0^t \eta_{\ell_1,0}(s)\widetilde{UV}_{\bar{I}}^j(s)\, ds + A_I^j \\
&+(-1)^{k-1} \sum_{\hat{n} \in \Omega_2(k-1)} j^{\frac{1}{2}-\alpha_{\hat{n}}} \sum_{\substack{(\omega_1,\hat{\omega}) \in \Omega(2,\ell_1)\times\Omega(\hat{n},\bar{I}) \\ \frac{1}{2}-\alpha_{\hat{n}} \le 0}} \int_0^t \frac{\eta_{\omega_1,\ell_1}(s)\eta_{\hat{\omega}}(s)e^{ij(\omega_1+\sum\hat{\omega})s}}{i^{k-1}h(\hat{\omega})}\, ds \\
&+(-1)^{k} \sum_{\hat{n} \in \Omega_2(k-1)} j^{-\frac{1}{2}-\alpha_{\hat{n}}} \sum_{\substack{(\omega_1,\hat{\omega}) \in \Omega(2,\ell_1)\times\Omega(\hat{n},\bar{I}) \\ \frac{1}{2}-\alpha_{\hat{n}} > 0}} \left\{ \frac{\eta_{\omega_1,\ell_1}(0)\eta_{\hat{\omega}}(0)}{i^k(\omega_1+\sum\hat{\omega})h(\hat{\omega})} \right. \\
&\qquad \left. + \int_0^t \frac{(\eta_{\omega_1,\ell_1}(s)\eta_{\hat{\omega}}(s))' e^{ij(\omega_1+\sum\hat{\omega})s}}{i^k(\omega_1+\sum\hat{\omega})h(\hat{\omega})}\, ds \right\} \\
&+(-1)^{k-1} \sum_{n_1=3}^{r} \sum_{\hat{n} \in \Omega_1(k-1)} j^{\frac{n_1-1}{n_1}-\alpha_{\hat{n}}} \sum_{\substack{(\omega_1,\hat{\omega}) \in \Omega(n_1,\ell_1)\times\Omega(\hat{n},\bar{I}) \\ 1-\frac{1}{n_1}-\alpha_{\hat{n}} \le 0}} \int_0^t \frac{\eta_{\omega_1}(s)\eta_{\hat{\omega}}(s)e^{ij(\omega_1+\sum\hat{\omega})s}}{i^{k-1}h(\hat{\omega})}\, ds
\end{aligned}$$

$$+(-1)^k \sum_{n_1=3}^{r} \sum_{\hat{n}\in\Omega_1(k-1)} j^{-\frac{1}{n_1}-\alpha_{\hat{n}}} \sum_{\substack{(\omega_1,\hat{\omega})\in\Omega(n_1,\ell_1)\times\Omega(\hat{n},\bar{I})\\ 1-\frac{1}{n_1}-\alpha_{\hat{n}}>0}} \left\{ \frac{\eta_{\omega_1}(0)\eta_{\hat{\omega}}(0)}{i^k(\omega_1+\sum\hat{\omega})h(\hat{\omega})} \right.$$

$$\left. + \int_0^t \frac{(\eta_{\omega_1}(s)\eta_{\hat{\omega}}(s))' e^{ij(\omega_1+\sum\hat{\omega})s}}{i^k(\omega_1+\sum\hat{\omega})h(\hat{\omega})}\, ds \right\}$$

$$+(-1)^k \sum_{n_1=3}^{r} \sum_{\hat{n}\in\Omega_2(k-1)} j^{-\frac{1}{n_1}-\alpha_{\hat{n}}} \sum_{\substack{(\omega_1,\hat{\omega})\in\Omega(n_1,\ell_1)\times\Omega(\hat{n},\bar{I})\\ \omega_1+\sum\hat{\omega}\neq 0\\ 1-\frac{1}{n_1}-\alpha_{\hat{n}}>0}} \left\{ \frac{\eta_{\omega_1}(0)\eta_{\hat{\omega}}(0)}{i^k(\omega_1+\sum\hat{\omega})h(\hat{\omega})} \right.$$

$$\left. + \int_0^t \frac{(\eta_{\omega_1}(s)\eta_{\hat{\omega}}(s))' e^{ij(\omega_1+\sum\hat{\omega})s}}{i^k(\omega_1+\sum\hat{\omega})h(\hat{\omega})}\, ds \right\}$$

$$+(-1)^{k-1} \sum_{n_1=3}^{r} \sum_{\hat{n}\in\Omega_2(k-1)} \sum_{\substack{(\omega_1,\hat{\omega})\in\Omega(n_1,\ell_1)\times\Omega(\hat{n},\bar{I})\\ \omega_1+\sum\hat{\omega}\neq 0\\ 1-\frac{1}{n_1}-\alpha_{\hat{n}}=0}} \int_0^t \frac{\eta_{\omega_1}(s)\eta_{\hat{\omega}}(s)e^{ij(\omega_1+\sum\hat{\omega})s}}{i^{k-1}h(\hat{\omega})}\, ds$$

$$+(-1)^{k-1} \sum_{n_1=3}^{r} \sum_{\hat{n}\in\Omega_2(k-1)} j^{1-\frac{1}{n_1}-\alpha_{\hat{n}}} \sum_{\substack{(\omega_1,\hat{\omega})\in\Omega(n_1,\ell_1)\times\Omega(\hat{n},\bar{I})\\ 1-\frac{1}{n_1}-\alpha_{\hat{n}}<0}} \int_0^t \frac{\eta_{\omega_1}(s)\eta_{\hat{\omega}}(s)e^{ij(\omega_1+\sum\hat{\omega})s}}{i^{k-1}h(\hat{\omega})}\, ds.$$

We then define $V_I^j = V_I + R_I^j$. So we have

$$\widetilde{UV}_I^j(t) = V_I^j(t) - \int_0^t u_{\ell_1}^j(s)\widetilde{UV}_{\bar{I}}^j(s)\, ds$$

$$= (-1)^k \sum_{\hat{n}\in\Omega_2(k-1)} j^{-\frac{1}{2}-\alpha_{\hat{n}}} \sum_{\substack{(\omega_1,\hat{\omega})\in\Omega(2,\ell_1)\times\Omega(\hat{n},\bar{I})\\ \frac{1}{2}-\alpha_{\hat{n}}>0}} \frac{\eta_{\omega_1,\ell_1}(t)\eta_{\hat{\omega}}(t)e^{ij(\omega_1+\sum\hat{\omega})t}}{i^k(\omega_1+\sum\hat{\omega})h(\hat{\omega})}$$

$$+(-1)^k \sum_{n_1=3}^{r} \sum_{\hat{n}\in\Omega_1(k-1)} j^{-\frac{1}{n_1}-\alpha_{\hat{n}}} \sum_{\substack{(\omega_1,\hat{\omega})\in\Omega(n_1,\ell_1)\times\Omega(\hat{n},\bar{I})\\ 1-\frac{1}{n_1}-\alpha_{\hat{n}}>0}} \frac{\eta_{\omega_1}(t)\eta_{\hat{\omega}}(t)e^{ij(\omega_1+\sum\hat{\omega})t}}{i^k(\omega_1+\sum\hat{\omega})h(\hat{\omega})}$$

$$+(-1)^k \sum_{n_1=3}^{r} \sum_{\hat{n}\in\Omega_2(k-1)} j^{-\frac{1}{n_1}-\alpha_{\hat{n}}} \sum_{\substack{(\omega_1,\hat{\omega})\in\Omega(n_1,\ell_1)\times\Omega(\hat{n},\bar{I})\\ 1-\frac{1}{n_1}-\alpha_{\hat{n}}>0\\ \omega_1+\sum\hat{\omega}\neq 0}} \frac{\eta_{\omega_1}(t)\eta_{\hat{\omega}}(t)e^{ij(\omega_1+\sum\hat{\omega})t}}{i^k(\omega_1+\sum\hat{\omega})h(\hat{\omega})}.$$

The $\widetilde{UV}_I^j$ are clearly of the form (30).

To finish the inductive step, we have to show that the V_I defined above have the form (29). Since

$$V_I(t) = (-1)^{k-1} \sum_{n_1=3}^{r} \sum_{\hat{n}\in\Omega_2(k-1)} \sum_{\substack{(\omega_1,\hat{\omega})\in\Omega(n_1,\ell_1)\times\Omega(\hat{n},\bar{I}) \\ \omega_1+\sum\hat{\omega}=0 \\ 1-\frac{1}{n_1}-\alpha_{\hat{n}}=0}} \int_0^t \frac{\eta_{\omega_1}(s)\eta_{\hat{\omega}}(s)}{i^{k-1}h(\hat{\omega})}\,ds,$$

for exactly the same reason as above, the two equalities $1-\frac{1}{n_1}-\alpha_{\hat{n}}=0$, $\omega_1+\sum\hat{\omega}=0$ imply that each $(\omega_1,\hat{\omega})$ is canceled out either purely by pairs or purely by some F. The mixed cancelation cannot happen. If the $(\omega_1,\hat{\omega})$ in the summation is canceled out purely by pairs, then k must be even and $(-\omega_1,-\hat{\omega})$ is also in the summation, so they add up to zero. So only those terms whose frequencies are canceled out purely by F will contribute to V_I. Moreover, if the frequencies are canceled purely by an $F\in Q_{E,\rho}$, then $\delta(E)$ must be equal to $|I|=k$. Precisely for $I=(\ell_1,\dots,\ell_k)$, V_I is given by

$$V_I = (-1)^{k-1} \sum_{E\in\mathcal{E}_k} \sum_{\rho=1}^{|E|} \sum_{F\in Q_{E,\rho}} \sum_{\hat{\omega}\in\Omega(F,I)} \int_0^t \frac{\eta_{\hat{\omega}}(s)}{i^{k-1}\omega_k(\omega_k+\omega_{k-1})\cdots(\omega_k+\cdots+\omega_2)}\,ds, \tag{31}$$

where $\Omega(F,I)$ is a subset of $\Omega(k,\ell_1)\times\cdots\times\Omega(k,\ell_k)$ such that $\hat{\omega}=(\omega_1,\dots,\omega_k)\in\Omega(F,I)$ iff as a set $\{\omega_1,\dots,\omega_k\}=F$, so, V_I can be written in form (29) for $3\le|I|\le r$. This finishes the recursive definition of V_I^j and $\widetilde{UV}_I^j$.

Clearly (A) and (B) imply that the $\mathbf{u}^j$ EI(r)-converge to $\mathbf{v}^\infty=\sum_{0<|I|\le r} v_I X_I$, where $v_I=\dot{V}_I$. To finish the proof of Theorem 5.1, we need to show that $\mathbf{u}^\infty=\mathbf{v}^\infty$. For each I and F, let $g(F,I)$ be the number $g(F,I)=\sum_{\hat{\omega}\in\Omega(F,I)}\frac{1}{\omega_k(\omega_k+\omega_{k-1})\dots(\omega_k+\cdots+\omega_2)}$. Then we have

$$v_I = (-1)^{k-1} \sum_{E\in\mathcal{E}_k} \sum_{\rho=1}^{|E|} \sum_{F\in Q_{E,\rho}} \frac{g(F,I)}{i^{k-1}} \prod_{\omega\in F} \eta_\omega(t). \tag{32}$$

In order to get $\mathbf{v}^\infty$ explicitly we need to determine the constants $g(F,I)$. From the expression of V_I in (31) we notice that the contributions to $\mathbf{v}^\infty$ from the terms in u_k^j with frequencies in different $F\in Q_{E,\rho}$ are *independent*. Precisely, for $\delta(E)\ge 3, F\in Q_{E,\rho}$, let

$$u_{k,F,E,\rho}^j(t) = j^{\frac{\delta(E)-1}{\delta(E)}} \sum_{\omega\in F\cap\Omega_{E,\rho,k}} \eta_\omega(t) e^{ij\omega t}, \qquad k=1,\dots,m,$$
$$u_{F,E,\rho}^j = (u_{1,F,E,\rho}^j,\dots,u_{m,F,E,\rho}^j),$$

and for $\delta(E)=2$, let $u_{k,E}^j(t)=j^{\frac{1}{2}}\sum_{\omega\in\Omega_{E,1,k}}\eta_{\omega,k}(t)e^{ij\omega t}, k=1,\dots,m$, $u_E^j=(u_{1,E}^j,\dots,u_{m,E}^j)$. Let $\mathbf{v}_{F,E,\rho}^\infty$ and $\mathbf{v}_E^\infty$ be the limiting extended inputs of the $u_{F,E,\rho}^j$ and u_E^j, respectively. We have

$$\mathbf{v}^\infty(t) = \sum_{\ell=1}^{m} \eta_{\ell,0}(t)\,X_\ell + \sum_{E\in\mathcal{E}_2} \mathbf{v}_E^\infty(t) + \sum_{n=3}^{r} \sum_{E\in\mathcal{E}_n} \sum_{\rho=1}^{|E|} \sum_{F\in Q_{E,\rho}} \mathbf{v}_{F,E,\rho}^\infty(t).$$

This is the high-frequency superposition principle we mentioned in the beginning of section 4. So in order to figure out the explicit formula of $\mathbf{v}^\infty$, we need only find out each $\mathbf{v}^\infty_{F,E,\rho}$ and $\mathbf{v}^\infty_E$.

For $\delta(E)=2$ we have $\mathbf{v}^\infty_E = -v_{\ell_1,\ell_2}\ X_{\ell_1}X_{\ell_2} - v_{\ell_2,\ell_1}\ X_{\ell_2}X_{\ell_1}$, if $E=[X_{\ell_1},X_{\ell_2}]$. Since

$$v_{\ell_1,\ell_2}(t) = -\sum_{\substack{(\omega_1,\omega_2)\in\Omega(2,\ell_1)\times\Omega(2,\ell_2)\\ \omega_1+\omega_2=0}} \frac{\eta_{\omega_1,\ell_1}(t)\eta_{\omega_2,\ell_2}(t)}{i\omega_2},$$

we see that

$$\mathbf{v}^\infty_{[X_{\ell_1},X_{\ell_2}]}(t) = v_{[X_{\ell_1},X_{\ell_2}]}(t)[X_{\ell_1},X_{\ell_2}], \tag{33}$$

where

$$v_{[X_{\ell_1},X_{\ell_2}]}(t) = \frac{1}{i\omega_{\ell_1,\ell_2}}\left(\eta_{\omega_{\ell_1,\ell_2},\ell_1}(t)\eta_{-\omega_{\ell_1,\ell_2},\ell_2}(t) - \eta_{-\omega_{\ell_1,\ell_2},\ell_1}(t)\eta_{\omega_{\ell_1,\ell_2},\ell_2}(t)\right).$$

For $F\in Q_{E,\rho}$, $\delta(E)\ge 3$, from (32) we have

$$\mathbf{v}^\infty_{F,E,\rho} = (-1)^{\delta(E)-1}\sum_{0<|I|\le r}\frac{g(F,I)}{i^{|I|-1}}\prod_{\omega\in F}\eta_\omega(t)\,X_I.$$

It is easy to see that for each $F\in Q_{E,\rho}$ all the $g(F,I)=0$ except for $I\in\Omega(E)$, where $\Omega(E)$ is the set of multi-indices I such that $\delta_k(I)=\delta_k(E)$ for $k=1,\dots,m$ (here $\delta_k(I)$ denotes the number of occurrence of k in I). (For $F\in Q_{E,\rho}$, if $I\notin\Omega(E)$, then $\Omega(F,I)$ is empty.) Let us write

$$\mathbf{v}^\infty_{F,E,\rho} = (-1)^{\delta(E)-1}\sum_{I\in\Omega(E)}\frac{g(F,I)}{i^{|I|-1}}\prod_{\omega\in F}\eta_\omega(t)\ X_I\ = \sum_{B\in E}\left(v_B^{F,E,\rho}\prod_{\omega\in F}\eta_\omega(t)\right)\ B.$$

In order to figure out the limit $\mathbf{v}^\infty_{F,E,\rho}$ of the $u^j_{F,E,\rho}$ for $\delta(E)>2$, we need only figure out $v_B^{F,E,\rho}$, so we may assume that $\eta_\omega=1$. Namely, we need only find out the limit of $\bar u^j_{F,E,\rho}(t)$, where $\bar u^j_{F,E,\rho}=(\bar u^j_{1,F,E,\rho},\dots,\bar u^j_{m,F,E,\rho})$ are given by $\bar u^j_{k,F,E,\rho}(t) = j^{\frac{\delta(E)-1}{\delta(E)}}\sum_{\omega\in F\cap\Omega_{E,\rho,k}}e^{ij\omega t}$, $\delta(E)\ge 3$. From the discussion of the limit of (15) defined in section 4 we know that, if we let $\xi^F_{B,\rho}=g_B(F\cap\Omega_{E,\rho,1},F\cap\Omega_{E,\rho,2},\dots,F\cap\Omega_{E,\rho,m})$, then $v_B^{F,E,\rho}=i^{1-\delta(E)}\xi^F_{B,\rho}$ for $B\in E$. So we get

$$\mathbf{v}^\infty_{F,E,\rho}(t) = \sum_{B\in E} i^{1-\delta(E)}\xi^F_{B,\rho}\prod_{\omega\in F}\eta_\omega(t)\,B. \tag{34}$$

Combining (33) and (34) we get

$$\begin{aligned}\mathbf{v}^\infty(t) = &\sum_{k=1}^m \eta_{k,0}(t)\ X_k + \sum_{\ell_1<\ell_2}\frac{1}{i\omega_{\ell_1,\ell_2}}\xi_{\ell_1,\ell_2}(t)\ [X_{\ell_1},X_{\ell_2}]\\ &+\sum_{n=3}^r\sum_{E\in\mathcal{E}_n}\sum_{B\in E}\sum_{\rho=1}^{|E|}\sum_{F\in Q_{E,\rho}} i^{1-\delta(E)}\xi^F_{B,\rho}\prod_{\omega\in F}\eta_\omega(t)\ B,\end{aligned}$$

where $\xi_{\ell_1,\ell_2}=\eta_{\omega_{\ell_1,\ell_2},\ell_1}\eta_{-\omega_{\ell_1,\ell_2},\ell_2}-\eta_{\omega_{-\ell_1,\ell_2},\ell_1}\eta_{\omega_{\ell_1,\ell_2},\ell_2}$. So $\mathbf{u}^\infty=\mathbf{v}^\infty$. Now the proof of Theorem 5.1 is complete.

7. The linear independence of g_B. In the previous section we proved Theorem 5.1. In order to justify the algorithm we need to show that it is always possible to choose the frequency sets so that condition (11) holds in addition to the other conditions. If $E \in \mathcal{E}_2$, this is obvious. Take an $E \in \cup_{n=3}^{r}\mathcal{E}_n$. Assume that $\delta(E) = n$, $|E| = N$, and $E = \{B_1, \dots, B_N\}$. By definition we know that each $\hat{\xi}_{B,\rho}$ is equal to $\xi^F_{B,\rho} = g_B(F \cap \Omega_{E,\rho,1}, \dots, F \cap \Omega_{E,\rho,m})$ for some fixed $F \in Q_{E,\rho}$. For each $Q_{E,\rho}$, assume that the fixed set $F \in Q_{E,\rho}$ contains the numbers $\omega_1^{E,\rho}, \dots, \omega_n^{E,\rho}$. Assume also that we have listed the $\omega_1^{E,\rho}, \dots, \omega_n^{E,\rho}$ such that

$$\{\omega_1^{E,\rho}, \dots, \omega_{\delta_1(E)}^{E,\rho}\} = F \cap \Omega_{E,\rho,1}, \dots, \{\omega^{E,\rho}_{\delta_1(E)+\cdots+\delta_{m-1}(E)+1}, \dots, \omega_n^{E,\rho}\} = F \cap \Omega_{E,\rho,m}.$$

Now if we think of $\omega^{E,\rho} = (\omega_1^{E,\rho}, \dots, \omega_n^{E,\rho})$ as a point in $\mathbb{R}^n$, we see that $\xi^F_{B,\rho} = g_B(\omega_1^{E,\rho}, \dots, \omega_n^{E,\rho})$. Take any N points $w^k = (w_1^k, \dots, w_n^k), k = 1, \dots, N$, in $\mathbb{R}^n$. Let $M(w^1, \dots, w^n)$ be the matrix $M(w^1, \dots, w^N) = (g_B(w^\rho))_{B\in E, 1\le\rho\le N}$. We will think of M as a matrix-valued function depending on nN variables $(w^1, \dots, w^N)$, i.e., as a matrix-valued function on $\mathbb{R}^{nN}$. Now let H be the subset of $\mathbb{R}^{nN}$ consisting of the points $(w^1, \dots, w^N)$ such that $w_1^k + \cdots + w_n^k = 0, k = 1, \dots, N$. Then $\det(\hat{\xi}^F_{B,\rho}) \neq 0$ just means that we need to take the point $(\omega^{E,1}, \dots, \omega^{E,N}) \in H$ to be not on the surface $\{\det(M) = 0\} \bigcap H$. If we can show that the determinant $\det(M)$ of M, as a rational function of nN variables, is not identically 0 on H, then $\det(M)$ is not zero on a relatively open dense subset of H. With this fact, if we think of the frequencies as taking from one very large dimensional Euclidean space $\mathbb{R}^\kappa$, then we can regard the conditions such as $Q_{E,\rho}$ being SMC, independent, the invertibility of each matrix $(\hat{\xi}_{B,\rho})_{B\in E, 1\le\rho\le|E|}$, etc., as taken some points in $\mathbb{R}^\kappa$ that are not on some finite number of surfaces which have measure 0 in $\mathbb{R}^\kappa$, so it is possible to choose the sets $\Omega(n,k)$ such that all the conditions 1, 2, 6, 7, 8, 9, 10, 11 hold simultaneously.

So all that is needed is to prove that $\det(M) \not\equiv 0$ on the set H. For a matrix-valued function like $M = (g_B(w^\rho))_{B\in E, 1\le\rho\le N}$, $\det(M) \not\equiv 0$ on H iff the functions $g_B(w), B \in E$, are linearly independent on the subset $\{(w_1, \dots, w_n) \mid w_1 + \cdots + w_n = 0\}$ of $\mathbb{R}^n$. (In the following we will simply call this the subset $w_1 + \cdots + w_n = 0$.) From the observation made about g_B in Remark 4.1 we know that each g_B is a linear combination of some $g_{B^{\mathbf{Y}}}$ with $B^{\mathbf{Y}} \in E^{\mathbf{Y}} = \{B^{\mathbf{Y}} \in \mathcal{B}^{\mathbf{Y}}, \delta_k(B^{\mathbf{Y}}) = 1, k = 1, \dots, n\}$. In order to show that the $g_B, B \in E$, are linearly independent on the set $w_1 + \cdots + w_n = 0$, we first show that the functions $g_B, B \in E^m$, are linearly independent on the set $w_1 + \cdots + w_m = 0$, where $E^m = \{B \in \mathcal{B}_m, \delta_k(B) = 1, k = 1, \dots, m\}$.

We recall that two brackets $B_1, B_2 \in \mathcal{B}$ are equivalent if $\delta_k(B_1) = \delta_k(B_2), k = 1, \dots, m$. We will write $B_1 \simeq B_2$ if B_1 is equivalent to B_2. Let $\mathcal{E}$ be the set of equivalence classes. For $k = 1, \dots, m$, let $\mathcal{E}^k$ be the subset of $\mathcal{E}$ such that $E \in \mathcal{E}^k$ iff $\delta(E) = k, \delta_\ell(E) \le 1, \ell = 1, \dots, m$. In particular, $\mathcal{E}^m$ has only one element, namely E^m. Let $\bar{\mathcal{E}} = \cup_{k=2}^m \mathcal{E}^k$. Let $\mathcal{B}^m$ be the subset of $\mathcal{B}$ that contains all the $B \in \mathcal{B}$ of degree $\delta(B) \le m$ with $\delta_k(B) \le 1, k = 1, \dots, m$. If $B \in \mathcal{B}^m$ is a bracket with $\Sigma_B = X_{\ell_1} \dots X_{\ell_k}$, and $w = (w_1, \dots, w_m) \in \mathbb{R}^m$ is a point, for simplicity, we introduce the following notations: $w_B = (w_{\ell_1}, \dots, w_{\ell_k})$, $\sum w_B = w_{\ell_1} + \cdots + w_{\ell_k}$, $\hat{g}_B(w_B) = \hat{g}_B(w_{\ell_1}, \dots, w_{\ell_k})$. We let Σ_E^0 denote the set $\{w_B \mid \sum w_B = 0\}$ of $\delta(E)$-tuples of numbers. It is clear that the set Σ_E^0 depends only on E and can be identified with the subset $\{(w_1, \dots, w_{\delta(E)}) \mid w_1 + \cdots + w_{\delta(E)} = 0\}$ of $\mathbb{R}^{\delta(E)}$.

Now for each $B \in E^m$, by definition we know that $g_B(w_1, \dots, w_m) = \hat{g}_B(w_B)$. So in order to show that the functions $g_B, B \in E^m$, are linearly independent on the

set $w_1 + \cdots + w_m = 0$, we need only prove that the $\hat{g}_B(w_B), B \in E^m$, are linearly independent on $\Sigma^0_{E^m}$. We will show the following:

(IND) For any $E \in \bar{\mathcal{E}}$, the functions $\hat{g}_B(w_B), B \in E$, are linearly independent on the set Σ^0_E.

For each $B \in \mathcal{B}$ of degree $\delta(B) > 1$, B can be written uniquely as $[B_1, B^2]$, where $B_1, B^2 \in \mathcal{B}$, $B_1 \preceq B^2$. We will call B_1, B^2 the *left* and the *right* factors of B and use $\lambda(B)$ and $\rho(B)$ to denote them, i.e., $B = [\lambda(B), \rho(B)]$. (The left and right factors have been defined for formal brackets, cf. section 2.) If $\delta(B^2) > 1$, then we can write the unique decomposition $B^2 = [B_2, B^3]$, $B_2, B^3 \in \mathcal{B}, B_2 \preceq B^3$. Continuing this way, each $B \in \mathcal{B}^m$ can be written uniquely as $B = [B_1, [B_2, \ldots, [B_{s_B}, B^{s_B+1}] \cdots]$, where $B_1, B_2, \ldots, B_{s_B}, B^{s_B+1} \in \mathcal{B}$, $B_1 \succeq B_2 \succeq \cdots \succeq B_{s_B}, B^{s_B+1} \in \{X_1, \ldots, X_m\}$. So we have associated with each $B \in \mathcal{B}$ an ordered set $\{B_1, \ldots, B_{s_B}\}$. Each of the brackets $B_1, \ldots, B_{s_B}$ is called a *principal L-factor* of B. For each $B \in \mathcal{B}$ with $B = [B_1, [B_2, \ldots, [B_{s_B}, B^{s_B+1}] \cdots]$, we will write $B^\ell = [B_\ell, [B_{\ell+1}, \ldots, [B_{s_B}, B^{s_B+1}] \cdots]$ for $\ell = 1, \ldots, s_B$, so $B = [B_1, [B_2, \ldots, [B_{\ell-1}, B^\ell] \cdots]$. Let $B \in \mathcal{B}$ be a bracket of degree $\delta(B) > 1$. We define $LF(B)$, the set of *L-factors* of B, recursively as follows. If $B = [X_{k_1}, X_{k_2}]$, then $LF(B)$ contains only one element X_{k_1}. If $\delta(B) > 2$, write the unique decomposition $B = [B_1, B_2], B_1, B_2 \in \mathcal{B}, B_1 \preceq B_2$. If $\delta(B_1) = 1$, we define $LF(B) = \{B_1\} \cup LF(B_2)$. If $\delta(B_1) > 1$, we define $LF(B) = \{B_1\} \cup LF(B_1) \cup LF(B_2)$. The elements of $LF(B)$ are called the L-factors of B.

Now we prove (IND). We use induction on $\delta(E)$ to prove it. If $E \in \bar{\mathcal{E}}$, $\delta(E) = 2$, then E has only one element, i.e., $[X_{k_1}, X_{k_2}]$ for some $k_1 < k_2$. Then by definition $\hat{g}_{[X_{k_1}, X_{k_2}]}(w_{k_1}, w_{k_2}) = 1/w_{k_1}$ which is not identically 0 on the set $w_{k_1} + w_{k_2} = 0$. Assume that (IND) is true for all $E \in \bar{\mathcal{E}}$ with $\delta(E) \le k-1$, $3 \le k \le r$. Let $E \in \mathcal{E}^k$. Assume that the functions $\hat{g}_B(w_B), B \in E$, are linearly dependent on Σ^0_E. Then there exist constants l_B such that $\sum_{B \in E} l_B \hat{g}_B(w_B) \equiv 0$ on Σ^0_E. Let $\bar{E}$ be the subset of E that contains all the $B \in E$ with $l_B \neq 0$. Then

$$\sum_{B \in \bar{E}} l_B \hat{g}_B(w_B) \equiv 0 \quad \text{on } \Sigma^0_E. \tag{35}$$

Note that the set $\{\lambda(B), B \in \bar{E}\}$ is an ordered set. Let $\bar{B} \in \bar{E}$ be a bracket such that $\lambda(\bar{B}) \succeq \lambda(B)$ for all $B \in \bar{E}$. Let $\bar{B}_1 = \lambda(\bar{B}), \bar{B}_2 = \rho(\bar{B})$. Let $E_{\bar{B}}$ be the subset of $\bar{E}$ such that $B \in E_{\bar{B}}$ iff there is an L-factor of B, denoted by B', that is either equivalent to $\bar{B}_1$ or equivalent to $\bar{B}_2$. (It is obvious that if such an L-factor exists, it is unique.) For each $B \in E_{\bar{B}}$, if B' is the L-factor of B that is either equivalent to $\bar{B}_1$ or equivalent to $\bar{B}_2$, then B' has to be a principal L-factor of B. Write the unique decomposition $B = [B_1, [B_2, \ldots, [B_{\kappa-1}, [B', B^{\kappa+1}] \cdots]]]$. We let $\tilde{B}$ be the bracket $\tilde{B} = [B_1, [B_2, \ldots, [B_{\kappa-1}, B^{\kappa+1}] \cdots]]$ and define

$$\hat{g}_{\tilde{B}}(w_{\tilde{B}}) = \frac{\hat{g}_{B_1}(w_{B_1})}{\sum w_{B_1}} \frac{\hat{g}_{B_2}(w_{B_2})}{\sum w_{B_2}} \cdots \frac{\hat{g}_{B_{\kappa-1}}(w_{B_{\kappa-1}})}{\sum w_{B_{\kappa-1}}} \hat{g}_{B^{\kappa+1}}(w_{B^{\kappa+1}}),$$

where, if as a formal bracket, $\Sigma_{\tilde{B}} = X_{\ell_1} \ldots X_{\ell_s}$, we write $w_{\tilde{B}} = (w_{\ell_1}, \ldots, w_{\ell_s})$. So we get

$$\hat{g}_B(w_B) = \frac{\hat{g}_{\tilde{B}}(w_{\tilde{B}}) \hat{g}_{B'}(w_{B'})}{\sum w_{B'}}.$$

Now it is clear that for each $B \in \bar{E} - E_{\bar{B}}$, $\hat{g}_B(w_B)$ does not contain the factors $\frac{1}{\sum w_{\bar{B}_1}}$ or $\frac{1}{\sum w_{\bar{B}_2}}$. (By this we mean that the multivariable polynomial $\frac{1}{\hat{g}_B(w_B)}$ does

not contain the factors $\sum w_{\bar{B}_1}$ or $\sum w_{\bar{B}_2}$.) From (35) we get

$$\sum_{B\in E_{\bar{B}}} \frac{l_B \hat{g}_{\tilde{B}}(w_{\tilde{B}})\hat{g}_{B'}(w_{B'})}{\sum w_{B'}} + \sum_{B\in \bar{E}-E_{\bar{B}}} l_B \hat{g}_B(w_B) \equiv 0 \text{ on } \Sigma^0_E. \tag{36}$$

Now on Σ^0_E, for any $B \in E_{\bar{B}}$, we have either $\sum w_{B'} = \sum w_{\bar{B}_1}$ or $\sum w_{B'} = \sum w_{\bar{B}_2} = -\sum w_{\bar{B}_1}$ (according to whether B' is equivalent to $\bar{B}_1$ or $\bar{B}_2$). Let $q_B = 1$ if $B' \simeq \bar{B}_1$ and $q_B = -1$ if $B' \simeq \bar{B}_2$. From (36) we have

$$\sum_{B\in E_{\bar{B}}} l_B q_B \hat{g}_{\tilde{B}}(w_{\tilde{B}})\hat{g}_{B'}(w_{B'}) \equiv 0 \tag{37}$$

if $\sum w_{B'} = 0$, $\sum w_{\tilde{B}} = 0$. Note that for any $B \in E_{\bar{B}}$, if B' is equivalent to $\bar{B}_2$, then $B' = \lambda(B)$, so $\rho(B) = \tilde{B}$ is equivalent to $\bar{B}_1$. (By the definition of $\bar{B}_1$ and $\bar{B}_2$ we know that $\bar{B}_1 \preceq \bar{B}_2$, $\bar{B}_1 \succeq \lambda(B)$ for all $B \in \bar{E}$. If there is a $B \in E_{\bar{B}}$ such that $B' \simeq \bar{B}_2$, then $\delta(\bar{B}_1) = \delta(\bar{B}_2)$. It follows from this that if B' is equivalent to $\bar{B}_2$, then $\rho(B)$ must be equivalent to $\bar{B}_1$.) Let $E(\bar{B}_1)$, $E(\bar{B}_2)$ be the equivalence classes determined by $\bar{B}_1$ and $\bar{B}_2$, respectively. Then $E(\bar{B}_2) \in \bar{\mathcal{E}}$ and either $\delta(\bar{B}_1) = 1$ or $E(\bar{B}_1) \in \bar{\mathcal{E}}$. Let us assume first that $\delta(\bar{B}_1) > 1$. Let $E^1_{\bar{B}} = \{B \in E_{\bar{B}} \mid B' \simeq \bar{B}_1\}$ and $E^2_{\bar{B}} = \{B \in E_{\bar{B}} \mid B' \simeq \bar{B}_2\}$. Then $\{B'\}_{B\in E^1_{\bar{B}}} \cup \{\tilde{B}\}_{B\in E^2_{\bar{B}}} \subset E(\bar{B}_1)$. (Here $\tilde{B} = \rho(B)$ for $B \in E^2_{\bar{B}}$.) Now (37) can be rewritten as

$$\sum_{B\in E^1_{\bar{B}}} \left(l_B q_B \hat{g}_{\tilde{B}}(w_{\tilde{B}})\right) \hat{g}_{B'}(w_{B'}) + \sum_{B\in E^2_{\bar{B}}} \left(l_B q_B \hat{g}_{B'}(w_{B'})\right) \hat{g}_{\tilde{B}}(w_{\tilde{B}}) \equiv 0. \tag{38}$$

The left-hand side of (38) is a linear combination of $\hat{g}_B, B \in E(\bar{B}_1)$. From the linear independence of the $\hat{g}_B(w_B), B \in E(\bar{B}_1)$, on $\Sigma^0_{E(\bar{B}_1)}$, we see that in particular the coefficient of $\hat{g}_{\bar{B}_1}(w_{\bar{B}_1})$ in the left-hand side of (38) is zero. Let $\bar{E}'$ be the subset of $\bar{E}$ that contains the brackets $B \in \bar{E}$ such that either $\lambda(B) = \bar{B}_1$ or $\rho(B) = \bar{B}_1$. Then the coefficient of $\hat{g}_{\bar{B}_1}(w_{\bar{B}_1})$ in the left-hand side of (38) is equal to

$$\sum_{B\in \bar{E}'\cap E^1_{\bar{B}}} l_B q_B \hat{g}_{\tilde{B}}(w_{\tilde{B}}) + \sum_{B\in \bar{E}'\cap E^2_{\bar{B}}} l_B q_B \hat{g}_{B'}(w_{B'}).$$

So we get

$$\sum_{B\in \bar{E}'\cap E^1_{\bar{B}}} l_B q_B \hat{g}_{\tilde{B}}(w_{\tilde{B}}) + \sum_{B\in \bar{E}'\cap E^2_{\bar{B}}} l_B q_B \hat{g}_{B'}(w_{B'}) \equiv 0 \tag{39}$$

on $\Sigma^0_{E(\bar{B}_2)}$. Notice that for $B \in \bar{E}' \cap E^1_{\bar{B}}$, the brackets $\tilde{B}$ are in $E(\bar{B}_2)$ (since $\tilde{B} = \rho(B) \simeq \bar{B}_2$). By the linear independence of the $\hat{g}_B(w_B), B \in E(\bar{B}_2)$, on $\Sigma^0_{E(\bar{B}_2)}$ we get that all the coefficients $q_B l_B$ in the summations in the left-hand side of (39) are 0. Therefore we get $l_{\bar{B}} = 0$, which is a contradiction. If $\delta(\bar{B}_1) = 1$, assume that $\bar{B}_1 = X_\ell$. Then by the definition of $\bar{B}$, $E_{\bar{B}} = E^1_{\bar{B}} = \{B \in \bar{E} \mid \lambda(B) = X_\ell\}$. In this case $l_{\bar{B}} = 0$ follows from the linear independence of the $\hat{g}_B(w_B), B \in E(\bar{B}_2)$, on $\Sigma^0_{E(\bar{B}_2)}$, which still contradicts with $l_{\bar{B}} \neq 0$. This finishes the proof of (IND). In particular, we get that the $g_B(w_1, \ldots, w_m), B \in E^m$, are linearly independent on the hyperplane $w_1 + \cdots + w_m = 0$ in $\mathbb{R}^m$.

In the general case, we fix an $E \in \mathcal{E}_n$. Take another group of indeterminates $\mathbf{Y} = \{Y_1, \ldots, Y_n\}$. Let us fix a choice of a P. Hall basis $\mathcal{B}^{\mathbf{Y}}$ of $L(\mathbf{Y})$. We know

that each $g_B, B \in E$, is a linear combination of some $g_{B^{\mathbf{Y}}}$ with $B^{\mathbf{Y}} \in E^{\mathbf{Y}} = \{B^{\mathbf{Y}} \in \mathcal{B}^{\mathbf{Y}}, \delta_k(B^{\mathbf{Y}}) = 1, k = 1, \dots, n\}$. Let θ_E be the algebra homomorphism from $A(\mathbf{Y})$ to $A(\mathbf{X})$, cf. Remark 4.1, defined by $\theta_E(Y_k) = X_\ell$ if $\delta_1(E) + \cdots + \delta_{\ell-1}(E) + 1 \le k \le \delta_1(E) + \cdots + \delta_\ell(E)$. Let $L_E(\mathbf{X})$, $L_{E^{\mathbf{Y}}}(\mathbf{Y})$ be the subsets of $L(\mathbf{X})$ and $L(\mathbf{Y})$, respectively, that are linearly spanned by the brackets $B \in E$ and $B^{\mathbf{Y}} \in E^{\mathbf{Y}}$. Let us use $\mathcal{I}$ to denote a multi-index $\mathcal{I} = (\ell_1, \dots, \ell_k)$ with entries $1 \le \ell_s \le n$. Then $L_E(\mathbf{X})$ is the subset of $L(\mathbf{X})$ linearly spanned by the brackets $[X_I]$ with $\delta_k(X_I) = \delta_k(E), k = 1, \dots, m$, and $L_{E^{\mathbf{Y}}}(\mathbf{Y})$ is the subset of $L(\mathbf{Y})$ spanned by the $[Y_{\mathcal{I}}]$ with $\delta_k(Y_{\mathcal{I}}) = 1, k = 1, \dots, n$. (Here $\delta_k(X_I)$ denotes the degree of X_k in X_I and $\delta_k(Y_{\mathcal{I}})$ denotes the degree of Y_k in $Y_{\mathcal{I}}$.) Clearly the restriction of θ_E to $L_{E^{\mathbf{Y}}}(\mathbf{Y})$ maps $L_{E^{\mathbf{Y}}}(\mathbf{Y})$ onto $L_E(\mathbf{X})$. Let us still use θ_E to denote the restriction of θ_E to $L_{E^{\mathbf{Y}}}(\mathbf{Y})$. The $L_E(\mathbf{X})$, $L_{E^{\mathbf{Y}}}(\mathbf{Y})$ can be regarded as vector spaces over $\mathbb{R}$. It is obvious that E is a basis of $L_E(\mathbf{X})$ and $E^{\mathbf{Y}}$ spans $L_{E^{\mathbf{Y}}}(\mathbf{Y})$. Assume $|E| = N$, $|E^{\mathbf{Y}}| = \bar{N}$, $E = \{B_1, \dots, B_N\}$, and $E^{\mathbf{Y}} = \{B_1^{\mathbf{Y}}, \dots, B_{\bar{N}}^{\mathbf{Y}}\}$. Let $M = (\alpha_{\ell,k})_{1 \le \ell \le \bar{N}, 1 \le k \le N}$ be the matrix with $\alpha_{\ell,k}$ determined by $\theta_E(B_\ell^{\mathbf{Y}}) = \sum_{k=1}^N \alpha_{\ell,k} B_k$, $\ell = 1, \dots, \bar{N}$. Because the map $\theta_E : L_{E^{\mathbf{Y}}}(\mathbf{Y}) \to L_E(\mathbf{X})$ is onto, the matrix has rank N. From Remark 4.1 we know that each of the $g_B, B \in E$, is a linear combination of some $g_{B^{\mathbf{Y}}}$ with $B^{\mathbf{Y}} \in E^{\mathbf{Y}}$. More precisely, we have the following $g_{B_k}(w_1, \dots, w_n) = \sum_{\ell=1}^{\bar{N}} \alpha_{\ell,k} g_{B_\ell^{\mathbf{Y}}}(w_1, \dots, w_n)$ for all $(w_1, \dots, w_n) \in \mathbb{R}^n$, i.e., $(g_{B_1}, \dots, g_{B_N}) = (g_{B_1^{\mathbf{Y}}}, \dots, g_{B_{\bar{N}}^{\mathbf{Y}}})M$. From this it follows that the linear independence of the $g_{B^{\mathbf{Y}}}, B^{\mathbf{Y}} \in E^{\mathbf{Y}}$, on the set $w_1 + \cdots + w_n = 0$ implies that the $g_B, B \in E$, are linearly independent on that set too. Now the judgment of the approximation algorithm is complete.

8. Some examples and variations. We give some examples and variations of the approximation algorithm. From conditions CH1, CH2, CH3 we know that the u_k^j in (22) are linear combinations of $j^\alpha \eta_\omega(t)\cos(\omega j t)$ and $j^\alpha \eta_\omega(t)\sin(\omega j t)$. From the proof of Theorem 5.1 we see that, for the control sequence defined in (22), the limiting extended input has form (23) mainly because of the fact that the frequencies are MC and independent. If we require that the frequencies satisfy some other kind of MC properties (this will be clear in the example below) and the frequencies associated with each part in the extended input be independent, then the limiting extended input can also be calculated explicitly and can be made equal to any prescribed extended input of finite order if the frequencies satisfy some additional conditions like the invertibility of the matrices $\{\hat{\xi}_{B,\rho}\}$, etc.

Example 8.1. Let us consider the case of a two-input system in $\mathbb{R}^5$; i.e., a system

$$\dot{x}(t) = u_1(t) f_1(x(t)) + u_2(t) f_2(x(t)),$$

where f_1, f_2 are smooth vector fields on $\mathbb{R}^5$. Assume that the vectors

$$f_1(x), f_2(x), [f_1, f_2](x), [f_1, [f_1, f_2]](x), [f_2, [f_1, f_2]](x)$$

span $\mathbb{R}^5$ everywhere. Let $t \to \gamma(t) \in \mathbb{R}^5$ on $[0, 1]$ be a smooth curve. Then, by the span condition, there exist smooth functions $v_1, \dots, v_5$ on $[0, 1]$ such that $t \to \gamma(t)$ is a solution of the equation

$$\dot{x} = v_1(t) f_1(x) + v_2(t) f_2(x) + v_3(t)[f_1, f_2](x) + v_4(t)[f_1, [f_1, f_2]](x) + v_5(t)[f_2, [f_1, f_2]](x)$$

with initial condition $x(0) = \gamma(0)$.

In this case the extended input $\mathbf{v}$ that we want to approximate is given by

$$\mathbf{v} = v_1 X_1 + v_2 X_2 + v_3[X_1, X_2] + v_4[X_1, [X_1, X2]] + v_5[X_2, [X_1, X_2]]. \tag{40}$$

Take three groups $\Omega_1 = \{\omega_{1,1}, \omega_{1,2}\}$, $\Omega_k = \{\omega_{k,1}, \omega_{k,2}, \omega_{k,3}\}$, $k = 2, 3$ of real numbers. Assume that (1) each Ω_k is MC; (2) the sets $\Omega_1, \Omega_2, \Omega_3$ are independent with respect to 3.

For any C^1 functions $\eta_1, \eta_2, \eta_{\omega_{1,1}}, \eta_{\omega_{1,2}}$, $\eta_{\omega_{k,l}}, k = 2, 3, l = 1, 2, 3$, on $[0, 1]$, let

$$\begin{aligned} u_1^j(t) &= \eta_1(t) + j^{\frac{1}{2}}\eta_{\omega_{1,1}}(t)\cos\omega_{1,1}jt + j^{\frac{2}{3}}\eta_{\omega_{2,1}}(t)\cos\omega_{2,1}jt \\ &\quad + j^{\frac{2}{3}}\eta_{\omega_{2,2}}(t)\cos\omega_{2,2}jt + j^{\frac{2}{3}}\eta_{\omega_{3,1}}(t)\cos\omega_{3,1}jt, \\ u_2^j(t) &= \eta_2(t) + j^{\frac{1}{2}}\eta_{\omega_{1,2}}(t)\sin\omega_{1,2}jt + j^{\frac{2}{3}}\eta_{\omega_{2,3}}(t)\cos\omega_{2,3}jt \\ &\quad + j^{\frac{2}{3}}\eta_{\omega_{3,2}}(t)\cos\omega_{3,2}jt + j^{\frac{2}{3}}\eta_{\omega_{3,3}}(t)\cos\omega_{3,3}jt. \end{aligned}$$

From Theorem 5.1 we have the following.

PROPOSITION 8.1. *The sequence* $\{u^j = (u_1^j, u_2^j)\}$ *of inputs defined above EI*(3)*-converges to*

$$\mathbf{u}^\infty(t) = \eta_1(t)X_1 + \eta_2(t)X_2 - \frac{\eta_{\omega_{1,1}}(t)\eta_{\omega_{1,2}}(t)}{2\omega_{1,1}}[X_1, X_2]$$

$$-\frac{\eta_{\omega_{2,1}}(t)\eta_{\omega_{2,2}}(t)\eta_{\omega_{2,3}}(t)}{4\omega_{2,1}\omega_{2,2}}[X_1, [X_1, X_2]] + \frac{\eta_{\omega_{3,1}}(t)\eta_{\omega_{3,2}}(t)\eta_{\omega_{3,3}}(t)}{4\omega_{3,2}\omega_{3,3}}[X_2, [X_1, X_2]].$$

Now it is easy to see that in order to get $\mathbf{u}^\infty$ to be equal to $\mathbf{v}$ we can simply let $\eta_1 = v_1$, $\eta_2 = v_2$, $\eta_{\omega_{1,1}} = 1$, $\eta_{\omega_{1,2}} = -2\omega_{1,1}v_3$, $\eta_{\omega_{2,1}} = 1$, $\eta_{\omega_{2,2}} = 1$, $\eta_{\omega_{2,3}} = -4\omega_{2,1}\omega_{2,2}v_4$, $\eta_{\omega_{3,1}} = 1$, $\eta_{\omega_{3,2}} = 1$, $\eta_{\omega_{3,3}} = 4\omega_{3,2}\omega_{3,3}v_5$. There are five terms in u_1^j and u_2^j, respectively. Next we give another control sequence which has fewer terms in u_1^j and u_2^j. Take three groups $\Omega_k = \{\omega_{k,1}, \omega_{k,2}\}$, $k = 1, 2, 3$ of nonzero frequencies. Assume that the $\omega_{k,l}$ satisfy (1) $\omega_{1,1} + \omega_{1,2} = 0$, $2\omega_{2,1} + \omega_{2,2} = 0$, $\omega_{3,1} + 2\omega_{3,2} = 0$ and that (2) $\Omega_1, \Omega_2, \Omega_3$ are independent with respect to 3.

For any functions $\eta_1, \eta_2, \eta_{\omega_{k,l}}, k = 1, 2, 3, l = 1, 2$, of class C^1 on $[0, 1]$, let

$$\begin{aligned} u_1^j(t) &= \eta_1(t) + j^{\frac{1}{2}}\eta_{\omega_{1,1}}(t)\cos\omega_{1,1}jt + j^{\frac{2}{3}}\eta_{\omega_{2,1}}(t)\cos\omega_{2,1}jt + j^{\frac{2}{3}}\eta_{\omega_{3,1}}(t)\cos\omega_{3,1}jt, \\ u_2^j(t) &= \eta_2(t) + j^{\frac{1}{2}}\eta_{\omega_{1,2}}(t)\sin\omega_{1,2}jt + j^{\frac{2}{3}}\eta_{\omega_{2,2}}(t)\cos\omega_{2,2}jt + j^{\frac{2}{3}}\eta_{\omega_{3,2}}(t)\cos\omega_{3,2}jt. \end{aligned}$$

Then we have the following.

PROPOSITION 8.2. *For any frequencies* $\omega_{k,l}, k = 1, 2, 3, l = 1, 2$, *satisfying* (1) *and* (2) *and any functions* $\eta_1, \eta_2, \eta_{\omega_{k,l}}, k = 1, 2, 3, l = 1, 2$, *of class* C^1 *on* $[0, 1]$, *the* $u^j = (u_1^j, u_2^j)$ *EI*(3)*-converge to*

$$\begin{aligned} \mathbf{u}^\infty = \eta_1 X_1 + \eta_2 X_2 &- \frac{\eta_{\omega_{1,1}}\eta_{\omega_{1,2}}}{2\omega_{1,1}}[X_1, X_2] - \frac{\eta_{\omega_{2,1}}^2\eta_{\omega_{2,2}}}{8\omega_{2,1}^2}[X_1, [X_1, X_2]] \\ &- \frac{\eta_{\omega_{3,1}}\eta_{\omega_{3,2}}^2}{4\omega_{3,1}\omega_{3,2}}[X_2, [X_1, X_2]]. \end{aligned}$$

Again, if we want $\mathbf{u}^\infty = \mathbf{v}$ we can let $\eta_1 = v_1$, $\eta_2 = v_2$, $\eta_{\omega_{1,1}} = 1$, $\eta_{\omega_{1,2}} = -2\omega_{1,1}v_3$, $\eta_{\omega_{2,1}} = 1$, $\eta_{\omega_{2,2}} = -8\omega_{2,1}^2 v_4$, $\eta_{\omega_{3,1}} = -4\omega_{3,1}\omega_{3,2}v_5$, $\eta_{\omega_{3,2}} = 1$.

The Lie brackets in the above example are of multiplicity 1. For a simple example with brackets of > 1, cf. [18].

Next we give some remarks about the approximation algorithm. First, using Theorem 6 in [20], we have the following.

THEOREM 8.1. *Let r be a positive integer. Let $\{u^j\}_{j=1}^{\infty}$ be the sequence defined in (22), where the functions $\eta_{0,k}$, η_ω and the sets $\Omega(n,k)$, $\Omega_{E,\rho}$, $\Omega_{E,\rho,k}$ satisfy conditions (1)–(10). Let $f_k : \mathbb{R}^n \times [0,T] \to \mathbb{R}^n$, $k = 1,\dots,m$ be time-varying vector fields on $\mathbb{R}^n$. Assume that $f_1,\dots,f_m$ are of class C^{r-1} in x and of class C^1 jointly in (x,t). Let $f_0 : \mathbb{R}^n \times [0,T] \to \mathbb{R}^n$ be continuous. Let $x_0 \in \mathbb{R}^n$ be a point and let x^j be maximal solutions of*

$$\dot{x} = f_0(x,t) + \sum_{k=1}^{m} u_k^j(t) f_k(x,t),\ x(0) = x_0.$$

Assume that the initial value problem

$$\begin{aligned}\dot{x}(t) = {} & f_0(x,t) + \sum_{k=1}^{m} \eta_{0,k}(t) f_k(x,t) + \sum_{1\le k_1<k_2\le m} \frac{1}{i\omega_{k_1,k_2}} \zeta_{k_1,k_2}[f_{k_1},f_{k_2}](x,t) \\ & + \sum_{n=3}^{r} \sum_{E\in\mathcal{E}_n} \sum_{B\in E} \sum_{\rho=1}^{|E|} \left(\sum_{F\in Q_{E,\rho}} i^{1-\delta(E)} \xi_{B,\rho}^{F} \prod_{\omega\in F} \eta_\omega(t) \right) \mathrm{Ev}(\mathbf{f})(B)(x,t), \\ x(0) = {} & x_0,\end{aligned}$$

where $\zeta_{k_1,k_2} = (\eta_{\omega_{k_1,k_2},k_1}\eta_{-\omega_{k_1,k_2},k_2} - \eta_{-\omega_{k_1,k_2},k_1}\eta_{\omega_{k_1,k_2},k_2})$, has a unique solution x^∞ which is defined on the whole interval $[0,T]$. (Here for each B, $\mathrm{Ev}(\mathbf{f})(B)$ is the vector field obtained by plugging in the vector field f_k for X_k in B.) Then the x^j are defined on $[0,T]$ for j large enough and converge to x^∞ uniformly on $[0,T]$ as $j \to \infty$.

From this theorem we see that exactly the same algorithm works for the time-varying vector field case even with a drift term. But we note that the vector fields $\mathrm{Ev}(\mathbf{f})(B)$ are Lie brackets of f_k for $k = 1,\dots,m$ only.

As suggested to us by Gurvits, it would be nicer if the approximation algorithm could be used to produce *feedback controls*. Suppose that for a suitable Lie bracket extension $\dot{x} = f_0(x,t) + \sum_{k=1}^{r} v_k f_k(x,t)$ of

$$\dot{x} = f_0(x,t) + \sum_{k=1}^{m} u_k f_k(x,t), \tag{41}$$

there exists a time-dependent feedback law $v_k = v_k(x,t)$, $k = 1,\dots,r$, such that the closed loop system

$$\dot{x} = f_0(x,t) + \sum_{k=1}^{r} v_k(x,t) f_k(x,t) \tag{42}$$

has some desired properties. Then one may try to produce time-dependent feedback controls $u_k^j = u_k^j(x,t)$ that generate trajectories of (41) that, as $j \to \infty$, converge to those of the closed loop system (42), at least on some fixed time interval $[0,T]$. It is easy to see that this problem can be reduced to a special case of Theorem 8.1. Our algorithm makes it possible to produce such an approximation as follows. Assume that the functions v_k are sufficiently smooth. Then, by introducing some new vector fields $g_1,\dots,g_\rho$, we can rewrite system (42) into a new system of the form

$$\dot{x} = f_0(x,t) + \sum_{k=1}^{\tau} g_k(x,t) + \sum_{k=\tau+1}^{\rho} g_k(x,t), \tag{43}$$

where the first τ functions $g_1, \ldots, g_\tau$ are linear combinations of $f_1, \ldots, f_m$ with sufficiently smooth coefficients, and the $g_k, k = \tau + 1, \ldots, \rho$, are Lie brackets of the $g_k, k \in \{1, \ldots, \tau\}$. So system (43) can be viewed as a special Lie bracket extension of

$$\dot{x} = f_0(x,t) + \sum_{k=1}^{\tau} w_k g_k(x,t), \tag{44}$$

with constant coefficients (0, or 1, cf. the example below). We then apply our algorithm to system (44) to get a control sequence $\{w^j = (w_1^j, \ldots, w_\tau^j)\}$ that generates trajectories converging to solutions of (43). To see how this can be done, let us examine the following example.

Example 8.2. Suppose $m = 2$ and (42) is given by

$$\dot{x} = f_0(x,t) + v_1(x,t)f_1(x,t) + v_2(x,t)f_2(x,t) + v_3(x,t)[f_1, f_2](x,t). \tag{45}$$

We rewrite $v_3[f_1, f_2] = [f_1, v_3 f_2] - (L_{f_1} v_3) f_2$ and let $g_1 = v_1 f_1, g_2 = v_2 f_2$, $g_3 = -(L_{f_1} v_3) f_2, g_4 = f_1, g_5 = v_3 f_2$. Then system (45) can be rewritten as $\dot{x} = f_0(x,t) + g_1(x,t) + g_2(x,t) + g_3(x,t) + [g_4, g_5](x,t)$, which arises from the extension $\dot{x} = f_0(x,t) + \sum_{k=1}^{5} \bar{w}_k g_k(x,t) + \bar{w}_6 [g_4, g_5](x,t)$ of

$$\dot{x} = f_0(x,t) + \sum_{k=1}^{5} w_k g_k(x,t) \tag{46}$$

by specializing the controls $\bar{w}_4 = \bar{w}_5 = 0$, $\bar{w}_1 = \bar{w}_2 = \bar{w}_3 = \bar{w}_6 = 1$. If we apply our algorithm to produce a control sequence $\{w^j\}$ for (46) that EI(2)-converges to the extended input $\mathbf{v} = X_1 + X_2 + X_3 + [X_4, X_5]$, then the time-dependent feedbacks $u_1^j(x,t) = w_1^j(t) v_1(x,t) + w_4^j(t)$, $u_2^j(x,t) = w_2^j(t) v_2(x,t) - w_3^j(t)(L_{f_1} v_3)(x,t) + w_5^j(x,t) v_3(x,t)$ have the desired properties.

We conclude this section by giving an estimate for the rate of how fast the trajectories generated by the u^j in (22) converge to trajectories generated by $\mathbf{u}^\infty$ in (23).

PROPOSITION 8.3. *Let $\mathbf{f} = (f_1, \ldots, f_m)$ be an $m$tuple of vector fields of class C^r on $\mathbb{R}^n$. Let x_0 be a point in $\mathbb{R}^n$. Let $\{u^j\}$ and $\mathbf{u}^\infty$ be defined in* (22) *and* (23)*, respectively, with the functions $\eta_\omega's$ and the frequency sets $\Omega(n,k)$ satisfying conditions* (1)–(10)*. Assume that the solution x^∞ generated by $\mathbf{u}^\infty$ and $\mathbf{f}$ with initial condition $x(0) = x_0$ is defined on $[0,T]$. Then there exists a constant K such that the solutions x^j with initial condition $x(0) = x_0$, generated by the u^j and $\mathbf{f}$, are defined on $[0,T]$ for j large enough and $\|x^j - x^\infty\|_{\sup} \le K j^{-\frac{1}{r}}$. Let Δ be a compact set in $\mathbb{R}^n$ that contains x^∞ in its interior. Then the constant K depends on $\mathbf{f}$ restricted on Δ, the ω, and the η_ω.*

Acknowledgment. The author would like to thank Professor H. J. Sussmann for proposing to him the topic and for his help and many fruitful discussions.

REFERENCES

[1] N. BOURBAKI, *Elements of Mathematics: Lie Groups and Lie Algebras*, Hermann, Paris, 1975.

[2] R. W. BROCKETT AND LIYI DAI, *Non-holonomic Kinematics and the Role of Elliptic Functions in Constructive Controllability*, in Nonholonomic Motion Planning, Z. X. Li and J. F. Canny, eds., Kluwer Academic Publishers, Boston, 1993, pp. 1–22.

[3] A. COLE, J. HAUSER, AND S. SASTRY, *Kinematics and control of a multifingered robot hand with rolling contact*, IEEE Trans. Automat. Control, 34 (1989), pp. 398–404.

[4] C. Fernandes, L. Gurvits, and Z. X. Li, *Foundations of Nonholonomic Motion Planning*, Technical Report, Robotics Research Laboratory, Courant Institute of Mathematics Science, New York, 1991.

[5] L. Gurvits and Z. X. Li, *Theory and Application of Nonholonomic Motion Planning*, Technical Report, Courant Institute of Mathematical Sciences, New York, 1990.

[6] G. W. Haynes and H. Hermes, *Nonlinear controllability via Lie theory*, SIAM J. Control Optim., 8 (1970), pp. 450–460.

[7] P. Jacobs and J. Canny, *Planning via the canonical trajectory approach*, in Proc. 1991 IEEE R&A Workshop on Nonholonomic Motion Planning, Sacramento, CA, 1991.

[8] G. Lafferriere and H. J. Sussmann, *Motion Planning for Controllable Systems without Drift: A Preliminary Report*, Rutgers University Systems and Control Center Report SYCON-90-04, Rutgers University, New Brunswick, NJ, 1990.

[9] J. P. Laumond, *Singularities and topological aspects in nonholonomic motion planning*, in Nonholonomic Motion Planning, Z. X. Li and J. F. Canny, eds., Kluwer Academic Publishers, Boston, 1993, pp. 149–200.

[10] R. M. Murray, and S. S. Sastry, *Steering nonholonomic control systems using sinusoids*, in Nonholonomic Motion Planning, Z. X. Li and J. F. Canny, eds., Kluwer Academic Publishers, Boston, 1993, pp. 109–148.

[11] R. M. Murray and S. S. Sastry, *Grasping and Manipulations Using Multifingered Robot Hands*, Memorandum UCB/ERL M90/24, Electronics Research Laboratory, University of California, Berkeley, CA, 1990.

[12] R. M. Murray and S. S. Sastry, *Steering nonholonomic systems in chain form*, in Proc. 30th IEEE CDC, Brighton, UK, 1991.

[13] R. M. Murray and S. S. Sastry, *Steering controllable systems*, in Proc. 29th IEEE CDC, Honolulu, HI, 1990.

[14] S. S. Sastry and Z. Li, *Robot motion planning with nonholonomic constraints*, in Proc. 28th IEEE CDC, Tampa, FL, 1989, pp. 211–216.

[15] H. J. Sussmann, *A product expansion for the Chen series*, in Theory and Applications of Nonlinear Control Systems, C. Byrnes and A. Lindquist, eds., North-Holland, Amsterdam, 1986, pp. 323–335.

[16] H. J. Sussmann, *A general theorem on local controllability*, SIAM J. Control Optim. 25 (1987), pp. 158–194.

[17] H. J. Sussmann and W. Liu, *Limits of highly oscillatory controls and approximation of general paths by admissible trajectories*, in Proc. 30th IEEE CDC, Brighton, UK, 1991.

[18] H. J. Sussmann and W. Liu, *Lie bracket extensions and averaging: The single-bracket case*, in Nonholonomic Motion Planning, Z. X. Li and J. F. Canny, eds., Kluwer Academic Publishers, Boston, 1993, pp. 109–148.

[19] W. Liu, *Averaging theorems for highly oscillatory differential equations and iterated Lie brackets*, SIAM J. Control Optim., 37 (1997), to appear.

[20] W. Liu, *Limiting process of control-affine systems with Hölder continuous inputs*, submitted.

[21] W. Liu, *Averaging Theorems for Highly Oscillatory Ordinary Differential Equations and the Approximation of General Paths by Admissible Trajectories for Nonholonomic Systems*, Ph.D. thesis, Rutgers University, New Brunswick, NJ, 1992.

SIAM J. CONTROL OPTIM.
Vol. 35, No. 4, pp. 1366–1386, July 1997

012

REGULARITY PROPERTIES OF THE PHASE FOR MULTIVARIABLE SYSTEMS*

KEVIN A. GRASSE† AND JONATHAN R. BAR-ON‡

Abstract. For multivariable, input-output systems that are represented as rational, transfer-function matrices, the most frequently used measures of relative stability are gain based. However, there are a number of important physical applications where the phase of a perturbation can also have a significant effect on relative stability. Such applications led J. R. Bar-on and E. A. Jonckheere [J. R. Bar-on, *Phase and Gain Margins for Multivariable Control Systems*, Ph.D. thesis, University of Southern California, 1990; J. R. Bar-on and E. A. Jonckheere, *Internat. J. Control*, 52 (1990), pp. 485–498] to define precisely the notions of phase, minimum-phase mapping, and phase margin for multivariable systems. The objective of this paper is to establish conditions under which the phase and minimum-phase mappings have certain desired regularity properties (e.g., continuity or differentiability). After a review of the definitions of the phase concepts under consideration, we collect a few well-known results about set-valued maps that have direct applications to parametrized families of constrained optimization problems. Using these results we show that, under very mild conditions, the minimum-phase mapping is lower semicontinuous as a function of frequency; as a consequence, the phase margin (initially defined as the infimum of the phase of all destabilizing unitary perturbations in the range of frequencies where destabilizing perturbations can occur) is achieved as the phase of a specific destabilizing unitary perturbation. We then establish sufficient conditions of gradually increasing strength for the minimum-phase mapping to be continuous and real analytic as a function of frequency. The proof of the real analyticity of the minimum-phase mapping relies on the implicit function theorem and the Lagrange multiplier theorem.

Key words. multivariable system, robust stability, unitary matrix, phase margin, parametric optimization problem, value function, marginal function

AMS subject classifications. 93C35, 93D09, 93D22

PII. S0363012994279243

1. Introduction.

A desirable feature of a multivariable control system is its ability to maintain stability in the presence of unknown perturbations. Such perturbations can be attributed to modeling errors (i.e., linearization or the neglect of high-frequency terms) as well as to external environmental factors that may impact the system. Consequently, it is important to have quantitative measurements of relative stability, or robustness, of control systems. For multi-input, multioutput (MIMO) control systems there currently exist a number of different measurements of relative stability. Most of these measurements are *gain* based; i.e., they focus only on the modulus (size) of the perturbation (see [9, 11, 12, 18, 19, 21]). However, the *phase* (or "angle of rotation") of the perturbation can also have a significant effect in some systems (see [3, 4, 7] for a more detailed discussion), and such effects are essentially ignored by gain-based stability measurements. These considerations motivated J. R. Bar-on and E. A. Jonckheere to define the notions of the *phase*, *minimum-phase mapping*, and *phase margin* for multivariable control systems [3, 4, 5]. We review the precise mathematical definitions of these terms in the next section.

*Received by the editors December 29, 1994; accepted for publication (in revised form) May 22, 1996. This research was supported in part by National Science Foundation grant ECS–9301196.
http://www.siam.org/journals/sicon/35-4/27924.html
†Department of Mathematics, University of Oklahoma, Norman, OK 73019–0315 (kgrasse@uoknor.edu).
‡School of Electrical Engineering, University of Oklahoma, Norman, OK 73019. Current address: Hughes Missile Systems Company, 805, M/S K6, P. O. Box 11337, Tucson, AZ 85734–1337 (jrbar-on@CCGATE.HAC.COM).

Our objective here is to establish some important mathematical properties of these phase-related concepts, which should further enhance their usefulness. For input-output systems that are representable as transfer-function matrices $L(s)$ and are closed-loop stable, we will consider the effect of unitary (pure phase) perturbations on their stability. In generic situations, there is a specific set of frequencies at which it is possible to destabilize the system via unitary perturbations; the set of these frequencies is called the *gain-crossover region*. To each frequency in the gain-crossover region one can assign the minimum value of the phase for the unitary perturbations that destabilize the system at that frequency. This results in the what is called the *minimum-phase mapping* of the system. The *phase margin* of the system is by definition the minimum value of the minimum-phase mapping as it ranges over all frequencies in the gain-crossover region. In order to ensure that the minimum is achieved, it is necessary that the minimum-phase mapping be at least lower semicontinuous, and we prove that this is always the case under very mild conditions on the transfer function $L(s)$. However, there is ample numerical evidence that the minimum-phase mapping is far better than merely lower semicontinuous. Thus we will also state specific sufficient conditions for the minimum-phase mapping to be continuous, differentiable, and even real analytic as a function of the frequency. The continuity results will be obtained from standard results on parametrized optimization problems. The differentiability results will follow from the Lagrange multiplier theorem and the implicit function theorem.

The fact that the minimum-phase mapping is (under reasonable assumptions) a differentiable function of frequency is important for at least two reasons. First, as we will see, the values of the minimum-phase mapping are computed as the solutions of a family of constrained optimization problems parametrized by frequency. For a general parametrized optimization problem one cannot even expect continuous dependence of the optimal value on the parameter, to say nothing of differentiable dependence. There is ample numerical evidence to suggest that the minimum-phase mapping is smooth in generic cases, and this poses a purely mathematical question of why it should be so. Second, it turns out that the regularity of the minimum-phase mapping is closely tied to the stability of the numerical methods that are employed to compute the phase margin. We will return to this point at the end of the next section.

The remainder of the paper is organized as follows. In section 2 we review the definitions of the phase, minimum-phase mapping, and phase margin as given by Bar-on and Jonckheere [3, 4, 5]. We also review Bar-on's formulation of the computation of the minimum-phase mapping as the minimum value of a constrained optimization problem, since this formulation is very useful in the theoretical development. Section 3 presents a concise review of a few well-known results in parametrized optimization problems. These results are applied in section 4 to show that the phase and, under appropriate conditions, the minimum-phase mapping are continuous on their domains. Finally, differentiability of the minimum-phase mapping is discussed in section 5.

2. Phase concepts for multivariable systems. We begin with a review of phase concepts for multivariable systems. To do this, we adopt the frequency domain approach and view a control system as a transfer function $L(s)$. More precisely, $L(s)$ denotes an $n \times n$ matrix function of $s \in \mathbb{C}$ whose entries are in the field of rational functions of s with real coefficients. Here $\mathbb{C}$ denotes the field of complex numbers, and we let $j \in \mathbb{C}$ denote the imaginary unit. It is assumed throughout that $n \geq 2$ since we wish to focus on multivariable systems. The assumption that $L(s)$ is square means that the number of system inputs equals the number of system outputs, which

is typical for feedback systems. We always assume that $L(s)$ has no poles on the imaginary axis $s = j\omega$ and is *proper*; i.e., $\lim_{s\to\infty} L(s) = L_\infty$, where L_∞ is a real $n \times n$ constant matrix. If $L_\infty = 0$, then we call $L(s)$ *strictly proper*. Since $L(s)$ is assumed to be proper, we can extend the continuous mapping $\omega \mapsto L(j\omega)$ of $\mathbb{R}$ into $\mathbb{C}^{n\times n}$ (the set of all $n \times n$ matrices with entries in $\mathbb{C}$) to a continuous mapping defined on the extended real numbers $\overline{\mathbb{R}} = \mathbb{R} \cup \{\infty\}$ (recall that $\overline{\mathbb{R}}$ is the one-point compactification of $\mathbb{R}$ and is homeomorphic to the unit circle $S^1 \subseteq \mathbb{C}$).

The multivariable Nyquist criterion [8, 10] states that $L(s)$ is closed-loop stable (in the sense that $L(s)(I + L(s))^{-1}$ is stable) if and only if

(N1) $\det(I + L(j\omega)) \neq 0 \ \forall\ \omega \in \overline{\mathbb{R}}$;

(N2) the number of counterclockwise encirclements of the origin by the curve $\gamma : \overline{\mathbb{R}} \to \mathbb{C}$ given by $\gamma(\omega) = \det(I + L(j\omega))$ equals the number of right-half-plane poles of $L(s)$ (here I denotes the $n \times n$ identity matrix and the curve γ is given the orientation induced from the standard positive orientation on $\overline{\mathbb{R}}$).

In addition to being proper and having no poles on the imaginary axis, we shall always assume that $L(s)$ is closed-loop stable in the sense of (N1) and (N2).

The perturbations that act on the system will be assumed to occur in the feedback path as discussed in [3, 4]. Since we wish to focus on the phase of such perturbations, it will also be assumed that the perturbations come from the group $U(n, \mathbb{C})$ of complex unitary $n \times n$ matrices defined by

$$U(n, \mathbb{C}) = \{\Delta \in \mathbb{C}^{n\times n} \mid \Delta^* = \Delta^{-1}\},$$

where the $*$ denotes the complex-conjugate transpose. If we give $U(n, \mathbb{C})$ the topology that it inherits as a subspace of the complex vector space $\mathbb{C}^{n\times n}$, then $U(n, \mathbb{C})$ is compact. The *phase* of a unitary matrix $\Delta \in U(n, \mathbb{C})$ is defined as

$$\Pi(\Delta) = \max\{|\arg(\lambda)| : \lambda \text{ is an eigenvalue of } \Delta\},$$

where the argument of a complex number $s \in \mathbb{C}$ is chosen to satisfy $-\pi < \arg(s) \leq \pi$. We recall that since Δ is unitary, every eigenvalue λ of Δ satisfies $|\lambda| = 1$. An equivalent definition of the phase (see [3, 4]) is given by

$$\Pi(\Delta) = \max\{\cos^{-1}(\mathrm{Re}(z^*\Delta z)) \mid z \in \mathbb{C}^n \text{ and } z^*z = 1\}, \tag{1}$$

where Re denotes the real part of a complex number and we select the branch of $\cos^{-1}$ that returns angles in the interval $[0, \pi]$. In section 4 we give a short proof that the function $\Pi : U(n, \mathbb{C}) \to [0, \pi]$ is continuous. Suppose now that the system is affected by perturbations $\Delta \in U(n, \mathbb{C})$ in the feedback path as indicated previously. The *stability set* $\mathcal{S}$ is by definition the set of $\Delta \in U(n, \mathbb{C})$ such that the perturbed system is closed-loop stable. Using the Nyquist criterion, we see that $\Delta \in \mathcal{S}$ if and only if (N1) and (N2) are satisfied with $L(j\omega)$ replaced by $L(j\omega)\Delta$. From this characterization of $\mathcal{S}$ it follows that $\mathcal{S}$ is an open subset of $U(n, \mathbb{C})$. Moreover, $\mathcal{S}$ is nonempty since $I \in \mathcal{S}$ by the assumed closed-loop stability of the unperturbed system.

Let $g : \mathbb{R} \times U(n, \mathbb{C}) \to \mathbb{C}$ denote the mapping given by

$$g(\omega, \Delta) = \det(I + L(j\omega)\Delta).$$

Since $L(s)$ is assumed to be proper, we can extend g to a mapping defined on $\overline{\mathbb{R}} \times U(n, \mathbb{C})$. The preimage $g^{-1}(0)$ will be of fundamental importance in our work. This preimage is a closed (hence compact) subset of $\overline{\mathbb{R}} \times U(n, \mathbb{C})$. Furthermore, if $\det(I +$

$L_\infty\Delta) \neq 0$ for every $\Delta \in U(n,\mathbb{C})$, then it is easy to see that $g^{-1}(0)$ is actually compact as a subset of $\mathbb{R} \times U(n,\mathbb{C})$. We will call such transfer functions *nicely proper*. This is the case, for example, if $||L_\infty|| < 1$ (the norm is the standard matrix norm).

From the Nyquist criterion, it is easy to see that the boundary of the stability set, $\partial\mathcal{S}$, is contained in the set

$$D = \{\Delta \in U(n,\mathbb{C}) \mid \exists\, \omega \in \overline{\mathbb{R}} \text{ such that } g(\omega,\Delta) = 0\}.$$

We also define the *gain-crossover region* by

$$\Omega = \{\omega \in \overline{\mathbb{R}} \mid \exists\, \Delta \in U(n,\mathbb{C}) \text{ such that } g(\omega,\Delta) = 0\}.$$

It is evident that D is compact (hence closed) in $U(n,\mathbb{C})$ and Ω is compact in $\overline{\mathbb{R}}$; furthermore, if $L(s)$ is nicely proper, then Ω is actually a compact subset of $\mathbb{R}$. If D is empty, then we are in the pleasant situation where no unitary perturbation destabilizes the system. On the other hand, if D is nonempty, then we define the *phase margin* of the system $L(s)$ by

$$\mathrm{PM}(L) = \min\{\Pi(\Delta) \mid \Delta \in D\}. \tag{2}$$

Observe that the minimum is attained since D is compact and, as we will see, the phase is a continuous function of its matrix argument. Consequently, closed-loop stability of the perturbed system is guaranteed if the perturbation Δ satisfies $\Pi(\Delta) < \mathrm{PM}(L)$. For computational purposes, it is convenient to redefine the phase margin in the following manner. Let $\mu\colon\Omega \to [0,\pi]$ be given by

$$\mu(\omega) = \min\{\Pi(\Delta) | \Delta \in U(n,\mathbb{C}) \text{ and } g(\omega,\Delta) = 0\}. \tag{3}$$

We will call μ the *minimum-phase mapping* of the system $L(s)$. Clearly the minimum-phase mapping and phase margin are related by

$$\mathrm{PM}(L) = \inf\{\mu(\omega) \mid \omega \in \Omega\}. \tag{4}$$

We will see later in section 4 that μ is always lower semicontinuous, so the compactness of $\Omega \subseteq \overline{\mathbb{R}}$ allows us to replace "inf" by "min" (see [13, p. 277]).

An effective means of computing the values of the minimum-phase mapping at a particular frequency ω in the gain crossover region Ω has been developed by J. R. Bar-on and E. A. Jonckheere in [3, 4]. They show that for $\omega \in \Omega$

$$\mu(\omega) = \cos^{-1}\left(-\frac{1}{2}\phi(\omega)\right), \tag{5}$$

where $\phi(\omega)$ is the minimum value of the constrained optimization problem:

$$\text{(CMP}\omega\text{)} \qquad \begin{array}{ll} \text{minimize} & z^*(L(j\omega)^* + L(j\omega))z, \quad z \in \mathbb{C}^n \quad (\omega \in \Omega \text{ fixed}) \\ \text{subject to} & \begin{cases} z^*z = 1, \\ z^*L(j\omega)^*L(j\omega)z = 1. \end{cases} \end{array}$$

Observe that the constraint set is always closed and bounded (hence compact) in $\mathbb{C}^n$. The continuity of the functional being minimized ensures a solution to the problem if the constraint set is nonempty. Thus the existence of solutions of CMPω depends on the feasibility of the constraints, which in turn is governed by the magnitude of

the (real) eigenvalues of the nonnegative definite (Hermitian) matrix $L(j\omega)^* L(j\omega)$; the nonnegative square roots of these eigenvalues are precisely the *singular values* of $L(j\omega)$. Specifically, letting $\overline{\sigma}(L(j\omega))$ (resp., $\underline{\sigma}(L(j\omega))$) denote the maximum (resp., minimum) singular value of $L(j\omega)$, we see that the constraint set in CMPω will be empty if and only if either $\overline{\sigma}(L(j\omega)) < 1$ or $\underline{\sigma}(L(j\omega)) > 1$. Consequently, the constraint set of CMPω is nonempty if and only if $\underline{\sigma}(L(j\omega)) \leq 1 \leq \overline{\sigma}(L(j\omega))$, in which case we say that the singular values of $L(j\omega)$ are "spread across one." It is shown in [3, 4] that the frequencies $\omega \in \overline{\mathbb{R}}$ for which the singular values of $L(j\omega)$ are spread across one coincide with the frequencies in the gain crossover region Ω; i.e.,

$$\Omega = \{\omega \in \overline{\mathbb{R}} \mid \underline{\sigma}(L(j\omega)) \leq 1 \leq \overline{\sigma}(L(j\omega))\}. \tag{6}$$

The continuity of the singular values as a function of the matrix (see, e.g., [20, p. 330]) gives an alternative proof of the fact that Ω is a closed (hence compact) subset of $\overline{\mathbb{R}}$ (and is a compact subset of $\mathbb{R}$ if $L(s)$ is nicely proper).

The problem CMPω is well suited to numerical computation because, after conversion to an equivalent real problem via a standard "decomplexification" process, it is a purely quadratic minimization problem whose global minima admit a precise characterization (we will have more to say on this in section 5). The phase margin can be effectively approximated by computing $\mu(\omega)$ on a suitably fine discrete subset of Ω and then selecting the smallest value of μ so obtained. While this method works well in practice, its success requires that μ be continuous as a function of ω. However, stronger regularity of μ yields additional desirable features of the numerical scheme by which PM(L) is computed. Specifically, conditions which guarantee the differentiability of μ coincide with conditions which guarantee the appropriate nonsingularity of the Jacobian when one attempts to solve CMPω (or its decomplexified equivalent) via the Lagrange multiplier technique and Newton's method, so one can interpret the differentiability of μ as a guarantee of the numerical stability of the routines by which PM(L) is computed. These considerations form the primary motivation for our study of the regularity properties of the minimum-phase mapping.

3. Generalities on parametrized optimization problems. The phase of a unitary matrix $\Delta \in U(n, \mathbb{C})$ is the maximum value of a constrained optimization problem in which the matrix Δ appears as a parameter (cf. equation (1)). Likewise, the value of the minimum-phase mapping at a specific frequency ω in the gain-crossover region ω is the minimum value of a constrained optimization problem CMPω in which the frequency ω appears as a parameter. Thus, the study of the regularity properties of the phase and the study of the regularity properties of the minimum-phase mapping are both special cases of the study of regularity properties of the "value function" for a parametrized family of constrained optimization problems. In this section we will state for future reference two results about the behavior of the value function for a certain class of parametrized constrained optimization problems.

Notation 3.1. Given a metric space X with metric d, we let $\mathcal{C}(X)$ denote the family of all nonempty compact subsets of X. For metric spaces X, Y, we recall that a *set-valued* function on Y is map $G : Y \to \mathcal{C}(X)$. As with single-valued functions, one can make precise what it means for a set-valued function to be continuous, upper semicontinuous (usc), and lower semicontinuous (lsc). We assume that these notions are familiar to the reader (see, e.g., [1, Chap. 1] and [13, Chap. 3] for details).

We will have occasion to refer to the following well-known proposition (see, e.g., [17]), which gives a simple sufficient condition for the lower semicontinuity of a set-valued function.

PROPOSITION 3.2. *Let X, Y be metric spaces, let $G: Y \to \mathcal{C}(X)$ be a set-valued function, and let $y_0 \in Y$ be such that for every $x \in G(y_0)$ there exist an open neighborhood U_0 of y_0 and a continuous function $\alpha_x : U_0 \to X$ such that $\alpha_x(y_0) = x$ and $\alpha_x(y) \in G(y)$ for every $y \in U_0$. Then G is lsc at y_0.*

The subsequent theorem and its corollary summarize the specific results that we will need concerning the continuity properties of the value function for parametrized constrained optimization problems. For the proofs, we refer the reader to [1, section 1.2] or [2, section 1.4]. (Note that in these references the value function is referred to as the "marginal function.")

THEOREM 3.3. *Let X, Y be metric spaces, let $f: X \times Y \to \mathbb{R}$ be continuous, let $G: Y \to \mathcal{C}(X)$ be a set-valued function, and define $\phi: Y \to \mathbb{R}$ by*

$$\phi(y) = \min \{f(x, y) \mid x \in G(y)\}.$$

Then

(a) *G usc $\Rightarrow$ ϕ lsc;*

(b) *G lsc $\Rightarrow$ ϕ usc;*

(c) *if G is continuous, then ϕ is continuous.*

COROLLARY 3.4. *Let X, Y, Z be metric spaces with X and Y compact, and let $\bar{z} \in Z$ be a fixed element. Let $f: X \times Y \to \mathbb{R}$ and $g: X \times Y \to Z$ be continuous mappings, and suppose that g has the additional property that for every $y \in Y$ there exists $x \in X$ such that $g(x, y) = \bar{z}$. Then for each fixed $y \in Y$ the constrained minimization problem*

$$\begin{array}{ll} \textit{minimize} & f(x, y) \quad x \in X \ (y \in Y \textit{ fixed}) \\ \textit{subject to} & g(x, y) = \bar{z} \end{array}$$

has a solution. Furthermore, if $\phi(y)$ denotes the constrained minimum value, then the mapping $\phi: Y \to \mathbb{R}$ is lower semicontinuous. In particular, there exists $\tilde{y} \in Y$ such that $\phi(\tilde{y}) \leq \phi(y)$ for every $y \in Y$, so $\min \{\phi(y) \mid y \in Y\}$ exists.

Proof. Define $G: Y \to \mathcal{C}(X)$ by $G(y) = \{x \in X \mid g(x, y) = \bar{z}\}$, verify G is usc, and apply the theorem. ∎

4. Continuity of the phase and minimum-phase mappings. The results of the previous section will now be applied to the study of the continuity properties of the phase and minimum-phase mappings. First we will show that phase of a unitary matrix, as defined in (1), is a continuous function of its matrix argument. Next we will consider the minimum-phase mapping (3) and show that, under very general conditions, it is an lsc function of frequency. A useful consequence of this lower semicontinuity is that we can replace "inf" by "min" in the definition (4) of the phase margin. We then formulate slightly more restrictive conditions which guarantee that the minimum-phase mapping is a continuous function of frequency.

PROPOSITION 4.1. *The phase mapping $\Pi: U(n, \mathbb{C}) \to [0, \pi]$ is continuous.*

Proof. The function $\cos^{-1}: [-1, 1] \to [0, \pi]$ is decreasing, so for each $\Delta \in U(n, \mathbb{C})$ we have

$$\Pi(\Delta) = \max \{\cos^{-1}(\mathrm{Re}(z^* \Delta z)) \mid z \in \mathbb{C}^n \text{ and } z^* z = 1\} = \cos^{-1}(\phi(\Delta)),$$

where $\phi: U(n, \mathbb{C}) \to \mathbb{R}$ is defined by

$$\phi(\Delta) = \min \{\mathrm{Re}(z^* \Delta z) \mid z \in \mathbb{C}^n \text{ and } z^* z = 1\}.$$

Thus to show that Π is continuous, it suffices to show that ϕ is continuous. Define mappings $f : \mathbb{C}^n \times U(n,\mathbb{C}) \to \mathbb{R}$ by $f(z,\Delta) = \mathrm{Re}(z^*\Delta z)$ and $G : U(n,\mathbb{C}) \to \mathcal{C}(\mathbb{C}^n)$ by $G(\Delta) = \{z \in \mathbb{C}^n \mid z^*z = 1\}$. It is clear that f is continuous. In addition, G is a constant set-valued function and is obviously continuous. With the above definitions we have

$$\phi(\Delta) = \min\{f(z,\Delta) \mid z \in G(\Delta)\},$$

so the continuity of ϕ is an immediate consequence of Theorem 3.3(c). ∎

THEOREM 4.2. *Let $L(s)$ be a $n \times n$ matrix function of $s \in \mathbb{C}$ whose entries are in the field of rational functions of s with real coefficients and assume that $L(s)$ satisfies the following conditions:*

(i) *$L(s)$ has no poles on the $j\omega$ axis;*

(ii) *$L(s)$ is proper in the sense that $\lim_{s\to\infty} L(s) = L_\infty$, where L_∞ is a real $n \times n$ matrix satisfying $\det(I + L_\infty) \neq 0$;*

(iii) *the gain-crossover region*

$$\Omega = \{\omega \in \overline{\mathbb{R}} \mid \exists \Delta \in U(n,\mathbb{C}) \text{ such that } \det(I + L(j\omega)\Delta) = 0\}$$

is nonempty. Then the minimum-phase mapping $\mu : \Omega \to [0,\pi]$ defined by

$$\mu(\omega) = \min\{\Pi(\Delta) \mid \Delta \in U(n,\mathbb{C}),\ \det(I + L(j\omega)\Delta) = 0\}$$

is lsc.

Proof. Let $g : U(n,\mathbb{C}) \times \Omega \to \mathbb{C}$ be defined by

$$g(\Delta,\omega) = \det(I + L(j\omega)\Delta),$$

where it is understood that we set $g(\Delta,\infty) = \det(I + L_\infty\Delta)$ in case $\infty \in \Omega$. Assumption (ii) ensures that g is continuous on its domain $U(n,\mathbb{C}) \times \Omega$, and this domain is obviously compact. The definition of Ω guarantees that for each $\omega \in \Omega$ there exists a $\Delta \in U(n,\mathbb{C})$ such that $g(\Delta,\Omega) = 0$. Moreover, the map $f : U(n,\mathbb{C}) \times \Omega \to [0,\pi]$ defined by $f(\Delta,\omega) = \Pi(\Delta)$ is continuous by Proposition 4.1. For each $\omega \in \Omega$ our definitions imply that $\mu(\omega)$ is the minimum value of the constrained optimization problem

$$\begin{aligned} &\text{minimize} && f(\Delta,\omega), \quad \Delta \in U(n,\mathbb{C}) \ (\omega \in \Omega \text{ fixed}), \\ &\text{subject to} && g(\Delta,\omega) = 0. \end{aligned}$$

Hence $\mu : \Omega \to [0,\pi]$ is lsc by Corollary 3.4. ∎

We next develop conditions under which the minimum-phase mapping is actually continuous as a function of frequency. These conditions involve a more subtle analysis of the singular values of $L(j\omega)$.

Remark 4.3. Let $L(s)$ be an $n \times n$ transfer-function matrix that satisfies the conditions of Theorem 4.2. By assumption, $L(s)$ has no poles on the $j\omega$-axis, so the matrix function $T(s) = L(-js)^T L(js)$ is holomorphic on a domain containing the real axis $s = \omega$. Furthermore, since the coefficients of the rational entries of $L(s)$ are real, it is clear that $T(s)$ is Hermitian for real s; i.e., $T(\omega)^* = T(\omega)$ for every $\omega \in \mathbb{R}$. From [14, Thm. S6.3] or [15, Chap. 2] we infer that there exists an $n \times n$ matrix function $P(s)$ which is holomorphic on a domain containing the real axis $s = \omega$, unitary on the real axis (i.e., $P(\omega)^* = P(\omega)^{-1}$ for $\omega \in \mathbb{R}$) and satisfies

$$P(\omega)^* T(\omega) P(\omega) = \mathrm{diag}\,[\lambda_1(\omega), \ldots, \lambda_n(\omega)] \tag{7}$$

for every $\omega \in \mathbb{R}$. In particular, (7) exhibits the (real and nonnegative) eigenvalues, $\lambda_i(\omega), i = 1, \ldots, n$, of $T(\omega)$ as real-analytic functions of ω. (Note, however, that no ordering of the eigenvalues is implied by the indexing.) We adopt the usual conventions regarding real analyticity at $\omega = \infty$: $\lambda_i(\omega)$ is real analytic at $\omega = \infty$ if and only if $\lambda_i(1/\omega)$ is real analytic at $\omega = 0$. Since $T(\omega)$ is nonnegative definite, we have $\lambda_i(\omega) \geq 0$ for each $i = 1, \ldots, n$. The singular values of $L(j\omega)$ are given by $\sigma_i(\omega) = \sqrt{\lambda_i(\omega)}, i = 1, \ldots, n$ (note that these too are not necessarily ordered); consequently, the singular values are real analytic on the open set where they are positive. It is clear that $\sigma_i(\omega) < 1 \iff \lambda_i(\omega) < 1$ and $\sigma_i(\omega) > 1 \iff \lambda_i(\omega) > 1$.

DEFINITION 4.4. *Let $L(s)$ be an $n \times n$ transfer-function matrix having the properties listed in Theorem* 4.2*, and let $\Omega \subseteq \overline{\mathbb{R}}$ denote its gain-crossover region. A frequency point $\omega_0 \in \Omega$ is said to be*

(a) Type I *if there exist $i, k \in \{1, \ldots, n\}$ such that $\sigma_i(\omega_0) < 1 < \sigma_k(\omega_0)$;*

(b) Type II *if it is not of Type* I*, the set of indices $I_1 = \{i \in \{1, \ldots, n\} \mid \sigma_i(\omega_0) = 1\}$ is a proper subset of $\{1, \ldots, n\}$, and one of the following conditions holds.*

(C1) $\overline{\sigma}(L(j\omega_0)) > 1$ *and there exists $\delta > 0$ such that $\sigma_i(\omega) \leq 1$ for every $i \in I_1$ and $\omega \in (\omega_0 - \delta, \omega_0 + \delta) \cap \Omega$.*

(C2) $\underline{\sigma}(L(j\omega_0)) < 1$ *and there exists $\delta > 0$ such that $\sigma_i(\omega) \geq 1$ for every $i \in I_1$ and $\omega \in (\omega_0 - \delta, \omega_0 + \delta) \cap \Omega$.*

Remark 4.5. (a) A frequency point $\omega_0 \in \Omega$ is Type I if and only if the singular values of $L(j\omega_0)$ are *strictly* spread across one in the sense that $\underline{\sigma}(L(j\omega_0)) < 1 < \overline{\sigma}(L(j\omega_0))$. Thus, the continuity of the maximum and minimum singular values as functions of frequency implies that a Type I frequency point $\omega_0 \in \Omega$ is an interior point of Ω. However, it can also be the case that Type II frequency points are interior points of Ω.

(b) If $\omega_0 \in \Omega$ is not Type I, then the set of indices I_1 is necessarily nonempty. This follows directly from the fact that $\omega_0 \in \Omega \iff \underline{\sigma}(L(j\omega_0)) \leq 1 \leq \overline{\sigma}(L(j\omega_0))$.

(c) If the transfer function $L(j\omega)$ has the property that for each $\omega \in \Omega$ at most one singular value is equal to one, then every $\omega \in \Omega$ that is not Type I will automatically be Type II.

(d) We refer the reader to Example 4.7 for an example of a transfer function whose gain-crossover region has an interior point that is neither Type I nor Type II.

THEOREM 4.6. *Let $L(s)$ be an $n \times n$ matrix function of $s \in \mathbb{C}$ whose entries are in the field of rational functions of s with real coefficients. Assume that $L(s)$ satisfies conditions* (i)*,* (ii)*, and* (iii) *of Theorem* 4.2 *and in addition has the property that every frequency ω in the gain-crossover region Ω is either Type* I *or Type* II*. Then the minimum-phase mapping $\mu : \Omega \to [0, \pi]$ is continuous.*

Proof. In light of Theorem 4.2 it suffices to prove that μ is usc. Recall that for $\omega \in \Omega$ the minimum-phase mapping $\mu(\omega)$ is given by (5), where $\phi(\omega)$ is the solution of the constrained minimization problem CMPω. Since $\cos^{-1} : [-1, 1] \to [0, \pi]$ is continuous and decreasing, it is easy to see that μ is usc if and only if ϕ is usc, so we will focus on proving the upper semicontinuity of ϕ on Ω. Let $X = \{z \in \mathbb{C}^n \mid z^* z = 1\}$ and define maps $\tilde{f}, \tilde{g} : X \times \Omega \to \mathbb{R}$ by

$$\tilde{f}(\zeta, \omega) = \zeta^* \big(L(j\omega)^* + L(j\omega)\big)\zeta,$$
$$\tilde{g}(\zeta, \omega) = \zeta^* L(j\omega)^* L(j\omega)\zeta.$$

Referring back to CMPω, we see that the first constraint is incorporated in the definition of X. Thus, for $\omega \in \Omega$ we have

$$\phi(\omega) = \min \{\tilde{f}(\zeta, \omega) \mid \zeta \in X \text{ and } \tilde{g}(\zeta, \omega) = 1\}.$$

As in Remark 4.3, we let $T(s) = L(-js)^T L(js)$, and we choose an $n \times n$ matrix function $P(s)$ which is holomorphic on a domain containing the real axis $s = \omega$, unitary on the real axis (i.e., $P(\omega)^* = P(\omega)^{-1}$ for $\omega \in \mathbb{R}$) and satisfies (7) for every $\omega \in \mathbb{R}$. Next we define maps $f, g : X \times \Omega \to \mathbb{R}$ by

$$\begin{aligned} f(z,\omega) &= \tilde{f}(P(\omega)z, \omega), \\ g(z,\omega) &= \tilde{g}(P(\omega)z, \omega) = z^* P(\omega)^* T(\omega) P(\omega) z. \end{aligned}$$

Observe that since $P(\omega)$ diagonalizes $T(\omega)$ we can rewrite $g(z,\omega)$ as

$$g(z,\omega) = \sum_{i=1}^{n} \lambda_i(\omega) |z_i|^2,$$

where $z_1, \ldots, z_n$ are the coordinates of z in $\mathbb{C}^n$. Since for each $\omega \in \Omega$ the unitary matrix $P(\omega)$ is an isometric bijection of the unit sphere X in $\mathbb{C}^n$ with itself, it is easy to see that

$$\phi(\omega) = \min\{f(z,\omega) \mid z \in X \text{ and } g(z,\omega) = 1\} = \min\{f(z,\omega) \mid z \in G(\omega)\},$$

where $G : \Omega \to \mathcal{C}(X)$ is the set-valued function given by

$$G(\omega) = \{z \in X \mid g(z,\omega) = 1\}.$$

By Theorem 3.3(b) to show that ϕ is usc, it suffices to show that G is lsc, and for this we will appeal to the criterion for lower semicontinuity given in Proposition 3.2.

Let ω_0 be an arbitrary frequency in Ω and let z_0 be an arbitrary point in $G(\omega_0)$. We must show that there exist a $\delta > 0$ and a continuous function $\alpha : (\omega_0 - \delta, \omega_0 + \delta) \cap \Omega \to X$ such that $\alpha(\omega_0) = z_0$ and $\alpha(\omega) \in G(\omega)$ for every $\omega \in (\omega_0 - \delta, \omega_0 + \delta) \cap \Omega$. Let $z_0 = (z_{01}, \ldots, z_{0n}) \in X \subseteq \mathbb{C}^n$ and note that

$$z_0 \in G(\omega_0) \iff g(z_0, \omega_0) = 1 \iff \sum_{i=1}^{n} (\lambda_i(\omega_0) - 1)|z_{0i}|^2 = 0. \tag{8}$$

(We have used the fact that $\sum_{i=1}^{n} |z_{0i}|^2 = (z_0)^* z_0 = 1$.) First let us suppose that there exists an index $k \in \{1, \ldots, n\}$ such that $\lambda_k(\omega_0) - 1 \neq 0$ and $z_{0k} \neq 0$. Then (8) yields

$$(\lambda_k(\omega_0) - 1)|z_{0k}|^2 = -\sum_{i \neq k} (\lambda_i(\omega_0) - 1)|z_{0i}|^2,$$

from which we obtain

$$|z_{0k}|^2 = -\sum_{i \neq k} \frac{\lambda_i(\omega_0) - 1}{\lambda_k(\omega_0) - 1} |z_{0i}|^2 > 0. \tag{9}$$

Choose $\delta > 0$ small enough so that for $\omega \in (\omega_0 - \delta, \omega_0 + \delta) \cap \Omega$ the function

$$\Lambda(\omega) = -\sum_{i \neq k} \frac{\lambda_i(\omega) - 1}{\lambda_k(\omega) - 1} |z_{0i}|^2$$

is defined and positive. The existence of such a δ follows from (9) and the continuity of the functions $\lambda_i(\omega),\ i = 1, \ldots, n$ with respect to ω. In particular, $\Lambda(\omega)$ is continuous

for $\omega \in (\omega_0 - \delta, \omega_0 + \delta)$. Choose $\theta_0 \in (-\pi, \pi]$ such that $z_{0k} = e^{j\theta_0}|z_{0k}|$ and for $\omega \in (\omega_0 - \delta, \omega_0 + \delta) \cap \Omega$ and $i \in \{1, \dots, n\}$ define complex-valued functions $z_i(\omega)$ by

$$z_i(\omega) = \begin{cases} z_{0i}, & i \neq k, \\ e^{j\theta_0}\sqrt{\Lambda(\omega)}, & i = k. \end{cases}$$

Observe that the function $z_k(\omega)$ is defined and continuous for $\omega \in (\omega_0 - \delta, \omega_0 + \delta) \cap \Omega$ since $\Lambda(\omega) > 0$ for these values of ω. The functions $z_i(\omega)$ for $i \neq k$ are obviously continuous since they are constant. Hence it is easy to see that the function α: $(\omega_0 - \delta, \omega_0 + \delta) \cap \Omega \to X$ defined by

$$\alpha(\omega) = \frac{1}{\sqrt{z(\omega)^* z(\omega)}} z(\omega), \quad z(\omega) = (z_1(\omega), \dots, z_n(\omega)), \tag{10}$$

is continuous and satisfies $\alpha(\omega_0) = z_0$ and $g(\alpha(\omega), \omega) = 1$ (or equivalently $\alpha(\omega) \in G(\omega)$) for $\omega \in (\omega_0 - \delta, \omega_0 + \delta)$. This completes the proof of the lower semicontinuity of G at the frequency ω_0 under the assumption that there exists an index $k \in \{1, \dots, n\}$ such that $\lambda_k(\omega_0) - 1 \neq 0$ and $z_{0k} \neq 0$. The reader will note that we have not yet made use of the assumption that every frequency point in the gain-crossover region is of Type I or II.

It remains to consider the case where for each $i \in \{1, \dots, n\}$ either $\lambda_i(\omega_0) - 1 = 0$ or $z_{0i} = 0$. In this case each term in the summation in (8) is zero. We set

$$I_1 = \big\{ i \in \{1, \dots, n\} \mid \lambda_i(\omega_0) - 1 = 0 \big\}, \quad I_2 = \{1, \dots, n\} \setminus I_1.$$

Observe that since $z_0 \in X \subseteq \mathbb{C}^n \setminus \{0\}$, not all of the coordinates z_{0i} can be zero; thus $I_1 \neq \emptyset$. On the other hand, the assumption that ω_0 is either of Type I or Type II forces $I_2 \neq \emptyset$. Moreover for $i \in I_2$ we have $z_{0i} = 0$. Equation (8) can be rewritten as

$$\sum_{i \in I_1} (\lambda_i(\omega_0) - 1)|z_{0i}|^2 + \sum_{i \in I_2} (\lambda_i(\omega_0) - 1)|z_{0i}|^2 = 0,$$

and this suggests that we define a real-analytic function $h(\omega)$ by

$$h(\omega) = \sum_{i \in I_1} (\lambda_i(\omega) - 1)|z_{0i}|^2. \tag{11}$$

Clearly $h(\omega_0) = 0$, so the real analyticity of h implies either that ω_0 is an isolated zero of h or that h is identically zero for all $\omega \in \mathbb{R}$. In either case there exists $\delta_0 > 0$ such that h does not change sign on the intervals $(\omega_0 - \delta_0, \omega_0]$ and $[\omega_0, \omega_0 + \delta_0)$; i.e., either $h(\omega) \leq 0$ for every $\omega \in (\omega_0 - \delta_0, \omega_0]$ or $h(\omega) \geq 0$ for every $\omega \in (\omega_0 - \delta_0, \omega_0]$, with a similar statement for the interval $[\omega_0, \omega_0 + \delta_0)$.

If ω_0 is of Type I, then let $i, k \in I_2$ be such that $\sigma_i(\omega_0) < 1 < \sigma_k(\omega_0)$. The continuity of the singular values yields a $\delta > 0$ such that $\delta \leq \delta_0$ and $\omega \in (\omega_0 - \delta, \omega_0 + \delta) \Rightarrow \sigma_i(\omega) < 1 < \sigma_k(\omega)$; in particular $(\omega_0 - \delta, \omega_0 + \delta) \subseteq \Omega$ and $\lambda_i(\omega) - 1 < 0 < \lambda_k(\omega) - 1$ for $\omega \in (\omega_0 - \delta, \omega_0 + \delta)$. Since h has constant sign on the interval $(\omega_0 - \delta, \omega_0]$, we can select an index $\ell \in I_2$ for which

$$\frac{1}{\lambda_\ell(\omega) - 1} h(\omega) \leq 0 \quad \forall\, \omega \in (\omega_0 - \delta, \omega_0]. \tag{12}$$

For $\omega \in (\omega_0 - \delta, \omega_0] \subseteq \Omega$ and $i \in \{1, \dots, n\}$ we then define complex-valued functions $z_i^-(\omega)$ by

$$z_i^-(\omega) = \begin{cases} z_{0i}, & i \in I_1, \\ 0, & i \in I_2, i \neq \ell, \\ \sqrt{-\frac{1}{\lambda_\ell(\omega) - 1} h(\omega)}, & i = \ell. \end{cases} \tag{13}$$

Observe that $z_\ell^-(\omega)$ is defined and continuous and has nonnegative real values by (12) (note, however, that $z_\ell^-(\omega)$ fails to have a one-sided derivative at ω_0 since $h(\omega_0) = 0$). The remaining functions $z_i^-(\omega)$ for $i \neq \ell$ are constant and thus obviously continuous. If we define $\alpha^- : (\omega_0 - \delta, \omega_0] \to X$ by the formula in (10) with the functions $z_i(\omega) = z_i^-(\omega)$ defined in (13), then it is easy to see that α^- is continuous and satisfies $\alpha^-(\omega_0) = z_0$ and $g(\alpha^-(\omega), \omega) = 1$. Since h also has constant sign on the interval $[\omega_0, \omega_0 + \delta)$, a similar argument yields a continuous map $\alpha^+ : \Omega \cap [\omega_0, \omega_0 + \delta) \to X$ such that $\alpha^+(\omega_0) = z_0$ and $g(\alpha^+(\omega), \omega) = 1$. We can then define $\alpha : (\omega_0 - \delta, \omega_0 + \delta) \to X$ by

$$\alpha(\omega) = \begin{cases} \alpha^-(\omega), & \omega_0 - \delta < \omega \leq \omega_0, \\ \alpha^+(\omega), & \omega_0 \leq \omega < \omega_0 + \delta. \end{cases}$$

It is evident that α is continuous, $\alpha(\omega_0) = z_0$, and $g(\alpha(\omega), \omega) = 1$ (or equivalently $\alpha(\omega) \in G(\omega)$) for every $\omega \in (\omega_0 - \delta, \omega_0 + \delta) \subseteq \Omega$. Consequently, G is lsc at ω_0 if ω_0 is of Type I.

The last case to consider is when ω_0 is of Type II. Then one of conditions C1 or C2 in Definition 4.4(b) holds. For concreteness we will carry out the remainder of the proof under the assumption that C1 holds (the proof when C2 holds is entirely analogous). For $\delta > 0$ as given in C1 and for $\omega \in (\omega_0 - \delta, \omega_0 + \delta) \cap \Omega$ the function h defined in (11) satisfies $h(\omega) \leq 0$. Furthermore, C1 yields an $\ell \in I_2$ such that $\sigma_\ell(\omega_0) > 1$ (in fact just choose ℓ so that $\sigma_\ell(\omega_0) = \overline{\sigma}(L(j\omega_0))$), which implies $\lambda_\ell(\omega_0) - 1 > 0$. The continuity of λ_ℓ enables us to shrink $\delta > 0$ if necessary so as to obtain $\lambda_\ell(\omega) - 1 > 0$ for every $\omega \in (\omega_0 - \delta, \omega_0 + \delta) \cap \Omega$. If we define $\alpha : (\omega_0 - \delta, \omega_0 + \delta) \cap \Omega \to X$ by the formula in (10), where the functions $z_i(\omega) = z_i^-(\omega)$ are as defined in (13), then once again α is continuous, $\alpha(\omega_0) = z_0$, and $g(\alpha(\omega), \omega) = 1$; equivalently, $\alpha(\omega) \in G(\omega)$ for every $\omega \in (\omega_0 - \delta, \omega_0 + \delta) \cap \Omega$. Consequently, G is lsc at ω_0 if ω_0 is of Type II and the proof is complete. □

Examples of systems with a continuous minimum-phase mapping are quite common; see, e.g., [4, section 7]. The graph of the minimum-phase mapping exhibited in the indicated reference gives numerical evidence of its continuity. However, continuity can also be deduced a priori via Theorem 4.6, since this example satisfies the conditions of Remark 4.5(c).

Example 4.7. For an example where the minimum-phase mapping fails to be continuous throughout the gain-crossover region, we consider the 3×3 transfer function

$$L(s) = \begin{bmatrix} \frac{80\kappa(s+9)(s+55)}{495(s+1)(s+80)} & 0 & 0 \\ 0 & \frac{\omega_n^2}{m_r(s^2+2\zeta\omega_n s+\omega_n^2)} & 0 \\ 0 & 0 & \frac{0.3(s+40)}{(s+20)} \end{bmatrix},$$

where $\kappa = 8.0953\ldots, \omega_r = 27.6002\ldots$ (the values are approximate), $\zeta = .05$, and

$$\omega_n = \frac{\omega_r}{\sqrt{1-2\zeta^2}}, \qquad m_r = \frac{1}{2\zeta\sqrt{1-\zeta^2}}.$$

Plots of the singular values of $L(j\omega)$ and the minimum-phase mapping $\mu(\omega)$ over the frequency range $0 \leq \omega \leq 60$ are shown in Figs. 1 and 2, respectively. Observe that although μ is an lsc function of ω as predicted by Theorem 4.2, μ has a discontinuity at $\omega = \omega_r = 27.6002\ldots$. An examination of the graph of the singular values of $L(j\omega)$ shows that the frequency ω_r is neither Type I nor Type II as specified in Definition 4.4; note that $\underline{\sigma}(L(j\omega_r)) < 1$, but it is not the case that both of the remaining singular

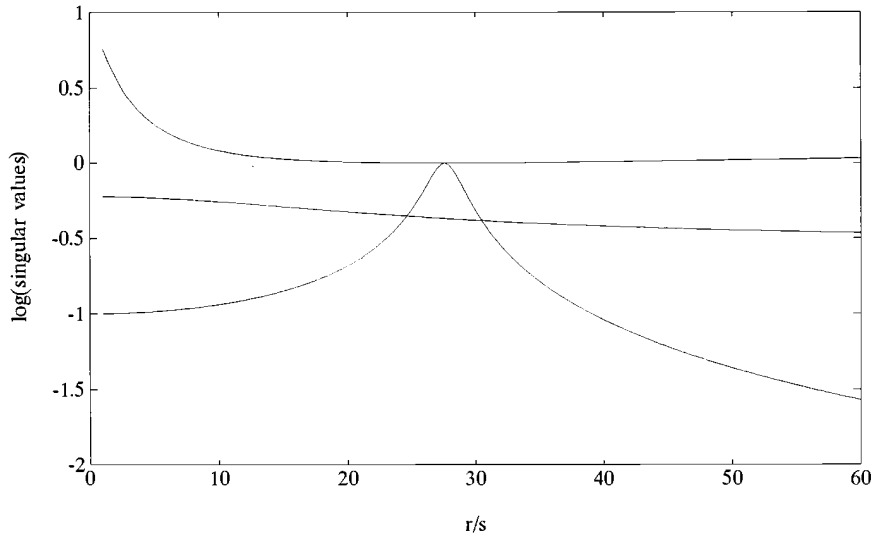


FIG. 1. *Plot of the* (*base* 10) *logarithm of the singular values against frequency for the transfer function in Example* 4.7.

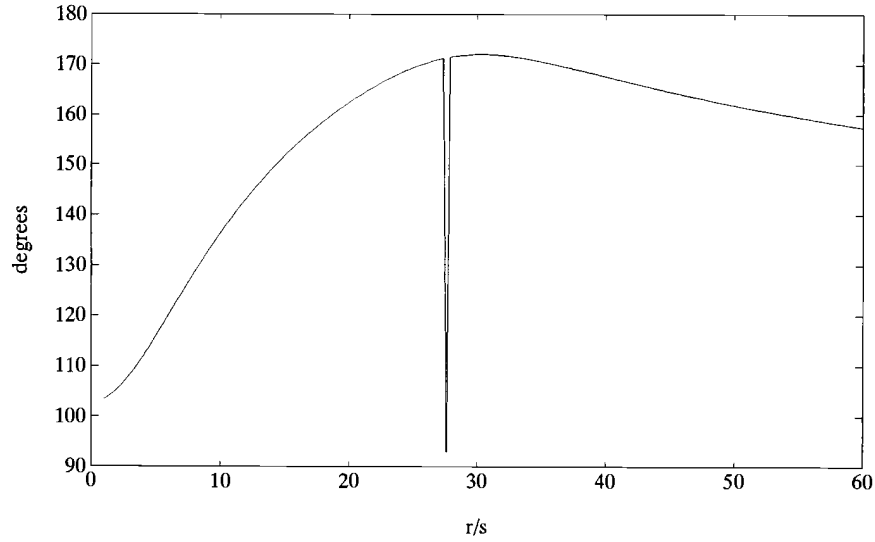


FIG. 2. *Plot of* $\mu(\omega)$ *for the transfer function in Example* 4.7.

values are ≥ 1 on an open interval containing ω_r, so C2 is violated. While simpler examples of such transfer functions could be constructed, the above example has the additional property of being closed-loop stable (and thus conforms to our standing assumptions).

5. Differentiability of the minimum-phase mapping. We next take up the question of the differentiability of the minimum-phase mapping $\mu : \Omega \to [0, \pi]$. As with the continuity question discussed in the previous section, it is convenient to express the minimum-phase mapping in terms of the value function of a parametrized family

of constrained optimization problems (see equation (5) and CMPω). However, the constrained optimization problem CMPω is not well suited to handling differentiability questions, since neither the functional nor the constraints are holomorphic functions of z or ω. Thus, we resort to the standard technique of converting CMPω to an equivalent real constrained optimization problem by the process of "decomplexification," which makes use of a canonical isomorphism between $\mathbb{C}^n$ and $\mathbb{R}^{2n}$. This process is described in [3], but it will also be reviewed here as there are specific aspects of the structure of the matrices obtained in the decomplexification process that are essential for our results. For other results on the differentiability of the value function of parametrized constrained optimization problems, the reader is referred to [16].

A (column) vector $z \in \mathbb{C}^n$ has a unique expression as $z = u + jv$, where u, v are (column) vectors in $\mathbb{R}^n$. This enables us to define a bijection

$$\Theta : \mathbb{C}^n \to \mathbb{R}^{2n}, \qquad \Theta(z) = \begin{bmatrix} u \\ v \end{bmatrix} \quad (\text{where } z = u + jv).$$

Indeed, if we view $\mathbb{C}^n$ as a real vector space, then Θ is an $\mathbb{R}$-linear isomorphism. Observe that if $x = \Theta(z)$, then $z^* z = x^T x$, where we view $x \in \mathbb{R}^{2n}$ as a $2n$-dimensional column vector and the superscript T denotes transpose. As is the case with vectors, a complex matrix $M \in \mathbb{C}^{n \times n}$ has a unique expression as $M = R + jS$, where $R, S \in \mathbb{R}^{n \times n}$. The linear mapping $M = R + jS$ from $\mathbb{C}^n$ into $\mathbb{C}^n$ and the bijection Θ together induce a linear mapping from $\mathbb{R}^{2n}$ into $\mathbb{R}^{2n}$, which is represented by the matrix

$$M_\Theta = \begin{bmatrix} R & -S \\ S & R \end{bmatrix} \in \mathbb{R}^{2n \times 2n}. \tag{14}$$

It will be convenient to let $\mathbb{R}_\Theta^{2n \times 2n}$ denote the subset of $\mathbb{R}^{2n \times 2n}$ consisting of those matrices $\mathcal{M}$ such that $\mathcal{M} = M_\Theta$ for some $M \in \mathbb{C}^{n \times n}$. One easily checks that the correspondence $M \mapsto M_\Theta$ of $\mathbb{C}^{n \times n}$ into $\mathbb{R}_\Theta^{2n \times 2n}$ is an isomorphism which preserves matrix sums and products. Of special interest is the matrix

$$\mathcal{J} = \begin{bmatrix} 0 & -I_n \\ I_n & 0 \end{bmatrix} \in \mathbb{R}^{2n \times 2n}$$

(I_n is the $n \times n$ identity matrix), which corresponds to the matrix $jI_n \in \mathbb{C}^{n \times n}$. It is clear that $\mathcal{J}^T = -\mathcal{J}$ and $\mathcal{J}^2 = -I_{2n}$. Moreover, $\mathcal{M}\mathcal{J} = \mathcal{J}\mathcal{M}$ for every matrix $\mathcal{M} \in \mathbb{R}_\Theta^{2n \times 2n}$ (in fact $\mathcal{M} \in \mathbb{R}^{2n \times 2n}$ is in $\mathbb{R}_\Theta^{2n \times 2n}$ if and only if $\mathcal{M}$ commutes with $\mathcal{J}$). It is also evident that $M \in \mathbb{C}^{n \times n}$ is Hermitian if and only if its associated matrix $M_\Theta \in \mathbb{R}_\Theta^{2n \times 2n}$ is symmetric. We use $\mathbb{R}_{\Theta,\mathrm{sym}}^{2n \times 2n}$ to denote the set of symmetric matrices in $\mathbb{R}_\Theta^{2n \times 2n}$.

Another useful fact about matrices in $\mathbb{R}_\Theta^{2n \times 2n}$ is the following. If λ is a real eigenvalue of a complex matrix $M = R + jS \in \mathbb{C}^{n \times n}$, then λ is also an eigenvalue of the corresponding matrix M_Θ. Moreover, if $x \in \mathbb{R}^{2n}$ is a (nonzero) eigenvector of M_Θ corresponding to λ, then it is easy to see that $\mathcal{J}x$ is also an eigenvector of M_Θ corresponding to λ and $\mathcal{J}x$ is orthogonal to x; i.e., $x^T \mathcal{J}\, x = 0$. It follows that every real eigenvalue of a matrix $\mathcal{M} \in \mathbb{R}_\Theta^{2n \times 2n}$ must have algebraic multiplicity greater than one. A more precise argument shows that every real eigenvalue has even multiplicity.

It is now a routine matter to reformulate the complex constrained optimization problem CMPω as a real constrained optimization problem. For each frequency ω in the gain-crossover region Ω we let

$$\mathcal{A}(\omega) = \big(L(j\omega)^* + L(j\omega)\big)_\Theta \in \mathbb{R}_{\Theta,\mathrm{sym}}^{2n \times 2n}, \qquad \mathcal{B}(\omega) = \big(L(j\omega)^* L(j\omega)\big)_\Theta \in \mathbb{R}_{\Theta,\mathrm{sym}}^{2n \times 2n} \tag{15}$$

denote the real symmetric matrices induced by the isomorphism Θ. Observe that $\mathcal{B}(\omega)$ is positive semidefinite, its eigenvalues coincide with those of $L(j\omega)^*L(j\omega)$, and the multiplicity of each eigenvalue of $\mathcal{B}(\omega)$ is twice that of the multiplicity of the corresponding eigenvalue of $L(j\omega)^*L(j\omega)$. In particular, the eigenvalues of $\mathcal{B}(\omega)$ are precisely the squares of the singular values of $L(j\omega)$, and consequently the eigenvalues of $\mathcal{B}(\omega)$ are spread across one if and only if the singular values of $L(j\omega)$ are spread across one. Since the entries of $L(s)$ are rational functions of s and, by assumption, have no poles on the $j\omega$-axis, the entries of $\mathcal{A}(\omega)$ and $\mathcal{B}(\omega)$ will be everywhere defined, rational functions of ω (i.e., all of their poles are complex). In particular, the entries of $\mathcal{A}(\omega)$ and $\mathcal{B}(\omega)$ are globally defined, real-analytic functions of $\omega \in \mathbb{R}$.

The real constrained optimization problem equivalent to CMPω then takes the form

$$\text{(RCMP}\omega\text{)} \qquad \begin{array}{ll} \text{minimize} & x^T\mathcal{A}(\omega)x, \quad x \in \mathbb{R}^{2n} \quad (\omega \in \Omega \text{ fixed}) \\ \text{subject to} & \begin{cases} x^Tx = 1, \\ x^T\mathcal{B}(\omega)x = 1. \end{cases} \end{array}$$

It is evident that the minimum value $\phi(\omega)$ of RCMPω coincides with the minimum value of CMPω, so we still have the formula (5) for the minimum-phase mapping. Thus, μ will be differentiable at every interior point ω_0 of the gain-crossover region Ω where $\phi'(\omega_0)$ exists and $|\phi(\omega_0)| < 2$. These considerations motivate us to first take up the differentiability with respect to ω of the minimum value $\phi(\omega)$ of the constrained optimization problem RCMPω.

Our main theorem on the differentiability of the minimum value $\phi(\omega)$ of RCMPω can be stated for slightly more general situations than described above. Specifically, it is not necessary to assume that the matrix functions $\mathcal{A}(\omega)$ and $\mathcal{B}(\omega)$ in $\mathbb{R}^{2n\times 2n}_{\Theta,\mathrm{sym}}$ are directly related to a transfer function $L(s)$ via the decomplexification process. Thus the differentiability theorem itself will not make any reference to transfer functions, but its relevance to transfer functions will be explained in Corollary 5.5. Furthermore, to make the assumptions in the differentiability theorem palatable, we will first make a few comments about the necessary conditions that apply to the constrained optimization problem RCMPω.

Fix ω_0 in the gain-crossover region Ω and let $\phi(\omega_0)$ be the solution to RCMPω_0; that is, let

$$\phi(\omega_0) = \min\{x^T\mathcal{A}(\omega_0)x \mid x^Tx = 1, x^T\mathcal{B}(\omega_0)x = 1\}.$$

Since the minimum is achieved, there exists $x_0 \in \mathbb{R}^{2n}$ such that $\phi(\omega_0) = (x_0)^T\mathcal{A}(\omega_0)x_0$ and $(x_0)^Tx_0 = (x_0)^T\mathcal{B}(\omega_0)x_0 = 1$. If the constraints are independent in the sense that the vectors x_0 and $\mathcal{B}(\omega_0)x_0$ are linearly independent, then the Lagrange multiplier theorem implies that there exist real numbers ξ_0, η_0 such that both the first-order condition

$$\text{(FO)} \qquad \big(\mathcal{A}(\omega_0) - \xi_0 I - \eta_0\mathcal{B}(\omega_0)\big)x_0 = 0$$

(I denotes the $2n \times 2n$ identity matrix) and the second-order condition

$$\text{(SO)} \quad x^T\big(\mathcal{A}(\omega_0) - \xi_0 I - \eta_0\mathcal{B}(\omega_0)\big)x \geq 0 \quad \forall\, x \in \mathbb{R}^{2n} \text{ such that } x^Tx_0 = x^T\mathcal{B}(\omega_0)x_0 = 0$$

are satisfied. Furthermore, under these assumptions the authors have shown in [6] that the strengthened version of the second-order condition

$$\text{(SSO)} \qquad x^T\big(\mathcal{A}(\omega_0) - \xi_0 I - \eta_0\mathcal{B}(\omega_0)\big)x \geq 0 \quad \forall\, x \in \mathbb{R}^{2n}$$

must hold for this problem. Conversely, if x_0 and $\mathcal{B}(\omega_0)x_0$ are linearly independent vectors that satisfy $(x_0)^T x_0 = (x_0)^T \mathcal{B}(\omega_0) x_0 = 1$ and if ξ_0 and η_0 are real numbers such that (FO) and (SSO) hold, then it is easy to see that $\phi(\omega_0) = \xi_0 + \eta_0$.

The following technical lemma will be used in the proof of the differentiability theorem. The proof of the lemma is straightforward and thus is omitted (a simple proof can be built around Corollary 3.4).

LEMMA 5.1. *Let $(\alpha, \beta) \subseteq \mathbb{R}$ be an open interval, and let $\eta, \nu : (\alpha, \beta) \to \mathbb{R}^q$ and $P : (\alpha, \beta) \to \mathbb{R}^{q \times q}$ be continuous maps. Suppose that for some $\omega_0 \in (\alpha, \beta)$ we have*

$$x^T P(\omega_0) x > 0 \quad \forall\, x \in \mathbb{R}^q \setminus \{0\} \text{ such that } x^T \eta(\omega_0) = x^T \nu(\omega_0) = 0.$$

Then there exists $\delta > 0$ such that $(\omega_0 - \delta, \omega_0 + \delta) \subseteq (\alpha, \beta)$ and for every $\omega \in (\omega_0 - \delta, \omega_0 + \delta)$ we have

$$x^T P(\omega) x > 0 \quad \forall\, x \in \mathbb{R}^q \setminus \{0\} \text{ such that } x^T \eta(\omega) = x^T \nu(\omega) = 0.$$

THEOREM 5.2. *Let $(\alpha, \beta) \subseteq \mathbb{R}$ be an open interval, and let*

$$\mathcal{A}, \mathcal{B} : (\alpha, \beta) \to \mathbb{R}^{2n \times 2n}_{\Theta,\mathrm{sym}}$$

be C^k, $1 \le k \le \infty$, (resp., real-analytic) functions such that $\mathcal{B}(\omega)$ is positive semidefinite and its eigenvalues are spread across one for each $\omega \in (\alpha, \beta)$. Define $\phi : (\alpha, \beta) \to \mathbb{R}$ by

$$\phi(\omega) = \min \{ x^T \mathcal{A}(\omega) x \mid x^T x = x^T \mathcal{B}(\omega) x = 1 \}.$$

Let $\omega_0 \in (\alpha, \beta)$ be such that there exists $x_0 \in \mathbb{R}^{2n}$ satisfying

(i) $(x_0)^T x_0 = (x_0)^T \mathcal{B}(\omega_0) x_0 = 1$;

(ii) *x_0 and $\mathcal{B}(\omega_0)x_0$ are linearly independent in $\mathbb{R}^{2n}$;*

(iii) *there exist real numbers ξ_0 and η_0 such that* (FO) *is satisfied and*

$$x^T \big(\mathcal{A}(\omega_0) - \xi_0 I - \eta_0 \mathcal{B}(\omega_0) \big) x > 0 \quad \forall\, x \in \mathbb{R}^{2n} \setminus \{0\} \text{ such that } x^T x_0 = x^T \mathcal{J}\, x_0 = 0.$$

Then $\phi(\omega_0) = (x_0)^T \mathcal{A}(\omega_0) x_0 = \xi_0 + \eta_0$ and ϕ is C^k (resp., real analytic) in a neighborhood of ω_0.

Remark 5.3. The hypotheses of Theorem 5.2 are quite reasonable in the sense that if $x_0 \in \mathbb{R}^{2n}$ is such that $\phi(\omega_0) = (x_0)^T \mathcal{A}(\omega_0) x_0$ and assumptions (i) and (ii) are satisfied, then by the discussion preceding Lemma 5.1 there will necessarily exist scalars ξ_0, η_0 such that the necessary conditions (FO) and (SSO) hold, so in particular the matrix

$$\mathcal{A}(\omega_0) - \xi_0 I - \eta_0 \mathcal{B}(\omega_0) \tag{16}$$

is positive semidefinite. Assumption (iii) demands more than (SSO) as it requires that the matrix (16) has a nontrivial null space of dimension 2 and is positive definite on the orthogonal complement of its null space. As was noted previously, the minimum positive dimension of the null space of (16) is 2 (the algebraic and geometric multiplicities of the eigenvalue 0 coincide since (16) is real symmetric).

Proof of Theorem 5.2. Define a mapping $F : \mathbb{R}^{2n} \times \mathbb{R} \times \mathbb{R} \times \mathbb{R} \times (\alpha, \beta) \to \mathbb{R}^{2n} \times \mathbb{R} \times \mathbb{R} \times \mathbb{R}$ by

$$F(x, \xi, \eta, \zeta, \omega) = \big((\mathcal{A}(\omega) - \xi I - \eta \mathcal{B}(\omega) - \zeta \mathcal{J}\,) x, x^T x, x^T \mathcal{B}(\omega) x, x_0^T \mathcal{J}\, x \big).$$

It is clear that F is C^k (resp., real analytic) with respect to the group of variables $x, \xi, \eta, \zeta, \omega$, and we have

$$F(x_0, \xi_0, \eta_0, 0, \omega_0) = (0, 1, 1, 0).$$

(Note that $(x_0)^T \mathcal{J} x_0 = 0$ by the skew symmetry of $\mathcal{J}$.) If we fix ω at the value ω_0 and let

$$\overline{D}F(x_0, \xi_0, \eta_0, 0, \omega_0) : \mathbb{R}^{2n+3} \to \mathbb{R}^{2n+3}$$

denote the Fréchet derivative of the map

$$(x, \xi, \eta, \zeta) \mapsto F(x, \xi, \eta, \zeta, \omega_0)$$

at $(x_0, \xi_0, \eta_0, 0)$, then this Fréchet derivative has a representation as a $(2n+3)\times(2n+3)$ matrix of the form

$$\overline{D}F(x_0, \xi_0, \eta_0, 0, \omega_0) = \begin{bmatrix} \mathcal{A}(\omega_0) - \xi_0 I - \eta_0 \mathcal{B}(\omega_0) & -x_0 & -\mathcal{B}(\omega_0) x_0 & -\mathcal{J}\, x_0 \\ 2(x_0)^T & 0 & 0 & 0 \\ 2(x_0)^T \mathcal{B}(\omega_0) & 0 & 0 & 0 \\ x_0^T \mathcal{J} & 0 & 0 & 0 \end{bmatrix}. \tag{17}$$

(We use the convention that vectors $x \in \mathbb{R}^{2n}$ are viewed as $2n$-dimensional column vectors in the matrix representation.) We claim that the linear map $\overline{D}F(x_0, \xi_0, \eta_0, 0, \omega_0)$ is a linear isomorphism of $\mathbb{R}^{2n+3}$ with itself. To prove this it suffices to prove that $\overline{D}F(x_0, \xi_0, \eta_0, 0, \omega_0)$ is one to one. Indeed, if $[x, \xi, \eta, \zeta]^T \in \mathbb{R}^{2n+3}$ is in the null space of the matrix (17), then

$$\begin{aligned} &\big(\mathcal{A}(\omega_0) - \xi_0 I - \eta_0 \mathcal{B}(\omega_0)\big) x - \xi x_0 - \eta \mathcal{B}(\omega_0) x_0 - \zeta \mathcal{J}\, x_0 = 0, \\ &(x_0)^T x = 0, \quad (x_0)^T \mathcal{B}(\omega_0) x = 0, \quad (x_0)^T \mathcal{J}\, x = 0. \end{aligned} \tag{18}$$

Multiply the first equation on the left by x^T and use the remaining equations, along with the symmetry of $\mathcal{B}(\omega_0)$ and the skew symmetry of $\mathcal{J}$, to obtain

$$x^T \big(\mathcal{A}(\omega_0) - \xi_0 I - \eta_0 \mathcal{B}(\omega_0)\big) x = 0.$$

Since $x^T x_0 = x^T \mathcal{J}\, x_0 = 0$, assumption (iii) implies that $x = 0$. Using this in the first of the equations (18), we obtain

$$\xi x_0 + \eta \mathcal{B}(\omega_0) x_0 + \zeta \mathcal{J}\, x_0 = 0. \tag{19}$$

Multiplication of (19) on the right by $x_0^T \mathcal{J}^{\,T} = -x_0^T \mathcal{J}$ results in

$$-\xi x_0^T \mathcal{J}\, x_0 - \eta x_0^T \mathcal{J} \mathcal{B}(\omega_0) x_0 - \zeta x_0^T \mathcal{J}^{\,2} x_0 = 0. \tag{20}$$

It has already been pointed out that $\mathcal{J}$ is skew symmetric, and we note in addition that $\mathcal{J}\mathcal{B}(\omega_0)$ is skew symmetric, since $\mathcal{B}(\omega_0)$ is symmetric and commutes with $\mathcal{J}$. The skew symmetry of these matrices yields $x_0^T \mathcal{J}\, x_0 = x_0^T \mathcal{J} \mathcal{B}(\omega_0) x_0 = 0$, so (20) becomes

$$0 = -\zeta x_0^T \mathcal{J}^{\,2} x_0 = \zeta x_0^T x_0 = \zeta,$$

since $\mathcal{J}^{\,2} = -I_{2n}$ and x_0 satisfies the constraint (i). Put $\zeta = 0$ in (19) and use the assumed linear independence of x_0 and $\mathcal{B}(\omega_0) x_0$ to conclude that $\xi = \eta = 0$. Hence the derivative $\overline{D}F(x_0, \xi_0, \eta_0, 0, \omega_0)$ is a linear isomorphism of $\mathbb{R}^{2n+3}$ with itself. By

the implicit function theorem, there exists $\delta_1 > 0$ such that $(\omega_0 - \delta_1, \omega_0 + \delta_1) \subseteq (\alpha, \beta)$ and there exist C^k (resp., real-analytic) functions

$$\sigma : (\omega_0 - \delta_1, \omega_0 + \delta_1) \to \mathbb{R}^{2n}, \qquad \xi, \eta, \zeta : (\omega_0 - \delta_1, \omega_0 + \delta_1) \to \mathbb{R},$$

which satisfy $\sigma(\omega_0) = x_0, \xi(\omega_0) = \xi_0, \eta(\omega_0) = \eta_0, \zeta(\omega_0) = 0$ and

$$F(\sigma(\omega), \xi(\omega), \eta(\omega), \zeta(\omega), \omega) = (0, 1, 1, 0) \quad \forall\, \omega \in (\omega_0 - \delta_1, \omega_0 + \delta_1). \tag{21}$$

For fixed $\omega \in (\omega_0 - \delta, \omega_0 + \delta)$ equation (21) yields

$$\big(\mathcal{A}(\omega) - \xi(\omega) I - \eta(\omega)\mathcal{B}(\omega) - \zeta(\omega)\mathcal{J}\,\big)\sigma(\omega) = 0, \tag{22a}$$

$$\sigma(\omega)^T \sigma(\omega) = 1, \tag{22b}$$

$$\sigma(\omega)^T \mathcal{B}(\omega)\sigma(\omega) = 1. \tag{22c}$$

We claim that $\zeta(\omega) = 0$ for every $\omega \in (\omega_0 - \delta_1, \omega_0 + \delta_1)$. To see this note that equation (22a) yields

$$\big(\mathcal{A}(\omega) - \xi(\omega) I - \eta(\omega)\mathcal{B}(\omega)\big)\sigma(\omega) = \zeta(\omega)\mathcal{J}\,\sigma(\omega). \tag{23}$$

Apply $\sigma(\omega)^T \mathcal{J}^{\,T}$ to both sides of (23) to get

$$\sigma(\omega)^T \mathcal{J}^{\,T} \big(\mathcal{A}(\omega) - \xi(\omega) I - \eta(\omega)\mathcal{B}(\omega)\big)\sigma(\omega) = \zeta(\omega)\sigma(\omega)^T \mathcal{J}^{\,T} \mathcal{J}\,\sigma(\omega). \tag{24}$$

Since $||\mathcal{J}\,\sigma(\omega)|| = ||\sigma(\omega)|| = 1$, the right-hand side of (24) clearly equals $\zeta(\omega)$. Furthermore, since $\mathcal{J}$ is skew symmetric and commutes with the real-symmetric matrix

$$\mathcal{A}(\omega) - \xi(\omega) I - \eta(\omega)\mathcal{B}(\omega), \tag{25}$$

it follows that $\mathcal{J}^{\,T}(\mathcal{A}(\omega) - \xi(\omega) I - \eta(\omega)\mathcal{B}(\omega))$ is skew symmetric, whence the left-hand side of (24) is 0. Thus $\zeta(\omega) = 0$. Equation (23) now takes the form

$$\big(\mathcal{A}(\omega) - \xi(\omega) I - \eta(\omega)\mathcal{B}(\omega)\big)\sigma(\omega) = 0. \tag{26}$$

Multiplying (26) by $\mathcal{J}$ and using the fact that $\mathcal{J}$ commutes with (25), we also obtain

$$\big(\mathcal{A}(\omega) - \xi(\omega) I - \eta(\omega)\mathcal{B}(\omega)\big)\mathcal{J}\,\sigma(\omega) = 0. \tag{27}$$

By assumption (iii) and Lemma 5.1 there exists $\delta_2 > 0$ such that $(\omega_0 - \delta_2, \omega_0 + \delta_2) \subseteq (\alpha, \beta)$ and for $\omega \in (\omega_0 - \delta_2, \omega_0 + \delta_2)$ we have

$$x^T \big(\mathcal{A}(\omega) - \xi(\omega) I - \eta(\omega)\mathcal{B}(\omega)\big)x > 0 \quad \forall\, x \in \mathbb{R}^{2n} \setminus \{0\} \text{ such that } x^T \sigma(\omega) = x^T \mathcal{J}\,\sigma(\omega) = 0. \tag{28}$$

For $\delta = \min\{\delta_1, \delta_2\}$ and for $\omega \in (\omega_0 - \delta, \omega_0 + \delta)$, equations (26) and (27) show that zero is an eigenvalue of the real-symmetric matrix (25) and $\sigma(\omega), \mathcal{J}\,\sigma(\omega)$ are corresponding orthogonal, unit-length eigenvectors. If we extend the set $\{\sigma(\omega), \mathcal{J}\,\sigma(\omega)\}$ to an orthonormal basis of $\mathbb{R}^{2n}$ consisting of eigenvectors of the matrix (25), then (28) implies that the remaining $2n-2$ eigenvalues of (25) (each eigenvalue listed as often as its multiplicity indicates) are positive. We conclude that for every $\omega \in (\omega_0 - \delta, \omega_0 + \delta)$

$$x^T \big(\mathcal{A}(\omega) - \xi(\omega) I - \eta(\omega)\mathcal{B}(\omega)\big)x \geq 0 \quad \forall\, x \in \mathbb{R}^{2n},$$

or equivalently

$$x^T\mathcal{A}(\omega)x \geq \xi(\omega)x^Tx + \eta(\omega)x^T\mathcal{B}(\omega)x \quad \forall x \in \mathbb{R}^{2n}. \tag{29}$$

In particular, if $x \in \mathbb{R}^{2n}$ satisfies $x^Tx = x^T\mathcal{B}(\omega)x = 1$, then (29) yields

$$x^T\mathcal{A}(\omega)x \geq \xi(\omega) + \eta(\omega). \tag{30}$$

Furthermore, if we multiply equation (26) on the right by $\sigma(\omega)^T$ and make use of (22b) and (22c), we obtain

$$\sigma(\omega)^T\mathcal{A}(\omega)\sigma(\omega) = \xi(\omega) + \eta(\omega). \tag{31}$$

From (30) and (31) we infer that

$$\phi(\omega) = \min\{x^T\mathcal{A}(\omega)x \mid x^Tx = x^T\mathcal{B}(\omega)x = 1\} = \xi(\omega) + \eta(\omega).$$

Consequently, $\phi : (\omega_0 - \delta, \omega_0 + \delta) \to \mathbb{R}$ is C^k (resp., real analytic), since the same is true of the functions ξ and η. This completes the proof. ∎

The following example demonstrates that we need some sort of strengthening of the basic necessary conditions (FO) and (SSO), such as assumption (iii) of Theorem 5.2, if we are to expect differentiability of ϕ.

Example 5.4. Consider the parametrized constrained optimization problem

$$\text{minimize} \quad 2\omega(x_1x_3 + x_4x_6) = x^T\mathcal{A}(\omega)x,$$

where $\omega \in \mathbb{R}$, $x = \operatorname{col}[x_1, x_2, x_3, x_4, x_5, x_6] \in \mathbb{R}^6$ and

$$\mathcal{A}(\omega) = \begin{bmatrix} A(\omega) & 0_{3\times 3} \\ 0_{3\times 3} & A(\omega) \end{bmatrix} \quad \text{with} \quad A(\omega) = \begin{bmatrix} 0 & 0 & \omega \\ 0 & 0 & 0 \\ \omega & 0 & 0 \end{bmatrix}$$

subject to the constraints $x^Tx = x^T\mathcal{B}(\omega)x = 1$, where $\mathcal{B}(\omega)$ is the constant 6×6 diagonal matrix

$$\mathcal{B}(\omega) = \operatorname{diag}[2, 1, 0, 2, 1, 0].$$

The matrix functions $\mathcal{A}(\omega)$ and $\mathcal{B}(\omega)$ are evidently real analytic and take values in $\mathbb{R}^{6\times 6}_{\Theta,\mathrm{sym}}$. Elementary computations show that

$$\phi(\omega) = \min\{x^T\mathcal{A}(\omega)x \mid x^Tx = x^T\mathcal{B}(\omega)x = 1\} = -|\omega|,$$

so ϕ fails to be differentiable at $\omega = 0$. One can readily see that this example fails to satisfy condition (iii) of Theorem 5.2.

We conclude by applying Theorem 5.2 to the specific case of the minimum-phase mapping of a multivariable control system.

COROLLARY 5.5. *Let $L(s)$ be an $n \times n$ matrix function of $s \in \mathbb{C}$ whose entries are in the field of rational functions of s with real coefficients. Assume that $L(s)$ satisfies conditions* (i), (ii), *and* (iii) *of Theorem* 4.2, *and let ω_0 be a frequency in the gain-crossover region Ω such that*

(i) *$I - L(j\omega_0)^*L(j\omega_0)$ is invertible (equivalently, one is not an eigenvalue of $L(j\omega_0)^*L(j\omega_0)$),*

(ii) *all of the eigenvalues of the matrix*

$$\big(L(j\omega_0)^* + L(j\omega_0) - \phi(\omega_0)L(j\omega_0)^*L(j\omega_0)\big)\big(I - L(j\omega_0)^*L(j\omega_0)\big)^{-1} \tag{32}$$

have multiplicity one, where $\phi(\omega_0)$ *is the minimum value of the constrained optimization problem* CMPω *(or, equivalently,* RCMPω*) with* $\omega = \omega_0$.

Then ω_0 *is an interior point of* Ω *and the minimum-phase mapping* $\mu : \Omega \to [0, \pi]$ *is real analytic on some open subinterval of* Ω *containing* ω_0.

Proof. Observe that (i) forces the singular values of $L(j\omega_0)$ to be strictly spread across one, so ω_0 is a Type I frequency point as specified in Definition 4.4.(a). Thus ω_0 is an interior point of Ω (cf. Remark 4.5.(a)), and continuity of μ in a neighborhood of ω_0 is assured by Theorem 4.6. Let $\mathcal{A}(\omega)$ and $\mathcal{B}(\omega)$ be as defined in (15) and for $\omega \in \Omega$ let $\phi(\omega)$ be the common minimum value of the associated constrained optimization problems CMPω and RCMPω. The eigenvalues of $\mathcal{B}(\omega)$ are real and nonnegative, and coincide with the eigenvalues of $L(j\omega)^*L(j\omega)$. Note, however, that the algebraic multiplicity of each eigenvalue of $\mathcal{B}(\omega)$ is twice that of the corresponding eigenvalue of $L(j\omega)^*L(j\omega)$. In particular, one is not an eigenvalue of $\mathcal{B}(\omega_0)$. By continuity of the maximum and minimum eigenvalues of $\mathcal{B}(\omega)$ as functions of ω, we can find an open interval $(\alpha, \beta) \subseteq \Omega$ such that $\omega_0 \in (\alpha, \beta)$ and the eigenvalues of $\mathcal{B}(\omega)$ are spread across one for every $\omega \in (\alpha, \beta)$.

Choose $x_0 \in \mathbb{R}^{2n}$ such that $x_0^T x_0 = x_0^T \mathcal{B}(\omega_0) x_0 = 1$ and $\phi(\omega_0) = x_0^T \mathcal{A}(\omega_0) x_0$. The existence of x_0 is guaranteed because the minimum value of RCMPω_0 is achieved on the constraint set. We claim that vectors x_0 and $\mathcal{B}(\omega_0)x_0$ must be linearly independent. For otherwise, there exists a real scalar λ such that $\mathcal{B}(\omega_0)x_0 = \lambda x_0$ (i.e., λ is an eigenvalue of $\mathcal{B}(\omega_0)$). But then the constraints on x_0 yield

$$1 = x_0^T \mathcal{B}(\omega_0) x_0 = x_0^T (\lambda x_0) = \lambda x_0^T x_0 = \lambda,$$

which contradicts the fact that one is not an eigenvalue of $\mathcal{B}(\omega_0)$.

Since x_0 and $\mathcal{B}(\omega_0)x_0$ are linearly independent, we can invoke the standard necessary conditions for RCMPω_0 discussed just prior to the statement of Lemma 5.1. Thus, there exist real numbers ξ_0, η_0 such that (FO) and (SSO) are satisfied. Furthermore $\phi(\omega_0) = \xi_0 + \eta_0$. Condition (FO) says that zero is an eigenvalue of the matrix

$$\mathcal{A}(\omega_0) - \xi_0 I - \eta_0 \mathcal{B}(\omega_0), \tag{33}$$

and from the structure of this matrix we know that the algebraic multiplicity of the eigenvalue zero must be even. Moreover, $\mathcal{J} x_0$ is a second linearly independent eigenvector of (33) corresponding to zero.

Let $N_0 \subseteq \mathbb{R}^{2n}$ denote the null space of the matrix (33). Then $\{x_0, \mathcal{J}x_0\} \subseteq N_0$, and we claim that

$$N_0 = \operatorname{span}\{x_0, \mathcal{J}x_0\}. \tag{34}$$

Since $\dim N_0 \geq 2$, to show (34) it suffices to show that $\dim N_0 \leq 2$. Using the equation $\phi(\omega_0) = \xi_0 + \eta_0$, we obtain

$$\begin{aligned} x \in N_0 &\Leftrightarrow \big(\mathcal{A}(\omega_0) - \xi_0 I - (\phi(\omega_0) - \xi_0)\mathcal{B}(\omega_0)\big)x = 0 \\ &\Leftrightarrow \big(\mathcal{A}(\omega_0) - \phi(\omega_0)\mathcal{B}(\omega_0) - \xi_0(I - \mathcal{B}(\omega_0))\big)x = 0 \\ &\Leftrightarrow \big[\big(\mathcal{A}(\omega_0) - \phi(\omega_0)\mathcal{B}(\omega_0)\big)(I - \mathcal{B}(\omega_0))^{-1} - \xi_0 I\big](I - \mathcal{B}(\omega_0))x = 0. \end{aligned}$$

Observe that $I - \mathcal{B}(\omega_0)$ is invertible since one is not an eigenvalue of $\mathcal{B}(\omega_0)$ by assumption. It follows that ξ_0 is an eigenvalue of the matrix

$$\big(\mathcal{A}(\omega_0) - \phi(\omega_0)\mathcal{B}(\omega_0)\big)(I - \mathcal{B}(\omega_0))^{-1} \tag{35}$$

and $I - \mathcal{B}(\omega_0)$ sets up a linear isomorphism between N_0 and the eigenspace of (35) corresponding to the eigenvalue ξ_0. From the formulas in (15) it is clear that the matrix (35) corresponds to the matrix (32) under the decomplexification isomorphism Θ. Since the eigenvalues of (32) are all of (algebraic) multiplicity 1 by assumption, we infer from that all real eigenvalues of (35) have algebraic multiplicity 2. Hence the dimension of eigenspace of (35) corresponding to ξ_0, also known as the geometric multiplicity of ξ_0, is at most two. Consequently, $\dim N_0 \leq 2$ and we obtain (34).

From (SSO) and (34), we deduce that any eigenvalue of (33) that corresponds to an eigenvector orthogonal to N_0 must be positive, and thus (33) is positive definite on the orthogonal complement of N_0. Theorem 5.2 then implies that ϕ is real analytic in some open subinterval of Ω containing ω_0.

The formula (5) will immediately yield the real analyticity of the minimum-phase mapping μ in a neighborhood of ω_0 once we know that $|\phi(\omega_0)| < 2$ (since $\cos^{-1}$ is real analytic on the open interval $(-1, 1)$). Let $z_0 \in \mathbb{C}^n$ be such that $\Theta(z_0) = x_0$. Then

$$\phi(\omega_0) = x_0^T \mathcal{A}(\omega_0) x_0 \Rightarrow \phi(\omega_0) = (z_0)^* L(j\omega_0)^* z_0 + (z_0)^* L(j\omega_0) z_0$$

and

$$x_0^T x_0 = x_0^T \mathcal{B}(\omega_0) x_0 = 1 \Rightarrow (z_0)^* z_0 = (z_0)^* L(j\omega_0)^* L(j\omega_0) z_0 = 1,$$

so that $\|z_0\| = \|L(j\omega_0)z_0\| = 1$. The Cauchy–Schwarz inequality yields

$$|(z_0)^* L(j\omega_0) z_0| \leq \|z_0\| \, \|L(j\omega_0)z_0\| = 1,$$

so it follows that $|\phi(\omega_0)| = 2$ if and only if equality holds in the Cauchy–Schwarz inequality. This in turn will hold if and only if both $L(j\omega_0)z_0$ and $L(j\omega_0)^* z_0$ are multiples of z_0; i.e., $L(j\omega_0)z_0 = \lambda z_0$ and $L(j\omega_0)^* z_0 = \rho z_0$ for some $\lambda, \rho \in \mathbb{C}$. However, this would give

$$1 = (z_0)^* L(j\omega_0)^* L(j\omega_0) z_0 = (\lambda z_0)^* (\lambda z_0) = |\lambda|^2 (z_0)^* z_0 = |\lambda|^2,$$

so $|\lambda| = 1$. Similarly, $|\rho| = 1$. We infer that

$$L(j\omega_0)^* L(j\omega_0) z_0 = L(j\omega_0)^* (\lambda z_0) = \lambda\rho z_0,$$

so $\lambda\rho$ is a eigenvalue of $L(j\omega_0)^* L(j\omega_0)$ of modulus 1. Since all eigenvalues of the matrix $L(j\omega_0)^* L(j\omega_0)$ are real and nonnegative, we obtain $\lambda\rho = 1$, which contradicts the assumption that one is not an eigenvalue of $L(j\omega_0)^* L(j\omega_0)$. Thus $|\phi(\omega_0)| < 2$ and the proof of the corollary is complete. ∎

REFERENCES

[1] J. P. Aubin and A. Cellina, *Differential Inclusions*, Springer-Verlag, New York, 1984.

[2] J. P. Aubin and H. Frankowska, *Set-Valued Analysis*, Birkhäuser Boston, Cambridge, MA, 1990.

[3] J. R. Bar-on, *Phase and Gain Margins for Multivariable Control Systems*, Ph.D. thesis, University of Southern California, Los Angeles, CA, 1990.

[4] J. R. Bar-on and E. A. Jonckheere, *Phase margins for multivariable control systems*, Internat. J. Control, 52 (1990), pp. 485–498.
[5] J. R. Bar-on and E. A. Jonckheere, *The geometry of the multivariable phase margin*, IEEE Trans. Automat. Control, AC–37 (1992), pp. 798–800.
[6] J. R. Bar-on and K. A. Grasse, *Global optimization of a quadratic functional with quadratic equality constraints*, J. Optim. Theory Appl., 82 (1994), pp. 379–386.
[7] D. S. Bernstein and W. M. Haddad, *Is there more to robust control than small gain?* in Proc. of the American Control Conference, Chicago, IL, June 1992, pp. 83–84.
[8] F. M. Callier and C. A. Desoer, *Linear Systems Theory*, Springer-Verlag, New York, 1991.
[9] R. R. E. DeGaston, *Exact calculation of the multiloop stability margin*, IEEE Trans. Automat. Control, AC-33 (1988), pp. 156–171.
[10] C. A. Desoer and Y. T. Wang, *On the generalized Nyquist stability criterion*, IEEE Trans. Automat. Control, AC–25 (1980), pp. 187–196.
[11] J. C. Doyle, *Analysis of feedback systems with structured uncertainties*, IEE Proceedings, Pt. D, 129 (1982), pp. 242–250.
[12] J. C. Doyle and G. Stein, *Multivariable feedback design: Concepts for a classical/modern synthesis*, IEEE Trans. Automat. Control, AC–26 (1981), pp. 4–16.
[13] J. Dugundji, *Topology*, Allyn and Bacon, Boston, 1966.
[14] I. Gohberg, P. Lancaster, and L. Rodman, *Matrix Polynomials*, Academic Press, New York, 1982.
[15] T. Kato, *Perturbation Theory for Linear Operators*, Springer-Verlag, New York, 1966.
[16] E. S. Levitin, *On differential problems of the optimal value of parametric problems of mathematical programming*, Soviet Math. Dokl., 15 (1974), pp. 603–608.
[17] E. Micheal, *Continuous selections* I, Ann. of Math., 63 (1956), pp. 361–382.
[18] I. Postlethwaite, J. M. Edmonds, and A. G. J. Macfarlane, *Principal gains and principal phases in the analysis of linear multivariable feedback systems*, IEEE Trans. Automat. Control, AC–26 (1981), pp. 32–46.
[19] A. Sideris and R. S. S. Pena, *Fast computation of the multivariable stability margin for real interrelated uncertain parameters*, IEEE Trans. Automat. Control, AC–34 (1989), pp. 1271–1276.
[20] E. D. Sontag, *Mathematical Control Theory*, Springer-Verlag, New York, 1990.
[21] R. K. Yedavalli, *Improved measures of stability robustness for linear state space models*, IEEE Trans. Automat. Control, AC–30 (1985), pp. 577–579.

SIAM J. CONTROL OPTIM.
Vol. 35, No. 4, pp. 1387–1421, July 1997

013

A GENERAL STOCHASTIC OUTER APPROXIMATIONS METHOD*

Y. V. VOLKOV† AND S. K. ZAVRIEV†

Abstract. Optimization problems involving an infinite number of constraints are considered. This paper presents a general stochastic outer approximations method which incorporates mechanisms for active search of relevant constraints and for dropping of irrelevant constraints. The method extracts the characteristic features of several stochastic outer approximations algorithms suggested by Wardi [*J. Optim. Theory Appl.*, 56 (1988), pp. 285–311; *J. Optim. Theory Appl.*, 64 (1990), pp. 615–640] and furthermore develops the approach to get advantages of the Eaves–Zangwill scheme. Similarly to Gonzaga and Polak [*SIAM J. Control Optim.*, 17 (1979), pp. 477–493] the method is based on the use of quasi-optimality functions satisfying some general unrestricted assumptions. These functions are usually employed in the stopping criteria of numerical techniques for solving simpler problems. It is shown that the method's trajectories almost surely converge to the quasi-optimal set. Following the proposed approach a stochastic algorithm for solving the approximation problem is constructed and studied.

The proposed general method can be considered as a developed Eaves–Zangwill method applying the multistart technique at each iteration for the search of relevant constraints' parameters.

Key words. outer approximations methods, stochastic programming, multistart method, semi-infinite programming problem, approximation problem, global optimization

AMS subject classifications. 90C34, 90C15, 90C26

PII. S0363012994263202

1. Introduction.

Optimization problems involving continua of constraints appear in different areas of applications (see, for example, the conference proceedings edited by Hettich [9] and Fiacco and Kortanek [7]). A typical problem with an infinite number of constraints is the semi-infinite programming problem

$$\text{sip:} \qquad f(x) \to \min_x$$
$$\text{such that (s.t.) } g(x,y) \le 0 \quad \forall y \in Y^0,$$
$$x \in X^0,$$

where $f(\cdot)$ and $g(\cdot,\cdot)$ are assumed to be continuously differentiable on a neighborhood of $X^0 \times Y^0$, $X^0 \subset \Re^k$ and $Y^0 \subset \Re^l$ are convex and compact and $\Re^k$ is the k-dimensional real space.

There are various numerical techniques for solving problems with continua of constraints (see, for instance, [9, 10, 8, 17, 4, 26, 27, 28]; among them outer approximations methods are of great importance.

The outer approximations methods are intended to solve a problem

$$\mathcal{P}^0: \qquad \text{find } x \in \mathcal{X}^0_{opt},$$

where $\mathcal{X}^0_{opt}$ has a very complicated description, e.g., for the semi-infinite programming

*Received by the editors February 16, 1994; accepted for publication (in revised form) May 28, 1996. The research of the first author was supported by International Science Foundation grant NBY000. The research of the second author was supported by International Science Foundation grant NBY000, grant NBY300 from the International Science Foundation and the Russian Government, and Russian Foundation for Fundamental Research grant 01448.

http://www.siam.org/journals/sicon/35-4/26320.html

†Operations Research Department, Faculty of Computational Mathematics and Cybernetics, Moscow State University, Moscow, Russia 119899 (natasergei@glas.apc.org).

problem $\mathcal{X}^0_{opt} = \{x \in \mathcal{X}^0 \mid f(x) = \min_{x' \in \mathcal{X}^0} f(x')\}$, where the feasible set $\mathcal{X}^0$ is as follows:

$$\mathcal{X}^0 = \{x \in X^0 \mid g(x,y) \le 0 \ \forall y \in Y^0\}.$$

The approach is to substitute for $\mathcal{P}^0$ a sequence $\mathcal{P}_n$ of approximating problems

$$\mathcal{P}_n : \qquad \text{find } x \in \mathcal{X}^n_{opt},$$

where $\mathcal{X}^n_{opt}$ has relatively simple descriptions, e.g., for the semi-infinite programming problem

$$\text{sip.}\mathcal{P}_n: \qquad \text{find } x \in \mathcal{X}^n_{opt},$$

$$\mathcal{X}^n_{opt} = \{x \in \mathcal{X}_n \mid f(x) = \min_{x' \in \mathcal{X}_n} f(x')\},$$

where the feasible set $\mathcal{X}_n$ is defined by a finite set of inequalities, i.e.,

$$\mathcal{X}_n = \{x \in X^0 \mid g(x,y) \le 0 \ \forall y \in Y_n\}, \ \ \mid Y_n \mid < +\infty$$

($\mid Y_n \mid$ denotes the cardinality of Y_n), $n = 1, 2, \ldots$. Consequently solving the problems $\mathcal{P}_n, n = 1, 2, \ldots$, we get a trajectory $\{x^n\}$ which is intended to converge to the optimal set of the original problem $\mathcal{P}^0$.

The pioneer works in the outer approximations methods [2, 14, 15], provided the monotonic growth of the descriptions of $\mathcal{X}^n_{opt}, \ n = 1, 2, \ldots$ (e.g., for sip $Y_1 \subset Y_2 \subset \cdots$). And unfortunately in all these algorithms the complexity of the description of $\mathcal{X}^n_{opt}$ (i.e., $\mid Y_n \mid$) grew rapidly with n, and quite quickly the problems $\mathcal{P}_n, \ n = 1, 2, \ldots,$ become almost as difficult as the original problem $\mathcal{P}^0$. To avoid this disadvantage Topkis [24] and Eaves and Zangwill [5] proposed special adaptive rules for forming $\mathcal{X}^n_{opt}, \ n = 1, 2, \ldots,$ involving constraints-dropping schemes which broke the monotonic growth of the descriptions of $\mathcal{X}^n_{opt}, \ n = 1, 2, \ldots$. This approach has been developed by Hogan [13] and Gonzaga and Polak [8]. Heunis [12] suggested employing Monte Carlo simulations for forming simpler problems; this idea later was refined by Wardi [26, 27] via constraints-dropping schemes for reducing the size of the constraint set.

In application to the semi-infinite programming problem the main points of adaptive rules for forming Y_{n+1} at the nth iteration of an outer approximations method are as follows.

Adding of Relevant Constraints.

Obtain $y^n_1, \ldots, y^n_{S_n} \in Y^0$ such that x^n does not approach the optimal set of the problem

$$\text{sip.}\overline{\mathcal{P}}_n : \qquad f(x) \to \min_x$$

$$s.t. \ \ g(x,y) \le 0 \ \ \forall y \in Y_n \ \cup \ \{y^n_1, \ldots, y^n_{S_n}\}.$$

Add $y^n_1, \ldots, y^n_{S_n}$ to Y_n to form $\overline{Y}_n$.

Dropping of Irrelevant Constraints.

Drop some points from $\overline{Y}_n$ to extract a subset $\Delta Y_n \subset \overline{Y}_n$ such that the constraints

$$g(x,y) \le 0 \ \forall y \in \Delta Y_n$$

are relevant at x^n with respect to the problem sip.$\overline{\mathcal{P}}_n$.

Then drop some sets from $\{\Delta Y_i,\ i = 1, \dots, n\}$ *to form*

$$Y_{n+1} = \bigcup_{j \in J_n} \Delta Y_j,\ J_n \subset \{1, \dots, n\}.$$

The following are approaches to the search of relevant constraints.

SCHEME AS (active search [5, 8]).

Compute y_1^n *as an approximate solution of the inner maximization problem*

$$\text{sip.}\mathcal{JP}_n : \qquad g(x^n, y) \to \max_{y \in Y}.$$

(Then $y_2^n, \dots, y_{S_n}^n$ *may be chosen arbitrarily.)*

SCHEME RS (passive random search [26, 27]).

Step 1. *Set* $i := 0$.

Step 2. *Set* $i := i + 1$.

Step 3. *Determine* y_i^n *by using the uniform probability distribution on* Y^0.

If an optimality condition of the problem

$$f(x) \to \min_x$$
$$s.t.\ \ g(x, y) \le 0 \ \ \forall y \in Y_n \ \cup \ \{y_1^n, \dots, y_i^n\}$$

is not sufficiently violated at x^n*, then go to Step* 2.

Else go to Step 4.

Step 4. *Set*

$$S_n := i,$$
$$\overline{Y}_n := Y_n \ \cup \ \{y_1^n, \dots, y_{S_n}^n\},$$

and exit.

Scheme AS is quite effective in the case when the inner problems are unimodal, but if they are not, there exist only effective descent algorithms to obtain a local maximum (or a stationary point) of $\text{sip.}\mathcal{JP}_n$. Thus, the execution of an outer approximations method using Scheme AS needs to apply a global optimization technique at each iteration and becomes too laborious. On the other hand it is evident that if the dimension of Y^0 is not very small, methods using Scheme RS cannot be effective because the parameters drawn by the uniform probability distribution on Y^0 are not essentially relevant. To achieve a balance between simplicity of the execution and the relevance of the outget constraints we extend Wardi's scheme as follows.

SCHEME RS.ACTIV (activated random search).

Step 1. *Set* $i := 0$.

Step 2. *Set* $i := i + 1$.

Step 3. *Determine* y_i^n *by using the uniform probability distribution on* Y^0.

Apply a local descent method starting with y_i^n *to obtain a local maximum* $y_i^{n,*}$ *of* $\text{sip.}\mathcal{JP}_n$.

If an optimality condition of the problem

$$f(x) \to \min_x$$
$$s.t.\ \ g(x, y) \le 0 \ \ \forall y \in Y_n \ \cup \ \{y_1^n, y_1^{n,*}, \dots, y_i^n, y_i^{n,*}\}$$

is not sufficiently violated at x^n*, then go to Step* 2.

Else go to Step 4.

Step 4. *Set*

$$S_n := i,$$

$$\overline{Y}_n := Y_n \ \cup \ \{y_1^n, y_1^{n,*}, \ldots, y_{S_n}^n, y_{S_n}^{n,*}\}$$

and exit.

In the present paper we consider an outer approximations method using Scheme RS.ACTIV for the search of relevant parameters to form approximative problems $\mathcal{P}_n,\ n = 1, 2, \ldots$. The constructed method possesses the guaranteed convergence properties of Wardi's algorithms (i.e., of the algorithms based on Scheme RS), and the method's computational efforts needed to obtain $x^n,\ n = 1, 2, \ldots$, are relatively inexpensive. At the same time the practical convergence rate of our method often appears to be similar to the rates of the methods based on Scheme AS which compute trajectories $\{x^n\}$ in a much more laborious way.

In section 2 the master method for solving the general problem $\mathcal{P}^0$ is constructed. The method employs the activated random search scheme RS.ACTIV for forming approximative problems $\mathcal{P}_n,\ n = 1, 2, \ldots$, and it analyzes these problems with the use of a quasi-optimality function which is supposed to satisfy some general unrestricted assumptions. Examples of appropriate quasi-optimality functions for the semi-infinite programming problem and the problem of solving a system with continuum of inequalities are considered in section 3.

We call the proposed master method (using Scheme RS.ACTIV) *the activated method*, in contrast to the similar method using the nonactivated Scheme RS, which is called *the standard method*. (Note that the standard method directly generalizes Wardi's algorithms.) The efficiency of the activation is explored in section 4, where we consider realizations of these methods for solving the global optimization problem. It appears that in this case the standard method becomes the pure random search global optimization algorithm and the activated method is an algorithm of the well-known multistart method. The advantages of the multistart method over the pure random search are evident.

In sections 5 and 6 the convergence theorem is proven. It is shown that trajectories of the master method almost surely converge to the quasi-optimal set of the considered general problem $\mathcal{P}^0$.

In section 7 a version of the master method for solving the approximation problem is constructed and studied. Numerical examples are presented in section 8.

For the reader's convenience we provide a list of notation.

Let

$$\mathbf{N} := \{1, 2, \ldots\},$$

$$\Re^1_+ = \{x \in \Re^1 \mid x \geq 0\},$$

$$B_\delta(x) = \{x' \in \Re^k \mid\mid\mid x' - x \mid\mid\mid < \delta\},\ \delta > 0.$$

We denote by $\mathfrak{M}(Y^0)$ the set of all subsets of $Y^0 \subset \Re^k$; $\mathfrak{M}_c(Y^0)$ and $\mathfrak{M}_f(Y^0)$ denote the set of all compact subsets and the set of all finite subsets, respectively. It is obvious that

$$\mathfrak{M}_f(Y^0) \subset \mathfrak{M}_c(Y^0) \subset \mathfrak{M}(Y^0).$$

For any $Y, Y' \in \mathfrak{M}_c(Y^0)$ define

$$\rho(Y, Y') := \max_{y \in Y} \min_{y \in Y'} \| y - y' \|,$$

$$h(Y, Y') := \max(\rho(Y, Y'),\ \rho(Y', Y)).$$

$h(Y, Y')$ is said to be the *Hausdorff distance between* Y *and* Y'.

Let $\{x_n\}$ be a bounded sequence, $x_n \in \Re^k,\ n = 1, 2, \ldots$, the set of all limit points of $\{x_n\}$ is denoted by $\overline{lt}\{x_n\}$. We say that $\{x_n\}$ converges to $C \subset \Re^k$ if

$$\overline{lt}\{x_n\} \subset C.$$

For a set $C \subset \Re^k$ we denote by $K_C(x)$ the tangent cone at a point $x \in C$; let *conv* C and C^* denote the convex hull and the polar of C, respectively.

Consider the constrained optimization problem

$$\text{opt:} \qquad f(x) \to \min$$

$$\text{s.t.} \quad g_i(x) \le 0,\ i = 1, \ldots, m,$$

$$x \in X^0,$$

where $f(\cdot)$, $g_i(\cdot)$, $i = 1, \ldots, m$, are assumed to be locally Lipschitz continuous on $\Re^k$, the optimal set X_{opt} is assumed to be nonempty, and $X^0 \in \Re^k$ is closed and convex.

Let $\partial f(x)$ denote the set of all generalized gradients (in the sense of Clarke [3]) of $f(\cdot)$ at $x \in \Re^k$.

It is known [3] that for every $x \in X_{opt}$ satisfying the constraints qualification

$$conv \bigcup_{i:\ g_i(x)=0} \partial g_i(x) + K^*_{X^0}(x) \ni \mathbf{0}$$

(when $g_i(x) \ne 0$, $i = 1, \ldots, m$, we formally suppose that this constraints qualification holds) the following optimality conditions hold:

$$\text{there exist } \lambda_1, \ldots, \lambda_s \ge 0,\ s \le k+1,\ \text{s.t.}$$

$$\partial f(x) + \sum_{i=1}^{s} \lambda_i\ \partial g_i(x) + K^*_{X^0} \ni \mathbf{0}, \tag{1.1}$$

$$\lambda_i\ g_i(x) = 0,\ i = 1, \ldots, s.$$

This fact provides a number of corollaries on the first-order necessary optimality conditions for various optimization problems, e.g., for the semi-infinite programming problem, the minimax problem, the problem of solving a system of inequalities, etc. (See [18, 6, 3].) In the course of the present paper we shall use some of these results without specially referring to the sources.

For any $f(\cdot) :\ X \to \Re^1$ we denote

$$\operatorname{Arg\,min}_{x \in X} f(x) := \{x \in X \mid f(x) = \min_{x' \in X} f(x')\};$$

$\arg\min_{x \in X} f(x)$ denotes an arbitrary element of the set $\operatorname{Arg\,min}_{x \in X} f(x)$. Similarly we define $\operatorname{Arg\,max}_{x \in X} f(x)$ and $\arg\max_{x \in X} f(x)$.

For a convex and closed $X \subset \Re^k$ the orthogonal projector on X, $pr_X(\cdot) : \Re^k \to X$, is given by

$$pr_X(x) := \arg\min_{x' \in X} \| x - x' \| .$$

Note that for the problem

$$f(x) \to \min,\ x \in X^0,$$

where $f(\cdot)$ is assumed to be differentiable on a neighborhood of $x \in X_{opt}$, the optimality condition (1.1) can be rewritten in the following form:

$$pr_{X^0}(x - \nabla f(x)) - x = 0.$$

2. The master method. Let us consider our problem in the most general form:

$$\mathcal{P}^0 : \qquad \text{find}\ \ x \in \mathcal{X}_{opt},$$

where $\mathcal{X}_{opt} \subset X^0,\ X^0 \subset \Re^k$ is compact, and the description of $\mathcal{X}_{opt}$ involves an infinite number of parameters $y \in Y^0,\ Y^0 \subset \Re^l$ is compact. To show the dependence of $\mathcal{P}^0$ upon the parameter set Y^0, we shall also denote $\mathcal{P}^0$ by $\mathcal{P}[Y^0]$.

To construct an outer approximations method we consider simpler problems of the same type as $\mathcal{P}^0$. Let for any $Y \in \mathfrak{M}_c(Y^0)$ the problem $\mathcal{P}[Y]$ be defined similarly to $\mathcal{P}^0$ but subject to the smaller parameter set Y; i.e., the description of $\mathcal{P}[Y]$ involves only parameters $y \in Y$. The presented general stochastic outer approximation method provides the sequential solving of problems $\mathcal{P}[Y_n],\ n = 1, 2, \ldots$, where each problem $\mathcal{P}[Y_n]$ depends upon only a finite number $\mid Y_n \mid$ of parameters, $Y_n \in \mathfrak{M}_f(Y^0),\ n = 1, 2, \ldots$.

Let us introduce a quasi-optimality function, i.e., a scalar nonnegative criterion $\Theta(\cdot, \cdot) : X^0 \times \mathfrak{M}_c(Y^0) \to \Re^1_+$ s.t. for any compact $Y \subset Y^0$ and $x \in X^0$ the value $\Theta(x, Y)$ estimates the quality of x as an approximate local solution of the problem $\mathcal{P}[Y]$. In particular, we suppose that

$$x \in \mathcal{X}_{opt} \ \Rightarrow\ \Theta(x, Y^0) = 0. \tag{2.1}$$

Usually the inequality

$$\Theta(x, Y) \le \varepsilon \tag{2.2}$$

can be employed as a stopping criterion for an effective local descent technique for solving $\mathcal{P}[Y]$ and, thus, when $\mid Y \mid$ is not large, we can effectively solve (2.2) even for small $\varepsilon > 0$.

We define the $\Theta(\cdot, \cdot)$–quasi-optimal set of the problem $\mathcal{P}[Y]$ as follows:

$$\mathcal{X}_{qopt}[Y] := \{x \in X^0 \mid \Theta(x, Y) = 0\},\ Y \in \mathfrak{M}_c(Y^0).$$

Note that by (2.1)

$$\mathcal{X}_{opt} \subset \mathcal{X}^0_{qopt},$$

where $\mathcal{X}^0_{qopt} := \mathcal{X}_{qopt}[Y^0] = \{x \in X^0 \mid \Theta(x, Y^0) = 0\}$.

Let the following assumptions hold.

Assumption A1. Let

$$\mathcal{X}_{qopt}[Y] \neq \emptyset \ \forall Y \in \mathfrak{M}_c(Y^0). \qquad \square$$

Assumption A2. For every $x \in X^0$ and $Y \in \mathfrak{M}_c(Y^0)$ let the following properties hold:

(i) If $\Theta(x, Y) > 0$, then there exist $\theta, \delta > 0$ such that

$$\Theta(x', Y') \geq \theta > 0$$

for every $x' \in B_\delta(x) \bigcap X^0$ and $Y' \in \mathfrak{M}_c(Y^0),\ h(Y, Y') < \delta$.

(ii) If $\Theta(x, Y) = 0$, then for every $\varepsilon > 0$ there exists $\delta > 0$ such that

$$\Theta(x', Y') < \varepsilon$$

for every $x' \in B_\delta(x) \bigcap X^0$ and $Y' \in \mathfrak{M}_c(Y^0),\ h(Y, Y') < \delta$. $\square$

For any $x \in X^0$ and $Y \in \mathfrak{M}_c(Y^0)$ a finite set $A,\ A \in Y^0$ is called an *active parameters set at* (x, Y) *with respect to the quasi-optimality function* $\Theta(\cdot, \cdot)$ if

$$\Theta(x, A) = \Theta(x, Y') = \Theta(x, Y) \ \ \forall Y' \in \mathfrak{M}_c(Y) :\ A \subset Y' \subset Y$$

and, moreover, there are no $A' \subset A,\ A' \neq A$, satisfying the same property. Let $\mathcal{A}(x, Y)$ denote the set of all active parameters sets at (x, Y) with respect to $\Theta(\cdot, \cdot)$. For $\Delta Y \subset Y$ we denote

$$\Delta Y \succeq \mathcal{A}(x, Y)$$

if there exists $A \in \mathcal{A}(x, Y)$ such that $\Delta Y \supset A$.

Now we construct a general stochastic iterative method converging almost surely (a.s.) to the set $\mathcal{X}_{qopt}[Y^0]$ (note that $\mathcal{X}_{opt} \subset \mathcal{X}_{qopt}[Y^0]$). The method suggests solving (2.2) with $Y = Y_n$ and $\varepsilon = \varepsilon_n$ for $n = 1, 2, \ldots$, where $\{Y_n\}$ is a sequence of finite subsets of Y^0 and $\varepsilon_n > 0,\ n = 1, 2, \ldots,\ \varepsilon_n \searrow 0$, are the corresponding precision levels; it uses special mechanisms for limiting the growth of $\mid Y_n \mid$: an *activated random search mechanism* for choosing relevant parameters to include in Y_n and *a dropping mechanism* to exclude irrelevant parameters from Y_n. The method refines the Wardi's stochastic algorithms (see [26, 27]) and, furthermore, develops this approach getting features of the Eaves–Zangwill method.

The main points of the method are focused in the following iterative stochastic procedure.

Procedure SPROC.ACTIV.

Input. $x \in X,\ Y \in \mathfrak{M}_f(Y^0)$.

Output. $\theta \in \Re^1_+,\ \Delta Y \in \mathfrak{M}_f(Y^0),\ \overline{Y} \in \mathfrak{M}_f(Y^0),$ *and* $S \in \mathbf{N}$.

Parameter. $\gamma > 0$.

Step 0 (*initial step*). *Set*

$$\overline{Y}_1 := Y,$$

$$i := 1.$$

Step 1 (*passive search of a relevant parameter*).

Determine a point $y_i \in Y$ *by using the uniform probability distribution on* Y.

Include y_i *in* $\overline{Y}_i$.

Step 1.ACTIV (*activated search of a relevant parameter*).
Searching on Y *obtain a point* $y_i^* = y_i^*(x, \overline{Y}_i, y_i)$ (*e.g., using a local search algorithm starting with* y_i *at an inner problem* $\mathcal{JP}(x, Y)$).
Include y_i^* *in* $\overline{Y}_i$.
Step 2. *Set* $\theta_i = \Theta(x, \overline{Y}_i)$.
Step 3 (*control step*). *If*

$$i\ \theta_i \le \gamma, \tag{2.3}$$

then set $i := i + 1$ *and go to Step* 1.
Step 4. *Set*

$$\overline{Y} := \overline{Y}_i,$$

$$\theta := \theta_i,$$

$$S := i.$$

Step 5 (*dropping*). *Analyzing* $\Theta(x, \overline{Y})$ *obtain*

$$\Delta Y \succeq \mathcal{A}(x, \overline{Y}).$$

Then exit.

The procedure without Step 1.ACTIV (with $y_i^* = y_i,\ i = 1, \dots, S$) is similar to the original Wardi's procedures and will be denoted SPROC. We call SPROC and SPROC.ACTIV *the standard procedure* and *the activated procedure*, respectively.

Now we present the master method.
Method SMETH.ACTIV.
Data. $x^1 \in X$.
Parameters. *Sequences* $\{\varepsilon_n\},\ \{\sigma_n\}$,

$$\varepsilon_n,\ \sigma_n > 0,\ n = 1, 2, \dots,\ \ \varepsilon_n,\ \sigma_n \searrow 0. \tag{2.4}$$

Stage 0 (*initial step*). *Set*

$$n := 1,$$

$$Y_1 := \emptyset.$$

Stage 1. *Enter* SPROC.ACTIV *with input* x^n *and* Y_n. *Denote the outputs by* Θ_n, ΔY_n, $\overline{Y}_n$, S_n.
Stage 2. *Set*

$$Y_{n+1} := \Delta Y_n \cup \bigcup_{\substack{j:\theta_j > \sigma_j, \\ 1 \le j \le n-1}} \Delta Y_j.$$

Stage 3. *Find* $x^{n+1} \in X$ *satisfying*

$$\Theta(x^{n+1}, Y_{n+1}) < \varepsilon_{n+1}.$$

Stage 4. *Set*

$$n =: n + 1.$$

Go to Step 1.

The method using SPROC instead of SPROC.ACTIV will be denoted SMETH. The methods SMETH and SMETH.ACTIV are said to be *the standard method* and *the activated method*, respectively.

The constructed general stochastic outer approximations method possesses the following property.

PROPOSITION. *Every trajectory* $\{x^n\}$ *of SMETH.ACTIV a.s. converges to* $\mathcal{X}_{qopt}[Y^0]$.

The formal statement and the proof of the proposition will be given below (see Theorem 6.1).

3. Examples of quasi-optimality functions. Now we present some examples of appropriate quasi-optimality functions for several classes of optimization problems with continua of inequalities, and we consider the corresponding mechanisms for active search of relevant constraints.

Example 3.1. Consider the problem of solving a system with continuum of inequalities:

$$\text{ineq.}\mathcal{P}^0: \qquad \text{to find } x \in \mathcal{X}^0_{opt},$$

$$\mathcal{X}^0_{opt} = \{x \in X^0 \mid g(x,y) \le 0 \ \forall y \in Y^0\},$$

where $g(\cdot,\cdot)$ is assumed to be continuously differentiable on a neighborhood of $X^0 \times Y^0$ and $X^0 \subset \Re^k$, $Y^0 \subset \Re^l$ are convex and compact.

We define the simpler problems ineq.$\mathcal{P}[Y]$, $Y \in \mathfrak{M}_c(Y^0)$, as follows:

$$\text{ineq.}\mathcal{P}[Y]: \qquad \text{to find} \quad x \in \mathcal{X}_{opt}[Y],$$

$$\mathcal{X}_{opt}[Y] = \{x \in X^0 \mid g(x,y) \le 0 \ \forall y \in Y\},$$

it is clear that $\mathcal{X}_{opt}[Y^0] = \mathcal{X}^0_{opt}$.

Let us define a quasi-optimality function $\Theta(\cdot,\cdot): \ X^0 \times \mathfrak{M}_c(Y^0) \ \to \ \Re^1_+$ by

$$\Theta(x,Y) := \max(0; \ \max_{y' \in Y} g(x,y')).$$

Obviously, Assumptions A1 and A2 hold and

$$\mathcal{X}^0_{opt} = \mathcal{X}_{qopt}[Y^0].$$

Note that for any $A \subset Y$ satisfying

$$A \subset \begin{cases} \operatorname{Arg}\max_{y' \in Y} \ g(x,y') & \text{if } \Theta(x,Y) > 0, \\ Y & \text{if } \Theta(x,Y) = 0, \end{cases}$$

the following equalities hold:

$$\Theta(x,A) = \Theta(x,Y') = \Theta(x,Y) \ \ \forall Y' : A \subset Y' \subset Y.$$

Therefore, for the considered quasi-optimality function $\Theta(\cdot,\cdot)$ the active parameters sets are singletons and

$$\mathcal{A}(x,Y) := \begin{cases} \operatorname{Arg} \ \max_{y' \in Y} \ g(x,y') & \text{if } \Theta(x,Y) > 0, \\ Y & \text{if } \Theta(x,Y) = 0. \end{cases}$$

In particular, if $\Theta(x, Y) > 0$ then for any $\Delta Y \subset Y$ we have

$$\Delta Y \succeq \mathcal{A}(x, Y) \;\Leftrightarrow \Delta Y \cap \mathrm{Arg} \max_{y' \in Y} g(x, y') \neq \emptyset.$$

Thus, the criterion $\Theta(\cdot, \cdot)$ can be employed in the stochastic outer approximations method SMETH.ACTIV for solving ineq.$\mathcal{P}^0$; we shall denote by SMETH.ACTIV.ineq and SPROC.ACTIV.ineq, respectively, the corresponding versions of the master method SMETH.ACTIV and of the procedure SPROC.ACTIV.

Let us introduce some notions to construct a mechanism of the activated random search of a relevant parameter y_i^* within the procedure SPROC.ACTIV for the problem ineq.$\mathcal{P}^0$.

Consider the inner problem

ineq.$\mathcal{JP}(x)$: $$g(x, y) \to \max$$

$$\text{s.t.} \;\; y \in Y^0,$$

and define the stationary set (the set of all points satisfying the first-order necessary optimality conditions) of ineq.$\mathcal{JP}(x)$ by

$$Y_{stat}(x) := \{y \in Y^0 \mid \nabla_y g(x, y) - K^*_{Y^0}(y) \ni \mathbf{0}\}.$$

We also define the ε-stationary set $Y^{\varepsilon}_{stat}(x)$ of the problem ineq.$\mathcal{JP}(x)$ by

$$Y^{\varepsilon}_{stat}(x) := \{y \in Y^0 \mid \;\; \| \, pr_{Y^0}(y + \nabla_y g(x, y)) - y \, \| \leq \varepsilon\}, \;\; \varepsilon \geq 0;$$

it is clear that $Y^0_{stat}(x) = Y_{stat}(x)$.

We suggest the following mechanism for activation of a parameter y_i generated by a random experiment within the procedure SPROC.ACTIV.ineq (at Step 1).

Step 1.ACTIV.ineq.

Apply a local descent technique for solving ineq.$\mathcal{JP}(x)$ *(for instance, the gradient projection method) starting with* y_i *to obtain* $y_i^* \in Y$ *such that*

$$y_i^* \in Y^{\varepsilon}_{stat}(x), \;\; g(x, y_i) \leq g(x, y_i^*).$$

Here $\varepsilon > 0$ *is a parameter.*

Example 3.2. Consider the semi-infinite programming problem sip, which we present in the form $\mathcal{P}^0$:

sip.$\mathcal{P}^0$: $$find \; x \in \mathcal{X}^0_{opt},$$

$$\mathcal{X}^0_{opt} = \{x \in \mathcal{X}^0 \mid f(x) = \min_{x' \in \mathcal{X}^0} f(x')\},$$

$$\mathcal{X}^0 = \{x \in X^0 \mid g(x, y) \leq 0 \;\; \forall y \in Y^0\},$$

where $f(\cdot)$, $g(\cdot, \cdot)$ are assumed to be continuously differentiable on a neighborhood of $X^0 \times Y^0$; $\nabla f(\cdot)$, $\nabla g(\cdot) \in \mathbf{C}_{Lip}(X^0 \times Y^0, L)$, $L > 0$, $X^0 \subset \Re^k$, $Y^0 \subset \Re^l$ are convex and compact. Let us also suppose that the Slater constraints qualification holds; i.e., $g(\cdot, y)$ is convex on X for every $y \in Y$ and there exists $x^* \in X$ such that

$$g(x^*, y) < 0 \;\; \forall y \in Y^0.$$

Define

$$\mathcal{X}^0_{stat} = \Big\{ x \in X^0 \mid \text{ there exist } \lambda_1, \dots, \lambda_s \geq 0$$

$$\text{and } y_1, \dots, y_s \in Y^0,\ s \leq k+1, \text{s.t.}$$

$$\nabla f(x) + \sum_{i=1}^{s} \lambda_i\ \nabla_x g(x, y_i) + K^*_{X^0}(x) \ni \mathbf{0};$$

$$\lambda_i\ g(x, y_i) = 0,\ i = 1, \dots, s \Big\},$$

$\mathcal{X}^0_{stat}$ is the stationary set (the set of all points satisfying the first-order necessary optimality conditions) of the problem sip.$\mathcal{P}^0$.

To construct an outer approximations technique for solving sip.$\mathcal{P}^0$ consider the following simpler problems:

sip.$\mathcal{P}[Y]$: find $x \in \mathcal{X}_{opt}[Y]$,

$$\mathcal{X}_{opt}[Y] = \{x \in \mathcal{X}[Y] \mid f(x) = \min_{x' \in \mathcal{X}[Y]} f(x')\},$$

$$\mathcal{X}[Y] = \{x \in X^0 \mid g(x, y) \leq 0\ \ \forall y \in Y\},$$

where $Y \in \mathfrak{M}_c(Y^0)$. It is obvious that sip.$\mathcal{P}[Y^0]$ is the original problem sip.$\mathcal{P}^0$.

The following quadratic subproblem proves to be useful in both theoretical and numerical analysis of constrained optimization problems (see [1, 19]):

sip.$\mathcal{QP}(x, Y)$: $\langle \nabla f(x), p \rangle + \frac{1}{2} \parallel p \parallel^2 \to \min$

$$\text{s.t. } g(x, y) + \langle \nabla_x g(x, y), p \rangle \leq 0\ \ \forall y \in Y,$$

$$x + p \in X,$$

where $Y \in \mathfrak{M}_c(Y^0)$, $x \in X$. It is easy to see that for every $x \in X$, $Y \in \mathfrak{M}_c(Y^0)$ the feasible set of sip.$\mathcal{QP}(x, Y)$ is not empty. Actually, for $p^* = x^* - x$ we have

$$g(x, y) + \langle \nabla_x g(x, y), p^* \rangle = g(x, y) + \langle \nabla_x g(x, y), x^* - x \rangle$$

$$\leq g(x, y) + g(x^*, y) - g(x, y) = g(x^*, y) < 0\ \forall y \in Y^0,$$

$$x + p^* = x^* \in X.$$

Therefore, for every $x \in X^0$, $Y \in \mathfrak{M}_c(Y^0)$ there exists the unique solution $p_0(x, Y)$ of the problem sip.$\mathcal{QP}(x, Y)$. Moreover, the following optimality criterion holds.

LEMMA 3.1. *For every $x \in X^0$ and $Y \in \mathfrak{M}_c(Y^0)$ a vector $p \in \Re^k$ is the solution of* sip.$\mathcal{QP}(x, Y)$ *if and only if there exist $\lambda_1, \dots, \lambda_s \geq 0$ and $y_1, \dots, y_s \in Y$, $s \leq k+1$ such that the following properties hold:*

$$p + \nabla f(x) + \sum_{i=1}^{s} \lambda_i\ \nabla_x g(x, y_i) + K^*_{X^0}(x + p) \ni \mathbf{0}, \tag{3.1}$$

$$\lambda_i\ (g(x, y_i) + \langle \nabla_x g(x, y), p \rangle) = 0,\ i = 1, \dots, s, \tag{3.2}$$

$$g(x,y) + \langle \nabla_x g(x,y), p \rangle \le 0 \ \forall y \in Y, \tag{3.3}$$

$$x + p \in X^0. \tag{3.4}$$

From Lemma 3.1 we get immediately the following important corollaries. Let $\mathcal{X}_{stat}[Y]$ denote the stationary set of sip.$\mathcal{P}[Y]$, $Y \in \mathfrak{M}_c(Y^0)$.

COROLLARY 3.1. *For every $x \in X^0$ and $Y \in \mathfrak{M}_c(Y^0)$ the vector $p_0(x,Y) = 0$ if and only if $x \in \mathcal{X}_{stat}[Y]$.*

COROLLARY 3.2. *For every $x \in X^0$, $Y \in \mathfrak{M}_c(Y^0)$, and sequences $\{x_n\}$, $\{Y_n\}$ satisfying*

$$x_n \in X^0, \ n = 1,2,\ldots, \ \lim_{n\to\infty} x_n = x,$$

$$Y_n \in \mathfrak{M}_c(Y^0), \ n = 1,2,\ldots, \ \lim_{n\to\infty} h(Y_n, Y) = 0,$$

the following property holds:

$$\lim_{n\to\infty} p_0(x_n, Y_n) = p_0(x,Y).$$

Let us define a quasi-optimality function $\Theta(\cdot,\cdot) : \ X^0 \times \mathfrak{M}_c(Y^0) \to \Re^1_+$ as follows:

$$\Theta(x,Y) = \| \ p_0(x,Y) \ \|, \ x \in X, \ Y \in \mathfrak{M}_c(Y^0).$$

Applying Corollaries 3.1 and 3.2, we obtain that Assumptions A1 and A2 hold and

$$\mathcal{X}_{qopt}[Y] = \mathcal{X}_{stat}[Y] \ \ \forall Y \in \mathfrak{M}_c(Y^0),$$

$$\mathcal{X}^0_{opt} \subset \mathcal{X}^0_{stat} = \mathcal{X}_{qopt}[Y^0].$$

For any $x \in X^0$ and $Y \subset Y^0$ denote

$$\mathfrak{Y}(x,Y) := \big\{ \{y_1,\ldots,y_s\} \in \mathfrak{M}_f(Y) \mid \ s \le k+1 \text{ and there exist } \lambda_1,\ldots,\lambda_s \ge 0$$

$$\text{s.t. (3.1), (3.2), (3.3), and (3.4) hold for } p = p_0(x,Y) \big\}.$$

It is clear that for the considered quasi-optimality function $\Theta(\cdot,\cdot)$ the set of all active parameters sets is as follows:

$$\mathcal{A}(x,Y) = \{ Y' \in \mathfrak{Y}(x,Y) \mid \text{ there does not exist } Y'' \in \mathfrak{Y}(x,Y)$$

$$\text{s.t. } Y'' \subset Y', \ Y'' \ne Y' \}.$$

Obviously,

$$\Delta Y \in \mathfrak{Y}(x,Y) \Rightarrow \ \Delta Y \succeq \mathcal{A}(x,Y).$$

We see that $\Theta(\cdot,\cdot)$ can be employed in the stochastic outer approximations method SMETH.ACTIV for solving sip.$\mathcal{P}^0$. The mechanism for active search of relevant constraints within the corresponding procedure SPROC.ACTIV.sip (i.e., Step 1.ACTIV.sip) can be similar to Step 1.ACTIV.ineq.

Note that we may also employ various quasi-optimality functions from [8, 26, 27].

4. Comparative efficiency of the activated and the standard versions of the method. Due to the active search modification at Step 1.ACTIV in the procedure PROC, the efficiency of the standard method can be essentially improved. To clarify the advantages of the activation we consider the following important example.

Consider the global optimization problem

$$\text{glob:} \qquad F(x) \to \min\limits_{x \in X^0},$$

where the objective function $F(\cdot)$ is assumed to be continuously differentiable on $X^0 \subset \Re^k$, $\nabla F(\cdot) \in \mathbf{C}_{Lip}(X^0, L)$, X^0 is convex and compact. Let us denote

$$X^0_{opt} := \{x \in X^0 \mid F(x) = \min\limits_{x' \in X^0} F(x')\},$$

$$X^{\varepsilon}_{stat} := \{x \in X^0 \mid\mid pr_{X^0}(x - \nabla F(x)) - x \mid\mid \leq \varepsilon\}, \varepsilon \geq 0,$$

X^0_{opt} and X^{ε}_{stat} are the optimal set and the ε-stationary set of the considered problem, respectively. Obviously, the problem can be rewritten in the form of a problem with an infinite number of constraints as follows:

$$\text{glob.}\mathcal{P}^0: \qquad \text{find } x \in \mathcal{X}_{opt},$$

$$\mathcal{X}_{opt} := \{x \in X^0 \mid F(x) - F(y) \leq 0 \ \forall y \in X^0\}.$$

Following the proposed general scheme we construct a stochastic outer approximations method for solving the problem glob.$\mathcal{P}^0$. First define approximative problems glob.$\mathcal{P}[X]$, $X \subset \mathfrak{M}_c(X^0)$, by

$$\text{glob.}\mathcal{P}[X]: \qquad \text{find } x \in \mathcal{X}_{opt}[X],$$

$$\mathcal{X}_{opt}[X] := \{x \in X^0 \mid F(x) \leq F(y) \ \ \forall y \in X\};$$

clearly,

$$\mathcal{X}_{opt}[X] := \{x \in X^0 \mid F(x) \leq \min\limits_{y \in X} F(y)\} \ \ \forall X \in \mathfrak{M}_c(X^0).$$

We also define the quasi-optimality function $\Theta(\cdot, X) : X^0 \to \Re^1_+$ by

$$\Theta(x, X) := \max(0; F(x) - \min\limits_{y \in X} F(y)), \ x \in X^0.$$

It is easily seen that

$$\mathcal{X}_{qopt}[X] := \{x \in X^0 \mid \Theta(x, X) = 0\} = \mathcal{X}_{opt}[X] \ \ \forall X \in \mathfrak{M}_c(X^0),$$

$$\mathcal{X}^0_{qopt} := \mathcal{X}_{qopt}[X^0] = \mathcal{X}^0_{opt},$$

and, moreover, Assumptions A1 and A2 hold.

Note that every active parameters set at (x, X) with respect to the considered quasi-optimality function $\Theta(\cdot, \cdot)$ is a singleton and

$$\mathcal{A}(x, X) = \{x \in X^0 \mid F(x) \leq \min\limits_{y \in X} F(y)\}. \tag{4.1}$$

After some straightforward simplifications based on the specificity of the considered problems glob.$\mathcal{P}[X]$ and the quasi-optimality function $\Theta(\cdot, X)$, $X \in \mathfrak{M}_f(X^0)$, the standard procedure SPROC (for simplicity we set $\gamma = 0$) adapted for glob.$\mathcal{P}[X]$ takes the following form:

PROCEDURE SPROC.glob.

Input. $x \in X^0$, $X \in \mathfrak{M}_f(X^0)$ $(x = \arg\min_{y \in X} F(y))$.

Output. $\Delta X = \{\overline{x}\}$, $\overline{X} = X \cup \{\overline{x}\}$.

Step 0. *Set* $i := 1$.

Step 1. *Determine a point* $x_i \in X^0$ *by using the uniform probability distribution on* X^0.

Step 2-3-4-5. *If*

$$F(x_i) \geq \min_{y \in X} F(y),$$

then $i := i + 1$ *and go to Step* 1. *Else set*

$$\overline{x} := x_i;\ \overline{X} := X \cup \{\overline{x}\};$$
$$\Delta\overline{X} := \{\overline{x}\}$$

(*note that by* (4.1) $\Delta\overline{X} \in \mathcal{A}(x, X)$).

Then exit.

Activating the standard procedure SPROC.glob by the use of a local descent technique for solving the problem

glob.$\mathcal{JP}$: $$F(y) \to \min_{y \in X^0},$$

we arrive at the procedure SPROC.ACTIV in the following form:

PROCEDURE SPROC.ACTIV.glob.

Input. $x \in X^0$, $X \in \mathfrak{M}_f(X^0)$ $(x = \arg\min_{y \in X} F(y))$.

Output. $\Delta X = \{\overline{x}\}$, $\overline{X} = X \cup \{\overline{x}\}$.

Step 0. *Set* $i := 1$.

Step 1. *Determine a point* $x_i \in X^0$ *by using the uniform probability distribution on* X^0.

Step 1.ACTIV. *Using a local descent technique (for instance, the gradient projection method) for solving the problem glob.*$\mathcal{JP}$, *starting with* x_i, *obtain* x_i^*,

$$x_i^* \in X^{\varepsilon}_{stat},\ F(x_i^*) \leq F(x_i).$$

Step 2-3-4-5. *If*

$$F(x_i^*) \geq \min_{y \in X} F(y),$$

then $i := i + 1$ *and go to Step* 1. *Else set*

$$\overline{x} := x_i^*,\ \overline{X} := X \cup \{\overline{x}\},$$
$$\Delta\overline{X} := \{\overline{x}\}.$$

Then exit.

Here $\varepsilon > 0$ is a parameter.

Taking into account that in the considered case the run of the master method does not depend upon $\{\varepsilon_n\}$ and $\{\sigma_n\}$ and making some further simplifications we obtain the following versions of the standard and the activated stochastic outer approximations methods.

METHOD SMETH.glob.

Data. $x^1 \in X^0$.

Stage 0. *Set* $n := 1$.

Stage 1.

Step 0. *Set* $i := 1$.

Step 1. *Determine a point* $x_i^n \in X^0$ *by using the uniform probability distribution on* X^0.

Step 2. *If*

$$F(x_i^n) \geq F(x^n),$$

then $i := i + 1$ *and go to Step* 1. *Else set*

$$\overline{x}^n := x_i^n.$$

Stage 3. Set

$$x^{n+1} := \overline{x}^n,$$
$$n := n + 1.$$

Go to Stage 1.

The activated stochastic outer approximations method for solving the global optimization problem is as follows:

METHOD SMETH.ACTIV.glob.

Data. $x^1 \in X^0$.

Stage 0. *Set* $n := 1$.

Stage 1.

Step 0. *Set* $i := 1$.

Step 1. *Determine a point* $x_i^n \in X^0$ *by using the uniform probability distribution on* X^0.

Step 1.ACTIV. *Using a local descent technique, starting with* x_i^n, *obtain* $x_i^{n,*}$,

$$x_i^{n,*} \in X_{stat}^{\varepsilon},\ F(x_i^{n,*}) \leq F(x_i^n).$$

Step 2. *If*

$$F(x_i^{n,*}) \geq F(x^n),$$

then $i := i + 1$ *and go to Step* 1. *Else set*

$$\overline{x}^n := x_i^{n,*}.$$

Stage 3. *Set*

$$x^{n+1} := \overline{x}^n,$$
$$n := n + 1.$$

Go to Stage 1.

Now it is obvious that the standard method SMETH.glob is the classical pure random search technique for global optimization and the activated method SMETH. ACTIV.glob is a version of the well-known multistart method (see [22]). The practical advantages of the multistart method over the pure random search are evident.

5. The basic property of the method. Let us introduce some necessary notation.

Let $(Y^0, \mathfrak{B}, \mu)$ be the probability space, where $\mathfrak{B}$ is the σ-algebra of Borel subsets of Y^0 and μ is the Borel measure on Y, normalized in such a way that $\mu(Y) = 1$. Set

$$\mathfrak{B}_n = \overset{n}{\underset{i=1}{\otimes}} \mathfrak{B},\ P_n = \overset{n}{\underset{i=1}{\otimes}} \mu,\ n = 1, 2, \ldots,$$

$$\Omega = \overset{\infty}{\underset{i=1}{\otimes}} Y^0,\ \mathfrak{A} = \overset{\infty}{\underset{i=1}{\otimes}} \mathfrak{B},\ P = \overset{\infty}{\underset{i=1}{\otimes}} \mu.$$

(Note that the existence of the countable product $(\Omega, \mathfrak{A}, P)$ of $(Y^0, \mathfrak{B}, \mu)$ follows from [16, Thm. III.3].) We shall denote by $\omega = (y_1, \ldots, y_n, \ldots)$ a typical element of Ω.

By the definition of the countable product of probability spaces for every $\Omega' \in \mathfrak{A}$, $B_1 \in \mathfrak{B}_{n_1}$, and $B_2 \in \mathfrak{B}_{n_2}$ satisfying the property

$$\forall \omega = (y_1, \ldots, y_n, \ldots) \in \Omega'$$

$$\Rightarrow (y_1, \ldots, y_{n_1}) \in B_1 \ \wedge\ (y_{n_1+1}, \ldots, y_{n_1+n_2}) \in B_2$$

the following estimation holds:

$$P(\Omega') \le P_{n_1}(B_1) \times P_{n_2}(B_2)$$

$$= P(\overline{B}_1) P_{n_2}(B_2), \tag{5.1}$$

where $\overline{B}_1 = B_1 \times \otimes_{i=n_1+1}^{\infty} Y^0 \in \mathfrak{A}$.

The following lemma proves useful.

LEMMA 5.1. *For any $\varepsilon,\ \eta > 0$ there exist $\tilde{S} = \tilde{S}(\varepsilon, \eta) \in \mathbf{N}$ and $\tilde{B} = \tilde{B}(\varepsilon, \eta) \in \mathfrak{B}_{\tilde{S}}$, $P_{\tilde{S}}(\tilde{B}) < \varepsilon$, such that*

$$h(\{y_1, \ldots, y_{\tilde{S}}\}, Y^0) \ge \eta$$

$$\Rightarrow (y_1, \ldots, y_{\tilde{S}}) \in \tilde{B}, \tag{5.2}$$

where $(y_1, \ldots, y_{\tilde{S}}) \in \otimes_{i=1}^{\tilde{S}} Y^0$.

For any $\omega = (y_1, \ldots, y_n, \ldots) \in \Omega$ and $n, s \in \mathbf{N}$ denote

$$Y_n^s(\omega) := \{y_{n+1}, \ldots, y_{n+s}\} \subset Y^0.$$

LEMMA 5.2. *There exists $\Omega^{(1)} \subset \Omega$, $P(\Omega^{(1)}) = 0$, such that*

$$\lim_{s \to \infty} h(Y_n^s(\omega), Y^0) = 0,\ n = 1, 2, \ldots,$$

for every $\omega \in \Omega \setminus \Omega^{(1)}$.

Proof. Fix an arbitrary $\varepsilon,\ 0 < \varepsilon < 1$, and a sequence $\eta_m > 0$, $m = 1, 2, \ldots,\ \eta_m \searrow 0$. Let $S_m = \tilde{S}(\varepsilon, \eta_m)$ and $\tilde{B}_m = \tilde{B}(\varepsilon, \eta_m) \subset \mathfrak{B}_{S_m}$ be defined by Lemma 5.1 for every $m = 1, 2, \ldots$. Set

$$\tilde{\Omega}_{n,m} := \underbrace{Y^0 \times \cdots \times Y^0}_{n\text{-}times} \times \tilde{B}_m \times \tilde{B}_m \times \cdots,$$

$$\tilde{\Omega}_n := \bigcup_{m=1}^{\infty} \tilde{\Omega}_{n,m},$$

$$\Omega^{(1)} := \bigcup_{n=1}^{\infty} \tilde{\Omega}_n.$$

Since obviously $P(\tilde{\Omega}_{n,m}) = 0$ for every $n, m = 1, 2, \ldots$, then

$$P(\Omega^{(1)}) = 0.$$

Let us fix an arbitrary ω,

$$\omega = (y_1, \ldots, y_n, \ldots) \in \Omega \setminus \Omega^{(1)}, \tag{5.3}$$

and suppose that for some $n \in \mathbf{N}$

$$\lim_{s \to \infty} h(Y_n^s(\omega), Y^0) \neq 0. \tag{5.4}$$

Since

$$Y_n^s(\omega) \subset Y_n^{s+1}(\omega) \subset Y^0,\ s = 1, 2, \ldots,$$

the sequence $\{h(Y_n^s(\omega), Y^0),\ s = 1, 2, \ldots\}$ is convergent and by (5.4)

$$h(Y_n^s(\omega), Y^0) \geq \lim_{s \to \infty} h(Y_n^s(\omega), Y^0) = \eta_0 > 0,\ s = 1, 2, \ldots.$$

It is easily seen that for every $s = 1, 2, \ldots$

$$Y_{n'}^s(\omega) \subset Y_n^{s+(n'-n)}(\omega) \subset Y^0,\ n' = n, n+1, \ldots;$$

hence,

$$h(Y_{n'}^s(\omega), Y^0) \geq h(Y_n^{s+(n'-n)}(\omega), Y^0),\ n' = n, n+1, \ldots.$$

Thus, we get

$$h(Y_{n'}^s(\omega), Y^0) \geq \eta_0 > 0,\ n' = n, n+1, \ldots,\ s = 1, 2, \ldots. \tag{5.5}$$

Choose an arbitrary m satisfying

$$\eta_m < \eta_0.$$

By (5.5) we have

$$h(Y_{n+iS_m}^{S_m}(\omega), Y) \geq \eta_0 > \eta_m,\ i = 0, 1, 2, \ldots.$$

Therefore by the definition of S_m and $\tilde{B}_m$ obtain

$$(y_{n+iS_m+1}, \ldots, y_{n+iS_m+S_m}) \in \tilde{B}_m,\ i = 0, 1, 2, \ldots.$$

Hence,

$$\omega = (y_1, \ldots, y_n, y_{n+1}, \ldots, y_{n+S_m}, y_{n+S_m+1}, \ldots)$$

$$\in \underbrace{Y \times \cdots \times Y}_{n\text{-}times} \times \overset{\infty}{\underset{i=0}{\otimes}} \tilde{B}_m = \tilde{\Omega}_n \subset \Omega^{(1)},$$

which contradicts with (5.3). Thus, the supposition (5.4) is not valid.

The lemma is proven. □

Every trajectory $\{x^n\}$ generated by SMETH.ACTIV depends upon outcomes of random experiments $y_1, \ldots, y_n, \ldots$ and hence $\{x^n\}$ is $\omega = (y_1, \ldots, y_n, \ldots)$-dependent, $\{x^n = x^n(\omega),\ n = 1, 2, \ldots\}$. Let us consider outcomes $\{x^n(\omega),\ n = 1, 2, \ldots\}$, $\omega \in \Omega$, of the method's run. We assume that the mappings $x^n(\cdot)$, $n = 1, 2, \ldots$, are measurable. Thus, $\{x^n\}$ is a sequence of random vectors on the probability space $(\Omega, \mathfrak{A}, P)$, and we shall study P-almost surely (P-a.s.) convergence properties of $\{x^n\}$.

For any $\omega = (y_1, y_2, \ldots) \in \Omega$ consider the corresponding trajectory $\{x^n\}$ of the method SMETH.ACTIV. Let $R_n(\omega)$ denote the number of random experiment executed to obtain $x^1(\omega), \ldots, x^n(\omega)$, i.e.,

$$x^n(\omega) = x^n(y_1, y_2, \ldots) = x^n(y_1, y_2, \ldots, y_{R_n}),\ n = 2, 3, \ldots.$$

(It is clear that $R_1(\omega) = 0$.)

By the definition of the method $S_n(\omega)$ is the number of random experiment executed at the nth iteration of SMETH.ACTIV to obtain $x^{n+1}(\omega)$; hence

$$R_{n+1}(\omega) = R_n(\omega) + S_n(\omega),\ n = 1, 2, \ldots.$$

It is possible that at the n_0th iteration of SMETH.ACTIV the inequality (2.3) in the procedure SPROC.ACTIV does not hold for every $i = 1, 2, \ldots$. Thus, the method generates a finite trajectory $x^1(\omega), \ldots, x^{n_0}(\omega)$. In this case we set

$$S_{n_0}(\omega) = +\infty,\ S_n(\omega) = 0,\ n = n_0 + 1, n_0 + 2, \ldots$$

and

$$N_{stop}(\omega) = n_0;$$

otherwise, $S_n(\omega) < +\infty,\ n = 1, 2, \ldots$, and

$$N_{stop}(\omega) = +\infty.$$

For a subset D of X^0 and $s \in \mathbf{N}$ consider the sets

$$\{1 \le n \le s \mid x^n(\omega) \in D\},\ s = 1, 2, \ldots;$$

obviously,

$$\{n \in \mathbf{N} \mid x^n(\omega) \in D\} = \bigcup_{s=1}^{\infty} \{1 \le n \le s \mid x^n(\omega) \in D\}.$$

Let us define

$$\mathcal{N}(D, s \mid \omega) :=\mid \{1 \le n \le s \mid x^n(\omega) \in D\} \mid,\ s = 1, 2, \ldots,$$

$$\mathcal{N}(D, \infty \mid \omega) :=\mid \{n \in \mathbf{N} \mid x^n(\omega) \in D\} \mid,$$

and

$$m_n(D \mid \omega) := \min\{s \in \mathbf{N} \mid \mathcal{N}(D, s \mid \omega) = n\};$$

when $\{s \in \mathbf{N} \mid \mathcal{N}(D, s \mid \omega) = n\} = \emptyset$ (i.e., $\mathcal{N}(D, \infty \mid \omega) < n$), we set $m_n(D \mid \omega) := +\infty$. It is clear that $m_{n_0}(D \mid \omega) = n^*$ means that the inclusion

$$x^n(\omega) \in D$$

holds for exactly n_0 of the first n^* elements of the trajectory $\{x^n(\omega),\ n = 1, 2, \ldots\}$ and, moreover, it holds for $n = n^*$; i.e., there exist $j_1, \ldots, j_{n_0},\ 1 \le j_1 \le \cdots \le j_{n_0} = n^*$, such that

$$x^j(\omega) \in D \quad \forall j \in \{j_1, \ldots, j_{n_0}\}$$

and

$$x^j(\omega) \notin D \quad \forall j \notin \{j_1, \ldots, j_{n_0}\},\ 1 \le j \le n^*.$$

Let us denote

$$\overline{\Omega}(D) := \{\omega \in \Omega \mid m_n(D \mid \omega) < +\infty \quad \forall n = 1, 2, \ldots\}.$$

It follows from the definition of $m_n(D \mid \omega)$ that

$$\omega \in \overline{\Omega}(D) \Leftrightarrow x^n(\omega) \in D \text{ for an infinite number of } n \in \mathbf{N}.$$

Set

$$\overline{\Omega}_\infty = \{\omega \in \Omega \mid N_{stop}(\omega) < +\infty\},$$

$$\Omega_\infty = \Omega \setminus \overline{\Omega}_\infty = \{\omega \in \Omega \mid \{x^n(\omega)\} \text{ is infinite}\}.$$

Note that, for every $D \subset X^0$,

$$\overline{\Omega}(D) \subset \Omega_\infty.$$

Now we establish the basic property of the considered outer approximations scheme.

LEMMA 5.3. *For every $\omega \in \Omega_\infty$ the following property holds:*

$$\lim_{n \to \infty} S_n(\omega) = +\infty.$$

Proof. Let us fix an arbitrary $\omega \in \Omega_\infty$ and consider the corresponding outcome sequences $\{x^n(\omega)\}$, $\{S_n(\omega)\}$, $\{Y_n(\omega)\}$, $\{\Delta Y_n(\omega)\}$, $\{\overline{Y}_n(\omega)\}$, $\{\theta_n(\omega)\}$; for simplicity we shall denote them by $\{x^n\}$, $\{S_n\}$, $\{Y_n\}$, $\{\Delta Y_n\}$, $\{\overline{Y}_n\}$, and $\{\theta_n\}$. Note that by (2.4), (2.5)

$$Y_n \subset Y_{n+1},\ n = 1, 2, \ldots. \tag{5.6}$$

Suppose that

$$\lim_{n\to\infty} S_n(\omega) \neq +\infty. \tag{5.7}$$

Therefore, there exist a subsequence $\{x^{t_n}\}$ of $\{x^n\}$ and $S_0 \in \mathbf{N}$ such that

$$\lim_{n\to\infty} x^{t_n} = x_0, \tag{5.8}$$

$$S_{t_n} \leq S_0,\ n = 1, 2, \ldots.$$

It follows from the definition of SPROC.ACTIV that

$$\theta_{t_n} = \Theta(x^{t_n}, \overline{Y}_{t_n}) > \gamma/S_{t_n} > \gamma/S_0 > 0,\ n = 1, 2, \ldots. \tag{5.9}$$

By (2.4) there exists $N \in \mathbf{N}$ such that

$$\sigma_n < \gamma/S_0,\ n = N, N+1, \ldots;$$

and without loss of generality we assume that

$$\theta_{t_n} > \sigma_{t_n},\ n = 1, 2, \ldots.$$

Hence by the construction of SMETH.ACTIV we have

$$Y_n \supset \Delta Y_{t_i}\ \ \forall i :\ t_i \leq n$$

for every $n \geq t_1$.

Therefore

$$Y_{t_n} \supset \Delta Y_{t_{n'}}\ \ \forall n, n',\ n \geq n'. \tag{5.10}$$

Let us define a sequence $\{U_n\}$ by the following rule:

$$U_1 = Y_{t_1},\ U_2 = Y_{t_1} \cup \Delta Y_{t_1}, \ldots,$$

$$U_{2n-1} = Y_{t_n},\ U_{2n} = Y_{t_n} \cup \Delta Y_{t_n}, \ldots.$$

It follows from (5.6) and (5.10) that

$$U_{2n} = Y_{t_n} \cup \Delta Y_{t_n} \subset Y_{t_{n+1}} = U_{2n+1},\ n = 1, 2, \ldots;$$

thus,

$$U_n \subset U_{n+1},\ n = 1, 2, \ldots.$$

Set

$$U_0 = \bigcup_{n=1}^{\infty} U_n.$$

By the monotonicity of $\{U_n\}$ we obtain

$$\lim_{n\to\infty} h(U_n, U_0) = 0. \tag{5.11}$$

By (2.6) we get

$$\Theta(x^{t_n}, Y_{t_n}) = \Theta(x^{t_n}, U_{2n-1}) \le \varepsilon_{t_n},\ n = 1, 2, \dots;$$

hence

$$\lim_{n\to\infty} \inf \Theta(x^{t_n}, U_{2n-1}) = 0.$$

Therefore from (5.8), (5.11), and Assumption A2(i) we obtain

$$\Theta(x_0, U_0) = 0$$

and, furthermore, by Assumption A2(ii)

$$\lim_{n\to\infty} \Theta(x^{t_n}, U_{2n}) = 0. \tag{5.12}$$

It follows from the definition of SPROC.ACTIV (see the dropping step) that

$$\Delta Y_n \succeq \mathcal{A}(x^n, \overline{Y}_n),$$

$$\Delta Y_n \subset Y_n \cup \Delta Y_n \subset \overline{Y}_n,\ n = 1, 2, \dots;$$

thus,

$$\Theta(x^n, \Delta Y_n) = \Theta(x^n, Y_n \cup \Delta Y_n) = \Theta(x^n, \overline{Y}_n),\ n = 1, 2, \dots.$$

Applying (5.9) we obtain

$$\Theta(x^{t_n}, U_{2n}) = \Theta(x^{t_n}, Y_{t_n} \cup \Delta Y_{t_n}) = \Theta(x^{t_n}, \overline{Y}_{t_n}) \ge \gamma/S_0 > 0,\ n = 1, 2, \dots;$$

hence

$$\lim_{n\to\infty} \sup \Theta(x^{t_n}, U_{2n}) > 0,$$

which contradicts with (5.12).

Thus, (5.6) does not hold.

The lemma is proven. □

6. Convergence theorem. The following lemma plays the important role in the convergence analysis of the method SMETH.ACTIV.

LEMMA 6.1. *For every $x \notin X^0_{qopt}$ there exist $\delta_* = \delta_*(x) > 0$,*

$$B_{\delta_*}(x) \cap X^0 \subset X^0 \setminus X^0_{qopt}, \tag{6.1}$$

and $\Omega^* = \Omega^*(x) \subset \Omega_\infty$, $P(\Omega^*) = 0$, such that the following property holds:

$$\overline{lt}\{x^n(\omega)\} \cap B_{\delta_*}(x) = \emptyset \quad \forall \omega \in \Omega_\infty \setminus \Omega^*.$$

Proof. Let us fix an arbitrary $x \notin X^0_{qopt}$. By the definition of X^0_{qopt} we have

$$\Theta(x, Y^0) > 0;$$

hence by Assumption A2(i) there exist $\delta_* > 0$ satisfying (6.1) and $\eta_*,\ d_* > 0$ such that

$$\Theta(x', U) \ge d_* > 0 \tag{6.2}$$

for every $x' \in B_{\delta_*}(x) \cap X^0$ and $U \subset Y^0,\ h(U, Y^0) < \eta_*$.

Set

$$\Omega^* = \overline{\Omega}(D),\ D = B_{\delta_*}(x).$$

It is obvious that $\Omega^* \subset \Omega_\infty$.

Let us fix an arbitrary $\varepsilon > 0$ and show that

$$P(\Omega^*) \le \varepsilon.$$

By Lemma 5.1 there exist $\tilde{S} \in \mathbf{N}$,

$$\tilde{S} > \gamma/d_*, \tag{6.3}$$

and $\tilde{B} \in \mathfrak{B}_S$, $P_S(\tilde{B}) < \varepsilon$, such that (5.2) holds.

It follows from Lemma 5.3 that for every $\omega \in \Omega^*$

$$\lim_{n\to\infty} S_{m_n(D|\omega)}(\omega) = +\infty;$$

hence,

$$\Omega^* = \bigcup_{l=1}^{\infty} \Omega_l,$$
$$\Omega_l = \{\omega \in \Omega^* \mid S_{m_n(D|\omega)}(\omega) \ge S+1,$$
$$n = l, l+1, \ldots\}.$$

It is clear that

$$\Omega_l \subset \Omega_{l+1},\ l = 1, 2, \ldots;$$

therefore,

$$P(\Omega^*) = \lim_{l\to\infty} P(\Omega_l). \tag{6.4}$$

Let us fix an arbitrary $l \in \mathbf{N}$ and set

$$\Omega^i = \{\omega \in \Omega \mid R_{m_l(D|\omega)}(\omega) = i\},\ i = 1, 2, \ldots;$$

it is easy to see that there exist $B_i \in \mathfrak{B}_i$, $i = 1, 2, \ldots$, such that

$$\Omega^i = B_i \times \bigotimes_{j=i+1}^{\infty} Y,\ i = 1, 2, \ldots. \tag{6.5}$$

Note that

$$\Omega^i \cap \Omega^{i'} = \emptyset\ \ \forall i \ne i',\ i, i' = 1, 2, \ldots,$$
$$\Omega_l \subset \bigcup_{i=1}^{\infty} \Omega^i.$$

Set

$$\Omega_l^i = \Omega_l \cap \Omega^i,\ i = 1, 2, \ldots.$$

Then obtain

$$\Omega_l^{i'} \cap \Omega_l^{i''} = \emptyset\ \ \forall i' \ne i'',\ i', i'' = 1, 2, \ldots,$$
$$\Omega_l = \bigcup_{i=1}^{\infty} \Omega_l^i,$$

and hence

$$P(\Omega_l) = \sum_{i=1}^{\infty} P(\Omega_l^i). \tag{6.6}$$

Consider an arbitrary Ω_l^i.

For every $\omega \in \Omega_l^i$ we have

$$R_{m_l(D|\omega)}(\omega) = i,$$

$$x^{m_l(D|\omega)} \in D \cap X^0 = B_{\delta_*}(x) \cap X^0, \tag{6.7}$$

$$S_{m_l(D|\omega)}(\omega) \geq S + 1.$$

(For simplicity we shall use m_l instead of $m_l(D \mid \omega)$.) Therefore,

$$x^{m_l}(\omega) = x^{m_l}(y_1, \ldots, y_i),$$

$$\tilde{S}\ \Theta(x^{m_l}(\omega), Y_{m_l} \cup \{y_{i+1}, y_{i+1}^*, \ldots, y_{i+s}, y_{i+s}^*\}) \leq \gamma,$$

and, furthermore, by (6.3)

$$\Theta(x^{m_l}(\omega), Y_{m_l} \cup \{y_{i+1}, y_{i+1}^*, \ldots, y_{i+s}, y_{i+s}^*\}) \leq d_*.$$

Thus, by (6.2) and (6.7) we obtain

$$h((Y_{m_l} \cup \{y_{i+1}, y_{i+1}^*, \ldots, y_{i+s}, y_{i+s}^*\}), Y^0) \geq \eta_*,$$

and hence

$$h(\{y_{i+1}, \ldots, y_{i+s}\}, Y^0) \geq \eta_*.$$

Since by the choice of $\tilde{S}$ the property (4.2) holds,

$$(y_{i+1}, \ldots, y_{i+s}) \in \tilde{B} \in \mathfrak{B}_{\tilde{S}},\ P_{\tilde{S}}(\tilde{B}) < \varepsilon.$$

Taking into account (6.5) and applying (5.1) we get

$$P(\Omega_l^i) \leq P(\Omega^i) P_{\tilde{S}}(\tilde{B}) \leq \varepsilon P(\Omega^i),\ i = 1, 2, \ldots.$$

Turning back to (6.6) we obtain

$$P(\Omega_l) = \sum_{i=1}^{\infty} P(\Omega_l^i) \leq \varepsilon \sum_{i=1}^{\infty} P(\Omega^i) \leq \varepsilon,\ l = 1, 2, \ldots;$$

therefore by (6.4)

$$P(\Omega^*) \leq \varepsilon.$$

Since in this estimation ε is an arbitrary positive number,

$$P(\Omega^*) = 0.$$

By the definition of $\Omega^* = \overline{\Omega}(B_{\delta_*}(x))$ for every $\omega \in \Omega_\infty \setminus \Omega^*$ the inclusion

$$x^n(\omega) \in B_{\delta_*}(x)$$

holds only for a finite number of $n \in \mathbf{N}$.

Thus,

$$\overline{lt}\{x^n(\omega)\} \cap B_{\delta_*}(x) = \emptyset \quad \forall \omega \in \Omega_\infty \setminus \Omega^*.$$

The lemma is proven. □

We also need the property of finite trajectories of SMETH.ACTIV.

LEMMA 6.2. *For every* $\omega \in \overline{\Omega}_\infty \setminus \Omega^{(1)}$, $P(\Omega^{(1)}) = 0$, *the following property holds:*

$$x^{N_{stop}}(\omega) \in X^0_{qopt},$$

where $\Omega^{(1)}$ *is defined in Lemma* 5.2.

Proof. The desired result follows immediately from Assumption A2(i) and Lemma 5.2. □

THEOREM 6.1. *There exists* $\Omega^{(0)} \subset \Omega$, $P(\Omega^{(0)}) = 0$, *such that for every* $\omega \in \Omega \setminus \Omega^{(0)}$ *the trajectory* $\{x^n(\omega)\}$ *of SMETH.ACTIV satisfies one of the following properties:*

(i) $N_{stop} < +\infty$ and $x^{N_{stop}}(\omega) \in X^0_{qopt}$;

(ii) $N_{stop} = +\infty$ and $\{x^n(\omega)\}$ converges to X^0_{qopt}.

Proof. Let us fix an arbitrary sequence $\zeta_j \searrow 0$ and consider the following sets:

$$Z_j = X^0 \setminus B_{\zeta_j}(X^0_{qopt}),\ j = 1, 2, \ldots.$$

It is clear that $Z_j,\ j = 1, 2, \ldots$, are compact and

$$\bigcup_{j=1}^{\infty} Z_j = X^0 \setminus X^0_{qopt}. \tag{6.8}$$

For any $j = 1, 2, \ldots$ it follows from (6.1) that

$$X \setminus X_{qopt} \supset \bigcup_{x \in Z_j} B_{\delta_*(x)}(x) \supset Z_j,$$

where $\delta_*(x) > 0$ is defined in Lemma 6.1.

Since Z_j is compact, there is a finite set of points $x_{i,j},\ i = 1, 2, \ldots, I_j$, such that

$$X^0 \setminus X^0_{qopt} \supset \bigcup_{i=1}^{I_j} B_{\delta_*(x_{i,j})}(x_{i,j}) \supset Z_j,$$

for every $j = 1, 2, \ldots$.

Hence, by (6.8)

$$X^0 \setminus X^0_{qopt} = \bigcup_{j=1}^{\infty} \bigcup_{i=1}^{I_j} B_{\delta_*(x_{i,j})}(x_{i,j}). \tag{6.9}$$

Define

$$\Omega^*_j = \bigcup_{i=1}^{I_j} \Omega^*(x_{i,j}), \quad \Omega^{(2)} = \bigcup_{j=1}^{\infty} \Omega^*_j,$$

where $\Omega^*(x) \subset \Omega$, $P(\Omega^*(x)) = 0$, is defined in Lemma 6.1. It is obvious that

$$P(\Omega^{(2)}) = 0.$$

Set

$$\Omega^{(0)} = \Omega^{(1)} \cup \Omega^{(2)},\ P(\Omega^{(0)}) = 0,$$

where $\Omega^{(1)} \subset \Omega$, $P(\Omega^{(1)}) = 0$, is defined in Lemma 5.2 (see also Lemma 6.2).

Let us fix an arbitrary $\omega \in \Omega \setminus \Omega^{(0)}$.

If $\omega \in \overline{\Omega}_\infty$, then $\omega \in \overline{\Omega}_\infty \setminus \Omega^{(1)}$ and by Lemma 6.2 the trajectory $\{x^n\}$ satisfies the property:

$$N_{stop}(\omega) < \infty, \quad x^{N_{stop}}(\omega) \in X^0_{qopt}.$$

If $\omega \in \Omega_\infty$, then

$$\omega \in \Omega_\infty \setminus \Omega^*(x_{i,j}),\ i = 1, \ldots, I_j,\ j = 1, 2, \ldots$$

Therefore, from Lemma 6.1 we have

$$\overline{lt}\{x^n(\omega)\} \cap B_{\delta_*(x_{i,j})}(x_{i,j}) = \emptyset,\ i = 1, \ldots, I_j,\ j = 1, 2, \ldots,$$

i.e.,

$$\overline{lt}\{x^n(\omega)\} \cap \bigcup_{j=1}^{\infty} \bigcup_{i=1}^{I_j} B_{\delta_*(x_{i,j})}(x_{i,j}) = \emptyset;$$

and hence by (6.9) we obtain

$$\overline{lt}\{x^n(\omega)\} \cap (X^0 \setminus X^0_{qopt}) = \emptyset.$$

Thus,

$$\overline{lt}\{x^n(\omega)\} \subset X^0_{qopt}.$$

The theorem is proven. □

7. Stochastic algorithm for solving the approximation problem. Consider the following approximation problem:

apprx.$\mathcal{P}^0$: find $x \in \mathcal{X}^0_{opt}$,

$$\mathcal{X}^0_{opt} = \{x \in X^0 \mid \max_{y \in Y^0} \mid \Phi(x, y) - F(y) \mid = v^0_{opt}\},$$

$$v^0_{opt} = \min_{x \in X^0} \max_{y \in Y^0} \mid \Phi(x, y) - F(y) \mid,$$

where the functions $\Phi(\cdot, \cdot)$, $F(\cdot)$ are assumed to be continuous and continuously differentiable on $X^0 \times Y^0$, $\nabla\Phi(\cdot) = (\nabla_x\Phi(\cdot, \cdot), \nabla_y\Phi(\cdot, \cdot)) \in \mathbf{C}_{Lip}(X^0 \times Y^0, L)$, $\nabla F(\cdot) \in \mathbf{C}_{Lip}(Y^0, L)$, $L > 0$, $X^0 \subset \Re^k$ is polyhedral and compact, and $Y^0 \subset \Re^l$ is convex and compact.

To demonstrate a practical embodiment of the proposed activated random search technique for forming relevant parameter sets within the outer approximations method we consider a stochastic algorithm for solving the problem apprx.$\mathcal{P}^0$. The algorithm is constructed as a version of the general method SMETH.ACTIV specially adapted for apprx.$\mathcal{P}^0$.

We also introduce the stationary set $\mathcal{X}^0_{stat}$ of the problem apprx.$\mathcal{P}^0$, i.e., the set of all points satisfying the first-order necessary optimality conditions. For $x \in X^0$ denote

$$R^0(x) := \{y \in Y^0 \mid \ \mid \Phi(x,y) - F(y) \mid \ = \max_{y \in Y^0} \mid \Phi(x,y) - F(y) \mid\}.$$

The stationary set $\mathcal{X}^0_{stat}$ is given by

$$\begin{aligned}\mathcal{X}^0_{stat} := \{x \in X^0 \mid &\text{there exist } y_1, \dots, y_s \in R^0(x)\\ &\text{and } \lambda_1, \dots, \lambda_s \in \Re^1,\ \sum_{i=1}^{s} \mid \lambda_i \mid = 1,\ s \le k+1, \text{s.t.}\\ &\sum_{i=1}^{s} \lambda_i \nabla_x \Phi(x, y_i) + K_{X^0}(x) \ni \mathbf{0},\\ &\lambda_i(\Phi(x, y_i) - F(y_i)) \ge 0,\ i = 1, \dots, s\}.\end{aligned}$$

Note that if the function $\Phi(x,y)$ is linear with respect to x for every $y \in Y^0$, i.e.,

$$\Phi(x,y) = \langle a(y), x\rangle + b(y) \quad \forall x \in X^0,\ a(y) \in \Re^k,\ b(y) \in \Re^1,\ y \in Y^0, \tag{7.1}$$

then the objective function

$$\varphi(x) = \max_{y \in Y^0} \mid \Phi(x,y) - F(y) \mid$$

of the problem apprx.$\mathcal{P}^0$ is convex on X^0 and

$$\mathcal{X}^0_{stat} = \mathcal{X}^0_{opt}.$$

To construct the outer approximations algorithm for solving apprx.$\mathcal{P}^0$ we define the simpler problems apprx.$\mathcal{P}[Y]$, $Y \in \mathfrak{M}_c(Y^0)$, by

apprx.$\mathcal{P}[Y]$: to find $x \in \mathcal{X}_{opt}[Y]$,

$$\mathcal{X}_{opt}[Y] := \{x \in X^0 \mid \ \max_{y \in Y} \mid \Phi(x,y) - F(y) \mid = v_{opt}[Y]\},$$

$$v_{opt}[Y] := \min_{x \in X^0} \max_{y \in Y} \mid \Phi(x,y) - F(y) \mid .$$

It is clear that the problem apprx.$\mathcal{P}[Y^0]$ is the initial problem apprx.$\mathcal{P}^0$ and $\mathcal{X}_{opt}[Y^0] = \mathcal{X}^0_{opt}$. For $x \in X^0$ and $Y \in \mathfrak{M}_c(Y^0)$ denote

$$R(x,Y) := \{y \in Y \mid \ \mid \Phi(x,y) - F(y) \mid \ = \max_{y \in Y} \mid \Phi(x,y) - F(y) \mid\};$$

obviously $R(x, Y^0) = R^0(x)$. The stationary set $\mathcal{X}_{stat}[Y]$ of the problem apprx.$\mathcal{P}[Y]$ is as follows:

$$\begin{aligned}\mathcal{X}_{stat}[Y] := \Big\{ x \in X^0 \mid &\text{ there exist } y_1, \dots, y_s \in R(x,Y)\\ &\text{and } \lambda_1, \dots, \lambda_s \in \Re^1,\ \textstyle\sum_{i=1}^{s} \mid \lambda_i \mid = 1,\ s \le k+1, \text{s.t.}\\ &\sum_{i=1}^{s} \lambda_i \nabla_x \Phi(x, y_i) + K^*_{X^0}(x) \ni \mathbf{0},\\ &\lambda_i(\Phi(x, y_i) - F(y_i)) \ge 0,\ i = 1, \dots, s\Big\}.\end{aligned}$$

Let $V \subset \Re^1$ be a line segment, $V = [\underline{v}, \overline{v}]$, where

$$\underline{v} < 0, \ \max_{x \in X^0} \max_{y \in Y^0} \mid \Phi(x,y) - F(y) \mid + L \text{ diam } X^0 < \overline{v}.$$

It is clear that for every $Y \in \mathfrak{M}_c(Y^0)$ the problem apprx.$\mathcal{P}[Y]$ can be rewritten in the following form:

apprx.$\mathcal{P}[Y]'$:
$$v \to \min_{x,v}$$
$$\text{s.t. } \Phi(x,y) - F(y) - v \le 0 \ \ \forall y \in Y,$$
$$-\Phi(x,y) + F(y) - v \le 0 \ \ \forall y \in Y,$$
$$x \in X^0, \ v \in V;$$

i.e., the solution set of apprx.$\mathcal{P}[Y]'$ is equal to $\mathcal{X}_{opt}[Y] \times \{v_{opt}[Y]\}$.

Similarly to Example 3.2 we define the auxiliary quadratic programming problem:

apprx.$\mathcal{QP}(x, Y)$:
$$v + \frac{1}{2} \parallel p \parallel^2 \to \min_{p,v}$$
$$\text{s.t. } \Phi(x,y) + \langle \nabla_x \Phi(x,y), p \rangle - F(x) - v \le 0 \ \ \forall y \in Y,$$
$$-\Phi(x,y) - \langle \nabla_x \Phi(x,y), p \rangle + F(x) - v \le 0 \ \ \forall y \in Y,$$
$$x + p \in X^0, \ v \in V.$$

It is easy to show that for every $x \in X^0$ and $Y \in \mathfrak{M}_f(Y^0)$ there exists the unique solution $(p_0(x,Y), v_0(x,Y))$ of apprx.$\mathcal{QP}(x,Y)$ and, moreover,

$$v_0(x,Y) = \max_{y \in Y} \mid \Phi(x,y) + \langle \nabla_x \Phi(x,y), p_0(x,Y) \rangle - F(x) \mid .$$

Furthermore, the following optimality criterion holds.

LEMMA 7.1. *For every $x \in X^0$ and $Y \in \mathfrak{M}_c(Y^0)$ a vector $(p,v) \in \Re^k \times \Re^1$ is the solution of* apprx.$\mathcal{QP}(x,Y)$ *if and only if there exist $y_1, \dots, y_s \in Y$ and $\lambda_1, \dots, \lambda_s \in \Re^1$, $\sum_{i=1}^s \mid \lambda_i \mid = 1$, $s \le k+1$, such that the following properties hold:*

$$p + \sum_{i=1}^{s} \lambda_i \nabla_x \Phi(x, y_i) + K^*_{X^0}(x+p) \ni \mathbf{0}, \tag{7.2}$$

$$v = \max_{y \in Y} \mid \Phi(x,y) + \langle \nabla_x \Phi(x,y), p \rangle - F(x) \mid, \tag{7.3}$$

$$\mid \Phi(x, y_i) + \langle \nabla_x \Phi(x, y_i), p \rangle - F(x) \mid = v, \ i = 1, \dots, s, \tag{7.4}$$

$$\lambda_i (\Phi(x,y) + \langle \nabla_x \Phi(x,y), p \rangle - F(x)) \ge 0, \ i = 1, \dots, s, \tag{7.5}$$

$$x + p \in X. \tag{7.6}$$

From Lemma 7.1 we get immediately the following corollaries.

COROLLARY 7.1. *For every* $x \in X^0$ *and* $Y \in \mathfrak{M}_c(Y^0)$ *the following property holds:*

$$p_0(x, Y) = 0 \Leftrightarrow x \in \mathcal{X}_{stat}(Y).$$

COROLLARY 7.2. *For every* $x \in X^0$, $Y \in \mathfrak{M}_c(Y^0)$, *and sequences* $\{x_n\}$, $\{Y_n\}$ *satisfying*

$$x_n \in X^0,\ n = 1, 2, \ldots,\ \lim_{n\to\infty} x_n = x,$$

$$Y_n \in \mathfrak{M}_c(Y^0),\ n = 1, 2, \ldots,\ \lim_{n\to\infty} h(Y_n, Y) = 0,$$

the following property holds:

$$\lim_{n\to\infty} p_0(x_n, Y_n) = p_0(x, Y).$$

Let us define a quasi-optimality function $\Theta(\cdot,\cdot) :\ X^0 \times \mathfrak{M}_c(Y^0)\ \to \Re^1_+$ by

$$\Theta(x, Y) :=\| p_0(x, Y) \|,\ x \in X^0,\ Y \in \mathfrak{M}_c(Y^0).$$

Applying Corollary 7.1 we obtain that

$$\mathcal{X}_{qopt}[Y] = \mathcal{X}_{stat}[Y]\ \ \forall Y \in \mathfrak{M}_c(Y^0)$$

and, in particular,

$$\mathcal{X}_{qopt}[Y^0] = \mathcal{X}^0_{stat} \supset \mathcal{X}^0_{opt}. \tag{7.7}$$

From Corollary 7.2 and (7.7) we deduce the following important result.

LEMMA 7.2. *The considered quasi-optimality function* $\Theta(\cdot,\cdot)$ *satisfies Assumptions* A1 *and* A2 *and*

$$\mathcal{X}_{qopt}[Y^0] = \mathcal{X}^0_{stat}.$$

Moreover, if (7.1) *holds, then*

$$\mathcal{X}_{qopt}[Y^0] = \mathcal{X}^0_{opt}.$$

For any $x \in X^0$ and $Y \in \mathfrak{M}_c(Y^0)$ denote

$$\mathfrak{Y}(x, Y) = \Big\{ \{y_1, \ldots, y_s\} \in \mathfrak{M}_f(Y) \mid s \le k+1 \text{ and there exist}$$

$$\lambda_1, \ldots, \lambda_s \ge 0,\ \sum_{i=1}^{s} \mid \lambda_i \mid = 1, \text{s.t. (7.2), (7.3), (7.4),}$$

$$\text{(7.5), and (7.6) hold for } p = p_0(x, Y) \text{ and } v = v_0(x, Y)\Big\}.$$

It is clear that for the considered quasi-optimality function $\Theta(\cdot,\cdot)$ the set of all active parameter sets is as follows:

$$\mathcal{A}(x, Y) = \{Y' \in \mathfrak{Y}(x, Y) \mid \text{ there does not exist } Y'' \in \mathfrak{Y}(x, Y)$$

$$\text{s.t. } Y'' \subset Y', Y'' \ne Y'\}.$$

Hence, we have

$$\forall \Delta Y \in \mathfrak{Y}(x, Y) \Rightarrow \Delta Y \succeq \mathcal{A}(x, Y). \tag{7.8}$$

Note that for any $x \in X^0$ and $Y \in \mathfrak{M}_f(Y^0)$ calculating $\Theta(x, Y)$ (i.e., solving the problem apprx.$\mathcal{QP}(x, Y)$) with the use of the conjugate gradient method (see [19]) we obtain $p_0(x, Y)$ together with a finite set $\mathcal{Y}_0(x, Y) = \{y_1, \ldots, y_s\} \subset Y$, $s \leq k + 1$, such that

$$p_0(x, Y) = -\sum_{i=1}^{s} \lambda_i \nabla_x \Phi(x, y_i) - \eta,\ \eta \in K^*_{X^0}(x + p_0(x, Y)), \tag{7.9}$$

$$\mathcal{Y}_0(x, Y) = \{y_1, \ldots, y_s\} \in \mathfrak{Y}(x, Y), \tag{7.10}$$

and by (7.8)

$$\mathcal{Y}_0(x, Y) \succeq \mathcal{A}(x, Y). \tag{7.11}$$

Now we construct an activation mechanism for the search of relevant parameters. Let us fix an arbitrary $x \in X^0$ and consider the inner problem

apprx.$\mathcal{JP}(x)$:
$$\mid \Phi(x, y) - F(y) \mid \rightarrow \max_y$$
$$\text{s.t. } y \in Y^0;$$

for simplicity we denote

$$f(y \mid x) := \mid \Phi(x, y) - F(y) \mid .$$

For every $y \in Y^0$, $f(y \mid x) \neq 0$, define

$$h(y \mid x) := pr_{Y^0}(y + \ \text{sign}(\Phi(x, y) - F(y))(\nabla_y \Phi(x, y) - \nabla F(y))) - y$$

$$= pr_{Y^0}(y + \nabla_y f(y \mid x)) - y.$$

When $f(y \mid x) = 0$ we set

$$h(y \mid x) := \arg \max_{\overline{h} \in \overline{H}} \parallel \overline{h} \parallel,$$

$$\overline{H} = \{pr_{Y^0}(y + (\nabla_y \Phi(x, y) - \nabla F(y))) - y,$$

$$pr_{Y^0}(y - (\nabla_y \Phi(x, y) - \nabla F(y))) - y\}.$$

The ε-stationary set $Y^{\varepsilon}_{stat}(x)$ of the problem apprx.$\mathcal{JP}(x)$ is given by

$$Y^{\varepsilon}_{stat}(x) := \{y \in Y^0 \mid\mid\mid h(y \mid x) \parallel \leq \varepsilon\},\ \varepsilon \geq 0.$$

Obviously, $Y^0_{stat}(x)$ is the set of all points satisfying the standard first-order necessary optimality conditions for apprx.$\mathcal{JP}(x)$; in particular, if

$$\min_{y \in Y^0} f(y \mid x) > 0,$$

then we have

$$Y^0_{stat}(x) = \{y \in Y^0 \mid -\nabla_y f(y \mid x) + K^*_{Y^0}(y) \ni \mathbf{0}\}.$$

The gradient projection method for the local search in apprx.$\mathcal{JP}(x)$ is as follows:

$$\text{GPM}(x): \qquad y_{n+1} = y_n + a_n h(y_n \mid x),\ n = 1, 2, \ldots,\ y_1 \in Y^0,$$

where stepsizes $a_n \geq 0$, $n = 1, 2, \ldots$, can be chosen, for instance, by Armijo's rule:

$$a_n = \eta^{i_n},$$

$$i_n := \min\{i \in \{0, 1, 2, \ldots\} \mid f(y_n + \eta^i h(y_n \mid x) \mid x) - f(y_n \mid x)$$

$$\geq \kappa\eta^i \parallel h(y_n \mid x) \parallel^2\},$$

$0 < \eta$, $\kappa < 1$ are parameters; $y_1 \in Y^0$ is a starting point.

Based on the convergence properties of the gradient projection method we obtain the following result.

LEMMA 7.3. *Every trajectory* $\{y_n\}$ *of* GPM(x) *converges to* $Y^0_{stat}(x)$, *and for every* $\varepsilon > 0$ *there exists* $N \in \mathbf{N}$ *such that*

$$f(y_1 \mid x) \leq f(y_N \mid x),\ y_N \in Y^{\varepsilon}_{stat}.$$

Now following the proposed general approach we construct a stochastic algorithm for solving the approximation problem apprx.$\mathcal{P}^0$.

METHOD SMETH.ACTIV.apprx.

Data. $x^1 \in X^0$.

Parameters. $\gamma, \varepsilon > 0$, $0 < \eta$, $\kappa < 1$, *and sequences* $\{\varepsilon_n\}$, $\{\sigma_n\}$, $\varepsilon_n, \sigma_n > 0$, $n = 1, 2, \ldots$, $\varepsilon_n, \sigma_n \searrow 0$.

Stage 0. *Set*

$$n := 1,$$

$$Y_1 := \emptyset.$$

Stage 1.

Step 0. *Set*

$$i := 1,$$

$$\overline{Y_n} := Y_n.$$

Step 1. *Determine a point* $y^n_i \in Y^0$ *by using the uniform probability distribution on* Y^0. *Set*

$$\overline{Y_n} := \overline{Y_n} \cup \{y^n_i\}.$$

Step 1.ACTIV. *Applying* GPM(x^n) *(with parameters* η *and* κ*), starting with* y^n_i, *obtain* $y^{n,*}_i \in Y^0$ *satisfying*

$$\mid \Phi(x^n, y^{n,*}_i) - F(y^{n,*}_i) \mid \geq \mid \Phi(x^n, y^n_i) - F(y^n_i) \mid,$$

$$y^{n,*}_i \in Y^{\varepsilon}_{stat}(x^n).$$

Set

$$\overline{Y}_n := \overline{Y}_n \cup \{y_i^{n,*}\}.$$

Step 2. *Solving the quadratic problem* apprx.$\mathcal{QP}(x^n, \overline{Y_n})$, *obtain* $p_0(x^n, \overline{Y_n})$ *and* $\mathcal{Y}_0(x^n, \overline{Y_n}) \subset \overline{Y_n}$, $\mid \mathcal{Y}_0(x^n, \overline{Y_n}) \mid \leq k+1$ (*see* (7.9), (7.10)).
Set

$$\theta_n^i := \| \, h(x^n, \overline{Y}_n) \, \| \, .$$

Step 3. *If*

$$i\theta_n^i \leq \gamma,$$

then set

$$i := i + 1$$

and go to Step 1.

Step 4. *Set*

$$\theta_n := \theta_n^i,$$

$$S_n := i,$$

$$\Delta Y_n := \mathcal{Y}_0(x^n, \overline{Y_n}).$$

Go to Stage 2.

Stage 2. *Set*

$$Y_{n+1} := \Delta Y_n \cup \bigcup_{\substack{j:\theta_j > \sigma_j, \\ 1 \leq j \leq n-1}} \Delta Y_j.$$

Stage 3. *Solving the problem* apprx.$\mathcal{P}[Y_{n+1}]$ *(for instance, by Pschenichny's linearization method), find* $x^{n+1} \in X^0$ *satisfying*

$$\Theta(x^{n+1}, Y_{n+1}) \leq \varepsilon_{n+1}.$$

Stage 4. *Set*

$$n := n + 1.$$

Go to Stage 1.

Obviously, for every $x^1 \in X^0$ the method SMETH.ACTIV.apprx generates a sequence of random vectors $\{x^n(\omega)\}$ on the probability space $(\Omega, \mathfrak{A}, P)$ (see section 5). And it is possible that for some $\omega \in \Omega$ the trajectory $\{x^n(\omega)\}$ is finite, i.e., we get $x^1, x^2(\omega), \ldots, x^{N_{stop}}(\omega)$, $N_{stop}(\omega) < +\infty$; in this case setting

$$x^n(\omega) := x^{N_{stop}}(\omega), \; n = N_{stop}, N_{stop}+1, \ldots,$$

we shall formally consider $\{x^n(\omega)\}$ as an infinite trajectory.

Applying Theorem 6.1, Lemmas 7.2 and 7.3, and (7.11) we obtain the following result.

THEOREM 7.1. *Every trajectory* $\{x_n\}$ *of the method SMETH.ACTIV.apprx P-a.s. converges to the stationary set* $\mathcal{X}^0_{stat}$ *of the problem* apprx.$\mathcal{P}^0$.

Moreover, under the assumption (7.1) $\{x^n\}$ *P-a.s. converges to the optimal set* $\mathcal{X}^0_{opt}$.

Remark 7.1. Under assumption (7.1), for every $Y \in \mathfrak{M}_f(Y^0)$ we have that apprx.$\mathcal{P}[Y]$ is a linear programming problem (see its form apprx.$\mathcal{P}[Y]'$). Thus, at Stage 3 in the method SMETH.ACTIV.apprx linear programming techniques can be effectively employed to find $x^{n+1} \in X^0$,

$$\Theta(x^{n+1}, Y_{n+1}) = 0.$$

Remark 7.2. To simplify problems apprx.$\mathcal{P}[Y_n]$, $n = 1, 2, \ldots$ (i.e., to increase practical convergence properties of SMETH.ACTIV.apprx) it turns out to be useful to cluster elements in Y_n under the condition

$$| y' - y'' | < \varepsilon \ \Rightarrow y'' := y', \ y', y'' \in Y_n.$$

8. Numerical experiments for the Chebyshev approximation problem. Let us consider the linear Chebyshev approximation problem

$$\max_{y \in Y^0} \left| F(y) - \sum_{i=1}^{k} x_i z_i(y) \right| \to \min_{x \in \Re^k},$$

where Y^0 is an l-dimensional box. For the case $\dim Y^0 = 2$, we choose $z_i(\cdot)$, $i = 1, \ldots, k$, to be $k(d)$ functions

$$x_1^{i_1} x_2^{i_2}, \ i_1 + i_2 \le d, \tag{8.1}$$

or alternatively, the $k(t)$ functions

$$x_1^{i_1} x_2^{i_2}, \ 0 \le i_1, i_2 \le t. \tag{8.2}$$

For better presentation we employ the following definitions:

N: the number of iterations = number of solved problems apprx.$\mathcal{P}[Y_n]$, $n = 1, 2, \ldots, N$;

M: the number of elements in apprx.$\mathcal{P}[Y_N]$;

C: the number of elements in $\overline{Y}_n$;

v^*: $v^* := \max_{y \in Y^*} | \Phi(x^N, y) - F(y) |$, where Y^* is an equispace grid, $Y^* \subset Y^0$;

v: $v := \max_{y \in Y_N} | \Phi(x^N, y) - F(y) |$.

We consider the following examples (see [11, 20, 21]).

Example 8.1.

Data: $F(y_1, y_2) = \log(y_1 + y_2) \sin y_1$,
$Y^0 = [0, 1] \times [1, 2.5]$,
$z_i(\cdot)$ as in (8.1).

Parameters: $\gamma = 0.1$,
$\varepsilon = 0.01$, $\kappa = 0.3$, $\eta = 0.7$,
$\varepsilon_n = \max(\varepsilon_0 (1.2)^{-n}, 10^{-3})$, $n = 1, 2, \ldots$,
$\sigma_n = \varepsilon_0 (1.2)^{-n}$, $n = 1, 2, \ldots$,
$\varepsilon_0 = 2$.

Results: See Tables 8.1 and 8.2 ($| Y^* | = 10000$).

TABLE 8.1
Example 8.2, $d = 2$.

Method	N	M	C	v	v^*
SMETH.ACTIV.apprx	10	12	224	0.012091	0.012094
Nonactivated SMETH.apprx	205	218	499	0.016666	0.014092

TABLE 8.2

Example	N	M	C	v	v^*
8.2, $d = 2$	10	12	224	0.012091	0.012094
8.2, $d = 3$	14	13	226	0.001437	0.001701
8.3, $d = 2$	25	22	74	0.177449	0.178742
8.3, $d = 3$	16	15	58	0.036201	0.036202
8.4, $d = 2$	14	30	72	0.265867	0.263526

Example 8.2.
Data: $F(y_1, y_2) = (1 + y_1)^{y_2}$,
$Y^0 = [0, 1] \times [1, 2.5]$,
$z_i(\cdot)$ as in (8.1).
Parameters: See Example 8.1.
Results: See Table 8.2 ($\mid Y^* \mid = 10000$).
Example 8.3.
Data: $F(y_1, y_2, y_3) = \cos(y_3)(1 + y_1)^{y_2}$,
$Y^0 = [0, 1] \times [1, 2.5] \times [0, 1]$,
$z_i(\cdot)$ as in (8.1).
Parameters: See Example 8.1.
Results: See Table 8.2 ($\mid Y^* \mid = 1000000$).

Now we can provide some comparisons between SMETH.ACTIV.apprx and existing outer approximations methods.

Regarding methods based on passive random search procedures for finding relevant parameters we can note that the nonactivated SMETH.apprx does not work effectively even on the simplest test problem from Example 8.1. Certainly, advantages of SMETH.ACTIV.apprx over the nonactivated SMETH.ACTIV.apprx are already clear after the considerations of section 4.

Regarding methods based on active search procedures we consider results of numerical experiments reported in [21]. Recall that Reemtsen's method involves a refined active search of relevant parameters and at the nth iteration to form a simpler linear problem apprx.$\mathcal{P}[Y_{n+1}]$ it solves the following discrete inner maximization problem:

$$\mid \Phi(x^n, y) - F(y) \mid \rightarrow \max_{y \in \tilde{Y}_n},$$

where $\tilde{Y}_n$ is an equispace grid, $\tilde{Y}_n \subset Y^0$.

To present results from [21] we introduce the following definitions:

N: the number of iterations = number of solved problems $\mathcal{P}[Y_n],\ n = 1, 2, \ldots, N$ (for a run of Reemtsen's method);

M: average number of elements in $\mathcal{P}[Y_n],\ n = 1, 2, \ldots, N$ (for a run of Reemtsen's method);

v^*: $v^* := \max_{y \in \tilde{Y}} \mid \Phi(x^N, y) - F(y) \mid$, where $\tilde{Y}$ is an equispace grid, $\tilde{Y} \subset Y^0$.

TABLE 8.3

Example	Method	N	M	C	v
8.2, $d=2$	SMETH.ACTIV.apprx	10	12	224	0.012091
	Reemtsen's method	4	21	>32761	0.0280626
8.2, $d=3$	SMETH.ACTIV.apprx	14	13	226	0.001437
	Reemtsen's method	6	22	>32761	0.0034744
8.3, $d=2$	SMETH.ACTIV.apprx	25	22	74	0.177449
	Reemtsen's method	4	24	>32761	0.1776570
8.3, $d=3$	SMETH.ACTIV.apprx	16	15	58	0.036201
	Reemtsen's method	5	23	>32761	0.0365746
8.4, $d=2$	SMETH.ACTIV.apprx	14	30	72	0.265867
	Reemtsen's method	4	24	>68921	0.1524860

In the reported results

$$\mid \tilde{Y} \mid = \begin{cases} 32761 & \text{for Example 8.2,} \\ 32761 & \text{for Example 8.3,} \\ 68921 & \text{for Example 8.4} \end{cases}$$

(see [21]).
Then we set

$$\mid C \mid := \begin{cases} \mid \overline{Y}_N \mid & \text{for a run of SMETH.ACTIV.apprx,} \\ \mid \tilde{Y}_N \mid & \text{for a run of Reemtsen's method.} \end{cases}$$

The results of the comparison are presented in Table 8.3. We can see that SMETH.ACTIV.apprx was used to solve several times more simpler linear problems than Reemtsen's method (and these problems apprx.$\mathcal{P}[Y_n]$, $n = 1, 2, \ldots, N$, are somewhat more complicated for SMETH.ACTIV.apprx), but the overall efficiency of SMETH.ACTIV.apprx seems to be better due to less computational efforts paid at each iteration (note that $\mid \overline{Y}_n \mid << \mid \tilde{Y}_n \mid$, $n = 1, 2, \ldots, N$).

9. Conclusion. We suggest that important advantages of the proposed general stochastic outer approximations method consist of the following.

First, the method has been constructed for the general problem $\mathcal{P}^0$, but it can be easily essentially adapted for any special class of optimization problems involving an infinite number of constraints. Note that, for instance, SMETH.ACTIV.glob and SMETH.ACTIV.apprx (see also Remark 7.1) are specific such versions of the general method SMETH.ACTIV. Moreover, for any special class of problems some new effective optimization techniques for solving simpler problems or inner problems can be instantly employed within SMETH.ACTIV's scheme.

In forthcoming papers we intend to present versions of the general method SMETH.ACTIV for solving semi-infinite programming problems and minimax optimization problems.

Second, the construction of SMETH.ACTIV is open for further developments. Thus, since the method SMETH.ACTIV can be considered as a developed Eaves–Zangwill method applying at each iteration the multistart scheme for the search of relevant parameters (note that if the inner problems $\mathcal{JP}(x^n)$, $n = 1, 2, \ldots$, are unimodal, SMETH.ACTIV appears to be similar to the Eaves–Zangwill method), some recent advanced techniques of the multistart method can be employed to develop SMETH.ACTIV. We shall point out some of these possible refinements:

(i) To limit the growth of the descriptions of problems $\mathcal{P}[Y_n]$, $n = 1, 2, \ldots$, we can apply clustering techniques (see, for instance, [25]) in forming relevant parameter sets Y_n, $n = 1, 2, \ldots$.

(ii) To improve the efficiency of the active search of relevant parameters (see Step 1.ACTIV of SPROC.ACTIV) we can apply local descent techniques specially developed for use within the multistart method (see, for instance, [23, 29]).

REFERENCES

[1] D.P. Bertsekas, *Constrained Optimization and Lagrange Multiplier Methods*, Academic Press, New York, London, 1982.

[2] E.W. Cheney and A.A. Goldstein, *Newton's method for convex programming and Tchebycheff approximation*, Numerical Mathematics, 1 (1959), pp. 253–268.

[3] F.H. Clarke, *Optimization and Nonsmooth Analysis*, John Wiley & Sons, New York, 1983.

[4] I.D. Coope and G.A. Watson, *A projected Lagrangian algorithm for semi-infinite programming*, Math. Programming, 32 (1985), pp. 337–356.

[5] B.C. Eaves and W.I. Zangwill, *Generalized cutting plane algorithms*, SIAM J. Control Optim., 9 (1971), pp. 529–542.

[6] V.V. Fedorov, *Numerical Methods for Maxmin Problems*, Nauka, Moscow, 1979 (in Russian).

[7] A.V. Fiacco and K.O. Kortanek, *Semi-Infinite Programming and Applications*, Lecture Notes in Econom. and Math. Systems 215, Springer-Verlag, New York, 1983.

[8] C. Gonzaga and E. Polak, *On constraint dropping schemes and optimality functions for a class of outer approximations algorithms*, SIAM J. Control Optim., 17 (1979), pp. 477–493.

[9] R. Hettich, ed., *Semi-Infinite Programming*, Lecture Notes in Control and Inform. Sci. 15, Springer-Verlag, New York, 1979.

[10] R. Hettich, *A review of numerical methods for semi-infinite optimization*, in Semi–Infinite Programming and Applications, Lecture Notes in Econom. and Math. Systems 215, Springer-Verlag, New York, 1983, pp. 158–178.

[11] R. Hettich, *An implementation of a discretization method for semi–infinite programming*, Math. Programming, 34 (1986), pp. 354–361.

[12] A.J. Heunis, *Use of a Monte-Carlo method in an algorithm which solves a set of functional inequalities*, J. Optim. Theory Appl., 45 (1984), pp. 89–99.

[13] W.W. Hogan, *Applications of a general convergence theory for outer approximations algorithms*, Math. Programming, 5 (1973), pp. 151–168.

[14] J.E. Kelley, *The cutting–plane method for solving convex programs*, J. Soc. Indust. Appl. Math., 8 (1960), pp. 703–712.

[15] E.S. Levitin and B.T. Polyak, *Constrained minimization methods*, USSR Computational Mathematics and Mathematical Physics, 6 (1966), pp. 1–50.

[16] J. Neveu, *Bases Mathematiques du Calcul des Probabilites*, Masson, Paris, 1964.

[17] N.M. Novikova, *Iterative Stochastic Methods for Solving Variational Problems of Mathematical Physics and Operations Research*, Contemporary Mathematics and Its Applications 3, Plenum, New York, London, 1993.

[18] B.N. Pschenichny, *Necessary Conditions for an Extremum*, 2nd ed., Nauka, Moscow, 1982 (in Russian).

[19] B.N. Pschenichny, *Linearization Method*, Nauka, Moscow, 1983 (in Russian).

[20] R. Reemtsen, *Modifications of the first Remez algorithm*, SIAM J. Numer. Anal., 27 (1990), pp. 507–518.

[21] R. Reemtsen, *Discretization methods for the solution of semi-infinite programming problems*, J. Optim. Theory Appl., 71 (1991), pp. 85–103.

[22] A.H.G. Rinnooy Kan and G.T. Timmer, *Stochastic methods for global optimization*, Amer. J. Math. Management Sci., 4 (1984), pp. 7–40.

[23] J.A. Snyman and L.P. Fatti, *A multistart global minimization algorithm with dynamic search trajectories*, J. Optim. Theory Appl., 54 (1987), pp. 121–142.

[24] D.M. Topkis, *Cutting plane methods without nested constraints*, Oper. Res., 18 (1970), pp. 404–413.

[25] A. Torn and A. Zilinskas, *Global Optimization*, Springer-Verlag, Berlin, 1989.

[26] Y. Wardi, *A stochastic algorithm for optimization problems with continua of inequalities*, J. Optim. Theory Appl., 56 (1988), pp. 285–311.

[27] Y. Wardi, *Stochastic approximation algorithm for minimax problems*, J. Optim. Theory Appl., 64 (1990), pp. 615–640.

[28] S.K. Zavriev, *Subgradient methods for two–stage lexicographic optimization with an infinite number of constraints*, Computational Mathematics and Modeling, 1 (1990), pp. 383–394.

[29] S.K. Zavriev, *On the global optimization properties of finite–difference local descent algorithms*, Journal of Global Optimization, 3 (1993), pp. 67–78.

SIAM J. CONTROL OPTIM.
Vol. 35, No. 4, pp. 1422–1433, July 1997

014

A REMARK ON EXISTENCE OF SOLUTIONS OF INFINITE-DIMENSIONAL NONCOMPACT OPTIMAL CONTROL PROBLEMS*

H. O. FATTORINI†

Abstract. Without compactness of the semigroup, optimal controls for semilinear infinite dimensional control problems may not exist even under the usual convexity–lower semicontinuity assumptions. This is shown with two counterexamples, and two fixes for nonexistence are discussed.

Key words. distributed parameter systems, maximum principle

AMS subject classifications. 93E20, 93E25

PII. S036301299528788X

1. Introduction. Let A be the infinitesimal generator of a strongly continuous semigroup $S(t)$ in a Banach space E. It is well known that existence theorems for the semilinear control system

$$y'(t) = Ay(t) + f(t, y(t), u(t)), \qquad y(0) = \zeta, \tag{1.1}$$

with cost functional

$$y_0(t, u) = \int_0^t f_0(\tau, y(\tau), u(\tau))d\tau \tag{1.2}$$

require assumptions of two types: (a) compactness of the semigroup $S(\cdot)$ and (b) some condition such as convexity and closedness of

$$\{(z_0, z) \in \mathbb{R} \times E;\ z_0 \geq f_0(t, y, u),\ z = f(t, y, u)\ \ (u \in U)\} \tag{1.3}$$

for every t, y, where U is the control set (see [12]). Lack of convexity or closedness of (1.3) can be "repaired" using relaxed controls [4], [5], but lack of compactness of the semigroup cannot, as we show in two examples in section 2. This motivates an extension of the definition of solution discussed in section 3, a revival of Gamkrelidze's *sliding optimal states* [7]. As shown in section 4, sliding optimal states can be interpreted using measure-valued solutions of (1.1).

2. Two counterexamples. The systems are of the form

$$y'(t) = Ay(t) + f(t, y(t)) + u(t)\,, \qquad y(0) = \zeta, \tag{2.1}$$

in a separable Hilbert space H. The control set U is the unit ball of H, and the space $C_{\rm ad}(0, \bar{t}; U)$ of admissible controls consists of all strongly measurable U-valued functions defined in $0 \leq t \leq \bar{t}$. A *solution* $y(t, u) = y(t, u(\cdot))$ of (2.1) is a function in the space $C(0, \bar{t}; H)$ of H-valued continuous functions in $0 \leq t \leq \bar{t}$ satisfying the integral equation

$$y(t) = S(t)\zeta + \int_0^t S(t - \tau)\{f(\tau, y(\tau)) + u(\tau)\}d\tau. \tag{2.2}$$

*Received by the editors June 19, 1995; accepted for publication (in revised form) May 28, 1996. This work was supported by National Science Foundation grant DMS-9221819 Amendment 007.
http://www.siam.org/journals/sicon/35-4/28788.html
†University of California, Department of Mathematics, Los Angeles, CA 90024 (hof@math.ucla.edu).

(The definition includes the requirement that $\tau \to f(\tau, y(\tau))$ belong to $L^1(0, \bar{t}; H)$.) In the two examples, $y(t, u)$ exists globally and is unique for every $u(\cdot) \in C_{ad}(0, \bar{t}; U)$.

A cost functional $y_0(t, u)$ (not necessarily of the form (1.2)) is *weakly lower semicontinuous* if, for every sequence $\{u_n(\cdot)\} \subseteq C_{\mathrm{ad}}(0, \bar{t}; U)$ such that

$$u_n(\cdot) \to \bar{u}(\cdot) \in C_{\mathrm{ad}}(0, \bar{t}; U) \quad L^1(0, \bar{t}; H) \text{ -weakly} \quad \text{in } L^\infty(0, \bar{t}; H) \tag{2.3}$$

and such that $y(t, u_n)$ exists in $0 \le t \le \bar{t}$ and

$$y(\cdot, u_n) \to y(\cdot) \quad \text{in} \quad C(0, \bar{t}; H), \tag{2.4}$$

we have

$$y_0(\bar{t}, \bar{u}) \le \limsup_{n \to \infty} y_0(\bar{t}, u_n)\,. \tag{2.5}$$

Example 2.1. Consider the linear control system

$$y'(t) = Ay(t) + u(t), \quad y(0) = 0, \tag{2.6}$$

in the space $H = L^2(0, 2\pi)$. The semigroup is $S(t)y(x) = y(x + t)$ ($y(\cdot)$ continued 2π-periodically outside of $(0, 2\pi)$). This semigroup has the infinitesimal generator $Ay(x) = y'(x)$, with domain consisting of all $y(\cdot) \in H$ absolutely continuous, with square integrable derivative and $y(0) = y(2\pi)$. The operator A is skew-adjoint, and $S(t)$ is a unitary group: $S(t)^* = S(-t)$. In particular, $\|S(t)y\| = \|y\|$ for all $y \in H$. The (fixed) control interval is $0 \le t \le \pi$, there are no state constraints or target condition, and the cost functional is

$$y_0(\pi, u) = \int_0^\pi \{(S(-t)\eta, y(t, u))^2 + (t^2 - \|y(t, u)\|^2)^2\} dt, \tag{2.7}$$

where η is a fixed element of H. Weak lower semicontinuity of this functional is obvious. (In fact, (2.5) holds with equality and lim instead of lim sup.)

We construct a minimizing sequence $\{u^n(\cdot)\}$. Let

$$u^n(t) = S(t)y_n\,, \qquad y_n(x) = \frac{1}{\sqrt{\pi}} \cos nx\,. \tag{2.8}$$

Trajectories of the system are

$$y(t, u) = \int_0^t S(t - \tau)u(\tau) d\tau \tag{2.9}$$

so that

$$y(t, u^n) = tS(t)y_n\,. \tag{2.10}$$

Since $y_n \to 0$ in H weakly, the same is true of $y(t, u^n)$ for all t. On the other hand, $\|y(t, u^n)\| = t$ so that the integrand of $y_0(\pi, u_n)$ tends to zero almost everywhere (a.e.). Thus $y_0(\pi, u^n) \to 0$ by the dominated convergence theorem. Since the functional is nonnegative, $\{u^n\}$ is a minimizing sequence. However, we show below that if

$$\eta(x) = \sum_{n=0}^{\infty} e^{-n^2} \cos nx, \tag{2.11}$$

there is no optimal control.

LEMMA 2.2. *Let H be a Hilbert space and $f(\tau)$ a strongly measurable H-valued function in a measurable set e with $\|f(\tau)\| \leq 1$. Assume that*

$$\left\| \int_e f(\tau) d\tau \right\| = |e| < \infty$$

($|\cdot|$ = Lebesgue measure). Then there exists a one-dimensional subspace H_0 of H such that $f(t) \in H_0$ a.e. in e.

Proof. If this is not true and $y \neq 0$ is arbitrary, we can write

$$f(\tau) = f_0(\tau) y + g(\tau), \tag{2.12}$$

where $g(\tau)$ belongs to the orthogonal complement of the subspace generated by y and $g(t) \neq 0$ in a subset of e of positive measure. We apply (2.12) to $y = \int_e f(\tau) d\tau$. Then $(y, f(\tau)) = (y, y) f_0(\tau)$ so that $|e|^2 = (y, y) = (y, y) \int_e f_0(\tau) d\tau$. It follows that $f_0(\tau) = 1$ a.e. in e, hence $g(\tau) = 0$ a.e. in e, absurd. This ends the proof.

Assume that there is an optimal control $\bar{u}(\cdot)$. Then we must have $\|y(t, \bar{u})\| = t$, in particular $\|y(\pi, \bar{u})\| = \pi$. Since $\|S(t - \tau)\bar{u}(\tau)\| \leq 1$, we may apply Lemma 2.2 to the integral (2.9) in $e = [0, 2\pi]$ and obtain that $S(\pi - \tau)\bar{u}(\tau) = \rho(\tau) y$, where $\|y\| = 1$ and $\rho(\cdot)$ is a scalar function with $|\rho(\tau)| = 1$ a.e.; a fortiori, $\bar{u}(\tau) = S(\tau - \pi)\rho(\tau) y$. It then follows that

$$y(t, \bar{u}) = S(t - \pi) y \int_0^t \rho(\tau) d\tau \, .$$

Since $\|y(t, \bar{u})\| = t$, we must have $\rho(\tau) \equiv 1$ or $\rho(\tau) \equiv -1$. Hence

$$y(t, \bar{u}) = t S(t) z \tag{2.13}$$

with $\|z\| = 1$. Replacing in the cost functional,

$$0 = y_0(\pi, \bar{u}) = \int_0^\pi t^2 (S(-t)\eta, S(t)z)^2 dt = \int_0^\pi t^2 (S(-2t)\eta, z)^2 dt$$

so that

$$(S(-2t)\eta, z) = 0 \qquad (0 \leq t \leq \pi) \, . \tag{2.14}$$

We have $S(-2t)(\cos nx) = \cos n(x - 2t) = \cos nx \cos 2nt + \sin nx \sin 2nt$; hence (2.14) is

$$\phi(t) = \sum_{n=0}^{\infty} e^{-n^2} \left(\cos 2nt \int_0^{2\pi} z(x) \cos nx dx - \sin 2nt \int_0^{2\pi} z(x) \sin nx dx \right) = 0$$

in $0 \leq t \leq \pi$. The function $\phi(t/2)$ is then identically zero in $0 \leq t \leq 2\pi$. It has a uniformly convergent Fourier series; thus all of its Fourier coefficients must be zero, and we have

$$\int_0^{2\pi} z(x) \cos nx \; dx = \int_0^{2\pi} z(x) \sin nx \; dx = 0$$

for $n = 1, 2, \ldots$, which implies that $z(x) = 0$ a.e., a contradiction since $\|z\| = 1$.

What fails in this example is not the functional, which is weakly lower semicontinuous; the problem is that (2.4) does not hold because of lack of compactness of

the semigroup. In fact, (2.3) is trivially satisfied. (Note also that (1.3) is convex and closed for every t, y.)

Example 2.1 raises the idea of weakening the definition of weak lower semicontinuity in such a way that *weak* rather than *strong* convergence is required in (2.4). We assume below a cost functional of the form $y_0(t, y, u)$ defined for $u(\cdot) \in C_{\rm ad}(0, T; U)$ and $y(\cdot) \in C(0, T; E)$. What makes it different from the cost functionals at play until now is that $y(\cdot)$ is not necessarily the trajectory $y(\cdot, u)$. We call $y_0(t, y, u)$ *weakly-weakly lower semicontinuous* if, for every sequence $\{u_n(\cdot)\} \subseteq C_{\rm ad}(0, \bar{t}; U)$ such that (2.3) holds, $y(t, u_n)$ exists in $0 \le t \le \bar{t}$, the $y(t, u_n)$ are uniformly bounded, and

$$(y, y(t, u_n)) \to (y, \bar{y}(t)) \quad (0 \le t \le \bar{t},\ y \in H) \tag{2.15}$$

for some $\bar{y}(\cdot) \in C(0, \bar{t}; U)$, we have the inequality corresponding to (2.5):

$$y_0(\bar{t}, \bar{y}, \bar{u}) \le \limsup_{n \to \infty} y(\bar{t}, y(u_n), u_n).$$

(The definition can be generalized to a Banach space E taking $y \in E^*$.) Under this new definition, we can show an existence result for optimal problems for a linear system

$$y'(t) = Ay(t) + Bu(t)\,, \tag{2.16}$$

with A the infinitesimal generator of a strongly continuous semigroup $S(\cdot)$ in an arbitrary Banach space E. We take $B : X^* \to E$, where X is another Banach space with X^* separable, so that $L^\infty(0, T; X^*) = L^1(0, T; X)^*$ (see Remark 4.5) and assume that $B^* : E^* \to X$ and that $U \subseteq X^*$ is such that $C_{ad}(0, T; U)$ is $L^1(0, T; X)$-weakly compact in $L^\infty(0, T; X^*)$. If $\{u^n(\cdot)\}$ is a minimizing sequence, we may always assume that $\{u^n(\cdot)\}$ is $L^1(0, \bar{t}; X)$-weakly convergent in $L^\infty(0, \bar{t}; X^*)$ so that if $y^* \in E^*$ we have

$$\langle y^*, y(t, u^n)\rangle = \langle y^*, S(t)\zeta\rangle + \int_0^{\bar{t}} \langle B^* S(t - \tau)^* y^*, u^n(\tau)\rangle d\tau\,. \tag{2.17}$$

It follows that (2.15) holds; hence we can take lim sup in the cost functional. The argument is much the same as that of Lemma 1.1 in [2] for the time-optimal problem. However, the result is not generalizable to nonlinear systems, as the following example shows.

Example 2.3. Consider the control system

$$y'(t) = Ay(t) + \phi(t^2 - \|y(t)\|^2)\xi + u(t)\,, \quad y(0) = 0, \tag{2.18}$$

with the same space H, operator A, and admissible control space $C_{ad}(0, \bar{t}; U)$ as in Example 2.1. The function $\phi(s)$ is infinitely differentiable, bounded, positive for $s \ne 0$ and $\phi(0) = 0$ (for instance, $\phi(s) = s^2/(1 + s^2)$), and ξ is an arbitrary nonzero element of H. It is well known that under these assumptions solutions exist globally and are uniformly bounded for $u(\cdot) \in C_{ad}(0, T; U)$.

We consider the optimal control problem in the fixed interval $0 \le t \le 2\pi$ with cost functional

$$y_0(2\pi, u) = \int_0^{2\pi} \{(S(-t)\eta, y(t, u))^2 + (S(t)\eta, y(t, u))^2 + (S(t)\eta, u(t))^2\} dt\,, \tag{2.19}$$

and no target condition or state constraints: $\eta \in H$ is again given by (2.11). The functional $y_0(2\pi, u)$ is weakly-weakly lower semicontinuous. That this is true for the first two terms is a consequence of the dominated convergence theorem. For the third term, we note that if $\{u_n(\cdot)\}$ is a sequence in $L^\infty(0,\bar t; H)$ such that $u_n(\cdot) \to \bar u(\cdot)$ $L^1(0,\bar t; H)$-weakly in $L^\infty(0,\bar t; H)$, then $(S(\cdot)\eta, u_n(\cdot)) \to (S(\cdot)\eta, \bar u(\cdot))$ $L^1(0,\bar t)$-weakly in $L^\infty(0,\bar t)$. A fortiori, $(S(\cdot)\eta, u_n(\cdot)) \to (S(\cdot)\eta, \bar u(\cdot))$ $L^2(0,\bar t)$-weakly in $L^2(0,\bar t)$; hence,

$$\int_0^{\bar t} (S(\tau)\eta, \bar u(\tau))^2 d\tau \le \limsup_{n\to\infty} \int_0^{\bar t} (S(\tau)\eta, u_n(\tau))^2 d\tau\,.$$

The sequence (2.8) is a minimizing sequence also for this problem, that is,

$$y_0(2\pi, u_n) \to 0 \qquad \text{as} \quad n \to \infty\,. \tag{2.20}$$

That this is true for the first two terms follows from (2.13) (we have $\|y(t,u^n)\| = t^2$ for the function defined by (2.9) so that $y(t,u^n)$ solves (2.6) as well as (2.18)), the fact that $y_n \to 0$ weakly, and the dominated convergence theorem. For the third, we note that $(S(t)\eta, u^n(t)) = (S(t)\eta, S(t)y_n) = (\eta, y_n)$. These observations are of course independent of ξ.

We show below that there are no optimal controls. In fact, assume $\bar u(\cdot)$ is optimal. Then we must have

$$(S(-t)\eta, y(t,\bar u)) = (S(t)\eta, y(t,\bar u)) = (S(t)\eta, \bar u(t)) = 0 \qquad (0 \le t \le 2\pi)\,. \tag{2.21}$$

Thus

$$\begin{aligned} 0 = (S(t)\eta, y(t,\bar u)) &= \left(S(t)\eta, \int_0^t S(t-\tau)\{\phi(\tau^2 - \|y(\tau,\bar u)\|^2)\xi + \bar u(\tau)\}d\tau\right) \\ &= \int_0^t (S(t)\eta, S(t-\tau)\xi)\phi(\tau^2 - \|y(\tau,\bar u)\|^2)d\tau + \int_0^t (S(t)\eta, S(t-\tau)\bar u(\tau))d\tau \\ &= \int_0^t (S(\tau)\eta, \xi)\phi(\tau^2 - \|y(\tau,\bar u)\|^2)d\tau + \int_0^t (S(\tau)\eta, \bar u(\tau))d\tau \\ &= \int_0^t (S(\tau)\eta, \xi)\phi(\tau^2 - \|y(\tau,\bar u)\|^2)d\tau \qquad (0 \le t \le 2\pi), \end{aligned} \tag{2.22}$$

where the second integral drops out in view of the third equality (2.21). We deduce that

$$(S(t)\eta, \xi)\phi(t^2 - \|y(t,\bar u)\|^2) = 0 \qquad (0 \le t \le 2\pi). \tag{2.23}$$

Assume that

$$(S(t)\eta, \xi) = 0 \qquad (0 \le t \le 2\pi)\,. \tag{2.24}$$

Then the argument following (2.14), this time applied to ξ, reveals that $\xi = 0$, a contradiction. On the other hand, assume that (2.24) is false. Then, since the function $(S(\cdot)\eta, \xi)$ is analytic, we must have $(S(\tau)\eta, \xi) \ne 0$ a.e.; hence (2.23) and the fact that $\phi(s) > 0$ for $s \ne 0$ imply that $\|y(t,\bar u)\| = t$ a.e. By continuity, $\|y(t,\bar u)\| = t$ in $0 \le t \le 2\pi$. It follows that for the optimal control $\bar u(\cdot)$, (2.18) reduces to the linear equation (2.6) so that the trajectory corresponding to $\bar u(\cdot)$ is given by (2.13):

$y(t,\bar{u}) = tS(t)z$ with $\|z\| = 1$. The first equality (2.21) then yields $(S(-2t)\eta, z) = 0$ $(0 \le t \le 2\pi)$; we apply once again the argument after (2.14) to conclude that $z = 0$, again a contradiction.

This time, nonexistence of optimal controls is caused by the fact that the weak limit of the sequence $\{y(\cdot, u^n)\}$ is *not* the solution $y(\cdot,\bar{u})$ of (2.18), a phenomenon typical in nonlinear equations. Note that the function $u \to (y,u)^2$ is continuous and convex so that the set (1.3) is closed and convex for every t, y.

3. Sliding trajectories. Paying heed to a celebrated dictum of Hilbert [8] (in the very free translation of [14]) such problems as those in Example 2.1 and 2.3 should have a solution if "solution" is suitably defined. One way to a suitable definition is to generalize Gamkrelidze's finite-dimensional definition of "sliding optimal states" in [7]. We do this for the system (2.1) in a reflexive, separable space E with a control set $U \subseteq E$. The admissible control space $C_{ad}(0,T;U)$ is the subspace of $L^\infty(0,T;E) = L^1(0,T;E^*)^*$ (see Remark 4.5) defined by $u(t) \in U$ a.e., and we assume that $C_{ad}(0,T;U)$ is $L^1(0,T;E^*)$-weakly compact in $L^\infty(0,T;E)$; this implies, among other things, that U must be closed and bounded. (All assumptions on the control space are satisfied, for instance, if $U =$ unit ball of E.) Finally, we assume that $f(t,y)$ is continuous in y for t fixed, strongly measurable in t for y fixed and that for every $r > 0$ there exists a constant $K(r)$ such that

$$\|f(t,y)\| \le K(r) \qquad (0 \le t \le T,\ \|y\| \le r). \tag{3.1}$$

This is enough to define solutions by (2.2), although it does not guarantee even local existence or uniqueness. In this situation, $y(t,u)$ means one of the solutions of (2.1) corresponding to $u(\cdot)$ if any exist. We assume a cost functional of the form $y_0(t,y,u)$, where $u = u(\cdot) \in C_{ad}(0,\bar{t};U)$ and $y(\cdot) \in C(0,\bar{t};E)$. The definition of weak-weak lower semicontinuity is the same in section 2.

A *sliding trajectory* of (2.1) in $0 \le t \le \bar{t}$ is any E-valued continuous function $y(t)$ such that there exists a sequence $\{u^n(\cdot)\} \subseteq C_{ad}(0,\bar{t};U)$ with $y(t,u^n)$ defined in $0 \le t \le \bar{t}$, the trajectories $\{y(\cdot,u^n)\}$ are uniformly bounded, and

$$y(t) = \lim_{n\to\infty} y(t,u^n) \tag{3.2}$$

E^*-weakly in $0 \le t \le \bar{t}$. If $y_0(\bar{t},u^n)$ approaches the minimum value m of the functional, we have a *sliding optimal trajectory*, not necessarily the trajectory corresponding to any admissible control (as is the case in Examples 2.1 and 2.3). To give interest to such a definition, one should be able (i) to define the cost functional for sliding trajectories and show these provide minimizing elements in cases where ordinary trajectories fail and (ii) to prove some version of Pontryagin's maximum principle for the elements $\{u^n(\cdot)\}$ of the minimizing sequence defining the sliding optimal trajectory. We do (i) in Theorem 3.1 below and (ii) (for a particular optimal problem) in Example 3.2. In part (b) of the result below, the optimal problem may include a target condition $y(\bar{t},u) \in Y$ or state constraints $y(t,u) \in M(t)$ $(0 \le t \le \bar{t})$ as long as Y and the $M(t)$ are E^*-weakly closed. Finally, we assume that $m > -\infty$.

THEOREM 3.1. (a) *Let $\{u^n(\cdot)\}$ be an arbitrary sequence in $C_{ad}(0,\bar{t};U)$ with $\{y(t,u^n)\}$ uniformly bounded in $0 \le t \le \bar{t}$. Then, if necessary passing to a subsequence,*

$$y(t,u^n) \to \bar{y}(t)\ E^*\text{-weakly in } 0 \le t \le \bar{t} \tag{3.3}$$

where $\bar{y}(t)$ is a sliding trajectory.

(b) *If* $\{u^n(\cdot)\} \subseteq C_{ad}(0,\bar{t};U)$ *is in addition a minimizing sequence, there exists* $\bar{u}(\cdot) \in C_{ad}(0,\bar{t};U)$ *such that* $\bar{u}(\cdot)$ *is the* $L^1(0,\bar{t};E^*)$*-weak limit of (a subsequence of)* $\{u^n(\cdot)\}$ *and*

$$y_0(\bar{t},\bar{y},\bar{u}) = m \tag{3.4}$$

so that $\bar{y}(\cdot)$ *is an optimal sliding trajectory.*

Proof. Selecting a subsequence and using (3.1) we may take for granted that $\{u^n(\cdot)\}$ is $L^1(0,\bar{t};E^*)$-weakly convergent to $\bar{u}(\cdot) \in C_{ad}(0,\bar{t};U)$ and that $\{f(\cdot,y(\cdot,u^n))\}$ is $L^1(0,\bar{t};E^*)$-weakly convergent to $\Phi(\cdot) \in L^\infty(0,\bar{t};E)$. We write (2.2) for each $u^n(t)$ and each $y(t,u^n)$, apply a functional $y^* \in E^*$ to both sides, and take limits. We deduce that $y(t,u^n) \to \bar{y}(t)$ E^*-weakly for $0 \le t \le \bar{t}$, $\bar{y}(t)$ given by

$$\bar{y}(t) = S(t)\zeta + \int_0^t S(t-\tau)\Phi(\tau)d\tau \tag{3.5}$$

and thus continuous. That (3.4) holds is obvious from the definitions.

Example 3.2. We consider the time-optimal problem for (2.1) with A the infinitesimal generator of a group $S(t)$ in a Hilbert space H. The nonlinear term is strongly measurable in t for y fixed and has a Fréchet derivative $\partial_y f(t,y) \in (H,H)$ with respect to y such that $\partial_y f(t,y)z$ is strongly measurable with respect to t for y,z fixed and continuous in y for t,z fixed. Finally, for every $r>0$ there exists constants $K(r)$, $L(r)$ such that

$$\|f(t,y)\|_H \le K(r), \quad \|\partial_y f(t,y)\|_{(H,H)} \le L(r) \quad (0 \le t \le T,\ \|y\| \le r). \tag{3.6}$$

((H,H) is the space of all linear bounded operators from H into itself equipped with the operator norm.) The target condition is $y(\bar{t},u) \in Y$, $Y \subseteq E$ closed, and there are no state constraints. We assume the existence of a minimizing sequence $\{u^n(\cdot)\}$ such that

$$\operatorname{dist}(y(t_n,u^n),Y) = \varepsilon_n \to 0 \qquad \text{as} \quad n \to \infty \tag{3.7}$$

with $t_n < \bar{t}$ = optimal time, and

$$\|y(t,u^n)\| \le C \qquad (0 \le t \le t_n)\,. \tag{3.8}$$

Using Theorem 5.2 in [6] we obtain a sequence $\{\tilde{u}^n\}$, $\tilde{u}^n \in C_{\mathrm{ad}}(0,t_n;U)$ with

$$d_n(u^n,\tilde{u}^n) = |\{t \in [0,t_n]; u^n(t) \ne \tilde{u}^n(t)\}| \le \sqrt{\varepsilon}_n \tag{3.9}$$

and sequences $\{\tilde{y}^n\} \subseteq Y, \{z_n\} \subseteq E$ such that

$$\|z_n\| = 1\,, (z_n, \xi^n - w^n) \le \sqrt{\varepsilon}_n(1 + \|w^n\|) \tag{3.10}$$

for w^n in the contingent cone $K_Y(\tilde{y}^n)$ and

$$\xi^n = \lim_{h \to 0+} \frac{1}{h}(y(\bar{t},\tilde{u}^n_{s,h,v}) - y(\bar{t}))$$

for s in a set of full measure of $[0,\bar{t}]$, where $u_{s,h,v}(t)$ is the spike perturbation $u_{s,h,v}(t) = v \in U$ $(s-h < t \le s), u_{s,h,v}(t) = u(t)$ elsewhere. Going to a subsequence we may assume that $\{z_n\}$ is weakly convergent, and we have

$$z = \text{weak} \lim_{n \to \infty} z_n \ne 0 \tag{3.11}$$

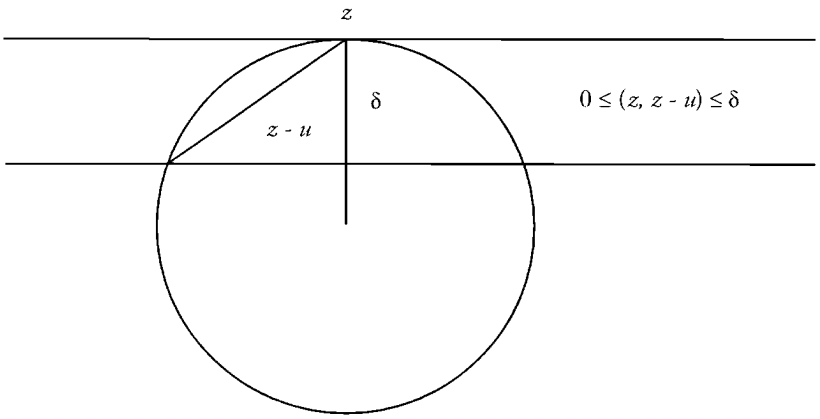


FIG. 1.

(see [3], [6]). Assumption (3.8) on global boundedness of $y(t, u^n)$, (3.9), and a simple application of Gronwall's lemma imply that the trajectories $y(t, \tilde{u}^n)$ exist globally and are uniformly bounded:

$$\|y(t, \tilde{u}^n)\| \le C \qquad (0 \le t \le t_n)\,. \tag{3.12}$$

Computing ξ^n as in [3], [6] and setting $w^n = 0$ in (3.10) we obtain

$$(S(t_n, s; \tilde{u}^n)^* z_n\,, v - \tilde{u}^n(s)) \le \sqrt{\varepsilon}_n \qquad (\|v\| \le 1) \tag{3.13}$$

a.e. in $0 \le t \le \bar{t}$, where $S(t, s; u)$ is the solution operator of the variational equation $z'(t) = \{A + \partial_y f(t, y(t, u))\}z(t)$. This inequality is exploited through the following result.

LEMMA 3.3. *Let* $\|z\| = 1$, $\|u\| \le 1$, $0 < \delta < 1$,

$$(z, v - u) \le \delta \qquad (\|v\| \le 1)\,. \tag{3.14}$$

Then

$$\|u - z\| \le \sqrt{2\delta} \tag{3.15}$$

Proof. Inequality (3.14) implies $(z, z - u) \le \delta$, while we have $(z, z - u) \ge 0$, so the proof can be read from Figure 1 and the fact that $1 - (1 - \delta)^2 + \delta^2 = 2\delta$.

We continue with Example 3.2. Since $S(t)$ is a group, the solution operator $S(t, s; u)$ is defined for $s \le t$ as well as for $s \ge t$. Moreover (3.12) implies that

$$\|S(t, s; \tilde{u}^n)\| \le C \qquad (0 \le s, t \le t_n)$$

and $S(t, s; \tilde{u}^n)^{-1} = S(s, t; \tilde{u}^n)$, thus the inverses are uniformly bounded as well. The same statements apply to the adjoint operator $S(t, s; \tilde{u}^n)^*$, so $\|S(t, s; \tilde{u}^n)^* z\| \ge c\|z\|$ $(0 \le s, t \le t_n)$ with $c > 0$. It follows from (3.10) that $\|z_n\|$ is bounded away from zero, hence

$$\|S(t_n, s; \tilde{u}^n)^* z_n\| \ge \rho > 0 \tag{3.16}$$

for n large enough.

We deduce from (3.13) divided by $\|S(t_n, s; \tilde{u}^n)^* z_n\|$, from (3.16), and from Lemma 3.3 that if we define

$$\bar{u}^n(s) = \frac{S(t_n, s; \tilde{u}^n)^* z_n}{\|S(t_n, s; \tilde{u}^n)^* z_n\|} \qquad (0 \le t \le t_n), \tag{3.17}$$

then

$$\|\tilde{u}^n(t) - \bar{u}^n(t)\| \to 0 \tag{3.18}$$

uniformly in $0 \le t \le t_n$ so that $\bar{u}^n(\cdot)$ is a minimizing sequence as well due to continuous dependence of the solution of (2.1) on the control. It finally results that although the time-optimal control problem for (2.1) may not have a solution, there exists a sequence $\{z_n\}$ with $\|z_n\| = 1$ and weak $\lim z_n \neq 0$ and such that (3.17) is a minimizing sequence; moreover, combining (3.9) with (3.18) we obtain

$$\|u^n - \bar{u}^n\|_{L^p(0,t_n)} \to 0 \tag{3.19}$$

for $p < \infty$. This may be considered "almost as good" as the maximum principle. Note, however, that the existence of a minimizing sequence with $t_n <$ optimal time is a serious restriction. Note also that it does not seem possible to extend (3.17) to $\bar{u}(s)$, the weak limit of the $u^n(s)$ in Theorem 3.1. In fact, to do this, one would have to take limits in (3.13). Weak convergence of $\{\tilde{u}^n(\cdot)\}$ could be exploited integrating against L^1 functions, but convergence of $S(t_n, s; \tilde{u}^n)^* z_n$ seems dubious in any sense.

Results for cost functionals other than time can be obtained by using the corresponding "approximate maximum principles" in [3], [6].

Applied to the systems in Examples 2.1 and 2.3 and to the minimizing sequence (2.8), the optimal sliding trajectory turns out to be $y(t) \equiv 0$ (the sequence (2.10) converges weakly to zero for all t).

4. Sliding trajectories as relaxed trajectories. In an equation like (2.1), measure-valued controls are not necessary, since the control appears linearly; what must be relaxed are the trajectories. This will be done averaging with time-dependent probability measures in accordance with the general idea of Young measures [13].

Let X be an arbitrary Banach space. $L^\infty_w(0, T; X^*)$ is the space of all X-weakly measurable X^*-valued functions $g(t)$ endowed with the norm $\|g\| =$ least c with $\langle y, g(t)\rangle \le c\|y\|$ a.e. in $0 \le t \le T$ for $y \in X$ ("a.e." depends on y). We have

$$L^1(0, T; X)^* = L^\infty_w(0, T; X^*) \tag{4.1}$$

with duality $\langle f(\cdot), g(\cdot)\rangle = \int \langle f(t), g(t)\rangle dt$. For more on the space $L^\infty_w(0, T; X^*)$ and on more general spaces see [9, p. 78]; the proof of (4.1) and more details on the duality are in [9, p. 94]. See also Remark 4.5 below, and [4], [5] for other control applications of $L^\infty_w(0, T; X^*)$.

Let B_r be the ball of center 0 and radius r in a Banach space E, and consider the space $BC(B_r)$ of bounded continuous functions in B_r equipped with the supremum norm and its dual $\Sigma_{rba}(B_r, \Phi_c)$ of all bounded, finitely additive regular measures $\mu = \mu(d\xi)$ defined in the field Φ_c generated by the closed subsets of B_r; $\Sigma_{rba}(B_r, \Phi_c)$ is furnished with the total variation norm. The duality map is $\langle f, \mu\rangle = \int_{B_r} f(\xi)\mu(d\xi)$. Applied to these spaces, (4.1) gives

$$L^1(0, T; BC(B_r))^* = L^\infty_w(0, T; \Sigma_{rba}(B_r, \Phi_c))\,. \tag{4.2}$$

Finally, let $\Pi_{rba}(B_r, \Phi_c)$ be the set of all probability measures η in $\Sigma_{rba}(B_r, \Phi_c)$ (that is, the set of measures satisfying $\eta \geq 0,\ \eta(B_r) = 1$), and let $L^\infty_w(0,T;\Pi_{rba}(B_r,\Phi_c))$ be the subspace of $L^\infty_w(0,T;\Sigma_{rba}(B_r,\Phi_c))$ consisting of all $\eta(\cdot)$ such that there exists an element $\mu(\cdot)$ in the equivalence class of $\eta(\cdot)$ in $L^\infty_w(0,T;\Sigma_{rba}(B_r,\Phi_c))$ with $\mu(t) \in \Pi_{rba}(B_r,\Phi_c)$ a.e. (For the technicalities associated with this definition see [9], [4], [5].)

We install the relaxed trajectories in (2.1) assuming that $t \to \langle y^*, f(t,\cdot)\rangle$ is a strongly measurable $BC(B_r)$-valued function for every $y^* \in E^*$ and satisfies (3.1). Below, $\eta(t, d\xi, u) = \eta(t, d\xi, u(\cdot))$ denotes an element of the space $\Pi_{rba}(B_r,\Phi_c)$ depending on $t \in [0,T]$ and $u = u(\cdot) \in C_{ad}(0,T;U)$.

Let $u = u(\cdot) \in C_{ad}(0,\bar t;U)$. Call $\eta(\cdot,u) = \eta(\cdot, d\xi, u) \in L^\infty_w(0,\bar t;\Pi_{rba}(B_r,\Phi_c))$ a *measure solution* of (2.1) in $0 \leq t \leq \bar t$ if

$$\begin{aligned}\int_{B_r} \langle y^*, \xi\rangle \eta(t, d\xi, u) &= \langle y^*, S(t)\zeta\rangle \\ &+ \int_0^t \int_{B_r} \langle S(t-\tau)^* y^*,\ f(\tau,\xi)\rangle \eta(\tau, d\xi, u) d\tau + \int_0^t \langle y^*, u(\tau)\rangle d\tau\end{aligned} \tag{4.3}$$

for all $y^* \in E^*$, $0 \leq t \leq \bar t$; for this definition to make sense, we need only the function $\tau \to \langle S(t-\tau)^* y^*, f(\tau,\xi)\rangle$ to belong to $L^1(0,t;BC(B_r))$, which follows from the hypotheses, strong continuity of the adjoint semigroup, and a simple approximation argument. The notation $\eta(t, d\xi, u)$ does not imply that, given $u(\cdot)$, there exists a unique measure solution satisfying (4.3); it indicates only the association of η and $u = u(\cdot)$ in the integral equation (4.3). The *relaxed trajectory* $y(t,\eta(u))$ corresponding to the measure solution $\eta(u) = \eta(\cdot\,, u) = \eta(\cdot, d\xi, u)$ is the E-valued function $y(t,\eta(u))$ defined by

$$\langle y^*, y(t,\eta(u))\rangle = \int_{B_r} \langle y^*, \xi\rangle \eta(t, d\xi, u) \tag{4.4}$$

for $y^* \in E^*$, $0 \leq t \leq \bar t$. Since $t \to \langle y^*, \cdot\rangle$ trivially belongs to $L^1(0,T;BC(B_r))$, $y(t,\eta(u))$ is E^*-weakly measurable and thus strongly measurable. (In fact, relaxed trajectories are continuous; see Remark 4.3.) We also have $|\langle y^*, y(t,\eta(u))\rangle| \leq r\|y^*\|$ since η is a probability measure, so that

$$\|y(t,\eta(u))\| \leq r \qquad (0 \leq t \leq \bar t)\,. \tag{4.5}$$

Usual solutions $y(t,u)$ of (2.1) with $\|y(t,u)\| \leq r$ correspond to measure solutions

$$\eta(t, d\xi, u) = \delta_{y(t,u)}(d\xi)\,. \tag{4.6}$$

The assumptions on the cost functional and the target set Y are the same in section 3; in particular, $m > -\infty$. We limit ourselves below to show that sliding trajectories are relaxed trajectories.

THEOREM 4.1. *Let $\bar y(t) = \lim y(t,u^n)$ be one of the sliding trajectories in Theorem 3.1. Then there exists a control $\bar u(\cdot)$ and a measure-valued solution $\eta(\cdot,\bar u)$ such that*

$$\bar y(t) = y(t,\eta(\bar u)).$$

Proof. Let $\eta(t, d\xi, u^n) = \delta_{y(t,u^n)}(d\xi)$. Using Alaoglu's theorem, select a (generalized) subsequence $\{u^k(\cdot), \eta(\cdot, u^k)\}$ such that

$$u^k(\cdot) \to \bar u(\cdot) \qquad L^1(0,\bar t;E^*)\text{-weakly} \quad \text{in} \quad L^\infty(0,T;E)\,, \tag{4.7}$$

$$\eta(\cdot,u^k) \to \bar\eta(\cdot) \quad L^1(0,T;BC(B_r))\text{-weakly} \quad \text{in} \quad L^\infty_w(0,T;\Sigma_{rba}(B_r,\Phi_c))\,, \tag{4.8}$$

where $\bar{\eta} \in L^\infty_w(0,T;\Pi_{rba}(B_r,\Phi_c))$; see [4]. Write (4.3) for $\eta(\cdot,u^k), u^k(\cdot)$. Convergence of the first (resp., the second) integral in (4.3) follows from (4.8) (resp., (4.7)), and we deduce that $\bar{u}(\cdot)$ and $\bar{\eta}(\cdot)$ satisfy (4.3), so $\eta(\cdot) = \eta(\cdot,\bar{u})$. This ends the proof.

Remark 4.2. The functional equation (4.3) is in some sense simpler than the original equation (2.1) (for instance, it is linear in η when $u = 0$), and the definition of solution is weaker; thus it would seem to be interesting (independently of control theory) to put conditions on $f(t,y)$ that would guarantee existence of measure-valued solutions, although perhaps not of ordinary solutions. We don't know of any such results; in fact, under the usual assumptions in control theory (those of Example 3.2) usual solutions exist and are unique (locally); measure-valued solutions only make their appearance to provide missing minima in certain problems lacking compactness (such as those in Examples 3.1 and 3.2).

Remark 4.3. Let $\varphi(\xi)$ be an E-valued function defined in B_r and such that $\langle y^*, \varphi(\xi)\rangle \in BC(B_r)$, and let $\eta(d\xi) \in \Pi_{rba}(B_r,\Phi_c)$. Define $\Phi = \int_{B_r} \varphi(\xi)\eta(d\xi) \in E$ as the only element of E satisfying

$$\langle y^*, \Phi\rangle = \int_{B_r} \langle y^*, \varphi(\xi)\rangle \eta(d\xi). \tag{4.9}$$

Then

$$\Phi = \int_{B_r} \varphi(\xi)\eta(d\xi) \in \overline{\text{conv}}(\varphi(B_r)), \tag{4.10}$$

where $\overline{\text{conv}}$ = closed convex hull (see [5]). Let

$$\Phi(\tau,u) = \int_{B_r} f(\tau,\xi)\eta(\tau,d\xi,u) \tag{4.11}$$

(the integral defined as in (4.9)). If $y^* \in E^*$, then $\tau \to \langle y^*, f(\tau,\cdot)\rangle$ belongs to $L^1(0,\bar{t};BC(B_r))$, so that $\Phi(t,u)$ is E^*-weakly measurable, hence (under the present assumptions on E) strongly measurable; moreover, $\|\Phi(\tau,u)\| \le K(r)$. It follows that y^* can be "simplified from" the integral equation (4.3) defining measure solutions; in other words, the integral equation can be written

$$y(t,\eta(u)) = S(t)\zeta + \int_0^t \{S(t-\tau)\Phi(\tau,u) + u(\tau)\}d\tau \tag{4.12}$$

and, in view of (4.10),

$$\Phi(\tau,u) \in \overline{\text{conv}}(f(\tau,B_r)) \tag{4.13}$$

a.e. in $0 \le t \le \bar{t}$ so that the role of the measure $\eta(\tau,d\xi,u)$ in the integral equation (4.3) is that of providing a (time-dependent) average of the values of $f(\tau,y)$. Not also that (4.12) shows that relaxed trajectories are continuous.

Remark 4.4. It is apparently unknown whether the relaxed trajectory corresponding to a measure-valued solution is a sliding trajectory, that is, whether it can be approximated weakly for all t by an uniformly bounded sequence of ordinary trajectories. A result of this type would be a "trajectory analogue" of the *relaxation theorems* (see [11] for finite-dimensional systems, [5] for infinite-dimensional generalizations) stating that trajectories driven by relaxed controls can be approximated by ordinary trajectories.

Remark 4.5. The space $L^\infty_w(0,T;X^*)$ has rather nonstandard properties when X is not separable; its elements $g(\cdot)$ may not be strongly measurable, even the norm function $t \to \|g(\cdot)\|$ may not be measurable, there are equivalent elements that do not coincide a.e., and the norm of $L^\infty_w(0,T;X^*)$ is not the same as the essential supremum norm. Even when X is separable, we may have $L^\infty_w(0,T;X^*) \neq L^\infty(0,T;X^*)$ (this happens for $X = C(U)$, the space of all continuous functions on a compact metric space U; the spaces $L^\infty_w(0,T;C(U)^*)$ are the basic spaces of relaxed controls in [10]). When X is separable, however, the norm function $t \to \|g(\cdot)\|$ is measurable and the essential supremum norm coincides with the norm of $L^\infty_w(0,T;X^*)$. We have $L^\infty_w(0,T;X^*) = L^\infty(0,T;X^*)$ when and only when X^* has the Radon−Nikodým property [1, Theorem 1, p. 98]; this happens for instance if X^* is separable [1, Theorem 1, p. 79] or if X is reflexive [1, Corollary 13, p. 76]. None of these "simplifications" applies here, since the space $\Pi_{rba}(B_r,\Phi_c)$ is not separable or reflexive, or even possesses the Radon−Nikodým property. Generally speaking, a space of finitely additive measures like $L^\infty_w(0,T;\Pi_{rba}(B_r,\Phi_c))$ is uncomfortably large and contains weird elements. For particular partial differential equations (rather than abstract differential equations like (2.1)), spaces of countably additive measures parameterized not only by time but by the space variables may be more suitable.

Acknowledgments. The author is most grateful to two anonymous referees for their constructive criticism, which resulted in substantial improvements to this paper.

REFERENCES

[1] J. Diestel and J. J. Uhl, *Vector Measures*, American Mathematical Society, Providence, RI, 1977.
[2] H. O. Fattorini, *Time-optimal control of solutions of operational differential equations*, SIAM J. Control, 2 (1964), pp. 54–59.
[3] H. O. Fattorini, *A unified theory of necessary conditions for nonlinear nonconvex control systems*, Appl. Math. Optim., 15 (1987), pp. 141–185.
[4] H. O. Fattorini, *Existence theory and the maximum principle for relaxed infinite dimensional optimal control problems*, SIAM J. Control Optim., 32 (1994), pp. 311–331.
[5] H. O. Fattorini, *Relaxation theorems, differential inclusions and Filippov's theorem for relaxed controls in semilinear infinite dimensional systems*, J. Differential Equations, 112 (1994), pp. 131–153.
[6] H. O. Fattorini and H. Frankowska, *Necessary conditions for infinite dimensional control problems*, Math. Control Signals Systems, 4 (1991), pp. 41–67.
[7] R. V. Gamkrelidze, *On sliding optimal states*, Dokl. Akad. Nauk SSSR, 143 (1962), pp. 1242–1245.
[8] D. Hilbert, *Über das Dirichletsche Prinzip*, Math. Ann., 59 (1904), pp. 161–186.
[9] A. Ionescu Tulcea and C. Ionescu Tulcea, *Topics in the Theory of Lifting*, Springer-Verlag, Berlin, 1969.
[10] J. Warga, *Relaxed variational problems*, J. Math. Anal. Appl., 4 (1962), pp. 129–145.
[11] J. Warga, *Optimal Control of Differential and Functional Equations*, Academic Press, New York, 1972.
[12] J. Yong, *Existence theory of optimal controls for distributed parameter systems*, Kodai Math. J., 15 (1992), pp. 193–220.
[13] L. C. Young, *Generalized curves and the existence of an attained minimum in the calculus of variations*, C. R. Sci. Lettres Varsovie C III, 30 (1937), pp. 212–234.
[14] L. C. Young, *Lectures in the Calculus of Variations and Optimal Control Theory*, W. B. Saunders, Philadelphia, 1969.

SIAM J. CONTROL OPTIM.
Vol. 35, No. 4, pp. 1434–1444, July 1997

015

SEQUENTIAL CONVEX SUBDIFFERENTIAL CALCULUS AND SEQUENTIAL LAGRANGE MULTIPLIERS*

LIONEL THIBAULT†

Abstract. The aim of this paper is to establish formulas for the subdifferentials of the sum and the composition of convex functions in terms of the subdifferentials of the data functions at nearby points. Applications to general optimization problems lead to a new notion of sequential Lagrange multipliers.

Key words. ε-subdifferential, sequential subdifferential calculus, sequential Lagrange multipliers

AMS subject classifications. 90C25, 52A41, 26E15

PII. S0363012995287714

Introduction. This paper has been motivated by the two important recent contributions by Hiriart-Urruty and Phelps [10] and Attouch, Baillon, and Théra [1] to the theory of subdifferential calculus of convex functions. In [10] a general formula has been established without any qualification condition for the sum of two proper convex lower semicontinuous functions over a locally convex topological vector space X. This formula is in terms of ε-approximate subdifferentials of the functions at the fixed point. As a consequence of the study undertaken about the variational sum of maximal monotone operators another formula has been obtained in [1] in terms of exact subdifferentials at nearby points but for Hilbert spaces. This latter formula also does not require any qualification condition.

The aim of this work is twofold. First we will show how the formula by Hiriart-Urruty and Phelps can be used to get a formula in terms of exact subdifferentials at nearby points. Second, with this formula at hand, we will consider a new notion of Lagrange multipliers and we will establish existence of multipliers in this sense for general constrained convex optimization problems where no constraint qualification is assumed. These generalized multipliers appear as sequences of multipliers at nearby points and the optimality conditions that they provide are both necessary and sufficient whenever all the data of the optimization problem are convex. Note that Hanson [9] has recently showed that the existence of a similar sequence of multipliers is a sufficient condition for a point to be optimal for minimization problems with differentiable *invex* functions defined on finite-dimensional spaces. However, he did not prove at all that this condition is necessary; that is, he did not show that such a sequence of multipliers does exist at any optimal solution which does not satisfy any constraint qualification.

In section 1 we will recall the formula by Hiriart-Urruty and Phelps, and we will establish some preliminary results. Section 2 is devoted to proving our sequential formula relative to the sum of convex functions. We will also show how the already known general formulas under general qualification conditions can be derived. The composition with a vector-valued convex mapping is considered in section 3. The formula obtained for such a composition allows us to derive in section 4 the existence

*Received by the editors June 14, 1995; accepted for publication (in revised form) May 28, 1996. http://www.siam.org/journals/sicon/35-4/28771.html

†Laboratoire d'Analyse Convexe, Case Courrier 051, Université Montpellier II, 34095 Montpellier, France (thibault@math.univ-montp2.fr).

of sequential Lagrange multipliers for any general constrained convex optimization problem without any constraint qualification. We have chosen in this article to avoid the use of nets and subnets and hence all the results will be established for reflexive Banach spaces. The more general spaces will be treated in other papers.

1. Preliminary results. In this section we are going to recall the main result of Hiriart-Urruty and Phelps [10], and we will give a new version of the Brøndsted–Rockafellar theorem which will be used in section 2.

Before stating the theorem by Hiriart-Urruty and Phelps let us recall that, for any convex function f from a topological vector space X into $\mathbb{R}\cup\{+\infty\}$ and for any real number $\varepsilon \geq 0$, the ε-subdifferential of f at any point $x \in \operatorname{dom} f := \{u \in X : f(u) < \infty\}$ is defined by

$$\partial_\varepsilon f(x) = \{x^* \in X^* : \langle x^*, u - x\rangle \leq f(u) - f(x) + \varepsilon \quad \text{for all } u \in X\}.$$

If $\varepsilon = 0$ one writes $\partial f(x)$, which is then called the subdifferential of f at x. When $\operatorname{dom} f \neq \emptyset$ one says that f is proper and for $x \notin \operatorname{dom} f$ one puts $\partial_\varepsilon f(x) = \emptyset$.

THEOREM 1.1 (see Hiriart-Urruty and Phelps [10]). *Let X be a locally convex vector space and $f_1, f_2 : X \longrightarrow \mathbb{R}\cup\{+\infty\}$ be two proper lower semicontinuous convex functions. Then for any $\overline{x} \in \operatorname{dom} f_1 \cap \operatorname{dom} f_2$, one has*

$$\partial(f_1 + f_2)(\overline{x}) = \bigcap_{\varepsilon>0} \operatorname{cl}_{w^*}(\partial_\varepsilon f_1(\overline{x}) + \partial_\varepsilon f_2(\overline{x})),$$

where cl_{w^*} denotes the closure with respect to the weak-star topology of X^*.

The version in Theorem 1.3 of the Brøndsted–Rockafellar theorem (which we have not found in the form given below in the literature, although the main idea of the proof was already used in Borwein [4]) will be needed in the next section. Note that it could be proved by the result of the version of Brøndsted–Rockafellar theorem by Borwein but the proof below is simple. It will be one of the keys of the proof of Theorem 2.1. Recall first the following form of the Ekeland variational principle.

THEOREM 1.2 (see Ekeland [8]). *Let (X, d) be a complete metric space and $f : X \longrightarrow \mathbb{R}\cup\{+\infty\}$ be a lower semicontinuous function. Let $\varepsilon > 0$ and $\overline{x} \in X$ such that*

$$f(\overline{x}) \leq \inf_X f + \varepsilon.$$

Then for any $\lambda > 0$ there exists $z \in X$ such that

$$d(z, \overline{x}) \leq \lambda, \qquad |f(z) - f(\overline{x})| \leq \varepsilon,$$

$f(z) < f(x) + \lambda^{-1}\varepsilon\, d(x, z)$ for all $x \neq z$ in X.

In the proof below and in what follows we will denote by $\mathbb{B}_X$ the closed unit ball centered at zero of a Banach space X.

THEOREM 1.3 (A version of the Brøndsted–Rockafellar theorem). *Let X be a Banach space and $f : X \longrightarrow \mathbb{R}\cup\{+\infty\}$ be a proper lower semicontinuous convex function. Then for any real number $\varepsilon > 0$ and any $x^* \in \partial_\varepsilon f(\overline{x})$ there exists $(x_\varepsilon, x^*_\varepsilon) \in X \times X^*$ such that*

$$\|x_\varepsilon - \overline{x}\| \leq \sqrt{\varepsilon},\ \|x^*_\varepsilon - x^*\| \leq \sqrt{\varepsilon}, \qquad |f(x_\varepsilon) - \langle x^*_\varepsilon, x_\varepsilon - \overline{x}\rangle - f(\overline{x})| \leq 2\varepsilon,$$

and

$$x^*_\varepsilon \in \partial f(x_\varepsilon).$$

Proof. By definition we have for all $x \in X$

$$\langle x^*, x - \overline{x}\rangle \leq f(x) - f(\overline{x}) + \varepsilon$$

and hence

$$f(\overline{x}) - \langle x^*, \overline{x}\rangle \leq f(x) - \langle x^*, x\rangle + \varepsilon.$$

By the Ekeland variational principle (see Theorem 1.2) applied to the function $f - \langle x^*, .\rangle$ (with $\lambda = \sqrt{\varepsilon}$) there exists an element $x_\varepsilon \in \overline{x} + \sqrt{\varepsilon}\ \mathbb{B}_X$ such that

$$|f(x_\varepsilon) - \langle x^*, x_\varepsilon\rangle - f(\overline{x}) + \langle x^*, \overline{x}\rangle| \leq \varepsilon \tag{1.1}$$

and

$$f(x_\varepsilon) - \langle x^*, x_\varepsilon\rangle \leq f(x) - \langle x^*, x\rangle + \sqrt{\varepsilon}\ \ \|x - x_\varepsilon\|$$

for all $x \in X$. Then

$$x^* \in \partial(f + \sqrt{\varepsilon}\ \|. - x_\varepsilon\|)(x_\varepsilon) = \partial f(x_\varepsilon) + \sqrt{\varepsilon}\ \mathbb{B}_{X^*}$$

and hence there exists $x^*_\varepsilon \in \partial f(x_\varepsilon)$ satisfying $\|x^*_\varepsilon - x^*\| \leq \sqrt{\varepsilon}$. It then follows from (1.1) that

$$|f(x_\varepsilon) - \langle x^*_\varepsilon, x_\varepsilon - \overline{x}\rangle - f(\overline{x})| \leq \varepsilon + |\langle x^*_\varepsilon - x^*, x_\varepsilon - \overline{x}\rangle| \leq 2\varepsilon$$

which completes the proof. □

Remark. The usual versions of the Brøndsted–Rockafellar theorem are given without the third inequality of the above theorem. As we will see later, this inequality will be used in force in the proof of Theorem 2.1.

2. Sequential calculus for sums of convex functions. If f_1 and f_2 are two convex functions from a Banach space X into $\mathbb{R} \cup \{+\infty\}$ and if $\overline{x} \in \operatorname{dom} f_1 \cap \operatorname{dom} f_2$, we will denote by $\|\quad\| - \limsup_{u_i \overset{f_i}{\longrightarrow} \overline{x}} [\partial f_1(u_1) + \partial f_2(u_2)]$ the set all limits $\|\quad\| - \lim_{n\longrightarrow\infty}(x^*_{1,n} + x^*_{2,n})$ for which there exists $x_{i,n} \xrightarrow[n\to\infty]{\|\ \|} \overline{x}$ such that $x^*_{i,n} \in \partial f_i(x_{i,n})$ and $f_i(x_{i,n}) - \langle x^*_{i,n}, x_{i,n} - \overline{x}\rangle \underset{n\longrightarrow\infty}{\longrightarrow} f_i(\overline{x})$.

In order to avoid the use of nets we will restrict ourselves to reflexive Banach spaces. The general case will considered in another paper.

THEOREM 2.1. *Let X be a reflexive Banach space and let $f_1, f_2 : X \longrightarrow \mathbb{R}\cup\{+\infty\}$ be two proper lower semicontinuous convex functions. Then for any $\overline{x} \in \operatorname{dom} f_1 \cap \operatorname{dom} f_2$*

$$\begin{aligned}
\partial(f_1 + f_2)(\overline{x}) &= \|\quad\| - \limsup_{u_i \overset{f_i}{\longrightarrow} \overline{x}} [\partial f_1(u_1) + \partial f_2(u_2)] \\
&= w^* - \limsup_{u_i \overset{f_i}{\longrightarrow} \overline{x}} [\partial f_1(u_1) + \partial f_2(u_2)] \\
&= \big\{\|\quad\| - \lim_n (x^*_{1,n} + x^*_{2,n}) : x^*_{i,n} \in \partial f_i(x_{i,n}), x_{i,n} \xrightarrow{\|\ \|} \overline{x}, \\
&\qquad \liminf_n \gamma(x_{1,n}, x_{2,n}, x^*_{1,n}, x^*_{2,n}) \leq 0\big\} \\
&= \big\{w^* - \lim_n (x^*_{1,n} + x^*_{2,n}) : x^*_{i,n} \in \partial f_i(x_{i,n}), x_{i,n} \xrightarrow{\|\ \|} \overline{x}, \\
&\qquad \liminf_n \gamma(x_{1,n}, x_{2,n}, x^*_{1,n}, x^*_{2,n}) \leq 0\big\},
\end{aligned}$$

where $\gamma(x, y, x^*, y^*) := f_1(\overline{x}) + f_2(\overline{x}) - f_1(x) - f_2(y) + \langle x^*, x - \overline{x}\rangle + \langle y^*, y - \overline{x}\rangle$.

Proof. (1) Let us prove that the first member is included in the second one. Fix any $x^* \in \partial(f_1 + f_2)(\overline{x})$. Since X is reflexive it follows from the theorem by

Hiriart-Urruty and Phelps (see section 1) that (for $\mathbb{N}^* := \mathbb{N} \setminus \{0\}$)

$$x^* \in \bigcap_{n\in\mathbb{N}^*} \operatorname{cl}_{w^*}\left[\partial_{1/n} f_1(\overline{x}) + \partial_{1/n} f_2(\overline{x})\right] = \bigcap_{n\in\mathbb{N}^*} \operatorname{cl}_{\|\ \|}\left[\partial_{1/n}\, f_1(\overline{x}) + \partial_{1/n}\, f_2(\overline{x})\right].$$

Therefore for each $n \in \mathbb{N}^*$ we have

$$x^* \in \partial_{1,n} f_1(\overline{x}) + \partial_{1/n} f_2(\overline{x}) + \frac{1}{n}\ \mathbb{B}_{X^*}$$

and hence there exist $u^*_{i,n} \in \partial_{1/n}\, f_i(\overline{x})$ and $b^*_n \in \mathbb{B}_{X^*}$ such that

$$x^* = u^*_{1,n} + u^*_{2,n} + \frac{1}{n}\ b^*_n. \tag{2.1}$$

By the Brøndsted–Rockafellar theorem in section 1 there exists $(x_{i,n}, x^*_{i,n}) \in \operatorname{graph} \partial f$ such that

$$\|x_{i,n} - \overline{x}\| \le \frac{1}{\sqrt{n}}, \qquad \|x^*_{i,n} - u^*_{i,n}\| \le \frac{1}{\sqrt{n}},$$

and

$$|f_i(x_{i,n}) - \langle x^*_{i,n}, x_{i,n} - \overline{x}\rangle - f_i(\overline{x})| \le \frac{2}{n}.$$

We may choose $b^*_{i,n} \in \mathbb{B}_{X^*}$ satisfying $u^*_{i,n} = x^*_{i,n} + \frac{1}{\sqrt{n}} b^*_{i,n}$.

It then follows from (2.1) that

$$x^* = x^*_{1,n} + x^*_{2,n} + \frac{1}{\sqrt{n}}\left(\frac{1}{\sqrt{n}} b^*_n + b^*_{1,n} + b^*_{2,n}\right)$$

and hence $x^* = \|\ \| - \lim_n (x^*_{1,n} + x^*_{2,n})$, which completes the proof of the first inclusion.

(2) Obviously the second member is included in the third and fourth ones and the third and fourth ones are included in the fifth one. So it remains to prove that the fifth member is included in the first one. Fix any x^* in the fifth member. There exist $x_{i,n} \xrightarrow[n\to\infty]{\|\ \|} \overline{x}$, $x^*_{i,n} \in \partial\, f_i(x_{i,n})$ such that $x^* = w^* - \lim(x^*_{1,n} + x^*_{2,n})$ and $\liminf \gamma_n \le 0$, where

$$\gamma_n := f_1(\overline{x}) + f_2(\overline{x}) - f_1(x_{1,n}) - f_2(x_{2,n}) + \langle x^*_{1,n}, x_{1,n} - \overline{x}\rangle + \langle x^*_{2,n}, x_{2,n} - \overline{x}\rangle.$$

Fix any $x \in X$. Then

$$\begin{aligned}\langle x^*_{i,n}, x - \overline{x}\rangle &= \langle x^*_{i,n}, x_{i,n} - \overline{x}\rangle + \langle x^*_{i,n} x - x_{i,n}\rangle \\ &\le \langle x^*_{i,n}, x_{i,n} - \overline{x}\rangle + f_i(x) - f_i(x_{i,n})\end{aligned}$$

and hence

$$\langle x^*_{1,n} + x^*_{2,n}, x - \overline{x}\rangle \le f_1(x) + f_2(x) - f_1(\overline{x}) - f_2(\overline{x}) + \gamma_n.$$

Taking the limit inferior of both members we get

$$\langle x^*, x - \overline{x}\rangle \le (f_1 + f_2)(x) - (f_1 + f_2)(\overline{x})$$

and hence $x^* \in \partial(f_1 + f_2)(\overline{x})$. So the proof of the theorem is complete. □

A similar result has been proved by Attouch, Baillon, and Théra (see Theorem 7.3 in [1]), where it is assumed that X is a Hilbert space. Their method is completely

different and depends heavily on the use of the theory of maximal monotone set-valued operators and of the Moreau–Yosida approximations of convex functions. In a subsequent paper we will show how one can obtain a formula where instead of

$$f_i(x_{i,n}) - \langle x^*_{i,n}, x_{i,n} - \overline{x}\rangle \underset{n\to\infty}{\longrightarrow} f_i(\overline{x})$$

one requires the stronger conditions

$$f_i(x_{i,n}) \underset{n\to\infty}{\longrightarrow} f_i(\overline{x}) \quad \text{and} \quad \langle x^*_{i,n}, x_{i,n} - \overline{x}\rangle \underset{n\to\infty}{\longrightarrow} 0. \tag{2.2}$$

With this we will extend the formula as given by Attouch, Baillon, and Théra [1] to reflexive Banach spaces. We also have to mention that this will inspire us to derive from Borwein and Ioffe [5], for example, similar formulas in equality form for nonconvex functions under some regularity assumptions.

Now we are going to show how the famous sum formula under the general Robinson qualification condition can be deduced from Theorem 2.1. Here we will restrict ourselves to the reflexive case. The general nonreflexive case will be treated elsewhere. The method below may also be applied to derive directly from Hiriart-Urruty and Phelps's formula, under any known qualification condition, the exact subdifferential calculus formula. This will appear elsewhere and will also be done for many other formulas.

Recall first that a point x is in the core of a convex subset C of X if for any $y \in X$ there exists a real number $t > 0$ such that $ty + (1-t)x \in C$. This is equivalent to $\mathbb{R}_+(C - x) = X$.

We can now prove the following corollary of Theorem 2.1. It has been proved for the first time by Rockafellar [18] for reflexive spaces, and then, for any Banach space, it was a consequence of the method of Rockafellar [18] and of Corollary 1 in Robinson [17]. Another proof has also been given in Aubin and Ekeland [3].

COROLLARY 2.2. *Let X be a reflexive Banach space and $f_1, f_2 : X \longrightarrow \mathbb{R}\cup\{+\infty\}$ be two proper lower semicontinuous convex functions. If*

$$0 \in \operatorname{core}(\operatorname{dom} f_1 - \operatorname{dom} f_2),$$

then for any $\overline{x} \in X$

$$\partial\,(f_1 + f_2)(\overline{x}) = \partial\, f_1(\overline{x}) + \partial\, f_2(\overline{x}).$$

Proof. We may obviously suppose that $\overline{x} \in \operatorname{dom} f_1 \cap \operatorname{dom} f_2$. We only have to prove that the first member is included in the second one. Fix any $x^* \in \partial(f_1+f_2)(\overline{x})$. By Theorem 2.1 there exist $x_{i,n} \xrightarrow[n\to\infty]{\|\ \|} \overline{x}$, $x^*_{i,n} \in \partial\, f_i(x_{i,n})$ such that

$$x^* = \lim_n (x^*_{1,n} + x^*_{2,n}) \quad \text{and} \quad \gamma_{i,n} := f_i(x_{i,n}) - \langle x^*_{i,n}, x_{i,n} - \overline{x}\rangle \underset{n\to\infty}{\longrightarrow} f_i(\overline{x}).$$

Consider any nonzero $v \in X$ and choose $s > 0$ and $u_i \in \operatorname{dom} f_i$ such that $sv = u_1 - u_2$. Then for $x^*_n := x^*_{1,n} + x^*_{2,n}$ we have

$$\begin{aligned}
\langle x^*_{1,n}, sv\rangle &= \langle x^*_{1,n}, u_1 - \overline{x}\rangle + \langle x^*_{1,n}, \overline{x} - u_2\rangle \\
&= \langle x^*_{1,n}, u_1 - x_{1,n}\rangle + \langle x^*_{1,n}, x_{1,n} - \overline{x}\rangle + \langle x^*_{1,n}, \overline{x} - u_2\rangle \\
&\le f_1(u_1) - f_1(x_{1,n}) + \langle x^*_{1,n}, x_{1,n} - \overline{x}\rangle + \langle x^*_{1,n}, \overline{x} - u_2\rangle \\
&\le f_1(u_1) - \gamma_{1,n} + \langle x^*_n, \overline{x} - u_2\rangle + \langle x^*_{2,n}, u_2 - x_{2,n}\rangle + \langle x^*_{2,n}, x_{2,n} - \overline{x}\rangle \\
&\le f_1(u_1) - \gamma_{1,n} + \langle x^*_n, \overline{x} - u_2\rangle + f_2(u_2) - f_2(x_{2,n}) + \langle x^*_{2,n}, x_{2,n} - \overline{x}\rangle \\
&= (f_1(u_1) - \gamma_{1,n}) + (f_2(u_2) - \gamma_{2,n}) + \langle x^*_n, \overline{x} - u_2\rangle.
\end{aligned}$$

It follows that the sequence $(x^*_{1,n})_n$ is w^*-bounded. Since $x^* = \lim_n (x^*_{1,n} + x^*_{2,n})$ and X is reflexive, we may suppose (extracting subsequences) that $x^*_{i,n} \xrightarrow[n\to\infty]{}^{w^*} x^*_i$, which ensures that $x^* = x^*_1 + x^*_2$. Moreover, as the graph of ∂f_i is sequentially $\| \ \| \times w^*$-closed, we have $x^*_i \in \partial f_i(\overline{x})$. So we have proved the inclusion $\partial (f_1 + f_2)(\overline{x}) \subset \partial f_1(\overline{x}) + \partial f_2(\overline{x})$ and hence the proof is complete. □

The formula could also be derived under the Attouch and Brézis [2] qualification condition or under the general condition in Rubinov [19] and Kutateladze [14] of general position of the domains.

3. Composition of convex functions. Let Y be a Banach space and Y_+ a convex cone of Y inducing a preorder $\leq_Y$ on Y defined by $y_1 \leq_Y y_2$ if and only if $y_2 - y_1 \in Y_+$. Let $+\infty$ be an abstract maximal element adjoined to Y.

Recall that a mapping $F : X \longrightarrow Y \cup \{+\infty\}$ is convex if for all $x, x' \in X$, and $t \in]0,1[$ one has

$$F(tx + (1-t)\,x') \leq_Y t\,F(x) + (1-t)\,F(x').$$

The set dom $F := \{x \in X : F(x) \in Y\}$ is the effective domain of F, $\mathrm{Im}F := F(X)$ the effective image of F and epi $F := \{(x,y) \in X \times Y : F(x) \leq_Y y\}$ the epigraph of F.

A function $f : Y \longrightarrow \mathbb{R} \cup \{+\infty\}$ is Y_+-nondecreasing on a subset S of Y if $f(y_1) \leq f(y_2)$ for all $y_1, y_2 \in S$ satisfying $y_1 \leq_Y y_2$. By convention one puts $f(+\infty) = +\infty$. If f is convex and Y_+-nondecreasing over $\mathrm{Im}F + Y_+$, then $f \circ F$ is convex.

THEOREM 3.1. *Let X, Y be two reflexive Banach spaces and Y_+ be a convex cone of Y. Suppose that $F : X \longrightarrow Y \cup \{+\infty\}$ is a convex mapping with closed epigraph and that $f : Y \longrightarrow \mathbb{R} \cup \{+\infty\}$ is a proper convex lower semicontinuous function which is nondecreasing over* $\mathrm{Im}(F) + Y_+$. *Then for $\overline{y} := F(\overline{x}) \in \mathrm{dom}\, f$ one has $x^* \in \partial(f \circ F)(\overline{x})$ if and only if there exist $x_n \xrightarrow{\| \ \|} \overline{x}$, $y_n \xrightarrow{\| \ \|} \overline{y}$, $x^*_n \xrightarrow{\| \ \|} x^*$, $e^*_n \xrightarrow{\| \ \|} 0$, $y'_n \in F(x_n) + Y_+$ with $y'_n \xrightarrow{\| \ \|} F(\overline{x})$ and $y^*_n \in Y^*_+ := \{y^* \in Y^* : \langle y^*, y\rangle \geq 0$ for all $y \in Y_+\}$ such that*

$$y^*_n + e^*_n \in \partial\, f(y_n),\ x^*_n \in \partial\,(y^*_n \circ F)(x_n),\ \langle y^*_n, y'_n\rangle = \langle y^*_n, F(x_n)\rangle$$

and

$$f(y_n) - \langle y^*_n, y_n - \overline{y}\rangle \longrightarrow f(\overline{y}) \quad \text{and} \quad \langle y^*_n, F(x_n) - \overline{y}\rangle \longrightarrow 0.$$

Proof. Put $f_1(x,y) := f(y)$ and $f_2(x,y) := \psi(x, y\,|\mathrm{epi}\, F)$ (the indicator function of epi F). One can easily verify that $x^* \in \partial\,(f \circ F)(\overline{x})$ if and only if $(x^*, 0) \in \partial\,(f_1 + f_2)(\overline{x}, \overline{y})$ and hence if and only if there exist

$$(0, y^*_n + e^*_n) + (x^*_n, -y^*_n) \xrightarrow{\| \ \|} (x^*, 0)$$

with

$$y^*_n + e^*_n \in \partial\, f(y_n),\ (x^*_n, -y^*_n) \in \partial\, \psi(.|\mathrm{epi}\, F)(x_n, y'_n),$$

$$y_n \xrightarrow{\| \ \|} \overline{y},\ (x_n, y'_n) \xrightarrow{\| \ \|} (\overline{x}, \overline{y}),$$

$$f(y_n) - \langle y^*_n + e^*_n, y_n - \overline{y}\rangle \longrightarrow f(\overline{y}), \tag{3.1}$$

$$f_2(x_n, y'_n) + \langle y^*_n, y'_n - \overline{y}\rangle - \langle x^*_n, x_n - \overline{x}\rangle \longrightarrow f_2(\overline{x}, \overline{y}) = 0. \tag{3.2}$$

Note that (3.2) is equivalent to

$$\langle y_n^*, y_n' - \overline{y}\rangle \longrightarrow 0 \tag{3.3}$$

since $x_n^* \xrightarrow{\|\ \|} x^*$, $x_n \xrightarrow{\|\ \|} \overline{x}$, and $f_2(x_n, y_n') = 0$. Note also that $(x_n^*, -y_n^*) \in \partial\,\psi(.\,|\,\mathrm{epi}\,F)(x_n, y_n')$ means that

$$\langle x_n^*, x - x_n\rangle - \langle y_n^*, y - y_n'\rangle \leq 0 \quad \text{for all } (x,y) \in \mathrm{epi}\,F. \tag{3.4}$$

On the one hand, taking for any $y' \in Y_+$, $x = x_n$, and $y = y_n' + y'$ in (3.4) we obtain $\langle y_n^*, y'\rangle \geq 0$ and hence $y_n^* \in Y_+^*$. On the other hand taking $x = x_n$ and $y = F(x_n)$ in (3.4) we obtain $\langle y_n^*, y_n' - F(x_n)\rangle \leq 0$ and hence $\langle y_n^*, y_n' - F(x_n)\rangle = 0$ since $y_n^* \in Y_+^*$ and $y_n' - F(x_n) \in Y_+$. Then (3.4) is equivalent to

$$\langle x_n^*, x - x_n\rangle \leq \langle y_n^*, y\rangle - \langle y_n^*, F(x_n)\rangle \quad \text{for all } (x,y) \in \mathrm{epi}\,F,$$

which is equivalent to

$$\langle x_n^*, x - x_n\rangle \leq y_n^* \circ F(x) - y_n^* \circ F(x_n) \quad \text{for all } x \in \mathrm{dom}\,F,$$

since $y_n^* \in Y_+^*$. As $y_n^* \circ F$ is convex (since $y_n^* \in Y_+^*$) (3.4) is then equivalent to $x_n^* \in \partial\,(y_n^* \circ F)(x_n)$. It also follows from the equality $\langle y_n^*, y_n' - F(x_n)\rangle = 0$ that (3.3) can be rewritten as

$$\langle y_n^*, F(x_n) - \overline{y}\rangle \longrightarrow 0.$$

Finally, since $e_n^* \xrightarrow{\|\ \|} 0$ and $y_n \xrightarrow{\|\ \|} \overline{y}$, (3.1) is equivalent to

$$f(y_n) - \langle y_n^*, y_n - \overline{y}\rangle \longrightarrow f(\overline{y})$$

and hence the proof is complete. □

In case the convex mapping F is assumed to be continuous, we have the following corollary. It generalizes the result proved by Levin in [15] and also that of [7].

COROLLARY 3.2. *Let X and Y be two reflexive Banach spaces and Y_+ be a closed convex cone in Y which is normal (see [16]). Suppose that $F : X \to Y$ is a continuous convex mapping and $f : Y \longrightarrow \mathbb{R} \cup \{+\infty\}$ is a proper convex lower semicontinuous function which is nondecreasing over $\mathrm{Im}(F) + Y_+$. Then for $\overline{y} := F(\overline{x}) \in \mathrm{dom}\,f$ one has $x^* \in \partial\,(f \circ F)(\overline{x})$ if and only if there exist*

$$y_n^* \in \partial\, f(y_n) \quad \text{and} \quad x_n^* \in \partial\,(y_n^* \circ F)(x_n),$$

with

$$x_n \xrightarrow{\|\ \|} \overline{x}, \quad y_n \xrightarrow{\|\ \|} F(\overline{x}), \quad x_n^* \xrightarrow{\|\ \|} x^*,$$

$$f(y_n) - \langle y_n^*, y_n - \overline{y}\rangle \to f(\overline{y}) \quad \text{and} \quad \langle y_n^*, F(x_n) - \overline{y}\rangle \longrightarrow 0.$$

Proof. Consider $x^* \in \partial\,(f \circ F)(\overline{x})$. Let x_n, y_n, x_n^*, e_n^*, y_n', and y_n^* be given by Theorem 3.1. Put $z_n^* := y_n^* + e_n^*$ and note that

$$x_n^* \in \partial\,(y_n^* \circ F)(x_n) \quad \text{and} \quad y_n^* \circ F = z_n^* \circ F + (y_n^* - z_n^*) \circ F$$

with $y_n^* - z_n^* = -e_n^* \xrightarrow{\|\ \|} 0$. Since Y_+ is normal, then F is k-Lipschitzian (for some real number $k \geq 0$) on a neighborhood V of $\overline{x}$ (see Theorem 5 in [17], for example),

and hence $(y_n^* - z_n^*) \circ F$ is $k\|y_n^* - z_n^*\|$-Lipschitzian on V. So there exists $v_n^* \xrightarrow{\|\ \|} 0$ with $u_n^* := x_n^* + v_n^* \in \partial\,(z_n^* \circ F)(x_n)$. Moreover $u_n^* \xrightarrow{\|\ \|} x^*$, $z_n^* \in \partial\, f(y_n)$, $\langle z_n^*, F(x_n) - \overline{y}\rangle \longrightarrow 0$, and

$$f(y_n) - \langle z_n^*, y_n - \overline{y}\rangle \longrightarrow f(\overline{y}).$$

Since it is not difficult to see that the converse also holds, the proof is complete. □

Remark. With the help of (2.2) another form will be established elsewhere.

In the following corollary, F is assumed to be a continuous linear mapping A.

COROLLARY 3.3. *Let X and Y be two reflexive Banach spaces and $A : X \longrightarrow Y$ be a continuous linear mapping. Suppose that $f : Y \longrightarrow \mathbb{R} \cup \{+\infty\}$ is a proper convex lower semicontinuous function. Then for $A\,\overline{x} \in \operatorname{dom} f$ one has $x^* \in \partial\,(f \circ A)(\overline{x})$ if and only if there exist*

$$y_n^* \in \partial\, f(y_n) \quad \text{with} \quad y_n^* \circ A \xrightarrow{\|\ \|} x^*,$$

$$y_n \xrightarrow{\|\ \|} A\,\overline{x} \quad \text{and} \quad f(y_n) - \langle y_n^*, y_n - \overline{y}\rangle \longrightarrow f(A\,\overline{x}),$$

which we will translate by

$$\partial\,(f \circ A)(\overline{x}) = \|\ \| - \limsup_{y \xrightarrow{f - \langle\,,\,\rangle} A\,\overline{x}} \big[\partial\, f(y) \circ A\big].$$

Proof. Consider the sequences given by Corollary 3.2 for $Y_+ = \{0\}$. Then

$$x_n^* \in \partial\,(y_n^* \circ A)(\overline{x}) = y_n^* \circ A \quad \text{and} \quad \langle y_n^*, A\,x_n - \overline{y}\rangle = \langle y_n^* \circ A, x_n - \overline{x}\rangle,$$

and hence the condition $\langle y_n^*, A\,x_n - \overline{y}\rangle \longrightarrow 0$ is superfluous. So the proof is complete. □

A mean-value result in the spirit of Zagrodny's mean value theorem (see [22]) can be deduced from the above corollary. An equality form (instead of the inequality one in [22]) can be reached here because of the convexity assumption. This form has been proved much earlier by Borwein [4] by another method.

COROLLARY 3.4. *Let X be a reflexive Banach space and $f : X \longrightarrow \mathbb{R} \cup \{+\infty\}$ be a proper convex lower semicontinuous function. Suppose that a and b are two different points of* $\operatorname{dom} f$. *Then there exists some c in the open line segment (a, b), (c_n) in X converging to c and $x_n^* \in \partial\, f(c_n)$ such that*

$$f(b) - f(a) = \lim_{n\to\infty} \langle x_n^*, b - a\rangle.$$

Proof. Consider the classical function g defined on $\mathbb{R}$ by

$$g(t) := f(a + t(b - a)) + t(f(a) - f(b)).$$

This convex function g is lower semicontinuous on [0,1] with $g(0) = g(1)$ and hence attains its minimum on [0,1] at some point $r \in]0, 1[$. This ensures that $O \in \partial\, g(r)$, which can be rewritten as

$$f(b) - f(a) \in \partial\,(h \circ A)(r),$$

where $h(x) := f(a + x)$ and $A(t) := t(b - a)$. By Corollary 3.3 one easily sees that there exists (c_n) in X converging to $c := a + r(b - a)$ and $x_n^* \in \partial\, f(c_n)$ such that

$$f(b) - f(a) = \lim_{n\to\infty} \langle x_n^*, b - a\rangle. \qquad \square$$

4. Sequential Lagrange multipliers. In all of what follows, X and Y are two reflexive Banach spaces, K a closed convex normal cone of Y, and C a closed convex subset of X. Consider the constrained minimization problem

$$(\mathcal{P}) \qquad \text{minimize } f(x) \quad \text{subject to } x \in C \quad \text{and } F(x) \in -K,$$

where f, F are mappings from X into $\mathbb{R}$ and Y, respectively.

Generally, under some constraint qualification, for example,

$$\mathbb{R}_+[K + F(C)] \text{ is a closed vector subspace of } Y, \tag{4.1}$$

one can prove (see [7]) that there exist a real number $\lambda > 0$, $y^* \in Y^*$ with $\langle y^*, y\rangle \geq 0$ for all $y \in K$ and $\langle y^*, F(\overline{x})\rangle = 0$ such that

$$0 \in \partial\,(\lambda f + y^* \circ F + \psi(., C))(\overline{x}) \tag{4.2}$$

whenever $\overline{x}$ is a solution of $(\mathcal{P})$ and f and F are continuous (or more generally lower semicontinuous) and convex. Note that the qualification condition (4.1) is much less stringent than the classical Slater condition ($F(x_0) \in -\operatorname{int} K$ for some $x_0 \in C$) which requires the nonemptiness of the topological interior of the cone K. However, even in finite-dimensional settings, when any constraint qualification is not satisfied, one only generally proved (up to now) the necessary optimality condition corresponding to (4.2) with $\lambda = 0$; that is,

$$0 \in \partial\,(y^* \circ F + \psi(., C))(\overline{x}). \tag{4.3}$$

This condition is not sufficient at all to ensure that $\overline{x}$ is a solution of $(\mathcal{P})$, and it does not take into account the objective function f. Moreover, when the space Y is not finite dimensional (i.e., infinitely many inequality and equality constraints are considered) one generally needs some extra (or compactness-like) condition on the mapping F or the cone K (see [21]) to obtain (4.3).

In this section we are going to show how the sequential subdifferential calculus can be used to establish new necessary and sufficient optimality conditions for the problem $(\mathcal{P})$. These conditions will make appear a new form of Lagrange multipliers that we call *sequential Lagrange multipliers*. We also have to mention that Hanson has proved very recently in [9] that a certain form of sequential Lagrange multipliers is a sufficient optimality condition whenever f and F are convex (more generally invex) differentiable functions and X and Y are finite dimensional. However, Hanson did not show that there exist sequential Lagrange multipliers for the solution points. So he did not establish any sequential Lagrange necessary optimality condition.

To prove the result we will employ a method that Jourani and the author have introduced (to our knowledge for the first time in optimization) in a common work which is the last chapter of the thesis [12].

This work, whose main result has been stated in [20], has constituted the article [13]. The method consisted of reducing a constrained optimization problem to the unconstrained minimization of a composition function and to apply an appropriate formula estimating the approximate subdifferential of a composition when the data nonlinear mappings are compactly Lipschitzian. There we used the distance function in the reduction procedure. In the proof below we will use a similar reduction procedure but with the indicator function.

THEOREM 4.1. *Assume that f and F are convex and continuous. Then a point $\overline{x} \in C \cap F^{-1}(-K)$ is a solution of the minimization problem $(\mathcal{P})$ if and only if there*

exist $x_n \longrightarrow \overline{x}$, $w_n \longrightarrow \overline{x}$, $y_n \longrightarrow F(\overline{x})$, $y_n^* \in Y^*$, $u_n^* \in \partial f(x_n)$, $v_n^* \in \partial (y_n^* \circ F)(x_n)$, *and* $w_n^* \in \partial \psi(., C)(w_n)$ *such that*

(i) $\langle y_n^*, y\rangle \geq 0$ *for all* $y \in K$ *and* $\langle y_n^*, y_n\rangle = 0$,

(ii) $0 = \| \ \| - \lim_{n\to\infty}(u_n^* + v_n^* + w_n^*)$,

(iii) $\langle y_n^*, y_n - F(\overline{x})\rangle + \langle w_n^*, w_n - \overline{x}\rangle \longrightarrow 0$ *and* $\langle y_n^*, F(x_n) - F(\overline{x})\rangle + \langle w_n^*, x_n - \overline{x}\rangle \longrightarrow 0$.

Proof. It is easily seen that $\overline{x}$ is a solution of the problem $(\mathcal{P})$ if and only if $\overline{x}$ minimizes the unconstrained function

$$f + \psi(., -K) \circ F + \psi(., C) := g \circ G,$$

where $G : X \longrightarrow \mathbb{R} \times Y \times X$ and $g : \mathbb{R} \times Y \times X \longrightarrow \mathbb{R}$ are defined via

$$G(x) = (f(x), F(x), x) \text{ and } g(r, y, x) = r + \psi(y, -K) + \psi(x, C),$$

and $\psi(., C)$ is the indicator function of C. Denote by Q the closed convex normal cone $[0, \infty[\times K \times \{0_X\}$ of $\mathbb{R} \times Y \times X$. Obviously g is convex lower semicontinuous and Q-nondecreasing and G is Q-convex and continuous. Then $\overline{x}$ is a solution of $(\mathcal{P})$ if and only if $0 \in \partial g \circ G(\overline{x})$, and hence if and only if (by Corollary 3.2) there exist

$$x_n \longrightarrow \overline{x}, \ (r_n, y_n, w_n) \longrightarrow (f(\overline{x}), F(\overline{x}), \overline{x}),$$

$$(y_n^*, w_n^*) \in \partial \psi(., -K)(y_n) \times \partial \psi(., C)(w_n),$$

$$x_n^* \in \partial (f + y_n^* \circ F + \langle w_n^*, .\rangle)(x_n) = \partial f(x_n) + \partial (y_n^* \circ F)(x_n) + w_n^*$$

(that is, $x_n^* = u_n^* + v_n^* + w_n^*$ for some $u_n^* \in \partial f(x_n)$ and $v_n^* \in \partial (y_n^* \circ F)(x_n)$) such that

$$\|x_n^*\| \longrightarrow 0, \ g(r_n, y_n, w_n) - (r_n - f(\overline{x})) - \langle y_n^*, y_n - F(\overline{x})\rangle - \langle w_n^*, w_n - \overline{x}\rangle \longrightarrow g(G(\overline{x}))$$

and

$$(f(x_n) - f(\overline{x})) + \langle y_n^*, F(x_n) - F(\overline{x})\rangle + \langle w_n^*, x_n - \overline{x}\rangle \longrightarrow 0.$$

The three last relations can be rephrased, respectively, as

$$\langle y_n^*, y_n - F(\overline{x})\rangle + \langle w_n^*, w_n - \overline{x}\rangle \longrightarrow 0$$

and

$$\langle y_n^*, F(x_n) - F(\overline{x})\rangle + \langle w_n^*, x_n - \overline{x}\rangle \longrightarrow 0.$$

So $\overline{x}$ is a solution of $(\mathcal{P})$ if and only if there exist (x_n) and (w_n) converging to $\overline{x}$, $y_n \xrightarrow{\| \ \|} F(\overline{x})$, $y_n^* \in \partial \psi(., -K)(y_n)$, $w_n^* \in \partial \psi(, C)(w_n)$, $u_n^* \in \partial f(x_n)$ and $v_n^* \in \partial (y_n^* \circ F)(x_n)$ such that

$$0 = \| \ \| - \lim(u_n^* + v_n^* + w_n^*)$$

and

$$\langle y_n^*, y_n - F(\overline{x})\rangle + \langle w_n^*, w_n - \overline{x}\rangle \longrightarrow 0 \text{ and } \langle y_n^*, F(x_n) - F(\overline{x})\rangle + \langle w_n^*, x_n - \overline{x}\rangle \longrightarrow 0.$$

So the proof is complete. □

The theorem makes clear that the sequence (y_n^*) may be considered as a generalized Lagrange multiplier. It is what we call a sequential Lagrange multiplier. It does not require any constraint qualification. Moreover, its applicability does not depend on the analytical forms of the mappings defining the constraints, whereas this may often be the case with the use of constraint qualifications (see comments and examples in [9]).

REFERENCES

[1] H. Attouch, J.-B. Baillon, and M. Théra, *Variational sum of monotone operators*, J. Conv. Anal., 1 (1994), pp. 1–29.

[2] H. Attouch and H. Brezis, *Duality for the sum of convex functions in general Banach spaces*, in Aspects of Mathematics and Its Applications, J. A. Barroso, ed., Elsevier, Amsterdam, 1986.

[3] J.-P. Aubin and I. Ekeland, *Applied Nonlinear Analysis*, Wiley, New York, 1994.

[4] J. M. Borwein, *A note on ε-subgradients and maximal monotonicity*, Pacific J. Math., 103 (1982), pp. 307–314.

[5] J. M. Borwein and A. D. Ioffe, *Proximal analysis in smooth spaces*, Set-Valued Analysis, 4 (1996), pp. 1–24.

[6] A. Brøndsted and R. T. Rockafellar, *On the subdifferential of convex functions*, Proc. Amer. Math. Soc., 16 (1965), pp. 605–611.

[7] C. Combari, M. Laghdir, and L. Thibault, *Sous-différentiels de fonctions convexes composées*, Ann. Sci. Math. Québec, 18 (1994), pp. 119–148.

[8] I. Ekeland, *Nonconvex minimization problems*, Bull. Amer. Math. Soc., 1 (1979), pp. 443–474.

[9] M. A. Hanson, *A generalization of the Kuhn-Tucker sufficiency conditions*, J. Math. Anal. Appl., 184 (1994), pp. 146–155.

[10] J.-B. Hiriart-Urruty and R. R. Phelps, *Subdifferential calculus using ε-subdifferentials*, J. Funct. Anal., 118 (1993), pp. 154–166.

[11] J.-B. Hiriart-Urruty, M. Moussaoui, A. Seeger, and M. Volle, *Subdifferential calculus, without qualification conditions, using approximate subdifferentials: A survey*, Nonlinear Anal. Theory Methods Appl., 24 (1995), pp. 1727–1754.

[12] A. Jourani, *Régularité métrique et ses applications en programmation mathématiques*, Ph.D. thesis, Université de Pau, France, 1989.

[13] A. Jourani and L. Thibault, *Approximations and metric regularity in mathematical programming in Banach space*, Math. Oper. Res., 18 (1993), pp. 390–401.

[14] S. S. Kutateladze, *Convex operators*, Russian Math. Surveys, 34 (1979), pp. 181-214.

[15] V. L. Levin, *On the subdifferential of a composite functional*, Soviet Math. Dokl., 11 (1970), pp. 1194–1195.

[16] A. L. Peressini, *Ordered Topological Vector Spaces*, Harper's Series in Modern Mathematics, Harper & Row, New York, 1967.

[17] S. M. Robinson, *Regularity and stability for convex multivalued functions*, Math. Oper. Res., 1 (1976), pp. 130–143.

[18] R. T. Rockafellar, *Conjugate Duality and Optimization*, SIAM, Philadelphia, 1974.

[19] A. M. Rubinov, *Sublinear operators and their applications*, Russian Math. Surveys, 32 (1977), pp. 115–175.

[20] L. Thibault, *Lagrange-Kuhn-Tucker multipliers for general mathematical programming problems*, in Optimization and Nonlinear Analysis (Conference held in Haifa in 1990), Pitman Research Notes in Mathematics Series, A. Ioffe, M. Marcus, and S. Reich, eds., Longman, Harlow, UK, 1992, pp. 311–315.

[21] L. Thibault, *Lagrange-Kuhn-Tucker multipliers for Pareto optimization problems*, to appear.

[22] D. Zagrodny, *Approximate mean value theorem for upper subderivatives*, Nonlinear Anal. Theory Methods Appl., 12 (1988), pp. 1413–1438.

SIAM J. CONTROL OPTIM.
Vol. 35, No. 5, pp. 1445–1469, September 1997

WEIGHTED SENSITIVITY MINIMIZATION IN SYSTEMS WITH A SINGLE INPUT DELAY: A STATE SPACE SOLUTION*

GILEAD TADMOR†

Abstract. A differential/algebraic matrix Riccati equation–based design method is developed. The solution is reached via a parallel, two-way analysis of an underlying differential game: in the framework of both the finite order delay system and a distributed parameter, ordinary evolution model.

Key words. input delay, Nehari problem, state space solutions, matrix Riccati equations

AMS subject classifications. 49K25, 93B36, 93D21

PII. S0363012995279651

1. Introduction.

This note concerns the weighted sensitivity minimization problem when the underlying plant has a single input lag. The usual solution approach is to reduce this control design problem to a model-matching Nehari problem [16]. An account of this reduction, of the subsequent state space solution, and of relevant references in the ordinary linear, time invariant (LTI) case can be found in [9]. Extensions to ordinary linear, time varying (LTV) systems are provided in [12, 32]. This note concerns the LTI problem when an ordinary plant is cascaded with the pure delay operator. The weight function is restricted to being rational. The LTI case is selected for simplicity; LTV extensions, following the ideas outlined in [12, 32], could be readily achieved.

Transform-based solutions of the (SISO) weighted sensitivity minimization problem in delay systems were obtained during the 1980s [7, 8, 36], including extensions to the case of multiple commensurate input lags [25] and a rather restricted class of distributed input delays [24]. The main tool in these solutions is the commutant lifting theorem which is described in [22, 23]. More general frameworks for the solution of distributed parameter H_∞problems include the skew Toeplitz approach and a state space approach in which abstract evolution models are used. Sample presentations of the two approaches are, respectively, [17] and [28, 33].

A drawback of state space solutions of distributed parameter linear-quadratic (LQ) optimization problems (LQR, LQG, H_∞, etc.) is that they require the often difficult solution of operator Riccati equations. A main contribution of this note is the reduction of the operator Riccati equations that arise in the context of the weighted sensitivity minimization problem to a set of algebraic and differential matrix Riccati equations. Similar results have been previously reported in [15], which inspired the current developments. The key idea in [15] is the use of a "lifting" technique and a reduction to a finite state space discrete time model with distributed inputs and outputs. [15] provides periodic solutions to a variety of delayed H_∞problems. The advantage of the method suggested here is that the generic design in the LTI case is LTI. The approach described here has been applied in [30, 29], with a similar type

*Received by the editors January 9, 1995; accepted for publication (in revised form) May 22, 1996. This research was supported by the Army Research Office and the National Science Foundation.
http://www.siam.org/journals/sicon/35-5/27965.html
†Department of Electrical and Computer Engineering, Northeastern University, Boston, MA 02115 (tadmor@cdsp.neu.edu).

of result, to a variant of the *standard H_∞ problem* and to the problem of robustness optimization in the *gap* metric in systems that involve a pure input lag.

2. Notation. Prime will denote an adjoint of a matrix or an operator. The notation of an "$L_2[a,b]$" space will include an indication of the value space of the functions considered (e.g., $L_2([a,b],\mathbb{R}^l)$) only when ambiguous. $W_2^1[a,b]$ stands for the usual Sobolev space of absolutely continuous $L_2[a,b]$ functions with $L_2[a,b]$ derivatives. The subscript "*loc*" will have the usual "local" meaning, e.g., $-L_{2\ loc}[0,\infty)=\cap_{t>0}L_2[0,t)$. The subscript "2" will indicate the $L_2[0,\infty)$ norm. In other cases inner products and norms will be distinguished by a subscript indicating the relevant space (e.g., $\|\phi\|_{L_2[-1,0]}$). The subscript "e" will indicate the Euclidean inner product and norm. H_∞ will stand for the Hardy space of uniformly bounded, analytic functions in the right half-plane with the $L_\infty(\jmath\mathbb{R})$ norm. The spectral radius of a matrix M will be denoted $\rho(M)$. The standard notation of "u_t" will be used in reference to the single-time unit history of the trajectory at the time t: $u_t(\theta)=u(t+\theta)$, $\theta\in[-1,0]$.

3. The problem. Consider a multivariable LTI system with a transfer function $G(s)=G_0(s)e^{-s}$ and a weight function $W(s)$, where $G_0(s)$ and $W(s)$ are rational. The weighted sensitivity minimization problem concerns a search for a selection of a stabilizing compensator C (i.e., such that $S=(I+GC)^{-1}\in H_\infty$) that minimizes the H_∞ norm $\|WS\|_\infty$. In the commonly treated suboptimal version of this problem, one seeks a characterization of suboptimal values $\gamma>\gamma_0$, where

$$\gamma_0=\inf\left\{\|WS\|_\infty\ :\ S\in H_\infty\right\}, \tag{1}$$

and a parameterization of the suboptimal set,

$$\mathbf{C}_\gamma=\left\{C\ :\ S\in H_\infty \text{ and } \|WS\|_\infty<\gamma\right\}. \tag{2}$$

A standard step is the reduction of the suboptimal weighted sensitivity minimization problem to an equivalent (or close to equivalent) Nehari problem. Details of the reduction, which remains valid in the presence of an input delay in $G(s)$, can be found in [9]. Consequently, the problem addressed is stated in the following form: given a rational transfer function $\Upsilon(s)\in L_\infty(\jmath\mathbb{R})$, characterize the optimal value

$$\gamma_0=\inf\left\{\|\Upsilon(s)-e^{-s}\Theta(s)\|_{L_\infty(\jmath\mathbb{R})}\ :\ \Theta\in H_\infty\right\} \tag{3}$$

and, given $\gamma>\gamma_0$, parameterize the suboptimal set

$$\mathbf{\Theta}_\gamma=\left\{\Theta\in H_\infty\ :\ \|\Upsilon(s)-e^{-s}\Theta(s)\|_{L_\infty(\jmath)}<\gamma\right\}. \tag{4}$$

Given the focus of this note on delay systems we shall limit our attention here to choices of Θ in the class of well-posed, neutral functional differential equations (FDEs).[1] This class has been identified [31] as containing all well-posed finite memory, Euclidean space evolutions, and the restriction is made merely for convenience and cohesion in the presentation. This restriction can be easily replaced by other classes of distributed parameter systems for certain closed-loop well-posedness properties to be assured, and, in particular, it effects neither the characterization of the optimal value γ_0 nor the form of the "central compensator" in this note's main result, Theorem 4.1.

[1] The issue of well-posedness of neutral FDEs is quite involved. For the purpose of this discussion well-posedness may be interpreted as the existence of an integral, variations-of-parameters solution formula, as described in [31, Cor. E].

Let $\Upsilon = [A, B, C, D]$ be a minimal realization over the state space $\mathbb{R}^n$. In view of the assumption that $\Upsilon(s)$ has no purely imaginary poles there exists a partition of the state space as the direct sum of a stable component and an antistable component:

$$\mathbb{R}^n = X_{st} \oplus X_{as}. \tag{5}$$

Associated with this partition are block structures of the coefficients A, B, C, and D:

$$A = \begin{bmatrix} A_{st} & 0 \\ 0 & A_{as} \end{bmatrix}, \quad B = \begin{bmatrix} B_{st} \\ B_{as} \end{bmatrix}, \quad \text{and} \quad C = [C_{st} \;\; C_{as}],$$

such that $\Upsilon = \Upsilon_{st} + \Upsilon_{as}$ with $\Upsilon_{st} = [A_{st}, B_{st}, C_{st}, D]$ stable and $\Upsilon_{as} = [A_{as}, B_{as}, C_{as}, 0]$ antistable.

Let X_{as} and Y_{as} be the positive definite, infinite time, controllability and observability Gramians of the antistable component, solving

$$A_{as}X_{as} + X_{as}A_{as}' = B_{as}B_{as}' \quad \text{and} \quad Y_{as}A_{as} + A_{as}'Y_{as} = C_{as}'C_{as}. \tag{6}$$

Denote

$$P = \begin{bmatrix} 0 & 0 \\ 0 & X_{as}^{-1} \end{bmatrix} \quad \text{and} \quad Q = \begin{bmatrix} 0 & 0 \\ 0 & Y_{as}^{-1} \end{bmatrix}.$$

It is a standard observation that P and Q are positive semidefinite, stabilizing solutions of the following deficient algebraic Riccati equations (AREs):

$$PA + A'P - PBB'P = 0 \quad \text{and} \quad AQ + QA' - QC'CQ = 0. \tag{7}$$

Left and right, stable co-prime factorizations of Υ with isometric denominators are provided in terms of these solutions: $\Upsilon = {M_l}^{-1}N_l = N_r{M_r}^{-1}$, where $N_l = [A - QC'C, B - QC'D, C, D]$, $M_l = [A - QC'C, -QC', C, I]$, $N_r = [A - BB'P, B, C - DB'P, D]$, and $M_r = [A - BB'P, B, -B'P, I]$. The following expressions for γ_0 and the set $\mathbf{\Theta}_\gamma$ are written in terms of N_l and M_l:

$$\gamma_0 = \inf\left\{\|N_l(s) - M_l(s)e^{-s}\Theta(s)\|_\infty \;:\; \Theta \in H_\infty\right\} \tag{8}$$

and

$$\mathbf{\Theta}_\gamma = \left\{\Theta \in H_\infty \;:\; \|N_l(s) - M_l(s)e^{-s}\Theta(s)\|_\infty < \gamma\right\}. \tag{9}$$

4. The main result. The following abbreviations and notations are used in the statement of the main result of this note, Theorem 4.1 below. They include the notation of "ρ_0" and are well defined for $\gamma^2 > \rho_0$.

$$\begin{aligned}
\rho_0 &= \max\left\{\rho(X_{as}Y_{as}),\ \rho(D'D)\right\}, & Z_{as} &= Y_{as}^{-1} - \tfrac{1}{\gamma^2}X_{as}, \\
R &= \begin{bmatrix} 0 & 0 \\ 0 & Z_{as}^{-1} \end{bmatrix}, & E_{22} &= -(\gamma^2 I - D'D)^{-1}D', \\
E_{12} &= -\gamma(\gamma^2 I - DD')^{-\frac{1}{2}}, & E_{21} &= (\gamma^2 I - D'D)^{-\frac{1}{2}}, \\
H_1 &= \gamma(\gamma^2 I - DD')^{-\frac{1}{2}}C, & H_2 &= (\gamma^2 I - D'D)^{-1}D'C, \\
G_1 &= (B - QC'D)(\gamma^2 - D'D)^{-\frac{1}{2}}, & G_2 &= (\gamma^2 QC' - BD')(\gamma^2 I - DD')^{-1},
\end{aligned} \tag{10}$$

where the block structure of R is consistent with the state space partition (5).

THEOREM 4.1. *$\gamma > \gamma_0$ if and only if the following two conditions hold.* (a) $\gamma^2 > \rho_0$. (b) *There exists a positive semidefinite, uniformly bounded solution of the following differential Riccati equation over the interval* $[0,1]$:

$$\dot{R}_0 + R_0(A - G_2C) + (A - G_2C)'R_0 + R_0G_1G_1'R_0 + H_1'H_1 = 0, \qquad R_0(1) = R. \tag{11}$$

Suppose that, indeed, $\gamma > \gamma_0$. Then a complete parameterization of the family $\mathbf{\Theta}_\gamma$ is provided in terms of the following realization of the mapping $u = \Theta w$ in members $\Theta \in \mathbf{\Theta}_\gamma$:

$$\begin{aligned} \dot{x}_c(t) &= A_c x_c(t) + B_{c1} w(t) + B_{c2} u(t-1), \\ u(t) &= C_{c1} x_c(t) + D_{c12}\phi(t) + D_{c13} u_t, \\ \psi(t) &= C_{c2} x_c(t) + D_{c21} w(t) + D_{c22} u(t-1) + D_{c23} u_t, \qquad \phi = \Theta_0 \psi, \end{aligned} \tag{12}$$

where the coefficients are as defined below and where the design parameter is the mapping Θ_0 that is defined as the I/O mapping in a stable, neutral FDE and is selected subject to the induced norm constraint $\|\Theta_0\| < 1$.

The following instructions concern the definition of the coefficients of (12). *Given the solution $R_0(t)$ of* (11), *define $A_0(t) = A - G_2C + G_1G_1'R_0(t)$, and let $\Phi_0(t,s)$ be the transition matrix generated by $A_0(t)$. In these terms define the matrices A_c, B_{c1}, B_{c2}, C_{c1}, D_{c12}, C_{c2}, D_{c21}, and D_{c22} via*

$$\begin{aligned} A_c &= A - QC'C, & B_{c1} &= B - QC'D, \\ B_{c2} &= QC', & C_{c1} &= \left(C(I - QR) + \tfrac{1}{\gamma^2} DB'R\right)\Phi_0(1,0), \\ D_{c12} &= (2E_{12}'E_{12})^{-\frac{1}{2}}, & C_{c2} &= -G_1'R_0(0) - \tfrac{1}{\gamma} D'H_1, \\ D_{c21} &= E_{21}^{-1}, & D_{c22} &= -\tfrac{1}{\gamma} D'E_{12}. \end{aligned} \tag{13}$$

Define a matrix-valued function

$$\Xi(r) = \int_0^r \Phi_0(r,s) G_1 G_1' \Phi_0(r,s)' ds, \tag{14}$$

and let D_{c13} and D_{c23} be the following finite rank operators over $L_2[-1,0]$:

$$\begin{aligned} D_{c13}u_t &= (C(I - QR) + \tfrac{1}{\gamma^2} DB'R) \\ &\quad \cdot \int_{-1}^0 \Phi_0(1, s+1)((I + \Xi(s+1)R_0(s+1)G_2 + \Xi(s+1)H_1'E_{12})u_t(s) ds, \\ D_{c23}u_t &= -G_1' \int_{-1}^0 \Phi_0(s+1,0)'(R_0(s+1)G_2 + H_1'E_{12})u_t(s) ds. \end{aligned} \tag{15}$$

The "central" solution that generates the suboptimal family $\mathbf{\Theta}_\gamma$ in (12) is a delay system, involving both discrete and distributed delay effects in the control function. The internal dynamics in (12) are governed by *both* the differential equation in $x_c(t)$ (the first equation) and the recursion formula for $u(t)$ (the second equation). As will become clear from the proof, this structure of the central solution fits into the general

framework of a neutral FDE, but it arises *regardless* of the explicit requirement that compensators be defined by neutral FDEs. That requirement, made merely to avoid issues of well-posedness, affects only the selection of Θ_0. It is noted that this form adheres to the pattern of other previously obtained solutions of the delayed Nehari problem.

5. The proof. For any selection of $\Theta \in H_\infty$, the transfer function $N_l(s) - M_l(s)e^{-s}\Theta(s)$ has the following stable realization:

$$\begin{aligned} \dot{x}(t) &= (A - QC'C)x(t) + (B - QC'D)w(t) + QC'u(t-1), \\ z(t) &= Cx(t) + Dw(t) - u(t-1), \\ y(t) &= w(t), \qquad u = \Theta y, \end{aligned} \tag{16}$$

which, ignoring the control delay, adheres to the form of a "standard H_∞problem," with Θ being the sought controller. As is well known, due to the input delay, initial data that are needed at any given time in order to determine the solution of (16) comprise both the momentary value of the trajectory x and the single time unit history of the control input. In that sense, the true state of (16) at the time t comprises the pair $(x(t), u_t)$. Focusing on the Euclidean and the L_2 topologies, these data will be embedded with the structure of the product Hilbert space $M_2 = \mathbb{R}^n \times L_2[-1, 0]$.

5.1. A differential game. Following an approach that became standard in the treatment of H_∞ optimization problems, the test of suboptimal model matching is reduced to an equivalent differential game. The following is the basic observation.

LEMMA 5.1. *If $\gamma > \gamma_0$, then there exists a unique solution to the following differential game, given any initial data $(x(0), u_0) \in M_2$ in* (16)*:*

$$\inf_{w \in L_2[0,\infty)} \left\{ \gamma^2 \|w\|_2^2 - \inf_{u \in L_2[0,\infty)} \|z\|_2^2 \right\}, \tag{17}$$

where the data for the internal optimization problem (over u) comprises both the initial data $(x(0), u_0)$ and the selection of w.

A brief outline of the proof. The basic ideas are the same as in the ordinary case. Details can be found in [27, 28]. The statement $\gamma > \gamma_0$ is equivalent to the existence of $\Theta \in H_\infty$ such that $\|N_l - M_l e^{-s}\Theta\|_\infty < \gamma$. This bound holds, in turn, if and only if there exists $\lambda > 0$ such that

$$\gamma^2 \|w\|_2^2 - \|z\|_2^2 \geq \lambda^2 \|w\|_2^2 \tag{18}$$

for any $w \in L_2[0, \infty)$, $u = \Theta w$, and with the zero initial data in (16). Assuming that $\gamma > \gamma_0$, the validity of (18) with $u = \Theta w$ (for some choice of Θ) leads to the following stronger, open-loop inequality:

$$\gamma^2 \|w\|_2^2 - \inf_{u \in L_2[0,\infty)} \|z\|_2^2 \geq \lambda^2 \|w\|_2^2. \tag{19}$$

As can be easily seen, the delay in u does not affect the solvability of the optimization problem in (19) and thus the arguments in [27, 28]. In particular, one can use (19) to show that the assumption "$\gamma > \gamma_0$" leads to existence and uniqueness of a solution of (17). □

Notice that in the open-loop context of (17), there is no requirement of causality in the dependence of the optimal selection of u on w.

5.2. Explicit solution of the differential game problem. The differential game (17) will be divided into two subproblems. One subproblem concerns input effects on $z(t)$ prior to $t = 1$. The other subproblem concerns such effects for $t > 1$.

It is observed that the selection of $u(t)$, $t \geq 0$, has no effect on the values of $z(t)$, $t \leq 1$. Once the initial data and $w(t)$, $t \in [0, 1]$, are fixed, the component of (17) that is determined by the selection of $w(t)$, $t > 1$, and of $u(t)$, $t > 0$, reduces to an equivalent problem that does not involve control delay, as follows. Denote $\tilde{u}(t) = u(t-1)$. Then, over the positive ray $t > 1$, there holds

$$\begin{array}{rcccccc} \dot{x} & = & (A - QC'C)x & + & (B - QC'D)w & + & QC'\tilde{u}, \\ z & = & Cx & + & Dw & - & \tilde{u}. \end{array} \tag{20}$$

The important fact is that the only constraint in (20) is the initial value of $x(1)$, as determined by $x(0)$, u_0, and the restriction of $w(t)$ to the interval $[0, 1]$. With (20) as the underlying system and given $x(1)$, the solution of (17) must define a solution of the following min-max problem:

$$\inf_{w \in L_2[1,\infty)} \left\{ \gamma^2 \|w\|^2_{L_2[1,\infty)} - \inf_{\tilde{u} \in L_2[1,\infty)} \|z\|^2_{L_2[1,\infty)} \right\}. \tag{21}$$

Problem (21) can be considered over the following alternative state space realization, where z is considered as the "control input":

$$\dot{x} = Ax + Bw - QC'z. \tag{22}$$

In this setting the optimal selection of $\tilde{u} \in L_2[1, \infty)$ is replaced by an optimal selection of $z \in L_2[1, \infty)$, subject to the additional qualitative requirement that the associated state of (22) be a member of $L_2[1, \infty)$.

This is precisely the framework to which one can reduce the ordinary version of the model-matching problem $\|N_l - M_l Q\|_\infty < \gamma$. Detailed analysis can be found, e.g., in [32]. The conclusions are summarized as follows.

THEOREM 5.2. *The infimum of values of γ for which the problem* (21) *has a finite solution is $\hat{\gamma} = \rho(X_{as}Y_{as})$. Given $\gamma > \hat{\gamma}$, define R as in* (10). *Then R is the unique, positive semidefinite solution of the deficient ARE,*

$$RA + A'R + R\left(\frac{1}{\gamma^2}B'B - QC'CQ\right)R = 0, \tag{23}$$

with the property that $A_1 = A + (\frac{1}{\gamma^2}B'B - QC'CQ)R$ is stable. Equivalently, $\gamma > \hat{\gamma}$ if and only if there exists a unique, $L_2[1, \infty)$ norm bounded solution to the following Hamilton–Jacobi–Bellman problem:

$$\begin{array}{rcccc} \dot{x} & = & Ax & + & \left(\frac{1}{\gamma^2}B'B - QC'CQ\right)p, \\ \dot{p} & = & & & -A'p \end{array} \tag{24}$$

given any specified initial value $x(1) \in \mathbb{R}^n$. That solution is characterized by the relation $p(t) = Rx(t)$, $t \geq 1$. A complete parameterization of the suboptimal set $\{\Theta \in H_\infty \ : \ \|N_l - M_l\Theta\|_\infty < \gamma\}$ is provided by the following realization:

$$\begin{array}{rcccccccc} \dot{x}_c & = & (A - QC'CQR)x_c & + & Bw & - & QC'\phi, & & \\ \tilde{u} & = & C(I - QR)x_c & + & Dw & - & \phi, & & \\ \psi & = & -\frac{1}{\gamma^2}B'Rx_c & + & w, & & & \phi = \Theta_0\psi, \end{array} \tag{25}$$

where the free design parameter, the stable system Θ_0*, is selected subject to the induced* L_2 *norm bound* $\|\Theta_0\| < \gamma$*. Moreover, given any selections* $w, \tilde{u} \in L_{2,loc}[1,\infty)$ *and* $t > r \geq 1$*, there holds*

$$\gamma^2\|w\|^2_{L_2[r,t]} - \|z\|^2_{L_2[r,t]} = \langle x(\alpha), Rx(\alpha)\rangle_e \Big|_{\alpha=r}^{\alpha=t} + \gamma^2\|w^{\triangle}\|^2_{L_2[r,t]} - \|\tilde{u}^{\triangle}\|^2_{L_2[r,t]}, \tag{26}$$

where, given the associated state x *in* (20)*, the trajectories* $w^{\triangle}$ *and* $\tilde{u}^{\triangle}$ *are defined as*

$$w^{\triangle} = w - \frac{1}{\gamma^2}B'Rx \qquad \text{and} \qquad \tilde{u}^{\triangle} = C(I-QR)x + Dw - \tilde{u}. \tag{27}$$

The solution to the open loop min-max game (21) *is given by the optimal selections* $w^{\triangle} = 0$ *and* $\tilde{u}^{\triangle} = 0$*; namely,*

$$w(t) = \frac{1}{\gamma^2}B'Rx(t) \quad \text{and} \quad u(t-1) = \tilde{u}(t) = \left(C(I-QR) + \frac{1}{\gamma^2}DB'R\right)x(t), \tag{28}$$

and the optimal value of the game is $-\langle x(1), Rx(1)\rangle_e$*.*

Obviously, the infimal value of γ for which a solution exists in (17) is at least the infimal value $\hat{\gamma}$ from (21). That is, $\gamma_0 \geq \hat{\gamma}$. Substituting $-\langle x(1), Rx(1)\rangle_e$ for the optimal value of the game (21) in (17), the original differential game (17) reduces to the following finite time optimization problem with the data $(x(0), u_0) \in M_2$:

$$\inf_{w\in L_2[0,1]} \left\{\gamma^2\|w\|^2_{L_2[0,1]} - \|z\|^2_{L_2[0,1]} - \langle x(1), Rx(1)\rangle_e\right\}. \tag{29}$$

When $\gamma > \rho_0$, the following equality is obtained by direct completion-of-squares computation and is used to simplify notation later on. There holds

$$\gamma^2\|w\|^2_{L_2[0,1]} - \|z\|^2_{L_2[0,1]} - \langle x(1), Rx(1)\rangle_e = \|\bar{w}\|^2_{L_2[0,1]} - \|\bar{z}\|^2_{L_2[0,1]} - \langle x(1), Rx(1)\rangle_e, \tag{30}$$

where

$$\bar{z}(t) = H_1x(t) + E_{12}u(t-1) \qquad \text{and} \qquad \bar{w}(t) = E_{21}^{-1}w(t) - \frac{1}{\gamma}D'\bar{z}(t). \tag{31}$$

Subject to these definitions, (16) becomes

$$\begin{array}{rcccccccc} \dot{x}(t) & = & (A-G_2C)x(t) & + & G_1\bar{w}(t) & + & G_2u(t-1), & & \\ \bar{z}(t) & = & H_1x(t) & + & & + & E_{12}u(t-1), & & \\ y(t) & = & H_2x(t) & + & E_{21}\bar{w}(t) & + & E_{22}u(t-1), & u = \Theta y, & \end{array} \tag{32}$$

where the abbreviated notation of (10) is used. Problem (29) can then be stated, equivalently, in the following form:

$$\inf_{\bar{w}\in L_2[0,1]} \left\{\|\bar{w}\|^2_{L_2[0,1]} - \|\bar{z}\|^2_{L_2[0,1]} - \langle x(1), Rx(1)\rangle_e\right\}. \tag{33}$$

Theorem 5.3, below, is based on results from [14, 26].

THEOREM 5.3. *Let* $\gamma > \hat{\gamma}$*. Then there exists a unique solution of the differential game* (17) *and, consequently, of the optimization problem* (29)*, given any data pair* $(x(0), u_0) \in M_2$*, if and only if conditions* (a) *and* (b) *in the statement of Theorem* 4.1 *are satisfied. Assume that to be the case. Then, for any initial data* $(x(0), u_0) \in M_2$*, there exists a unique* $L_2[0,1]$ *solution for the following Hamilton–Jacobi–Bellman problem:*

$$\begin{array}{rcccccc} \dot{x}(t) & = & (A-G_2C)x(t) & + & G_1G_1'p(t) & + & G_2u(t-1), \\ \dot{p}(t) & = & -H_1'H_1x(t) & - & (A-G_2C)'p(t) & - & H_1'E_{12}u(t-1) \end{array} \tag{34}$$

subject to the terminal constraint $p(1) = Rx(1)$. *In the particular case where* $u_0 = 0$, *the state and co-state of* (34) *are related via* $p(t) = R_0(t)x(t)$, *where* R_0 *is the solution of* (11). *The optimal selection of* w *in the problem* (29), *and the corresponding optimal selection of* $\bar{w}$ *in the problem* (33), *are then*

$$\bar{w}(t) = G_1'p(t) \qquad \textit{and} \qquad w(t) = H_2x(t) + E_{21}G_1'p(t) + E_{22}u(t-1). \tag{35}$$

The proof of necessity in Theorem 4.1 will be complete once the necessity claim in Theorem 5.3 is established. The detailed characterization of the solution of (29) is provided in preparation for the proof of the sufficiency claim in Theorem 4.1.

Proof. In the proof of necessity of the conditions (a) and (b) of Theorem 4.1 for solvability of (29), it certainly suffices to establish the same for the particular case of $u_0 = 0$. Referring to that case, it has been established in [14, 26] that a solution of (29) exists if and only if the induced norm of the mapping $\bar{w} \mapsto (R^{\frac{1}{2}}x(1), \bar{z})$: $L_2[0,1] \mapsto \mathbb{R}^n \times L_2[0,1]$ is smaller than 1, which, in turn, holds if and only if $\gamma^2 > \rho(D'D)$ and the differential Riccati equation (11) has a bounded solution. It is also a standard observation, used in [14, 26], that the state and co-state of the homogeneous (34) are related via $p = R_0x$ (when R_0 exists). The proof of necessity is therefore complete.

The fact that all solutions of (29) (hence, eventually, of (17)), should they exist, are characterized by (34) via (35), is a standard observation which can be established following the arguments in [14, 26] (or any other variant argument of the "Lagrange multiplier" type). Similarly, should a solution of (34) exist, it is a basic fact from optimal control theory that such a solution defines a solution of (29), as explained above. It remains to show that if $\gamma > \rho_0$ and a solution of (11) exists, then there also exists a unique solution of the inhomogeneous (34), given any initial data $(x(0), u_0) \in M_2$.

Indeed, let the definitions of the matrix function $A_0(t)$, and of the transition matrix $\Phi_0(t,s)$ that is generated by A_0, stand, as made in the statement of Theorem 4.1. Then let x and q be defined by the following boundary value problem:

$$\begin{array}{rcccccc} \dot{x}(t) & = & A_0(t)x(t) & - & G_1G_1'q(t) & + & G_2u(t-1), \\ \dot{q}(t) & = & & - & A_0(t)'q(t) & + & (R_0(t)G_2 + H_1'E_{12})u(t-1) \end{array} \tag{36}$$

subject to the specified initial value $x(0)$ and the terminal value $q(1) = 0$. Existence, uniqueness, and boundedness of the solution are guaranteed in (36) due to the fact that the second equation is independent of the first. All that remains is to observe that solutions of (34) and those of (36) are related by the equality $p = R_0x - q$. The proof is complete. □

This completes the proof of necessity of conditions (a) and (b) in Theorem 4.1 for γ to be a suboptimal value. From this point on the discussion continues under the assumption that conditions (a) and (b) in Theorem 4.1 are satisfied by a specified γ. In particular, solvability of both (23) and (11), of the differential game (17), and of the allied optimization problem (29) is assumed. The explicit goal of the ensuing analysis will be to establish the validity of the parameterization (12) of the suboptimal class $\mathbf{\Theta}_\gamma$, and thus, constructively, the sufficiency of the stated conditions for the inequality $\gamma > \gamma_0$.

LEMMA 5.4. *Fix* γ *satisfying conditions* (a) *and* (b) *in Theorem* 4.1, *and let the definitions made heretofore stand. Let* x, w, *and* u *be the optimal trajectories in the solution of* (17) *along* $[0,\infty)$; *let* p *be defined as the co-state trajectory in* (34), *along* $[0,1]$, *continued by the co-state trajectory of* (24) *along* $(1,\infty)$; *and let* $\bar{w}$ *be the*

allied trajectory, as defined in (31). *Then the following hold.* (a) *The co-state* $p(t)$ *is continuous.* (b) *The pair* (x,p) *satisfies the Hamilton–Jacobi–Bellman boundary value problem* (34) *when shifted from* $[0,1]$ *to any interval* $[t,t+1]$, $t>0$. (c) *At any given time* $t\geq 0$, *the values of* $p(t)$, $\bar{w}(t)$, *and* $u(t)$ *are determined by the pair* $(x(t),u_t)\in M_2$ *via*

$$p(t)=R_0(0)x(t)+\int_0^1\Phi_0(r,0)'(R_0(r)G_2+H_1'E_{12})u_t(r-1)dr, \tag{37}$$

$$\bar{w}(t)=G_1'\left(R_0(0)x(t)+\int_0^1\Phi_0(r,0)'(R_0(r)G_2+H_1'E_{12})u_t(r-1)dr\right), \tag{38}$$

$$\begin{aligned}u(t)=\left(C(I-QR)+\frac{1}{\gamma^2}DB'R\right)\cdot\Bigg(&\Phi_0(1,0)x(t)\\&+\int_0^1\Phi_0(1,r)((I+\Xi(r)R_0(r))G_2+\Xi(r)H_1'E_{12})u_t(r-1)dr\Bigg),\end{aligned} \tag{39}$$

where $\Xi(t)$ *is as defined in* (14). *For future reference we introduce the notations of matrices* K^0 *and* L^0 *and operators* K^1 *and* L^1 *over* $L_2[-1,0]$ *such that* (38) *and* (39) *are abbreviated as*

$$\begin{aligned}&\bar{w}(t)=L^0x(t)+L^1u_t,\qquad w(t)=(H_2+E_{21}L^0)x(t)+E_{21}L^1u_t+E_{22}u(t-1),\\&\qquad\qquad\qquad\qquad\text{and}\quad u(t)=K^0x(t)+K^1u_t.\end{aligned} \tag{40}$$

(d) *Combined, the complete state feedback expressions* (38) *and* (39) *are defining a stabilizing feedback policy in* (32) *and provide for exponential decay of the complete state*

$$\|f(t)\|_{M_2}\leq\alpha e^{-\beta t}\|f(0)\|_{M_2}, \tag{41}$$

where $f(t)=(x(t),u_t)$ *and where* $\alpha>0$ *and* $\beta>0$ *are appropriately selected and fixed constants.* (e) *The optimal value of* (17)–(29) *is given by the quadratic form* $-\langle f(0),\mathcal{R}f(0)\rangle_{M_2}$, *where* $f(0)=(x(0),u_0)$ *comprises the specified initial data and where* $\mathcal{R}$ *is the bounded, positive semidefinite operator on* M_2 *that is defined in terms of the solution of* (34), *as* $\mathcal{R}(x(0),u_0)=(p(0),\ E_{12}'(H_1x(\cdot+1)+E_{12}u_0(\cdot))+G_2'p(\cdot+1))$.

Proof. We have already established that the optimal state trajectory of (16) coincides with the state of (24) along $[1,\infty)$ and with the state of (34) along the ray $[0,1]$. The co-state in (24) is related to the state via $p=Rx$, and in (34) it is required to satisfy the terminal condition $p(1)=Rx(1)$. Hence the continuity of p, which is claim (a) in the statement of Lemma 5.4.

The purpose of the following observations is to show that the optimal trajectories x, u, and the associated p satisfy (34) when shifted to any interval $[t,t+1]$. First, using (31), the optimal selections in (28) translate to $\bar{w}(t)=G_1'Rx(t)$, $t\geq 1$. Thus the expression (35) for the optimal $\bar{w}$ is valid for all $t\geq 0$.

Second, using the expression in (28) for $u(t-1)$, substituting Rx for p, and taking into account the precise definitions of G_1 and G_2 in (10), the state equation in (34) becomes a stable, homogeneous ODE

$$\dot{x}=A_1x, \tag{42}$$

where $A_1 = A - (\frac{1}{\gamma^2}B'B - QC'CQ)R$, as defined in the statement of Theorem 5.2. Now substitute p for Rx in (42); this equation takes the shape of the state equation in (24). In short, we have established that the triplet x, p, and u satisfies the state equation in (34) throughout the positive ray $t \geq 0$.

Third, the combination of the expression (28) for $u(t-1)$ and of $p = Rx$ provides for the equality $C'G_2'p(t) - H_1'(H_1x(t) + E_{12}u(t-1)) = 0$, for all $t \geq 1$. Thus the co-state equation of (34) reduces to the co-state equation in (24). In particular, the triplet x, p, and u satisfies the co-state equation in (34) throughout the positive ray $t \geq 0$. This completes the proof of claim (b) in Lemma 5.4.

We shall now provide explicit expressions for the solution of (34), shifted to the interval $[t, t+1]$, in terms of the data $(x(t), u_t)$. These expressions are obtained by a straightforward manipulation of the variations-of-parameters formula, beginning with the second equation in (36), continuing with the first equation in the same system, and then substituting $p = R_0x - q$:

$$q(t+r) = -\int_r^1 \Phi_0(s,r)'(s)(R_0(s)G_2 + H_1'E_{12})u_t(s-1)ds, \tag{43}$$

$$\begin{aligned} x(t+r) = \Phi_0(r,0)x(t) &+ \int_0^r \Phi_0(r,s)((I + \Xi(s)R_0(s))G_2 + \Xi(s)H_1'E_{12})u_t(s-1)ds \\ &+ \ \Xi(r)\int_r^1 \Phi_0(s,r)'(R_0(s)G_2 + H_1'E_{12})u_t(s-1)ds, \end{aligned} \tag{44}$$

$$\begin{aligned} p(t+r) = R_0(r)\Phi_0(r,0)x(t) &\\ &+ \ R_0(r)\int_0^r \Phi_0(r,s)((I + \Xi(s)R_0(s))G_2 + \Xi(s)H_1'E_{12})u_t(s-1)ds \\ &+ \ (I + R_0(r)\Xi(r))\int_r^1 \Phi_0(s,r)'(R_0(s)G_2 + H_1'E_{12})u_t(s-1)ds. \end{aligned} \tag{45}$$

Substituting $r = 0$ in (45) one obtains the expression (37) for $p(t)$. The expression (38) is then a consequence of the relation $\bar{w} = G_1'p$. The expression (39) is obtained by substituting $r = 1$ in (44) and then using (28) to characterize $u(t)$ in terms of $x(t+1)$. This completes the proof of claim (c) in Lemma 5.4.

To establish the stability claim we first make note of the fact that the coupling of the complete state feedback expressions (38) and (39) with the state equation of (32) forms a well-posed integrodifferential equation. The uniqueness of its solution forces it to coincide with the trajectories that are associated with the solution of (17), as discussed above. Once again, it is also noticed that for $t \geq 1$ the trajectory of the optimal $x(t)$ is governed by the homogeneous, stable ODE (42). Invoking (28), there must exist positive α and β such that, for all $t \geq 1$, the following holds:

$$\|u(t-1)\|_e, \ \|x(t)\|_e \leq \alpha e^{-\beta(t-1)}\|x(1)\|_e. \tag{46}$$

Continuity of the mapping $(x(0), u_0) \mapsto (x(1), x(\cdot), u(\cdot)) : M_2 \mapsto \mathbb{R}^n \times L_2[0,1] \times L_2[0,1]$ is obvious from (44) and (39). Thus (with a possible need for a modification of α), the expression (46) leads to (41), and part (d) of Lemma 5.4 is established.

To compute the optimal value in (17)–(29) we first introduce notations of operators $\mathcal{M}$, $\mathcal{N}$, and $\mathcal{J}$. The definition $\mathcal{M}(x(0), p(0), u_0) = (x(1), x(\cdot), p(\cdot), u_0(\cdot + 1))$ is

made where the pair (x,p) defines the solution of (34), viewed as a causal initial value problem with the data $(x(0),p(0),u_0)$. The mapping $\mathcal{N}(x(0),u_0)=(x(0),p(0),u_0)$ translates the original data for the boundary value problem (34) to initial data, using (37) with $t=0$. The operator $\mathcal{J}$ is represented by the matrix

$$\mathcal{J}=\begin{bmatrix} R & 0 & 0 & 0\\ 0 & H_1'H_1 & 0 & H_1'E_{12}\\ 0 & 0 & -G_1'G_1 & 0\\ 0 & E_{12}'H_1 & 0 & E_{12}'E_{12}\end{bmatrix}. \tag{47}$$

In these terms the solution (x,p) of the boundary value problem (34) is satisfying $\mathcal{MN}(x(0),u_0)=(x(1),x(\cdot),p(\cdot),u_0(\cdot+1))$ and there holds $\mathcal{JMN}(x(0),u_0)=(Rx(1),H_1'\bar{z},-G_1'\bar{w},E_{12}'\bar{z})$, where $\bar{z}(t)=H_1x(t)+E_{12}u(t-1)$ and $\bar{w}(t)=G_1'p(t)$. It is a summary of our results heretofore that the optimal value of the cost functional in (33), hence in (29), is $-\langle\mathcal{MN}f(0),\,\mathcal{JMN}f(0)\rangle_{M_2}$. We thus denote $\mathcal{R}=\mathcal{N}'\mathcal{M}'\mathcal{JMN}$ and continue to show that $\mathcal{R}$ is realized by the solution of (34), as stated in the theorem. The following observations are made to that end.

The computation of the conjugate of the I/O operator that is defined in terms of a causal system is straightforward. In the case of $\mathcal{M}$, the relation $(\phi^1,\phi^2,\phi^3)=\mathcal{M}'(\psi^1,\psi^2,\psi^3,\psi^4)$ is characterized by $\phi^1=p_1(0)$, $\phi^2=-x_1(0)$, $\phi^3(\cdot)=E_{12}'H_1x_1(\cdot)+G_2'p_1(\cdot)+\psi^4(\cdot+1)$, where the pair (x_1,p_1) provides a solution of the following variant of (34):

$$\begin{aligned}\dot{x}_1 &= (A-G_2C)x_1 + G_1G_1'p_1 + \psi^3,\\ \dot{p}_1 &= -H_1'H_1x_1 - (A-G_2C)'p_1 - \psi^2,\end{aligned} \tag{48}$$

subject to the terminal conditions $x_1(1)=0$ and $p_1(1)=\psi^1$. In particular, when $(\psi^1,\psi^2,\psi^3,\psi^4)=\mathcal{JMN}(x(0),u_0)$, the pair (x_1,p_1) is defined in terms of the following system:

$$\begin{aligned}\dot{x}_1(t) &= (A-G_2C)x_1(t) + G_1G_1'(p_1-p)(t),\\ \dot{p}_1(t) &= -H_1'H_1(x+x_1)(t) - (A-G_2C)'p_1(t) - H_1'E_{12}u(t-1),\end{aligned} \tag{49}$$

with (x,p) from the original solution of (34), $x_1(1)=0$, and $p_1(1)=Rx(1)$. Denoting $x_2(t)=x(t)+x_1(t)-\Phi_0(t,0)x_1(0)$ and $p_2(t)=p_1(t)-R_0(t)\Phi_0(t,0)x_1(0)$ and taking derivatives, one observes that the pair (x_2,p_2) satisfies (34) with the same data and boundary conditions as (x,p). Following from the established uniqueness of the solution, $x_2=x$ and $p_2=p$.

Given $(\phi^1,\phi^2,\phi^3)\in\mathbb{R}^n\times\mathbb{R}^n\times L_2[-1,0]$, it follows directly from the definition of $\mathcal{N}$ that $\mathcal{N}'(\phi^1,\phi^2,\phi^3)=(\phi^1+R_0(0)\phi^2,\ (E_{12}'H_1+G_2'R_0(\cdot+1))\Phi_0(\cdot+1,0)\phi^2+\phi^3(\cdot))$. In particular, when $(\phi^1,\phi^2,\phi^3)=\mathcal{M}'\mathcal{JMN}(x(0),u_0)$, the value of $\mathcal{R}(x(0),u_0)=\mathcal{N}'\mathcal{M}'\mathcal{JMN}(x(0),u_0)$, as stated in the statement of part (e) of Lemma 5.4, is obtained. The proof of the lemma is complete. ∎

5.3. The abstract model. As mentioned earlier, the complete state of the process that is described by the delay differential equations (16)–(32) must include an account of the history of the control input, as provided by $f(t)=(f^0(t),f^1(t,s))=(x(t),u_t(s))\in M_2$. This section concerns a brief account of an abstract evolution model for the dynamics of the complete state $f(t)$. General background on c_0-semigroups and on distributed parameter systems can be found, e.g., in [4, 5, 10,

13, 18]. Abstract models for delay systems over the state space M_2 (and related spaces) have been studied extensively, and a sample of relevant sources and leads is [1, 2, 3, 6, 11, 19, 20, 21, 31, 34, 35]. As compared with most of these references, our system, and accordingly, the model we use, is relatively simple and dates back to the 1970s (see, e.g., [11]). We shall thus be content with a brief explanation of the model and refer the interested reader to existing literature for more details.

Starting with formal definitions, the abstract model for the complete state dynamics, input, and output in (16) will be written as follows:

$$\begin{array}{rcccccl} \dot f & = & \mathcal{A}f & + & \mathcal{B}_1 w & + & \mathcal{B}_2 u, \\ z & = & \mathcal{C}_1 f & + & \mathcal{D}_{11} w, & & \\ y & = & & & w, & & u = \Theta y, \end{array} \tag{50}$$

with the coefficients

$$\begin{array}{rclcrcl} \mathcal{A}f & = & \left((A-QC'C)f^0 + QC'f^1(-1), \frac{d}{ds}f^1\right), & & & & \\ \mathcal{B}_1 w & = & ((B-QC'D)w, 0), & & \mathcal{B}_2 u & = & (0, \delta_0 u), \\ \mathcal{C}_1 f & = & Cf^0 - f^1(-1), & & \mathcal{D}_{11} w & = & Dw, \end{array} \tag{51}$$

where δ_0 is the usual Dirac function, centered at the origin, and where the domain of the infinitesimal generator is

$$\mathcal{D}(\mathcal{A}) = \left\{ f \in M_2 \ : \ f^1 \in W_2^1[-1,0],\ f^1(0) = 0 \right\}. \tag{52}$$

The meaning of this model will be explained via the following series of observations.

LEMMA 5.5. *Let $\mathcal{S}(t)$ be the family of linear operators over M_2, as defined by the homogeneous dynamics in* (16) *(that is, with $w(t) = 0$ and $u(t) = 0$, $t > 0$) and the relation $(x(t), u_t) = \mathcal{S}(t)(x(0), u_0)$. Then $\mathcal{S}(t)$ is a c_0-semigroup, generated by $\mathcal{A}$, as defined above.*

Outline of the proof. The fact that the family $\mathcal{S}(t)$ adheres to the axioms of a c_0-semigroup ($\mathcal{S}(0) = \mathcal{I}$, $\mathcal{S}(t+s) = \mathcal{S}(t)\mathcal{S}(s)$ and strong continuity in t [10]) is clear. The stated forms of $\mathcal{A}$ and of its domain can be motivated by formal differentiation of $f(t) = (x(t), u_t)$ and by the fact that, in the homogeneous dynamics, $u_t(s) = 0$ for $t + s > 0$. A complete and rigorous proof of the validity of the stated forms of $\mathcal{A}$ and of its domain can be obtained, e.g., by adaptations of the arguments used in one of the following proofs: [5, Thm. 2.4.6], [3, Thm. 2.3], or [31, Thms. A and B].

The following remarks concern the adaptation of the frameworks of [3, 31], which originally concern neutral FDEs of the form

$$\frac{d}{dt}\mathcal{E}z_t = \mathcal{F}z_t, \qquad t > 0. \tag{53}$$

The current setting can be brought to this form with $z = (x, u)$, $\mathcal{E}(z_t) = z_t(0) = z(t)$, and $\mathcal{F}(x_t, u_t) = ((A - QC'C)x_t(0) + QC'u_t(-1),\ 0)$. In the framework of the cited papers this would have called for the use of the higher-dimensional "M_2" state space $\mathbb{R}^{n+l} \times L_2([-1,0], \mathbb{R}^{n+l})$ (where $C \in \mathbb{R}^{l \times n}$), with the complete state $(\mathcal{E}z_t, z_t)$. The current simplification is due to the following two facts. (a) The last l entries of $\mathcal{F}z_t$ vanish, making the subspace where the last l entries of $\mathcal{E}(x_t, u_t)$ are zero, an invariant subspace under the evolution of (53). Focusing on that subspace, the last l entries of $\mathcal{E}z_t = \mathcal{E}(x_t, u_t)$ can be removed from the state $(\mathcal{E}z_t, z_t)$. (b) The dependence on

x_t in both $\mathcal{E}z_t$ and $\mathcal{F}z_t$ is restricted to the value of $x(t) = x_t(0)$. Thus, no needed information is lost when the state component $x_t \in L_2[-1, 0]$ is replaced by $x(t) \in \mathbb{R}^n$ and the definitions of $\mathcal{E}$ and $\mathcal{F}$ are modified accordingly. Following these modifications, the current state choice of $f(t) = (x(t), u_t)$ is obtained, while the arguments in the cited articles concerning the form of the domain of the infinitesimal generator remain valid. □

In what follows, we shall encounter several other semigroups over M_2 and provide, without proof, the forms of their generators and their respective domains. Indeed, in each of these cases, one will be able to draw on arguments from [3, 31] to verify the association of the semigroup and the generator. In particular, in each of these cases, the dynamics under consideration will be generated by a neutral FDE of the form (53), to which the comments in the outline of the proof of Lemma 5.5 apply. This association will be key to our ability to fairly freely consider perturbations of $\mathcal{S}(t)$, an issue that is quite delicate in the general framework of c_0-semigroup theory.

For later reference we write down the explicit form of the relationship $f(t) = \mathcal{S}(t)f(0)$, as derived from the variation-of-parameters formula in (16):

$$
\begin{aligned}
f^0(t) &= e^{(A-QC'C)t}f^0(0) + \int_0^{\min(t,1)} e^{(A-QC'C)(t-s)}QC'f^1(0,\ s-1)ds, \\
f^1(t,\theta) &= \begin{cases} f^1(0, t+\theta) & -1 \le \theta \le -t, \quad 0 \le t < 1, \\ 0 & \text{else.} \end{cases}
\end{aligned} \tag{54}
$$

As is well known, a restriction of $\mathcal{S}(t)$ to the dense subspace $\mathcal{D}(\mathcal{A}) \subset M_2$ defines a c_0-semigroup over $\mathcal{D}(\mathcal{A})$ relative to the stronger graph($\mathcal{A}$) topology. Also, the definition of $\mathcal{S}(t)$ extends, by dense injection, to a c_0-semigroup over the larger space $\mathcal{D}(\mathcal{A}')' \supset M_2$. Such extensions and restrictions are used extensively, e.g., in the much more general discussions in [21, 19]. The definition of the restriction to $\mathcal{D}(\mathcal{A})$ is obvious. The following details concern the extension to $\mathcal{D}(\mathcal{A}')'$.

By direct computation one finds the form of $\mathcal{A}'$,

$$
\mathcal{A}'g = \left((A - QC'C)'g^0,\ -\ \frac{d}{ds}g^1 \right), \tag{55}
$$

and the domain,

$$
\mathcal{D}(\mathcal{A}') = \left\{ g \in M_2 \ :\ g^1 \in W_2^1[-1, 0],\ \ g^1(-1) = CQg^0 \right\}. \tag{56}
$$

Integration by parts and (56) yield the following equality for any $f \in M_2$ and $g \in \mathcal{D}(\mathcal{A}')$:

$$
\langle f, g \rangle_{M_2} = \left\langle \left(f^0 + QC' \int_{-1}^0 f^1(r)dr, \int_s^0 f^1(r)dr \right),\ \left(g^0, \frac{d}{ds}g^1(s) \right) \right\rangle_{M_2}. \tag{57}
$$

The emerging representation of elements $g \in \mathcal{D}(\mathcal{A}')$ by the associated pairs $(g^0, \frac{d}{ds}g^1) \in M_2$ defines an isomorphism between $\mathcal{D}(\mathcal{A}')$ and M_2, and the definition of a norm $\|g\|_{\mathcal{D}(\mathcal{A}')} = \|(g^0, \frac{d}{ds}g^1)\|_{M_2}$ is consistent with the graph($\mathcal{A}$) topology in $\mathcal{D}(\mathcal{A}')$. Using this representation, an element $h \in \mathcal{D}(\mathcal{A}')'$ must be determined by an associated element $(h^0, h^1) \in M_2$, via the pairing

$$
\langle h, g \rangle_{\mathcal{D}(\mathcal{A}')'}, \mathcal{D}(\mathcal{A}') = \left\langle (h^0, h^1)\ ,\ \left(g^0, \frac{d}{ds}g^1(s) \right) \right\rangle_{M_2}. \tag{58}
$$

Thus, the definition of a norm $\|h\|_{\mathcal{D}(\mathcal{A}')'} = \|(h^0, h^1)\|_{M_2}$ is consistent with the adjoint space topology of $\mathcal{D}(\mathcal{A}')'$.

Referring to the representation of elements of $\mathcal{D}(\mathcal{A}')$ and of $\mathcal{D}(\mathcal{A}')'$ by M_2 pairs, as explained above, one defines the following continuous injection $\imath : M_2 \mapsto \mathcal{D}(\mathcal{A}')'$, its unbounded inverse $\pi : \mathcal{D}(\mathcal{A}')' \mapsto M_2$, and their adjoints:

$$
\begin{aligned}
\imath(f) &= \left(f^0 + QC' \int_{-1}^{0} f^1(r)dr \ , \ \int_s^0 f^1(r)dr \right), \\
\pi(h) &= \left(h^0 - QC'h^1(-1) \ , \ -\frac{d}{ds}h^1(s) \right), \\
\imath'(\phi) &= \left(\phi^0 \ , \ CQ\phi^0 + \int_{-1}^{s} \phi^1(r)dr \right), \\
\pi'(\psi) &= \left(\psi^0 \ , \ \frac{d}{ds}\psi(s) \right).
\end{aligned}
\tag{59}
$$

The extended semigroup is defined as a continuous continuation of $\imath \circ \mathcal{S}(t) \circ \pi$ from the dense submanifold $\imath(M_2)$ to the entire $\mathcal{D}(\mathcal{A}')'$. For simplicity in what follows, we shall use the same notation for the original semigroup $\mathcal{S}(t)$ and for its extension. (The same notational policy will be applied to other semigroups that will be encountered later on.) The following expression for the relation $h(t) = \mathcal{S}(t)h(0)$, $h(0) \in \mathcal{D}(\mathcal{A}')'$, in the extended semigroup, is easily obtained, appealing to the original definition of $\mathcal{S}(t)$ over M_2 and to the definition of its extension:

$$
h^0(t) = e^{(A-QC'C)t}h^0(0) - \int_0^{\min(t,1)} e^{(A-QC'C)(t-s)}(A - QC'C)QC'h^1(0, s-1)ds,
$$

$$
h^1(t,\theta) = \begin{cases} h^1(0, t+\theta), & -1 \le \theta \le -t, \quad 0 \le t < 1, \\ \\ 0 & \text{otherwise.} \end{cases}
\tag{60}
$$

The input operator $\mathcal{B}_1$ takes values in M_2. The value $h = \mathcal{B}_1 w$ is interpreted as a member of the extended state space $\mathcal{D}(\mathcal{A}')'$ via the injection $\imath$. That is, $h = ((B - QC'D)w, 0)$. As defined, the operator $\mathcal{B}_2$ takes values in $\mathcal{D}(\mathcal{A}')'$. With the M_2 representation of elements of $\mathcal{D}(\mathcal{A}')'$, as described above, $h = \mathcal{B}_2 u$ is identified with the pair $h = (QC'u, \ 1(s)u)$, where "$1(s)$" is the $L_2[-1, 0]$ function that takes the unit value throughout. Thus interpreted, $\mathcal{B}_2$ is a bounded operator as well. The state response of the inhomogeneous (50) is defined in terms of "mild solutions"; that is, it is defined in terms of the variations-of-parameters formula, which is well defined over the extended state space $\mathcal{D}(\mathcal{A}')'$:

$$
h(t) = \mathcal{S}(t)h(0) + \int_0^t \mathcal{S}(t-r)(\mathcal{B}_1 w(r) + \mathcal{B}_2 u(r))dr.
\tag{61}
$$

LEMMA 5.6. *Given any initial data $f(0) = (x(0), u_0) \in M_2$, $h(0) = \imath(f(0)) \in \mathcal{D}(\mathcal{A}')'$, and inputs $w, u \in L_{2\ loc}[0, \infty)$, let $h(t)$ be the response of* (61) *and let $f(t) = (x(t), u_t)$ be defined in terms of the response of* (16). *Then $h(t) = \imath(f(t))$ throughout.*

Proof. The fact that the contribution of the initial data is as stated follows from the original definition of the semigroup $\mathcal{S}(t)$ over M_2 and its extension to $\mathcal{D}(\mathcal{A}')'$. As follows from the definition of $\mathcal{S}(t)$ and of $\mathcal{B}_1$, the term $\mathcal{S}(t-r)\mathcal{B}_1 w(r)$ has a

zero $L_2[-1,0]$ component, and the $\mathbb{R}^n$ component is identical to the one from the variations-of-parameters formula in (16). So the claim concerning the contribution of w is clear as well. Assuming $w=0$ and $f(0)=0$, we thus focus on the contribution of u.

As follows from (60), the $L_2[-1,0]$ component of $\mathcal{S}(t-r)\mathcal{B}_2u(r)$, denoted $\phi^1(t,r,\theta)$, is

$$\phi^1(t,r,\theta)=\begin{cases} u(r), & \max(0,t+\theta)\le r\le t,\quad \theta\in[-1,0],\\ 0, & \text{else,}\end{cases} \tag{62}$$

and the $\mathbb{R}^n$ component, denoted $\phi^0(t,r)$, is

$$\begin{aligned}\phi^0(t,r)&=\left(e^{(A-QC'C)(t-r)}-\int_0^{\min(t-r,1)}e^{(A-QC'C)(t-r-s)}(A-QC'C)ds\right)QC'u(r),\\ &=\begin{cases} QC'u(r), & \max(t-1,0)\le r\le t,\\ e^{A(t-r-1)}QC'u(r), & \text{else.}\end{cases}\end{aligned} \tag{63}$$

Integrating over the interval $r\in[0,t]$, we thus get

$$\begin{aligned}h^0(t)&=\int_0^{\max(t-1,0)}e^{(A-QC'C)(t-r-1)}QC'u(r)dr+QC'\int_{\max(t-1,0)}^t u(r)dr\\ &=\int_{\min(1,t)}^t e^{(A-QC'C)(t-r)}QC'u(r-1)dr+QC'\int_{\max(-1,-t)}^0 u_t(r)dr,\\ h^1(t,\theta)&=\int_{\max(0,t+\theta)}^t u(r)dr=\int_{\max(-t,\theta)}^0 u_t(r)dr.\end{aligned} \tag{64}$$

Comparing with the definition of $f(t)$ by the variations-of-parameters formula in (16) (with $x(0)=0$,$u_0=0$, and $w(t)=0$, $t>0$) and the definition of the injection $\imath$, it is obvious that, here too, $h(t)=\imath(f(t))$. □

It is noticed that while $\mathcal{C}_1$ is unbounded over M_2, it does define a bounded operator when restricted to the dense submanifold $M_2^1=\mathbb{R}^n\times W_2^1[-1,0]$. Following from the previous lemma, when $u\in W_{2\ loc}^1[-1,\infty)$, the state response in (50), as defined via (61), is $f(t)=(x(t),u_t)\in M_2^1$. Thus, $z(t)=\mathcal{C}_1f(t)+\mathcal{D}_{11}w(t)$ is a well-defined trajectory, coinciding with the value of the controlled output trajectory in (16). Furthermore, embedding trajectories of z with the $L_{2\ loc}[0,\infty)$ topology, the mapping $(f(0),u,w)\mapsto z$ extends to a continuous mapping over the entire $M_2\times L_{2\ loc}[0,\infty)\times L_{2\ loc}[0,\infty)$. The output equation in (50), as well as in other abstract models that will be considered hereafter, are to be understood in this sense. Similarly, for notational convenience, we shall identify the left-hand side of the following equality (and similar equalities that will be encountered) with the well-defined right-hand side (and its appropriate counterparts),

$$\left\langle f\ ,\ \int_0^t \mathcal{S}(r)'\mathcal{C}_1'\mathcal{C}_1\mathcal{S}(r)dr\ f\right\rangle_{M_2}=\int_0^t\langle\mathcal{C}_1\mathcal{S}(r)f\ ,\ \mathcal{C}_1\mathcal{S}(r)f\rangle_{M_2}dr.$$

5.4. Solution of the differential game: A complete state space description. The following two lemmas set the groundwork for the completion of the proof of Theorem 4.1.

LEMMA 5.7. *Assume that conditions* (a) *and* (b) *in Theorem* 4.1 *are satisfied, whereby a unique solution of the differential game* (17) *does exist, as shown above. Let a family of mappings* $f(t) = \mathcal{S}_1(t)f(0)$, $t \geq 0$, *be defined by shifts along solutions of* (17). *That is,* $f(t) = (x(t), u_t)$, *where* x *and* u *are the optimal trajectory and control in* (17), *given the initial data* $f(0) = (x(0), u_0) \in M_2$. *Then* $\mathcal{S}_1$ *defines an exponentially stable* c_0*-semigroup over* M_2, *generated by*

$$\mathcal{A}_1 f = \left((A - G_2C + G_1L^0)f^0 + G_1L^1f^1 + G_2f^1(-1),\ \frac{d}{ds}f^1 \right) \tag{65}$$

as defined over the domain

$$\mathcal{D}(\mathcal{A}_1) = \left\{ f \in M_2\ :\ f^1 \in W_2^1[-1,0],\quad f^1(0) = K^0f^0 + K^1f^1 \right\}, \tag{66}$$

where the notation of (40) *is used.*

Proof. The following facts imply that, indeed, $\mathcal{S}_1$ is a well-defined, exponentially stable c_0-semigroup. (a) The solution of (17) is unique, given the initial data. (b) The restriction of the solution of (17) to any positive ray $[t, \infty)$ must coincide with the solution of the restriction of (17) to that ray, with the initial data $f(t) = (x(t), u_t)$. (This follows from the observation in part (b) of Lemma 5.4.) (c) The stability observation in part (d) in Lemma 5.4.

An intuitive motivation of (65)–(66) as the correct form of the generator of $\mathcal{S}(t)$ and of its domain can be motivated by formal differentiation. Here too, an adaptation of the more general arguments concerning semigroup representations of neutral FDEs in [3, 31] will provide a complete and rigorous proof. The following are counterparts of the similar remarks in the outlined proof of Lemma 5.5, concerning that adaptation. As was established in Lemma 5.4, the optimal inputs in the solution of (17) are given by (40). Thus the evolution of the optimal trajectory $(x(t), u_t)$ is governed by the unique solution of the integrodifferential equation, coupling (16) and (40). The coupled system can be brought to the neutral FDE form of (53) with $z = (x, u)$ and with

$$\begin{aligned}
\mathcal{F}(x_t, u_t) &= ((A - G_2C + G_1L^0)x_t(0) + G_1L^1u_t + G_2u_t(-1),\ 0),\\
\mathcal{E}(x_t, u_t) &= (x_t(0),\ u_t(0) - K^0x_t(0) - K^1u_t).
\end{aligned}$$

As we focus on the invariant subspace where the last l entries of $\mathcal{E}(x_t, u_t)$ vanish, the definitions of the operators $\mathcal{E}$ and $\mathcal{F}$ allow us to simplify the state from the standard choice of $(\mathcal{E}z_t, z_t)$ to $f(t)$ and yet maintain the validity of the arguments from [3, 31]. ∎

With any initial state $f(0) = (x(0), u_0) \in M_2$, inputs $w, u \in L_{2\ loc}[0, \infty)$, and the associated response in (16), we associate trajectories

$$w^\nabla(t) = \bar{w}(t) - L^0x(t) - L^1u_t \quad \text{and} \quad u^\nabla(t) = (2E_{12}'E_{12})^{\frac{1}{2}}(u(t) - K^0x(t) - K^1u_t). \tag{67}$$

The state dynamics and the output of (16)–(50) can be represented by the following abstract equation:

$$\begin{aligned}
\dot{f} &= \mathcal{A}_1 f \quad + \quad \bar{\mathcal{B}}_1 w^\nabla \quad + \quad \bar{\mathcal{B}}_2 u^\nabla,\\
\bar{z} &= \bar{\mathcal{C}}_1 f,
\end{aligned} \tag{68}$$

where

$$\bar{\mathcal{B}}_1 w^\nabla = (G_1 w^\nabla, 0), \quad \bar{\mathcal{B}}_2 u^\nabla = (2E_{12}'E_{12})^{-\frac{1}{2}}\mathcal{B}_2 u^\nabla, \quad \bar{\mathcal{C}}_1 f = H_1 f^0 + E_{12} f^1(-1). \tag{69}$$

Here, again, the state equation is to be understood in the "mild" sense, in terms of the variations-of-parameters formula:

$$f(t) = \mathcal{S}_1(t)f(0) + \int_0^t \mathcal{S}_1(t-r)(\bar{\mathcal{B}}_1 w^\nabla(r) + \bar{\mathcal{B}}_2 u^\nabla(r))dr. \tag{70}$$

In complete analogy to Lemma 5.6 it is observed that the state $f(t)$ in (70) coincides with $(x(t), u_t)$ in the solution of (32) with $\bar{w}(t) = w^\nabla(t) + L^0x(t) + L^1u_t$ and with $u(t) = (2E_{12}'E_{12})^{-\frac{1}{2}}u^\nabla(t) + K^0x(t) + K^1u_t$. Using the notation of (69) and the explicit expression for $\mathcal{R}$ in part (e) of Lemma 5.4, it is useful to note that there holds

$$w^\nabla(t) = \bar{w}(t) - \bar{\mathcal{B}}_1'\mathcal{R}f(t) \qquad \text{and} \qquad u^\nabla(t) = 2\bar{\mathcal{B}}_2'\mathcal{R}f(t). \tag{71}$$

LEMMA 5.8. *Let the hypotheses of Lemma* 5.7 *stand. Then the following hold.* (a) *The operator* $\mathcal{R}$*, as defined in Lemma* 5.4*, satisfies the following operator Riccati equation, for each* $f \in M_2$*:*

$$\langle f, \mathcal{R}f\rangle_{M_2} = \int_0^\infty \langle f,\ \mathcal{S}_1(t)'\left(\bar{\mathcal{C}}_1'\bar{C}_1 - \mathcal{R}\bar{\mathcal{B}}_1\bar{\mathcal{B}}_1'\mathcal{R}\right)\mathcal{S}_1(t)f\rangle_{M_2}dt. \tag{72}$$

(b) *The following equality is satisfied over any finite time interval:*

$$\|\bar{w}\|^2_{L_2[0,t]} - \|\bar{z}\|^2_{L_2[0,t]} - \langle f(t), \mathcal{R}f(t)\rangle_{M_2} = \|w^\nabla\|^2_{L_2[0,t]} - \|u^\nabla\|^2_{L_2[0,t]} - \langle f(0), \mathcal{R}f(0)\rangle_{M_2}. \tag{73}$$

Proof. It will be convenient to use the abbreviation $\Delta = \mathcal{R}\bar{\mathcal{B}}_1\bar{\mathcal{B}}_1'\mathcal{R} - \bar{\mathcal{C}}_1'\bar{\mathcal{C}}_1$. Two facts are used in establishing (72). First, following from (71), the feedback expression for $\bar{w}$ in (40) is equivalent to $\bar{w} = \bar{\mathcal{B}}_1'\mathcal{R}f$. Second, by part (e) of Lemma 5.4, the optimal cost in (17) is $-\langle f(0), \mathcal{R}f(0)\rangle_{M_2}$. Combining these observations with the definition of $\mathcal{S}_1(t)$, we have

$$\begin{aligned} -\langle f(0), \mathcal{R}f(0)\rangle_{M_2} &= \|\bar{w}\|_2^2 - \|\bar{z}\|_2^2 \\ &= \|\bar{\mathcal{B}}_1'\mathcal{R}f\|_2^2 - \|\bar{\mathcal{C}}_1 f\|_2^2 = \textstyle\int_0^\infty \langle f(0), \mathcal{S}_1(t)'\Delta\mathcal{S}_1(t)f(0)\rangle_{M_2}dt, \end{aligned} \tag{74}$$

which is (72).

The following fact is an immediate consequence of (72):

$$\langle f, \mathcal{R}f\rangle_{M_2} - \langle f, \mathcal{S}_1(t)'\mathcal{R}\mathcal{S}_1(t)f\rangle_{M_2} + \int_0^t \langle f\ ,\ \mathcal{S}_1(r)'\Delta\mathcal{S}_1(r)f\rangle_{M_2}dt = 0, \tag{75}$$

holding for any $t > 0$ and $f \in M_2$. This fact will be used in these next derivations of (73):

$$\begin{aligned} &\|\bar{w}\|^2_{L_2[0,t]} - \|\bar{z}\|^2_{L_2[0,t]} - \langle f(t), \mathcal{R}f(t)\rangle_{M_2} + \langle f(0), \mathcal{R}f(0)\rangle_{M_2} \\ &\quad = \|w^\nabla + \bar{\mathcal{B}}_1'\mathcal{R}f\|^2_{L_2[0,t]} - \|\bar{\mathcal{C}}_1 f\|^2_{L_2[0,t]} - \langle f(t), \mathcal{R}f(t)\rangle_{M_2} + \langle f(0), \mathcal{R}f(0)\rangle_{M_2} \\ &\quad = \int_0^t \left\langle \left(\mathcal{S}_1(r)f(0) + \int_0^r \mathcal{S}_1(r-s)(\bar{\mathcal{B}}_1 w^\nabla(s) + \bar{\mathcal{B}}_2 u^\nabla(s))ds\right), \right. \end{aligned}$$

$$
\begin{aligned}
&\Delta\left(\mathcal{S}_1(r)f(0)+\int_0^r \mathcal{S}_1(r-q)(\bar{\mathcal{B}}_1 w^{\nabla}(q)+\bar{\mathcal{B}}_2 u^{\nabla}(q))dq\right)\Big\rangle_{M_2} dr\\
&+\ 2\int_0^t\Big\langle w^{\nabla}(s)\ ,\ \bar{\mathcal{B}}_1'\mathcal{R}\left(\mathcal{S}_1(s)f(0)+\int_0^s \mathcal{S}_1(s-q)(\bar{\mathcal{B}}_1 w^{\nabla}(q)+\bar{\mathcal{B}}_2 u^{\nabla}(q))dq\right)\Big\rangle_{M_2} ds\\
&+\ \|w^{\nabla}\|^2_{L_2[0,t]}\\
&-\ \Big\langle\left(\mathcal{S}_1(t)f(0)+\int_0^t \mathcal{S}_1(t-s)(\bar{\mathcal{B}}_1 w^{\nabla}(s)+\bar{\mathcal{B}}_2 u^{\nabla}(s))ds\right),\\
&\qquad\mathcal{R}\left(\mathcal{S}_1(t)f(0)+\int_0^t \mathcal{S}_1(t-q)(\bar{\mathcal{B}}_1 w^{\nabla}(q)+\bar{\mathcal{B}}_2 u^{\nabla}(q))dq\right)\Big\rangle_{M_2}\\
&+\ \langle f(0),\mathcal{R}f(0)\rangle_{M_2}\\
=\ &\left\langle f(0),\left(\mathcal{R}-\mathcal{S}_1(t)'\mathcal{R}\mathcal{S}_1(t)+\int_0^t \mathcal{S}_1(r)'\Delta\mathcal{S}_1(r)dr\right)f(0)\right\rangle_{M_2}\\
&+\ 2\left\langle f(0),\int_0^t \mathcal{S}_1(s)'\left(\Delta\int_0^s \mathcal{S}_1(s-q)\left(\bar{\mathcal{B}}_1 w^{\nabla}(q)+\bar{\mathcal{B}}_2 u^{\nabla}(q)\right)dq+\mathcal{R}\bar{\mathcal{B}}_1 w^{\nabla}(s)\right)ds\right\rangle_{M_2}\\
&+\ 2\left\langle f(0)\ ,\ \mathcal{S}_1(t)'\mathcal{R}\int_0^t \mathcal{S}_1(t-q)\left(\bar{\mathcal{B}}_1 w^{\nabla}(q)+\bar{\mathcal{B}}_2 u^{\nabla}(q)\right)dq\right\rangle_{M_2}\\
&+\ 2\int_0^t\int_0^r\Big\langle\mathcal{S}_1(r-s)\quad\left(\bar{\mathcal{B}}_1 w^{\nabla}(s)+\bar{\mathcal{B}}_2 u^{\nabla}(s)\right),\\
&\qquad\Delta\mathcal{S}_1(r-s)\int_0^s \mathcal{S}_1(s-q)\left(\bar{\mathcal{B}}_1 w^{\nabla}(q)+\bar{\mathcal{B}}_2 u^{\nabla}(q)\right)dq\Big\rangle_{M_2} ds\ dr\\
&+\ 2\int_0^t\left\langle w^{\nabla}(s)\ ,\ \bar{\mathcal{B}}_1'\mathcal{R}\int_0^s \mathcal{S}_1(s-q)\left(\bar{\mathcal{B}}_1 w^{\nabla}(q)+\bar{\mathcal{B}}_2 u^{\nabla}(q)\right)dq\right\rangle_{M_2} ds\\
&+\ \|w^{\nabla}\|^2_{L_2[0,t]}\\
&-\ 2\int_0^t\Big\langle\mathcal{S}_1(t-s)\quad\left(\bar{\mathcal{B}}_1 w^{\nabla}(s)+\bar{\mathcal{B}}_2 u^{\nabla}(s)\right),\\
&\qquad\mathcal{R}\mathcal{S}_1(t-s)\int_0^s \mathcal{S}_1(s-q)\left(\bar{\mathcal{B}}_1 w^{\nabla}(q)+\bar{\mathcal{B}}_2 u^{\nabla}(q)\right)dq\Big\rangle_{M_2} ds\\
=\ &2\left\langle f(0),\ \int_0^t \mathcal{S}_1(q)'\left(\int_0^{t-q}\mathcal{S}_1(r)'\Delta\mathcal{S}_1(r)dr+\mathcal{R}-\mathcal{S}_1(t-q)'\mathcal{R}\mathcal{S}_1(t-q)\right)\right.\\
&\qquad\left.\cdot\left(\bar{\mathcal{B}}_1 w^{\nabla}(q)+\bar{\mathcal{B}}_2 u^{\nabla}(q)\right)dq\right\rangle_{M_2} ds
\end{aligned}
\tag{76}
$$

$$
\begin{aligned}
&- \ 2\int_0^t \left\langle u^\nabla(q)\ ,\ \bar{\mathcal{B}}_2'\mathcal{R}\mathcal{S}_1(q)f(0)\right\rangle_{M_2} dq + \|w^\nabla\|^2_{L_2[0,t]} \\
&+ \ 2\int_0^t \left\langle w^\nabla(s)\ ,\ \bar{\mathcal{B}}_1'\left(\int_0^{t-s}\mathcal{S}_1(r)'\Delta\mathcal{S}_1(r)dr + \mathcal{R} - \mathcal{S}_1(t-s)'\mathcal{R}\mathcal{S}_1(t-s)\right)\right. \\
&\qquad \left.\cdot\int_0^s \mathcal{S}_1(s-q)\left(\bar{\mathcal{B}}_1 w^\nabla(q) + \bar{\mathcal{B}}_2 u^\nabla(q)\right)dq\right\rangle_{M_2} ds \\
&+ \ 2\int_0^t \left\langle u^\nabla(s),\ \bar{\mathcal{B}}_2'\left(\int_0^{t-s}\mathcal{S}_1(r)'\Delta\mathcal{S}_1(r)dr - \mathcal{S}_1(t-s)'\mathcal{R}\mathcal{S}_1(t-s)\right)\right. \\
&\qquad \left.\cdot\int_0^s \mathcal{S}_1(s-q)\left(\bar{\mathcal{B}}_1 w^\nabla(q) + \bar{\mathcal{B}}_2 u^\nabla(q)\right)dq\right\rangle_{M_2} ds \\
&= \ \|w^\nabla\|^2_{L_2[0,t]} - \int_0^t \left\langle u^\nabla(s)\ ,\ 2\bar{\mathcal{B}}_2'\mathcal{R}f(s)\right\rangle_{M_2} ds \\
&= \ \|w^\nabla\|^2_{L_2[0,t]} - \|u^\nabla\|^2_{L_2[0,t]},
\end{aligned}
$$

which establish the validity of (73) and complete the proof of Lemma 5.8. □

LEMMA 5.9. *The family of admissible, strictly γ suboptimal selections of the system Θ comprises those systems that can be realized by the following abstract models:*

$$
\begin{aligned}
\dot{f}_c &= \mathcal{A}_c f_c + \mathcal{B}_{c1} w + \mathcal{B}_{c2}\phi, \\
u &= \mathcal{C}_{c1} f_c + \mathcal{D}_{c12}\phi, \qquad (77)\\
\psi &= \mathcal{C}_{c2} f_c + \mathcal{D}_{c21} w, \qquad \phi = \Theta_0\psi,
\end{aligned}
$$

where the coefficients are defined as follows:

$$
\begin{aligned}
\mathcal{A}_c f_c &= \left((A - QC'C)f_c^0 + QC' f_c^1(-1)\ ,\ \frac{d}{ds}f_c^1\right), \\
\mathcal{B}_{c1} &= \mathcal{B}_1, & \mathcal{B}_{c2} &= \bar{\mathcal{B}}_2, \\
\mathcal{C}_{c1} f_c &= K^0 f_c^0 + K^1 f_c^1, & \mathcal{D}_{c12} &= (2E_{12}'E_{12})^{-\frac{1}{2}}, \\
\mathcal{C}_{c2} f_c &= -\left(L^0 + \frac{1}{\gamma}D'H_1\right)f_c^0 - L^1 f_c^1 - \frac{1}{\gamma}D'E_{12}f_c^1(-1), & \mathcal{D}_{c21} &= E_{21}^{-1},
\end{aligned} \qquad (78)
$$

where

$$
\mathcal{D}(\mathcal{A}_c) = \left\{f_c \in M_2\ :\ f_c^1 \in W_2^1[-1,0],\quad f_c^1(0) = K^0 f_c^0 + K^1 f_c^1\right\} \qquad (79)
$$

and where the free design parameter, the mapping Θ_0, is defined by the I/O mapping in a stable, neutral FDE with the $L_2[0,\infty)$ induced norm bound $\|\Theta_0\| < 1$.

Proof. The arguments are essentially the same as in the proof of a counterpart statement in the finite-dimensional framework of [32], where the equality (73) is playing the key role. For completeness, we shall outline the arguments in the current distributed parameter framework. These arguments are broken into several steps.

Step 1. *Well-Posedness and a Neutral FDE Interpretation of* (77). Again we draw on observations, similar to those made in section 5.3, concerning neutral FDEs and associated semigroups. Following from the definition of $\mathcal{A}_c$ (and of the rest of the coefficients) it is noted that (77) is defined in terms of a closed-loop interconnection of two well-posed, inhomogeneous neutral FDEs. The first subsystem is the stable neutral FDE that defines Θ_0. The second neutral FDE is defined in terms of the explicit state space formulas with Θ_0 removed. A close look reveals that this abstract model is a realization of an integrodifferential equation where the homogeneous part couples (16), say, with the state "x_c" and the regression $u(t) = K^0 x_c(t) + K^1 u_t$. The neutral FDE interpretation of this equation is similar to the one provided in the proof of Lemma 5.4 in the context of $\mathcal{S}_1(t)$, omitting the terms in L^0 and L^1. The inhomogeneous system includes the added input $(2E_{12}'E_{12})^{-\frac{1}{2}}\phi(t)$ in the regression for $u(t)$, and thus creates a well-posed inhomogeneous neutral FDE. (Details on inhomogeneous neutral FDEs can be found in [31].) Removing Θ_0, the open-loop mapping $\phi \mapsto \psi$ is strictly proper, whereby the interconnection results in a well-posed neutral FDE. (Indeed, the restriction of Θ_0 to the class of neutral FDEs is made expressly for this purpose, and the class of admissible "Θ_0" could easily be extended to any distributed parameter systems class that will provide for a meaningful well-posedness and state space realizability of (77).)

Since the state is defined in terms of trajectories of (16), the equality (73) remains valid, where f_c substitute for f throughout. In particular, it is noted that, with respect to the state f_c, $w^\nabla = \psi$, $u^\nabla = \phi$ and a counterpart $\bar{z}_c$ of $\bar{z}$ is defined as $\bar{\mathcal{C}}_1 f_c$. These observations will be our main tools in what follows.

Step 2. *Stability of* (77). Without reference to the detailed model, we shall use the notation f_0 in reference to the state of a stable abstract model of Θ_0. The bounded mapping from the initial state to the output, in that system, will be denoted $\Upsilon_0 \ : f_0 \mapsto \psi : M_2 \mapsto L_2[0,\infty)$. The associated I/O mapping will be denoted simply $\Theta_0 : \psi \mapsto \phi : L_2[0,\infty) \mapsto L_2[0,\infty)$. Thus, having started at a nonzero state, $\phi = \Upsilon_0 f_0(0) + \Theta_0 \psi$. The induced norm bound $\|\Theta_0\| < 1$ allows us to denote $\lambda^2 = 1 - \|\Theta_0\|^2$.

To establish internal stability one considers the case where an arbitrary, combined initial state $(f_c(0), f_0(0))$ is in place and where the external input is $w = 0$, whereby $w^\nabla = \psi = \mathcal{C}_{c2} f_c$. Invoking (73), then for each $t > 0$ there holds

$$
\begin{aligned}
0 &= \langle f_c(0), \mathcal{R} f_c(0)\rangle_{M_2} - \|\bar{z}_c\|^2_{L_2[0,t]} - \|\psi\|^2_{L_2[0,t]} + \|\phi\|^2_{L_2[0,t]} \\
&\le \langle f_c(0), \mathcal{R} f_c(0)\rangle_{M_2} - \|\psi\|^2_{L_2[0,t]} + \|\Theta_0\psi + \Upsilon_0 f_0(0)\|\|^2_{L_2[0,t]} \\
&\le \langle f_c(0), \mathcal{R} f_c(0)\rangle_{M_2} - \|\psi\|^2_{L_2[0,t]} + \big(\|\Theta_0\|\|\psi\|_{L_2[0,t]} + \|\Upsilon_0\|\|f_0\|_{M_2}\big)^2 \\
&= \langle f_c(0), \mathcal{R} f_c(0)\rangle_{M_2} + \|\Upsilon_0\|^2\|f_0\|^2_{M_2} + 2\|\Theta_0\|\|\Upsilon_0\|\|f_0\|_{M_2}\|\psi\|_{L_2[0,t]} - \lambda^2\|\psi\|^2_{L_2[0,t]}.
\end{aligned}
\tag{80}
$$

Viewing the rightmost term in (80) as a quadratic polynomial with a negative leading coefficient in the variable $\|\psi\|_{L_2[0,t]}$, the inequality forces $\|\psi\|_{L_2[0,t]}$ to take values between the two zeros of the polynomial. This leads to the following bound:

$$
\|\psi\|_{L_2[0,t]} \le \|\Theta_0\|\|\Upsilon_0\|\|f_0\|_{M_2} + \Big(\big(\|\Theta_0\|^2 + \lambda^2\big)\|\Upsilon_0\|^2\|f_0\|^2_{M_2} + \lambda^2\langle f_c(0), \mathcal{R} f_c(0)\rangle_{M_2}\Big)^{\frac{1}{2}},
\tag{81}
$$

which holds for each t. It thus extends to a similar bound, as $t \to \infty$, and can be abbreviated in the form

$$\|\psi\|_2 < \alpha \|(f_0(0), f_c(0))\|_{M_2 \times M_2}, \tag{82}$$

with some fixed $\alpha > 0$ and all initial data. Since Θ_0 is defined by a stable system, the mapping $(f_0(0), \psi) \mapsto (f_0, \phi) : M_2 \times L_2[0,\infty) \mapsto L_2[0,\infty) \times L_2[0,\infty)$ is bounded. Thus, with a possible modification of α, (82) translates into

$$\|\phi\|_2,\ \|f_0\|_2 < \alpha \|(f_0(0), f_c(0))\|_{M_2 \times M_2}. \tag{83}$$

We now use, once again, the observation that $f_c(t) = (x_c(t), u_t)$, ψ, and ϕ play in (16) the precise counterpart roles that $f = (x(t), u_t)$, w^∇, and u^∇, respectively, play, with x_c substituting for x. In particular, the following counterpart of the state equation from (68) is satisfied:

$$\dot{f}_c = \mathcal{A}_1 f_c + \bar{\mathcal{B}}_1 \psi + \bar{\mathcal{B}}_2 \phi. \tag{84}$$

The stability of $\mathcal{A}_1$ implies boundedness of the mapping $(f_c(0), \psi, \phi) \mapsto f_c : M_2 \times L_2[0,\infty) \times L_2[0,\infty) \mapsto L_2[0,\infty)$. Coupling this fact with (82) and (83) it turns out that, with the possible need to modify α, we have

$$\|(f_c, f_0)\|_2 < \alpha \|(f_0(0), f_c(0))\|_{M_2 \times M_2}. \tag{85}$$

It is a standard observation (see, e.g., [5, Lem. 5.1.2]) that the bound (85) implies uniform exponential stability in the combined system (77).

Step 3. *Strict γ Suboptimality.* We now consider the case where Θ is of the form (77), $w \neq 0$ in (16)–(50) and in (77), and where the initial data are all zero. Namely, $f(0) = 0 = f_c(0)$ and $f_0(0) = 0$. As noted above, then both $f(t) = (x(t), u_t)$ and $f_c(t) = (x_c(t), u_t)$ correspond to solutions of (16) with the same inputs u and w and the zero initial data, and with the $\mathbb{R}^n$ "states" $x(t)$ and $x_c(t)$. The uniqueness of the solution of (16) implies that $x(t) = x_c(t)$ throughout, whereby $f(t) = f_c(t)$ throughout. In particular, then $w^\nabla = \psi$ and $u^\nabla = \phi = \Theta_0 w^\nabla$. Using the established stability of (77), we can then let $t \to \infty$ in (73) and obtain

$$\gamma^2 \|w\|_2^2 - \|z\|_2^2 = \|w^\nabla\|_2^2 - \|\Theta_0 w^\nabla\|_2^2 \geq \lambda^2 \|w^\nabla\|_2^2. \tag{86}$$

The signals w and w^∇ are now related by a stable system, combining the state equation (68) (with the zero initial data), the relation $u^\nabla = \Theta_0 w^\nabla$, and the output equation

$$w = \mathcal{D}_{c21}^{-1}(w^\nabla - \mathcal{C}_{c2} f). \tag{87}$$

Thus defined, the mapping $w^\nabla \mapsto w$ is a bounded operator over $L_2[0,\infty)$ and $\lambda^2 \|w^\nabla\|_2^2$ can be bounded below by $\mu^2 \|w\|_2^2$, with some fixed positive μ. Substituting this lower bound in (86), it follows that the closed-loop induced norm of the mapping $w \mapsto z$ is, at most, $\sqrt{\gamma^2 - \mu^2}$.

This completes the proof of the fact that each selection of Θ via the parameterization (77) is indeed a stable and a strictly γ suboptimal selection. In particular, the family of these selections is nonempty, and the proof of the sufficiency claim in Theorem 4.1 is complete.

Step 4. *Completeness of the Parameterization* (77). Let Θ be any strictly γ suboptimal, stable neutral FDE. The fact that Θ is strictly γ suboptimal means that (with the zero initial data) there holds

$$\gamma^2\|w\|_2^2 - \|z\|_2^2 > \mu^2\|w\|_2^2 \tag{88}$$

in the closed-loop system, with any $w \in L_2[0,\infty)$ and some fixed $\mu > 0$. The system that combines (16) and Θ is defined in terms of a stable neutral FDE. Associating the state equation in this system with the second output mapping of (77),

$$w^\nabla = \mathcal{C}_{c2} f + \mathcal{D}_{c21} w, \tag{89}$$

one thus obtains a stable realization of the mapping $w \mapsto w^\nabla$ in the closed-loop system. In particular, with the zero initial state, $\mu^2\|w\|_2^2$ can be bounded below by $\lambda^2\|w^\nabla\|_2^2$, with some fixed, positive λ. Invoking (73) with the zero initial data and with $t \to \infty$, the following closed-loop inequality emerges:

$$\|w^\nabla\|_2^2 - \|u^\nabla\|_2^2 > \lambda^2\|w^\nabla\|_2^2, \tag{90}$$

indicating that the induced norm of the closed-loop mapping $w^\nabla \mapsto u^\nabla$ is smaller than 1.

It has to be verified that this mapping is realized by a stable system. The following is an abstract model for that mapping:

$$\begin{array}{rcl} \dot{f}_\nabla &=& \mathcal{A}_\nabla f_\nabla + \mathcal{B}_{\nabla 1} w^\nabla + \mathcal{B}_{\nabla 2} u, \\ u^\nabla &=& \mathcal{C}_{\nabla 1} f_\nabla + \mathcal{D}_{\nabla 12} u, \\ w &=& \mathcal{C}_{\nabla 2} f_\nabla + \mathcal{D}_{\nabla 21} w^\nabla, \qquad u = \Theta w, \end{array} \tag{91}$$

where the coefficients are defined as follows:

$$\begin{array}{ll} \mathcal{A}_\nabla f_\nabla = \left((A - G_2 C + G_1 L^0) f_\nabla^0 + G_1 L^1 u_t + G_2 f_\nabla^1(-1)\ ,\ \dfrac{d}{ds} f^1 \right), & \\ \mathcal{B}_{\nabla 1} = \bar{\mathcal{B}}_1, & \mathcal{B}_{\nabla 2} = \mathcal{B}_2, \\ \mathcal{C}_{\nabla 1} f_\nabla = -(2E_{12}' E_{12})^{\frac{1}{2}} (K^0 f_\nabla^0 + K^1 f_\nabla^1), & \mathcal{D}_{\nabla 12} = (2E_{12}' E12)^{\frac{1}{2}}, \\ \mathcal{C}_{\nabla 2} = -\mathcal{D}_{c21}^{-1} \mathcal{C}_{c2}, & \mathcal{D}_{c21} = \mathcal{D}_{c21}^{-1}, \end{array} \tag{92}$$

where

$$\mathcal{D}(\mathcal{A}_\nabla) = \mathcal{D}(\mathcal{A}) = \left\{ f_\nabla \in M_2 \ :\ f_\nabla^1 \in W_2^1[-1,0],\quad f_\nabla^1(0) = 0 \right\}. \tag{93}$$

As all other models encountered heretofore, this model too is associated with (16) (with the $\mathbb{R}^n$ "state" $x_\nabla(t)$) via $f_\nabla(t) = (x_\nabla(t), u_t)$, and subject to the input lows $w = \mathcal{C}_{\nabla 2} f_\nabla + \mathcal{D}_{\nabla 21} w^\nabla$ and $u = \Theta w$. Without specifying the details of the stable dynamic equations by which Θ is defined, we shall refer to the complete state of such a realization with the notation f_1. The proof of closed-loop stability follows the pattern of similar arguments in [27, 28, 32] and is now outlined for completeness.

Indeed, in contradiction, suppose that the closed-loop system in (92) is not uniformly exponentially stable and consider that system with the zero external input, $w^\nabla = 0$ (hence with $w = \mathcal{C}_{\nabla 2} f_\nabla$). We argue that this must imply that the ratios

$$\frac{\|w\|_{L_2[0,t]}}{\|(f_\nabla(0), f_1(0))\|_{M_2 \times M_2}} \tag{94}$$

are not uniformly bounded over all possible selections of the combined initial data and of $t > 0$. For suppose those ratios were uniformly bounded, whereby the closed-loop mapping $(f_\nabla(0), f_1(0)) \mapsto w = \mathcal{C}_{\nabla 2} f_\nabla : M_2 \times M_2 \mapsto L_2[0,\infty)$ were a bounded operator. To see that this is impossible (under the contradiction assumption), note that the closed loop (16) is stable anyway, whereby the mapping $(w, f_\nabla(0), f_1(0)) \mapsto (f_\nabla, f_1)$: $L_2[0,\infty) \times M_2 \times M_2 \mapsto L_2[0,\infty)$ defines a bounded operator. Combining these two facts we would have then concluded that the closed-loop mapping $(f_\nabla(0), f_1(0)) \mapsto (f_\nabla, f_1) :\ M_2 \times M_2 \mapsto L_2[0,\infty)$ is bounded. Yet this latter mapping is defined, by our assumption, in an unstable system, and by [5, Lem. 5.1.2], cannot be bounded.

Having selected t and the combined initial data, let the dynamics under consideration be determined over $[0,t]$ by the closed loop (91) with $w^\nabla = 0$, as above, and by the combined rules $w^\nabla = 0$ and $u^\nabla = 0$, over (t,∞). Over the latter ray the dynamics of f_∇ is generated by the stable $\mathcal{A}_1$. Let the definition $w = \mathcal{C}_{\nabla 2} f$ prevail throughout and let z be the associated output of (16), with x_∇ substituting for x.

Let κ be a uniform induced norm bound over the closed-loop mapping $(f(0), f_1(0)) \mapsto z : M_2 \times M_2 \mapsto L_2[0,\infty)$ in the stable closed-loop system (16). The same bound applies when f_∇ substitutes for f. We recall that $\sqrt{\gamma^2 - \mu^2}$ is an induced norm bound over the closed-loop mapping $w \mapsto z : L_2[0,\infty) \mapsto L_2[0,\infty)$, in (16). Applying (73) to the trajectories that were constructed above, we get

$$\begin{aligned}
&\langle f_\nabla(0), \mathcal{R} f_\nabla(0)\rangle_{M_2} + \|u^\nabla\|^2_{L_2[0,t]} = \|z\|_2^2 - \gamma^2 \|w\|_2^2 \\
&\quad \le \left(\sqrt{\gamma^2 - \mu^2}\|w\|_2 + \kappa \|(f_\nabla(0), f_1(0))\|_{M_2 \times M_2}\right)^2 - \gamma^2 \|w\|_2^2 \\
&\quad = -\mu^2 \|w\|_2^2 \left(1 - 2\tfrac{\kappa\sqrt{\gamma^2-\mu^2}}{\mu^2} \tfrac{\|(f_\nabla(0), f_1(0))\|_{M_2\times M_2}}{\|w\|_2} - \tfrac{\kappa^2}{\mu^2} \tfrac{\|(f_\nabla(0), f_1(0))\|^2_{M_2\times M_2}}{\|w\|_2^2}\right).
\end{aligned} \tag{95}$$

If indeed it were possible for the ratios (94) to be made arbitrarily large by a selection of t and of the combined initial state, then it would be possible to make the rightmost expression in (95) arbitrarily close to the negative value $-\mu^2\|w\|_2^2$. Yet, the leftmost expression of the same inequality is nonnegative, which is a contradiction. Thus the closed-loop system (91), governing the closed-loop mapping $\Theta_0 : w^\nabla \mapsto u^\nabla$, must be stable. We have already established that its induced norm is strictly smaller than unity. The fact that (77) is a realization of the mapping $w \mapsto u$ in terms of the mapping $w^\nabla \mapsto u^\nabla$ is clear.

This completes the proof of Lemma 5.9. ∎

As noted above, the abstract model (77) is a representation of a system, coupling (16), with $x_c(t)$ substituting for $x(t)$, the control selection that couples the regression formula $u(t) = K^0 x_c(t) + K^1 u_t + (2E_{12}' E12)^{-\frac{1}{2}} \phi(t)$, the feedback rule $\phi = \Theta_0 \psi$, and the output $\psi(t) = -(L^0 + \frac{1}{\gamma} D' H_1) x_c(t) - L^1 u_t - \frac{1}{\gamma} D' E_{12} u(t-1) + E_{21}^{-1} w(t)$. These facts are precisely those that are captured in the parameterization (12), in the statement of Theorem 4.1. The proof of the theorem is thus complete. ∎

REFERENCES

[1] C. Bernier and A. Manitius, *On semigroups in $\mathbb{R}^n \times l^p$ corresponding to differential equations with delays*, Canad. J. Math., 30 (1977), pp. 897–914.

[2] J. A. Burns, T. L. Herdman, and H. W. Stech, *The Cauchy problem for linear functional differential equations*, in Integral and Differential Equations, T. L. Herdman, S. M. Rankin, and H. W. Stech, eds., Marcel Dekker, New York, 1981, pp. 139–149.

[3] J. A. Burns, T. L. Herdman, and H. W. Stetch, *Linear functional differential equations as semigroups on product spaces*, SIAM J. Math. Anal., 14 (1983), pp. 98–116.
[4] R. F. Curtain and A. J. Pritchard, *Infinite Dimensional Linear Systems Theory*, Springer-Verlag, New York, 1978.
[5] R. F. Curtain and H. J. Zwart, *An Introduction to Infinite Dimensional Linear Systems Theory*, Springer-Verlag, New York, 1995.
[6] M. C. Delfour, *The largest class of hereditary systems defining a c_0-semigroup on the product space*, Canad. J. Math., 32 (1980), pp. 969–978.
[7] D. Flamm and S. Mitter, *H_∞ sensitivity minimization for delay systems*, System Control Lett., 9 (1987), pp. 17–24.
[8] C. Foiaş, A. Tannenbaum, and G. Zames, *Weighted sensitivity minimization for delay systems*, IEEE Trans. Automat. Control, AC 31 (1986), pp. 763–766.
[9] B.A. Francis, *A Course in H_∞ Control Theory*, Springer-Verlag, Berlin, 1986.
[10] E. Hille and R. S. Phillips, *Functional Analysis and Semigroups*, AMS Colloquium Publications XXXI, Providence, RI, 1957.
[11] A. Ichikawa, *Optimal Control and Filtering of Evolution Equations with Delays in Control and Observation*, Tech. Rep. 53, Control Theory Centre, University of Warwick, UK, 1977.
[12] M. A. Kaashoek and J. Kos, *The Nehari-Takagi problem for input-output operators of time varying continuous time systems*, Integral Equations Operator Theory, 18 (1994), pp. 435–467.
[13] T. Kato, *Perturbation Theory for Linear Operators*, Springer-Verlag, New York, 1972.
[14] P. P. Khargonekar, K. M. Nagpal, and K. R. Poolla, *H_∞ control with transients*, SIAM J. Control Optim., 29 (1991), pp. 1373–1393.
[15] K. M. Nagpal and R. Ravi, *H_∞ Control and estimations problems with delayed measurements: State space solutions*, in Proceedings of the American Control Conference, 1994, pp. 2379–2383.
[16] Z. Nehari, *On bounded bilinear forms*, Ann. of Math., 65 (1957), pp. 153–162.
[17] H. Özbay and A. Tannenbaum, *A skew Toeplitz approach to the H_∞ optimal control of multivariable distributed systems*, SIAM J. Control Optim., 28 (1990), pp. 653–670.
[18] A. Pazy, *Semigroups of Linear Operators and Relations to Differential Equations*, Springer-Verlag, New York, 1983.
[19] A. J. Pritchard and D. Salamon, *The linear-quadratic problem for retarded systems with delays in the control and observation*, IMA J. Math. Control Inform., (1985), pp. 335–362.
[20] A. J. Pritchard and D. Salamon, *The linear-quadratic problem for infinite dimensional systems with unbounded input and output operators*, SIAM J. Control Optim., 25 (1987), pp. 121–144.
[21] D. Salamon, *Control and Observation of Neutral Systems*, Pitman, Boston, 1984.
[22] D. Sarason, *Generalized interpolation in H_∞*, Trans. Amer. Math. Soc., 127 (1967), pp. 179–203.
[23] B. Sz.-Nagy and C. Foiaş, *Harmonic Analysis of Operators on Hilbert Spaces*, North Holland, Amsterdam, 1970.
[24] G. Tadmor, *An interpolation problem associated with H_∞ optimization in systems with distributed lags*, System Control Lett., 8 (1987), pp. 313–319.
[25] G. Tadmor, *H_∞ interpolation in systems with commensurate input lags*, SIAM J. Control Optim., 27 (1989), pp. 511–526.
[26] G. Tadmor, *I/O norms in general linear systems*, Internat. J. Control, 51 (1990), pp. 911–921.
[27] G. Tadmor, *Worst case design in the time domain: The maximum principle and the standard H_∞ problem*, Math. Control Signals Systems, 3 (1990), pp. 301–324.
[28] G. Tadmor, *The standard H_∞ problem and the maximum principle: The general linear case*, SIAM J. Control Optim., 31 (1993), pp. 831–846.
[29] G. Tadmor, *H_∞ control in systems with a single input lag*, in Proceedings of the American Control Conference, 1995, pp. 321–325.
[30] G. Tadmor, *Robust control in the gap: A state space solution in the presence of a single input delay*, IEEE Trans. Automat. Control, in press.
[31] G. Tadmor and J. Turi, *Neutral equations and associated semigroups*, J. Differential Equations, 116 (1995), pp. 59–87.
[32] G. Tadmor and M. Verma, *Factorization and the Nehari theorem in time varying systems*, Math. Control Signals Systems, 5 (1992), pp. 419–452.
[33] B. van Keulen, *H_∞ Control for Infinite Dimensional Systems: A State Space Approach*, Ph.D. thesis, University of Groningen, Groningen, The Netherlands, 1993.

[34] R. B. Vinter, *On the evolution of the state of linear differential equations in m^2: Properties of the generator*, J. Inst. Math. Appl., 21 (1978), pp. 13–23.

[35] R. B. Vinter and R. H. Kwong, *The infinite time quadratic control problem for linear systems with state and control delays: An evolution approach*, SIAM J. Control Optim., 19 (1981), pp. 139–153.

[36] K. Zhou and P. P. Khargonekar, *On the weighted sensitivity minimization problem for delay systems*, System Control Lett., 8 (1987), pp. 307–312.

SIAM J. CONTROL OPTIM.
Vol. 35, No. 5, pp. 1470–1486, September 1997

002

ON THE TIME-DISCRETIZATION OF CONTROL SYSTEMS*

VLADIMIR VELIOV†

Abstract. This paper develops an approach for obtaining discrete approximations to nonlinear (affine) control systems that are of higher than first order of accuracy with respect to the discretization step h. The approach consists of two parts: first the set $\mathcal{U}$ of measurable admissible controls is replaced by an appropriate finite-dimensional subset $\mathcal{U}_N$; then the differential equations corresponding to controls from $\mathcal{U}_N$ (which are in a reasonable sense "regular") are discretized by single step discretization methods. The main result estimates the accuracy in the first part, measured in terms of a prescribed collection of performance indexes. The result can be interpreted both in the context of approximation of optimal control problems and in the context of approximation of the reachable set. In the first case, accuracy $O(h^2)$ is proven for appropriate Runge–Kutta-type discretization methods, without explicitly or implicitly requiring any regularity of the optimal solutions. In the case of a convex reachable set we obtain $O(h^2)$ approximation with respect to the Hausdorff distance and $O(h^{1.5})$ accuracy in the nonconvex case. An application to the time-aggregation of discrete-time control systems is also presented.

Key words. control systems, differential inclusion, reachable set, discrete approximation, Runge–Kutta scheme, optimal control

AMS subject classifications. 49M25, 65L12

PII. S0363012995288987

1. Introduction.

In this paper we consider a control system

$$\dot{x} = f(t,x,u), \quad x(t_0) = x_0,\ u(\cdot) \in \mathcal{U},\ x \in \mathbf{R}^n,\ t \in [t_0,T], \tag{1}$$

with a given set $\mathcal{U}$ of admissible control functions $u(\cdot) : [t_0,T] \mapsto \mathbf{R}^r$ and fixed x_0, t_0, and T. In fact, the system under consideration will be linear with respect to u, but for notational convenience, we discuss the general case in the introduction.

For a prescribed collection of functions (performance indexes) $\mathcal{G} = \{g(\cdot)\,;\ g : \mathbf{R}^n \mapsto \mathbf{R}\}$, the set of real numbers

$$\{\inf_{x(\cdot)} g(x(T))\,;\ g(\cdot) \in \mathcal{G}\}$$

will be considered as a characterization of the performance of the system. Accordingly, if

$$\dot{y} = \tilde{f}(t,y,v), \quad y(t_0) = x_0,\ v(\cdot) \in \mathcal{V},\ y \in \mathbf{R}^n,\ t \in [t_0,T], \tag{2}$$

is another system, then

$$\varepsilon^-((2),(1)) = \sup_{g\in\mathcal{G}} \left\{ \inf_{y(\cdot)} g(y(T)) - \inf_{x(\cdot)} g(x(T)) \right\}$$

is a measure of the "disadvantage," with respect to the performance, of system (2) compared with (1). In particular, if $\tilde{f} = f$ and only the admissible control sets differ, $\varepsilon^-((2),(1))$ (which will be denoted in this case by $\varepsilon^-(\mathcal{V},\mathcal{U})$) is a measure of the "loss"

*Received by the editors July 12, 1995; accepted for publication (in revised form) May 28, 1996. http://www.siam.org/journals/sicon/35-5/28898.html

†Institute of Statistics, Informatics and Operations Research, University of Vienna, Universitätsstrasse 5, A-1010 Vienna, Austria, and Institute of Mathematics and Informatics, Bulgarian Academy of Sciences, P.O. Box 373, 1113 Sofia, Bulgaria (veliov@uranus.tuwien.ac.at).

of performance when the set $\mathcal{U}$ of admissible controls is replaced by $\mathcal{V}$. The principle result of the paper is an estimation of the last quantity in the case when

$$\mathcal{U} = \{u(\cdot)\,;\ u(\cdot) \in L_1^r[t_0, T],\ u(t) \in U \text{ for a.e. } t\}, \tag{3}$$

where $U \subset \mathbf{R}^r$ is a convex and compact set and

$$\mathcal{V} = \mathcal{U}_N \stackrel{def}{=} \{u(\cdot) \in \mathcal{U}\,;\ u(\cdot) \text{ is constant on } (t_i, t_{i+1}),\ i = 0, \ldots, N-1\}, \tag{4}$$

where N is a natural number (of jumps) and $t_i = t_0 + ih$, $h = (T - t_0)/N$. From Dontchev and Farkhi [8] or Wolenski [23], it follows that under mild conditions

$$\varepsilon^-(\mathcal{U}_N, \mathcal{U}) \leq C/N. \tag{5}$$

The problem of estimating $\varepsilon^-(\mathcal{U}_N, \mathcal{U})$ can be considered from the general point of view of the sensitivity analysis of extremal problems. It has been know since the 17th Century that a smooth function deviates from its value at an *extreme* point proportionally to the square of the distance of its argument to this point. This observation (of J. Kepler, in 1615) also has a counterpart for constrained optimization problems in normed spaces. In our case the extreme point in question is a minimizing control function $\hat{u}(\cdot)$ (subject to infinitely many linear constraints in $L_2^r(0, T)$, corresponding to the constraint $u(t) \in U$) for a particular performance index $g(\cdot) \in \mathcal{G}$. The general analysis, however, fails in this case, since the Hausdorff distance, with respect to any L_p-norm, between the sets $\mathcal{U}_N$ and $\mathcal{U}$ does not even tend to zero when $N \longrightarrow +\infty$. A more specific analysis is needed, in this sense, even for the above-mentioned first-order estimation (5).

On the other hand, the papers [22, 20] imply that a second-order estimation

$$\varepsilon^-(\mathcal{U}_N, \mathcal{U}) \leq C/N^2 \tag{6}$$

holds in two rather different cases (supposing, however, that the functions $g \in \mathcal{G}$ have equi-Lipschitz-continuous derivatives): in the case when system (1) is linear and in the case of system (1) for which $f(t, x, U)$ is a uniformly strongly convex set. The main result in the present paper, presented in section 2, extends the second-order estimation (6) to the case of nonlinear systems that are affine with respect to u and satisfy an additional structural condition. The subsequent sections present some applications. (The problem considered here can also be interpreted within the issue of approximation of differential inclusions by collections of differential equations that are finite-dimensionally parametrizable or finite; cf. [11]).

In section 3 the main result is applied for obtaining appropriate Runge–Kutta-type discrete approximations to nonlinear optimal control problems that ensure second-order approximation (relative to the discretization step) with respect to the performance value. In contrast to the known results this estimation is not based on any regularity assumptions for the optimal solution, and even the constant C in the estimation corresponding to (6) is in a reasonable sense *robust* with respect to the data. In particular, the cases of an optimal control function that is of unbounded variation, or one that is nonintegrable in the Riemann sense, are not excluded by the suppositions under which the second-order estimation holds.

Section 4 is devoted to the issue of discrete approximations of the reachable set of (1), (3). The main result, together with a duality argument, implies (under an appropriate convexity assumption) second-order error estimation in the Hausdorff

metric for certain *set-valued* Runge–Kutta-type discretization schemes. Notice that, according to [22], Runge–Kutta schemes of higher order of accuracy exist in the set-valued case only under conditions that are rather restrictive in the context of control theory (cf. [4]). For an extended bibliography concerning the issue of discretization of control systems and differential inclusions, see [6, 10, 15].

Section 5 concerns the following issue with obvious practical motivation: what is the "loss of performance" of a given N-stage discrete-time control system if the control value is allowed to change only at the stages kM, $k = 1, \dots, N/M$, rather than at each stage? The performance is defined as above, by a collection of indexes. The main result is used to obtain an exact estimation, as well as a constructive way of defining corresponding N/M-stage time-aggregated discrete-time systems.

2. Main result. It is supposed further that system (1) has the form

$$\dot{x} = f_0(t,x) + F(t,x)u, \quad x(t_0) = x_0, \tag{7}$$

where $f_0 : [t_0, T] \times \mathbf{R}^n \to \mathbf{R}^n$, $F : [t_0, T] \times \mathbf{R}^n \to \mathbf{R}^n \times \mathbf{R}^r$. The set $\mathcal{U}$ of admissible control functions is defined as in (3), where $U \subset \mathbf{R}^r$ is a convex and compact set.

The next theorem estimates the "loss of performance" of system (7) when the rN-dimensional control set $\mathcal{U}_N$ defined by (4) is used instead of the infinite-dimensional (in general) control set $\mathcal{U}$.

Denote by $R(\tau)$ the reachable set of (7) on $[t_0, \tau]$, that is, $x \in R(\tau)$ if and only if there is a $u(\cdot) \in \mathcal{U}$ such that a corresponding trajectory $x(\cdot)$ of (7) exists on $[t_0, \tau]$ and $x(\tau) = x$.

Assumption 1. There is a convex compact set $S \subset \mathbf{R}^n$ such that $R(t) \subset \operatorname{int} S$ for every $t \in [t_0, T]$.

Assumption 2. The components of f_0 and F are differentiable with respect to t and x, and all the first derivatives are Lipschitz continuous with respect to $(t,x) \in [t_0, T] \times S$.

Assumption 3 (structural condition). The columns $f_1(t,x), \dots, f_r(t,x)$ of $F(t,x)$ satisfy the relations

$$[f_i, f_j]_x(t,x) \stackrel{def}{=} \left(\frac{\partial f_i}{\partial x} f_j - \frac{\partial f_j}{\partial x} f_i \right)(t,x) = 0$$

for every $i, j = 1, \dots, r$ and $(t,x) \in [t_0, T] \times S$.

Assumption 4. There are constants L_g and L'_g such that each function $g \in \mathcal{G}$ is differentiable in the interior of S, and $\partial g / \partial x$ is bounded by L_g and is Lipschitz continuous at each point of $R(T)$ with Lipschitz constant L'_g.

We mention that Assumption 3 for the Lie brackets of the vector fields f_j is always fulfilled if F is independent of x, or $\dim u = 1$. Otherwise it poses a restriction on the interaction between the different control components. It is an open problem whether Assumption 3 is essential for the claim of Theorem 2.1 below. The author has some reason to expect that the answer might be (in general) affirmative, but a counterexample is not available.

Remark 1. The constant in the estimation (8) below does not depend directly on the particular function $f = f_0 + Fu$. Rather, it depends (and can be evaluated following the proof) on certain constants associated with f:

—the Lipschitz constants L_0, L_F, L'_0, and L'_F of f_0, and the columns of F and their first derivatives in the set $[t_0, T] \times S$ (denoted further briefly as L);

—the maximum M_0 and M_F of $|f_0(t,x)|$ and $|F(t,x)|$ in $[t_0, T] \times S$ and the maximum M'_0 and M'_F of the norms of the derivatives with respect to x of f_0 and the

columns of F in $[t_0, T] \times S$ (everywhere, the operator matrix norm is meant); these constants will be succinctly symbolized by M.

THEOREM 2.1. *Let Assumptions* 1–4 *be fulfilled. Then there exists a constant* C *depending only on* L, M, L_g, L'_g, $|U|$, *and* $T - t_0$, *such that for every natural number* N

$$\varepsilon^-(\mathcal{U}_N, \mathcal{U}) \stackrel{def}{=} \sup_{g \in \mathcal{G}} \left\{ \inf_{v(\cdot) \in \mathcal{U}_N} g(y(T)) - \inf_{u(\cdot) \in \mathcal{U}} g(x(T)) \right\} \leq \frac{C}{N^2}. \tag{8}$$

(Here $y(\cdot)$ *is the trajectory corresponding to* $v(\cdot)$, *and* $x(\cdot)$ *is the trajectory corresponding to* $u(\cdot)$, *according to* (7)*).*

This theorem is a direct consequence of the following one (corresponding to the case of a set $\mathcal{G}$ consisting of a single function g). For a particular $g(\cdot) \in \mathcal{G}$, consider the following terminal optimal control problem for system (7):

$$g(x(T)) \longrightarrow \min, \tag{9}$$

Assumptions 1–4 ensure existence of a solution in both class $\mathcal{U}$ and $\mathcal{U}_N$ of admissible control functions. The corresponding optimal values will be denoted by $\hat{g}$ and $\hat{g}_N$.

THEOREM 2.2. *Let Assumptions* 1–4 *be fulfilled. Then there exist constants* C_1 *and* C_2 *depending only on* L, M, $|U|$, *and* $T - t_0$ *such that for every natural number* N

$$0 \leq \hat{g}_N - \hat{g} \leq \frac{C_1 L_g + C_2 L'_g}{N^2}. \tag{10}$$

Notice that C_1 and C_2 do not depend on the properties of the optimal control, which can be even of unbounded variation, nonintegrable in the Riemann sense, or even discontinuous almost everywhere (Assumptions 1–4 do not exclude these possibilities; see Silin [19]).

The proof will be preceded by some auxiliary results.

Let $l(\cdot) : [t_0, T] \mapsto \mathbf{R}^n$ satisfy the conditions

(i) $l(\cdot)$ is Lipschitz continuous with constant L_l;

(ii) $\dot{l}(\cdot)$ is of bounded variation.

The following result, playing a crucial role in the forthcoming analysis, is a consequence of [5, Prop. 2].

PROPOSITION 2.3. *The function*

$$\hat{l}(t) = \sup_{u \in U} \langle l(t), u \rangle$$

is Lipschitz continuous and

$$\bigvee_{t_0}^{T} \dot{\hat{l}}(\cdot) \leq 2L_l + 2|U| \bigvee_{t_0}^{T} \dot{l}(\cdot).$$

The precise (natural) meaning of the variations of $\dot{l}$ and $\dot{\hat{l}}$—these functions are defined almost everywhere—is also given in [5].

By [2, Thm. 8.2.8] the set-valued mapping

$$\hat{U}(t) = \{u \in U \,;\, \langle l(t), u \rangle = \hat{l}(t)\}$$

is measurable. Let $\hat{u}(\cdot)$ be an arbitrary measurable selection of $\hat{U}(\cdot)$.

The proof of the following lemma is standard.

LEMMA 2.4. *If $\dot{l}(t)$ and $\hat{\dot{l}}(t)$ exist for some $t \in [t_0, T]$, then*

$$\dot{\hat{l}}(t) = \langle \dot{l}(t), \hat{u}(t) \rangle.$$

For the fixed selection $\hat{u}(\cdot)$ of $\hat{U}(\cdot)$ define

$$u_N^i = \frac{1}{h} \int_{t_i}^{t_{i+1}} \hat{u}(t)\, dt \tag{11}$$

and

$$u_N(t) = u_N^i \quad \text{for} \quad t \in [t_i, t_{i+1}),\ i = 0, \dots, N-1. \tag{12}$$

LEMMA 2.5.

$$0 \le \int_{t_0}^{T} \langle l(t), \hat{u}(t) - u_N(t) \rangle\, dt \le \left(2L_l + 3|U| \bigvee_{t_0}^{T} \dot{l}(\cdot) \right) h^2.$$

Proof. By definition, and using Lemma 2.4, we estimate

$$0 \le \int_{t_0}^{T} \langle l(t), \hat{u}(t) - u_N(t) \rangle\, dt = \frac{1}{h} \sum_{i=0}^{N-1} \int_{t_i}^{t_{i+1}} \int_{t_i}^{t_{i+1}} \langle l(t), \hat{u}(t) - \hat{u}(s) \rangle\, ds\, dt$$

$$= \frac{1}{h} \sum_{i=0}^{N-1} \int_{t_i}^{t_{i+1}} \int_{t_i}^{t_{i+1}} [\langle l(t), \hat{u}(t) \rangle - \langle l(s), \hat{u}(s) \rangle + \langle l(s) - l(t), \hat{u}(s) \rangle]\, ds\, dt$$

$$= \frac{1}{h} \sum_{i=0}^{N-1} \int_{t_i}^{t_{i+1}} \int_{t_i}^{t_{i+1}} \left[\int_s^t \dot{\hat{l}}(\tau)\, d\tau - \left\langle \int_s^t \dot{l}(\tau)\, d\tau, \hat{u}(s) \right\rangle \right] ds\, dt$$

$$= \frac{1}{h} \sum_{i=0}^{N-1} \int_{t_i}^{t_{i+1}} \int_{t_i}^{t_{i+1}} \int_s^t [\dot{\hat{l}}(\tau) - \langle \dot{l}(\tau), \hat{u}(s) \rangle]\, d\tau\, ds\, dt$$

$$= \frac{1}{h} \sum_{i=0}^{N-1} \int_{t_i}^{t_{i+1}} \int_{t_i}^{t_{i+1}} \int_s^t [\dot{\hat{l}}(\tau) - \dot{\hat{l}}(s) + \langle \dot{l}(s) - \dot{l}(\tau), \hat{u}(s) \rangle]\, d\tau\, ds\, dt$$

$$\le \left(\bigvee_{t_0}^{T} \dot{\hat{l}}(\cdot) + |U| \bigvee_{t_0}^{T} \dot{l}(\cdot) \right) h^2 \le \left(2L_l + 3|U| \bigvee_{t_0}^{T} \dot{l}(\cdot) \right) h^2,$$

where the last estimation is based on Proposition 2.3. ∎

Further, we denote for brevity

$$L_f = L_0 + |U| L_F, \quad M_f = M_0 + |U| M_F, \quad M_f' = M_0' + |U| M_F'.$$

A standard application of the Gronwall inequality gives the following estimation.

LEMMA 2.6. *Let $\hat{u}(\cdot) \in \mathcal{U}$ be arbitrary and let $\hat{x}(\cdot)$ be the corresponding solution of* (7). *Define $u_N(\cdot)$ as in* (11), (12), *and let $x_N(\cdot)$ be the trajectory of* (7) *corresponding to $u_N(\cdot)$. Then*

$$\|\hat{x}(\cdot) - x_N(\cdot)\| \le e^{L_f(T-t_0)} (T - t_0) L_F (1 + M_F) |U| h.$$

The proof of the following lemma goes along the line of the direct proof of the maximum principle for the problem (9) and therefore is omitted.

LEMMA 2.7. *Let $\hat{u}(\cdot) \in \mathcal{U}$ be an arbitrary optimal control of* (9) *and let $u(\cdot) \in \mathcal{U}$ be another admissible control. Let $\hat{x}(\cdot)$, $\hat{g}$ and $x(\cdot)$, g be the corresponding trajectories and values of the objective function. Denote by $\hat{\psi}(\cdot)$ the solution of the adjoined system*

$$\dot{\hat{\psi}} = -\frac{\partial f^*}{\partial x}(t, \hat{x}(t), \hat{u}(t))\hat{\psi}, \quad \hat{\psi}(T) = -\frac{\partial g}{\partial x}(\hat{x}(T)), \tag{13}$$

where $f(t,x,u) = f_0(t,x) + F(t,x)u$ and * denotes transposition. Then

$$0 \le g - \hat{g} \le \int_{t_0}^{T} \langle \hat{\psi}(t), F(t, \hat{x}(t))(\hat{u}(t) - u(t)) \rangle \, dt$$

$$+ \int_{t_0}^{T} \left\langle \hat{\psi}(t), \sum_{k=1}^{r} (\hat{u}^k(t) - u^k(t)) \frac{\partial f_k}{\partial x}(t, \hat{x}(t))(x(t) - \hat{x}(t)) \right\rangle dt$$

$$+ L'_g |\hat{x}(T) - x(T)|^2 + (L'_0 + L'_F|U|)\|\hat{\psi}\|_C \|\hat{x} - x\|_L^2,$$

where u^i is the ith component of u and the constants in the above expression are defined in Remark 1.

LEMMA 2.8. *Let $\hat{u}(\cdot)$, $\hat{x}(\cdot)$, and $\hat{\psi}(\cdot)$ be as in Lemma* 2.7. *Then the function*

$$l(t) = F^*(t, \hat{x}(t))\hat{\psi}(t) \tag{14}$$

satisfies conditions (i) *and* (ii) *above. Moreover, both the Lipschitz constant of $l(\cdot)$ and the variation of $\dot{l}(\cdot)$ can be estimated by CL_g, where C depends only on L, M, $|U|$, and $T - t_0$.*

Proof. A straightforward estimation of the Lipschitz constant L_l of $l(\cdot)$ gives

$$L_l \le L_F(1 + M_f)M_\psi + M_F L_\psi,$$

where M_ψ and L_ψ are the maximum of the norm and the Lipschitz constant of $\psi(\cdot)$, respectively. Moreover, having in mind (13), we can estimate

$$M_\psi \le e^{M'_f(T-t_0)} L_g, \qquad L_\psi \le e^{M'_f(T-t_0)} M'_f L_g = C' L_g, \tag{15}$$

where C' depends on the constants listed in the formulation of the lemma.

Let us estimate the variation of $\dot{l}(\cdot)$. Direct calculation shows that the derivative of the ith component of $l(\cdot)$ is

$$\dot{l}_i = \left\langle \hat{\psi}, \frac{\partial f_i}{\partial t} + [f_i, f_0] + \sum_{j=1}^{r} \hat{u}_j [f_i, f_j] \right\rangle,$$

where all the arguments in the above expression are either t or (where appropriate) $(t, \hat{x}(t))$. The last term, however, is zero, according to Assumption 3. Therefore, $\dot{l}(\cdot)$ turns out to be even Lipschitz continuous, and its Lipschitz constant can be explicitly estimated by $C'' L_g$, where C'' depends only on the constants listed in the formulation of the lemma. The same applies, therefore, to the variation of $\dot{l}(\cdot)$. ∎

Now we are ready to proceed with the proof of the theorem.

Proof of Theorem 2.2. All numbers c_1, $c_2, \dots$ appearing in the subsequent proof depend only on the constants listed in Theorem 2.2.

Let $\hat{u}(\cdot) \in \mathcal{U}$ be an arbitrary optimal control, $\hat{x}(\cdot)$ be the corresponding trajectory of (7) and $\hat{\psi}(\cdot)$ be the corresponding solution of the adjoined equation (13).

Define the control function $u_N(\cdot) \in \mathcal{U}_N$ as

$$u_N(t) = \frac{1}{h} \int_{t_i}^{t_{i+1}} \hat{u}(s)\, ds \ \text{ for } \ t \in [t_i, t_{i+1}),\ i = 0, \dots, N-1,$$

and let $x_N(\cdot)$ be the corresponding trajectory of (7). According to Lemma 2.7

$$0 \le \hat{g}_N - \hat{g} \le g(x_N(T)) - g(\hat{x}(T)) \le \int_{t_0}^{T} \langle l(t), \hat{u}(t) - u_N(t) \rangle\, dt \tag{16}$$

$$+ \sum_{k=1}^{r} \int_{t_0}^{T} \langle B_k(t), x_N(t) - \hat{x}(t) \rangle (\hat{u}^k(t) - u_N^k(t))\, dt + e \|\hat{x} - x_N\|_C^2,$$

where $l(t)$ is defined by (14),

$$B_k(t) = \frac{\partial f_k^*}{\partial x}(\hat{x}(t), t) \hat{\psi}(t),$$

and $e = L_g' + L_f' M_\psi \le L_g' + c_1 L_g$ (see (15)).

Denote $\Delta(t) = x_N(t) - \hat{x}(t)$. According to Lemma 2.6, $|\Delta(t)| \le c_2 h$. Since $\hat{u}(\cdot)$ is an optimal control to problem (9), (7), (3) and, therefore, satisfies the maximum principle, we have that

$$\langle l(t), \hat{u}(t) \rangle = \max_{u \in U} \langle l(t), u \rangle = \hat{l}(t).$$

Applying Lemmas 2.5 and 2.8, we obtain the inequality

$$0 \le \hat{g}_N - \hat{g} \le \sum_{k=1}^{r} \int_{t_0}^{T} \langle B_k(t), x_N(t) - \hat{x}(t) \rangle (\hat{u}^k(t) - u_N^k(t))\, dt + \frac{c_3 L_g + c_4 L_g'}{N^2}.$$

The first term in the right-hand side can be represented as

$$\delta = \frac{1}{h} \sum_{k=1}^{r} \sum_{i=0}^{N-1} \int_{t_i}^{t_{i+1}} \int_{t_i}^{t_{i+1}} \langle B_k(t), \Delta(t) \rangle (\hat{u}^k(t) - \hat{u}^k(s))\, ds\, dt.$$

Because of the Lipschitz continuity of $B_k(\cdot)$ (see (15)) and the estimation of $\Delta(\cdot)$ in Lemma 2.6 we have

$$\delta \le \frac{1}{h} \sum_{k=1}^{r} \sum_{i=0}^{N-1} \int_{t_i}^{t_{i+1}} \int_{t_i}^{t_{i+1}} \langle B_k(t_i), \Delta(t) \rangle (\hat{u}^k(t) - \hat{u}^k(s))\, ds\, dt + (c_5 L_g + c_6 L_g')/N^2.$$

Using the fact that the double integral of $\hat{u}^k(t) - \hat{u}^k(s)$ equals zero, we obtain

$$\delta \le \frac{1}{h} \sum_{k=1}^{r} \sum_{i=0}^{N-1} \int_{t_i}^{t_{i+1}} \int_{t_i}^{t_{i+1}} \langle B_k(t_i), \Delta(t) - \Delta(t_i) \rangle (\hat{u}^k(t) - \hat{u}^k(s))\, ds\, dt + (c_5 L_g + c_6 L_g')/N^2.$$

Thanks to Lemma 2.6, we have for $t \in [t_i, t_{i+1}]$

$$\Delta(t) - \Delta(t_i) = \int_{t_i}^{t} \dot{\Delta}(\tau)\, d\tau = \int_{t_i}^{t} (f_0(x_N(\tau),\tau) - f_0(\hat{x}(\tau),\tau))\, d\tau$$

$$+ \int_{t_i}^{t} \sum_{j=1}^{r} (f_j(x_N(\tau),\tau)u_N^j(\tau) - f_j(\hat{x}(\tau),\tau)\hat{u}^j(\tau))\, d\tau$$

$$= \int_{t_i}^{t} \sum_{j=1}^{r} \hat{f}_j(t_i)(u_N^j(\tau) - \hat{u}^j(\tau))\, d\tau + O(h^2),$$

where $\hat{f}_j(t) = f_j(\hat{x}(t), t)$ (further, we use similar notation for the derivative) and $O(h^2) \le c_7/N^2$. Hence,

$$\delta \le \frac{1}{h} \sum_{k=1}^{r} \sum_{i=0}^{N-1} \sum_{j=1}^{r} \int_{t_i}^{t_{i+1}} \int_{t_i}^{t_{i+1}} \int_{t_i}^{t} \left\langle \hat{\psi}(t_i), \frac{\partial \hat{f}_k}{\partial x} \hat{f}_j(t_i) \right\rangle$$

$$\times (u_N^j(\tau) - \hat{u}^j(\tau))(\hat{u}^k(t) - \hat{u}^k(s))\, d\tau\, ds\, dt + (c_7 L_g + c_8 L_g')/N^2$$

$$= \frac{1}{h^2} \sum_{i=0}^{N-1} \sum_{j=1}^{r} \sum_{k=1}^{r} \left\langle \hat{\psi}(t_i), \frac{\partial \hat{f}_k}{\partial x} \hat{f}_j(t_i) \right\rangle \int_{t_i}^{t_{i+1}} \int_{t_i}^{t_{i+1}} \int_{t_i}^{t_{i+1}} \int_{t_i}^{t}$$

$$\times (\hat{u}^j(\theta) - \hat{u}^j(\tau))(\hat{u}^k(t) - \hat{u}^k(s))\, d\tau\, d\theta\, ds\, dt + (c_8 L_g + c_9 L_g')/N^2.$$

Using Assumption 3 one can rewrite the above sums as

$$\frac{1}{h^2} \sum_{i=0}^{N-1} \sum_{k=1}^{r} \left\langle \hat{\psi}(t_i), \frac{\partial \hat{f}_k}{\partial x} \hat{f}_k(t_i) \right\rangle \int_{t_i}^{t_{i+1}} \int_{t_i}^{t_{i+1}} \int_{t_i}^{t_{i+1}} \int_{t_i}^{t} (\hat{u}^k(\theta) - \hat{u}^k(\tau))(\hat{u}^k(t) - \hat{u}^k(s))\, d\tau\, d\theta\, ds\, dt$$

$$+ \frac{1}{h^2} \sum_{i=0}^{N-1} \sum_{1 \le j < k \le r} \left\langle \hat{\psi}(t_i), \frac{\partial \hat{f}_k}{\partial x} \hat{f}_j(t_i) \right\rangle \int_{t_i}^{t_{i+1}} \int_{t_i}^{t_{i+1}} \int_{t_i}^{t_{i+1}} \int_{t_i}^{t} [(\hat{u}^j(\theta) - \hat{u}^j(\tau))(\hat{u}^k(t) - \hat{u}^k(s))$$

$$+ (\hat{u}^k(\theta) - \hat{u}^k(\tau))(\hat{u}^j(t) - \hat{u}^j(s))]\, d\tau\, d\theta\, ds\, dt.$$

It can be shown that each of the above fourtuple integrals equals zero, which completes the proof of the theorem. □

3. Discretization of optimal control problems. Theorem 2.2 provides a basis for obtaining discrete approximations of second-order accuracy (with respect to the length of the discretization step) to optimal control problems. The underlying idea is that thanks to Theorem 2.2 one can first restrict the consideration to control functions that are constant between each of two neighboring grid points t_i, t_{i+1}, sacrificing only the $O(h^2)$ value of the objective function. For any fixed control function from this class, one can then discretize the corresponding differential equation (7) by means of any *single-step* discretization scheme that provides at least $O(h^3)$ local accuracy. According to our smoothness assumptions and the fact that the control is constant in any particular subinterval $[t_i, t_{i+1})$, such schemes exist. In this way one

can obtain a variety of discrete-time problems, each of which provides an approximation of second-order accuracy (in the sense described in the next lines) to the original problem. To illustrate the approach we take a second-order symmetric Runge–Kutta scheme (known as Euler–Cauchy method).

The approach is directly applicable to terminal-type optimal control problems, but the well-known trick used below allows us to consider the following more general problem (where the criterion is nonlinear in u):

$$g(x(T)) + \int_{t_0}^{T} [e_0(t, x(t)) + \langle e(t, x(t)), \varphi(u(t))\rangle]\, dt \longrightarrow \min, \tag{17}$$

$$\dot{x} = f_0(t, x) + F(t, x)u, \quad x(t_0) \in X_0, \tag{18}$$

$$u(t) \in U \tag{19}$$

on the fixed time-interval $[t_0, T]$. Here $e_0(t, x)$, $e(t, x) = (e_1(t, x), \dots, e_p(t, x))$, and $\varphi(u) = (\varphi_1(u), \dots, \varphi_p(u))$ are given functions. Let $\hat{J}$ be the optimal value of the objective function. Given the natural number N we find below an N-stage discrete-time optimal control problem $\mathcal{P}_N$ such that for any optimal control $u_N = (u_N^0, \dots, u_N^{N-1})$ of $\mathcal{P}_N$, if $u_N(\cdot) \in \mathcal{U}$ is defined as $u_N(t) = u_N^k$ for $t \in [t_k, t_{k+1})$, then the corresponding performance value $J(u_N(\cdot))$ of (17), (18) satisfies

$$J(u_N(\cdot)) \le \hat{J} + \text{const}/N^2.$$

Thus, the meaning of accuracy that we employ does not imply any relation between the optimal controls or trajectories of the two problems; rather, it ensures that the control function for the continuous-time problem that we construct by solving the discrete-time one is $O(1/N^2)$-optimal. We refer to Dontchev [6, 7], Mordukhovich [17], and the papers quoted in [6, 7] for results concerning (first-order) approximation of the optimal control function via discretization of problem (17), (18), (19).

A number of papers explicitly or implicitly prove accuracy $O(1/N)$ of approximation if the discrete-time problem is obtained by formal application of the Euler or other discretization schemes to the differential equation and the integral involved in (17), (18) (see [6, 7] for a comprehensive bibliography). The known results for higher order accuracy (cf. [12, 18]), however, are based on certain regularity requirements for the optimal controls of (17), (18), (19), which are by no means implied by our suppositions below. On the contrary, it might happen that all optimal controls are of unbounded variation, nonintegrable in Riemann sense, or even discontinuous almost everywhere (cf. [19]). Nevertheless, we prove second-order accuracy, even with a constant (multiplying h^2 in the estimation) that is not "sensitive" with respect to the data, supposing, however, that a certain additional structural condition (corresponding to Assumption 3) is satisfied.

The next result extends the one obtained for linear systems in [22] (accuracy $O(h^2)$) and the one from [21], where accuracy $O(1/N^{1.5})$ is proven under similar suppositions.

Define the following discrete-time optimal control problem.

Minimize

$$g(x_N) + \frac{h}{2} \sum_{k=0}^{N-1} [e_0(t_k, x_k) + e_0(t_{k+1}, x_{k+1}) + \langle e(t_k, x_k) + e(t_{k+1}, x_{k+1}), \varphi(u_k)\rangle] \tag{20}$$

subject to

$$x_{k+1} = x_k + 0.5h[f(t_k, x_k, u_k) + f(t_{k+1}, x_k + hf(t_k, x_k, u_k), u_k)], \tag{21}$$

$$u_k \in U, \quad k = 0, \dots, N-1, \tag{22}$$

where, as in the previous section, $f(t, x, u) = f(t, x) + F(t, x)u$.

THEOREM 3.1. *Let Assumptions* 1–4 *from section* 2 *be fulfilled (for* $\mathcal{G} = \{g\}$*). Let, in addition, the functions* φ_j *be convex and continuous, the functions* e_j *satisfy the same differentiability conditions as* f_i*, and* $e_j(x, t) \geq 0$ *for every* (t, x)*,* $j = 1, \dots, p$*. Let, moreover,* $\langle \frac{\partial e_j}{\partial x}, f_i \rangle(t, x) = 0$ *for* $i = 1, \dots, r$*,* $j = 1, \dots, p$*, and* $(t, x) \in [t_0, T] \times \mathbf{R}^n$*. Then there are constants* C_1 *and* C_2 *such that for every natural number* N *and for every* ε*-optimal control* $(u_0, u_1, \dots, u_{N-1})$ *of problem* (20)–(22) *with corresponding performance value* J_N

(1)

$$|\hat{J} - J_N| \leq C_1/N^2 + \varepsilon; \tag{23}$$

(2) *the control function* $u_N(\cdot)$ *defined as* $u_N(t) = u_k$ *for* $t \in [t_k, t_{k+1})$*,* $k = 0, \dots, N-1$*, when applied to* (17), (18) *gives value* $J(u_N(\cdot))$ *of the objective function, that satisfies*

$$J(u_N(\cdot)) \leq \hat{J} + C_2/N^2 + \varepsilon.$$

("ε-optimal" in the above formulation means that $J_N \leq \hat{J}_N + \varepsilon$, where $\hat{J}_N$ is the optimal value in the problem (20)–(22).)

Remark 2. The constants C_1 and C_2 in the above theorem do not depend on the particular data. Rather, they depend on certain constants associated with the data, as in Theorem 2.2 (with $e_0(\cdot)$ and $e(\cdot)$ treated similarly as $f_0(\cdot)$ and $F(\cdot)$). In this sense C_1 and C_2 are not sensitive with respect to the data, nor to the properties of the optimal control.

Proof. One can reformulate problem (17), (18), (19) as follows:

$$g(x(T)) + x^0(T) \longrightarrow \min, \tag{24}$$

$$\dot{x}^0 = e_0(t, x) + \langle e(t, x), v \rangle, \quad x^0(t_0) = 0, \tag{25}$$

$$\dot{x} = f_0(t, x) + \sum_{i=1}^{p} f_i(t, x)u^i, \quad x(t_0) = x_0, \tag{26}$$

$$(u(t), v(t)) \in W, \tag{27}$$

where

$$W = \{(u, v)\,;\ u \in U,\ \varphi_j(u) \leq v^j \leq \bar{M},\ j = 1, \dots, p\}$$

is apparently convex and compact and the constant $\bar{M}$ is chosen so that

$$\bar{M} \geq \max\{\varphi_j(u)\,;\ u \in U,\ j = 1, \dots, p\}.$$

Thanks to the nonnegativity of $e_i(\cdot)$ it is easy to verify that the optimal values of problems (17), (18), (19) and (24)–(27) coincide. Moreover, if $\hat{u}(\cdot)$ is an optimal control of the former problem, then $(\hat{u}(\cdot), \varphi(\hat{u}(\cdot))$ is an optimal control to the latter. The same applies also to the pair consisting of the discrete-time problem (20), (21), (22) and the following one:

$$g(x_N) + x_N^0 \longrightarrow \min, \tag{28}$$

$$x_{k+1}^0 = 0.5h \sum_{k=0}^{N-1} [e_0(t_k, x_k) + e_0(t_{k+1}, x_{k+1}) + \langle e(t_k, x_k) + e(t_{k+1}, x_{k+1}), v_k \rangle], \tag{29}$$

$$x_{k+1} = x_k + 0.5h[f(t_k, x_k, u_k) + f(t_{k+1}, x_k + hf(t_k, x_k, u_k), u_k)], \tag{30}$$

$$(u_k, v_k) \in W, \quad k = 0, \dots, N-1. \tag{31}$$

Let $\hat{u}(\cdot)$ be an optimal control of (17), (18), (19). Since $(\hat{u}(\cdot), \hat{v}(\cdot)) = \varphi(\hat{u}(\cdot))$ is an optimal control of (24)–(27), we can apply Theorem 2.2. Notice that the condition $\langle \frac{\partial e_j}{\partial x}, f_i \rangle = 0$ implies that Assumption 3 of Theorem 2.1 is fulfilled for the extended system (29), (30). Hence,

$$J(\hat{u}_N(\cdot), \hat{v}_N(\cdot)) \le \hat{J} + \frac{C}{N^2}, \tag{32}$$

where $J(u, v)$ is the performance value of (24)–(27) corresponding to the control function (u, v), and $\hat{u}_N, \hat{v}_N$ is the piecewise constant control with jumps only at the grid points $t_1, \dots, t_{N-1}$, obtained by local averaging of (u, v), as in (11), (12). Discretizing equations (25), (26) by using the chosen Runge–Kutta scheme (and the fact that the control is constant in each subinterval $[t_i, t_{i+1})$) and thanks to the equivalence of problems (20), (21), (22) and (28)–(31), we obtain the estimation

$$\hat{J}_N \le \hat{J} + \frac{C_1}{N^2}. \tag{33}$$

Now let $(\tilde{u}_0, \dots, \tilde{u}_{N-1})$ be an ε-optimal control in problem (20)–(22). Then $((\tilde{u}_0), \dots, \tilde{u}_{N-1}, \tilde{v}_0 = \varphi(\tilde{u}_0), \dots, \tilde{v}_{N-1} = \varphi(\tilde{u}_{N-1}))$ is an ε-optimal control to problem (28)–(31). Therefore, the continuous-time control

$$\tilde{u}(t) = \tilde{u}_k, \quad \tilde{v}(t) = \tilde{v}_k, \quad t \in [t_k, t_{k+1})$$

gives the value

$$J(\tilde{u}, \tilde{v}) \le \hat{J}_N + \varepsilon + \frac{C_2}{N^2};$$

therefore,

$$J(\tilde{u}) \le \hat{J}_N + \varepsilon + \frac{C_2}{N^2}.$$

This, together with (33), gives on one hand

$$\hat{J} \le \hat{J}_N + \varepsilon + \frac{C_2}{N^2},$$

and on the other hand

$$J(\tilde{u}) \le \hat{J} + \varepsilon + \frac{C_1 + C_2}{N^2}.$$

The constants C_1 and C_2 result from the constants in Theorem 2.2 and from the constant in the local $O(h^3)$-error of the Runge–Kutta scheme, which can also be estimated by the quantities mentioned in Remark 2. □

Finally we mention that, as it is shown in [22], formal application of higher order Runge–Kutta schemes (even to time-invariant linear optimal control problems) does not provide better than $O(h^2)$ estimation in (23), excepting some "special" cases.

4. Approximation of the reachable set. The issue of approximation of the reachable set of a control system (or differential inclusion) is treated in many papers, but wherever discretization schemes are employed, typically at most first-order estimation of the error is obtained (cf. [8, 14, 11, 9, 1]). The paper [22] develops a second-order approximation scheme applicable to linear control systems. Below, we extend this result, as an application of Theorem 2.1.

For two compact sets $P, Q \subset \mathbf{R}^n$ the Hausdorff semidistance from P to Q is defined as

$$H^+(P,Q) = \inf\{\alpha \ge 0\,;\ P \subset Q + \alpha\mathbf{B}\}$$

($\mathbf{B}$ is the unit ball in $\mathbf{R}^n$), and the Hausdorff distance is

$$H(P,Q) = \max\{H^+(P,Q), H^+(Q,P)\}.$$

If the set Q is convex, then one can represent

$$H^+(P,Q) = \sup_{|l|=1}\{0, \inf_{x\in Q}\langle l,x\rangle - \inf_{x\in P}\langle l,x\rangle\}. \tag{34}$$

The above relation will be applied in the cases $P = R(T)$ (the reachable set of (7) in the set $\mathcal{U}$ of admissible controls) and $Q = R_N(T)$ (the reachable set of (7) in the set $\mathcal{U}_N$ of admissible controls). If $R_N(T)$ happens to be convex, then in view of (34), one can choose the collection of performance indexes

$$\mathcal{G} = \{g(\cdot) = \langle l, \cdot\rangle\,;\ |l| = 1\},$$

which obviously fulfills Assumption 4 with $L_g = 1$ and $L'_g = 0$. Supposing that Assumptions 1–3 are also fulfilled, one can apply Theorem 2.1 to get the estimation

$$H(R_N(T), R(T)) \le \frac{C}{N^2},$$

where C is as in Theorem 2.2. Then a similar argument as in section 3 leads to a *set-valued* version of any single-step discretization scheme for ordinary differential equations. The approach is illustrated by the next theorem, which corresponds to the Runge–Kutta scheme used in section 3. The same scheme, but interpreted in a different way in the set-valued case, is employed by Lempio [16].

THEOREM 4.1. *Let Assumptions* 1–3 *from section* 2 *be fulfilled. Given* N*, define the sequence of sets* X_k^N *by*

$$X_{k+1}^N = \bigcup_{x\in X_k^N}\{x + 0.5h\tilde{f}_h(k,x,U)\}, \quad X_0^N = \{x_0\},\ \ k = 0,\dots,N-1, \tag{35}$$

where

$$\tilde{f}_h(k,x,u) = f(t_k,x,u) + f(t_{k+1}, x + hf(t_k,x,u), u).$$

Then there exists a constant C (*as in Theorem* 2.2) *such that for any* N *for which* X_N^N *is convex*

$$H(X_N^N, R(T)) \le \frac{C}{N^2}.$$

In the case of a linear system (7), the set-valued dynamical system (35) has the form

$$X_{k+1}^N = A_k X_k^N + B_k U$$

(cf. [22], where A_k and B_k are given explicitly), and X_N^N is apparently convex. Computer implementations of (35) are described in [13, 3].

If the convexity assumption for $R_N(T)$ is not satisfied (or cannot be verified), then nonlinear functions $g(\cdot)$ should be involved in a dual representation like (34). However, Theorem 2.1 imposes restrictions on the set of "test" functions $\mathcal{G}$, namely, uniform boundedness of the Lipschitz constant of the derivative. This restriction results in specific geometric requirements for the set $Q = R_N(T)$, necessary to ensure that the separation argument (needed in the proof of the nonconvex/nonlinear version of (34)) still works. Here we skip this analysis, quoting only a result from [21] that is independent of the geometry of $R_N(T)$, but the order of the estimation is worse.

THEOREM 4.2. *Let Assumptions* 1–3 *be fulfilled. Then there exists a constant* C (*as in Theorem* 2.2) *such that for any* N,

$$H(X_N^N, R(T)) \le \frac{C}{N^{1.5}},$$

where X_N^N *is the end state of the set-valued dynamic system* (35).

5. Time-step aggregation of discrete-time control systems. In this section we apply Theorem 2.1 to the following problem with obvious practical motivation. Consider a discrete-time control system

$$x_{i+1} = \mathcal{F}(i, x_i, u_i), \quad i = 0, \dots, N-1, \;\; x_0 \text{ given}, \tag{36}$$

where $x_i \in \mathbf{R}^n$, $u_i \in U \subset \mathbf{R}^r$. Generally speaking, we shall discuss the issue of approximation of (36) by another discrete-time system with significantly smaller number of steps, eventually sacrificing performance value. A natural way to do this is to keep the input vector u_i constant on M successive steps, changing its value only at the "moments" $i = kM$. Presumably $N = KM$, so that from a control point of view, system (36) becomes a K-step system (and can be explicitly approximately rewritten as such, as shown at the end of the section). However, the main point will be the analysis of the loss of performance, resulting from the aggregation of the input values.

The function $\mathcal{F}$ is supposed to be in the form

$$\mathcal{F}(i,x,u) = x + h(p_0(i,x) + P(i,x)u), \tag{37}$$

where $p_0(i,\cdot) : \mathbf{R}^n \mapsto \mathbf{R}^n$, $P(i,\cdot) = [p_1(i,\cdot), \dots, p_r(i,\cdot)] : \mathbf{R}^n \mapsto \mathbf{R}^{n\times r}$. The increment factor h is supposed to be "small," while the number of steps N is "large," so that $T = hN$ is a "finite" number.

The time-aggregated system has the form

$$y_{k+1} = \mathcal{F}_M(k, y_k, v_k), \quad k = 0, \dots, K-1, \quad y_0 = x_0, \tag{38}$$

where

$$\mathcal{F}_M(k, y, v) = \mathcal{F}((k+1)M - 1, \cdot, v) \circ \cdots \circ \mathcal{F}(kM, \cdot, v)(y) \tag{39}$$

is the M-times iterated value of y with fixed v.

For any sequence $u_0, \dots, u_{N-1}$, $u_i \in U$, the corresponding (according to (36)) sequence $x_0, \dots, x_N$ is called a trajectory of (36). Similarly, for any sequence $v_0, \dots, v_{K-1}$, $v_i \in U$, the corresponding (according to (38)) sequence $y_0, \dots, y_K$ is a trajectory of (38). In order to compare the two sets of trajectories (denoted further by X_N and Y_K, respectively), we define the following criterion, as in section 1. Let $\mathcal{G} = \{g(\cdot)\}$ be a prescribed collection of functions $g : \mathbf{R}^n \mapsto \mathbf{R}$, which are Lipschitz continuous together with their first derivatives, with a common Lipschitz constant L_g. The value

$$\delta(N, K) = \sup_{g \in \mathcal{G}} \{ \inf_{(y_0, \dots, y_K) \in Y_K} g(y_K) - \inf_{(x_0, \dots, x_N) \in X_N} g(x_N) \}$$

will be considered as a measure of the loss of performance when passing from (36) to (38). In particular, if $\mathcal{G}$ consists of a single function g, then $\delta(N, K)$ is the difference between the optimal values of the problems $\min g(y_K)$ and $\min g(x_N)$ subject to (38) and (36), respectively. If $\mathcal{G}$ consists of all linear functionals with unit norm and the reachable sets of (36) and (38) are convex, then $\delta(N, K)$ is just the Hausdorff distance between the reachable sets.

Under the assumptions listed below, one can obtain in a standard way that for each $M > 1$

$$\delta(N, K) \leq \text{const} \frac{1}{K}. \tag{40}$$

Below we obtain the estimation

$$\delta(N, K) \leq C \left(\frac{1}{N} + \frac{1}{K^2} \right), \tag{41}$$

which is essentially better than (40). In particular, for the reasonable choice $K = M = \sqrt{N}$ we obtain estimation Ch, while (40) gives $C\sqrt{h}$. Moreover, the proof allows us to explicitly obtain aggregated systems of the type shown in (38) (without iterating the operator $\mathcal{F}$ as in (39)), for which (41) is satisfied.

We suppose the following.

Assumption $1'$. There is a constant R such that

$$|p_j(i, x)| \leq R(1 + |x|) \tag{42}$$

for every $x \in \mathbf{R}^n$, $j = 0, \dots, r$, $i = 0, \dots, N-1$.

Denote by S the unit ball in $\mathbf{R}^n$ centered at the origin and with the following radius: $(1 + |x_0|) \exp(3R(1 + r|U|))$.

Assumption $2'$. The components of $p_0(i, \cdot)$ and $P(i, \cdot)$ and all their first derivatives are Lipschitz continuous (with a constant $L_{\mathcal{F}}$) in the set S. Moreover, there are constants d_0 and d_1 such that

$$|p_j(i+1, x) - p_j(i, x)| \leq h d_0, \quad i = 0, \dots, N-1, \tag{43}$$

$$|p_j(i+1,x) - 2p_j(i,x) + p_j(i-1,x)| \le 2h^2 d_1, \quad i = 1,\dots,N-1, \tag{44}$$

for any $x \in S$ and $j = 0,\dots,r$.

Assumption 3′.

$$[p_\alpha(i,\cdot), p_\beta(j,\cdot)](x) = 0$$

for every $x \in S$, $\alpha,\beta = 1,\dots,r$ and for those $i,j \in \{0,\dots N-1\}$ for which $|i-j| \le 4$.

The growth condition (42) replaces the more general boundedness condition in Assumption 1 just for technical convenience. Obviously, Assumption 3′ is automatically fulfilled if P is independent of x or if $r = \dim u = 1$. Conditions (43) and (44) are the discrete analogs of the differentiability conditions with respect to t in Assumption 2.

THEOREM 5.1. *Let the right-hand side of* (36) *have the form of* (37). *Then there exists a constant C depending only on R, L_g, $L_{\mathcal{F}}$, M, d_0, d_1, $|U|$, and T, such that* (41) *is satisfied for any N, M (for which $K = N/M$ is natural), for any functions p_0 and P for which Assumptions* 1′–3′ *are fulfilled, and for any $h \le T/N$.*

In the proof we use the following spline interpolation lemma, which can be verified by direct inspection.

LEMMA 5.2. *Let $\xi_0,\dots,\xi_N \in \mathbf{R}^n$ be given, $t_i = ih$, $h = T/N$. Denote*

$$\Delta_1(i) = \frac{\xi_{i+1} - \xi_{i-1}}{2h},$$

$$\Delta_2(i) = \frac{\xi_{i+1} - 2\xi_i + \xi_{i-1}}{2h^2}, \quad i = 1,\dots,N-1$$

(by definition, $\Delta_2(0) = \Delta_2(N) = 0$, $\Delta_1(0) = (\xi_1 - \xi_0)/h$, $\Delta_1(N) = (\xi_N - \xi_{N-1})/h$). Assume that $|\Delta_1(i)| \le d_0$, $|\Delta_2(i)| \le d_1$. Then the function

$$\xi(t) = \begin{cases} \xi_i + (t-t_i)\Delta_1(i) + \frac{(t-t_i)^2}{2}(3\Delta_2(i) - \Delta_2(i+1)) \ \text{ for } \ t \in [t_i, t_i + \frac{h}{2}], \\ \xi_{i+1} + (t-t_{i+1})\Delta_1(i+1) + \frac{(t-t_{i+1})^2}{2}(3\Delta_2(i+1) - \Delta_2(i)), \quad t \in (t_i + \frac{h}{2}, t_{i+1}) \end{cases}$$

has the following properties:

(1) *$\xi(t_i) = \xi_i$, $i = 0,\dots,N$;*

(2) *$\xi(\cdot)$ is Lipschitz continuous with constant $d_0 + 2hd_1$;*

(3) *$\xi(\cdot)$ is Lipschitz continuous with constant $4d_1$;*

(4) *$t_0,\dots,t_N$ being fixed, for $t \in [t_i, t_{i+1})$ the value $\xi(t)$ depends only on ξ_{i-1}, ξ_i, ξ_{i+1}, ξ_{i+2}; the dependence is linear, with coefficients that depend only on $\Delta_0(i)$ and $\Delta_1(i)$;*

(5) *$|\xi(t)| \le 2.5 \max\{|\xi_0|,\dots,|\xi_N|\}$.*

Proof of Theorem 5.1. For each $j = 0,\dots,r$ we apply Lemma 5.2 for $\xi_i = p_j(i,x)$, where x is considered as a parameter. The suppositions of the lemma are satisfied according to (43) and (44) with the same d_0 and d_1. Denote by $f_j(\cdot,x)$ the resulting interpolation spline $\xi(\cdot)$. Thanks to property (4) of $\xi(\cdot)$ in Lemma 5.2 the function $f_j(t,\cdot)$ is differentiable with a Lipschitz continuous derivative whose Lipschitz constant can be estimated by d_0 and d_1. The same applies to $f_j(\cdot,x)$ for any $x \in S$, because of properties (2) and (3) of $\xi(\cdot)$. Thus, Assumption 2 from section 2 is fulfilled for f_0 and $F = (f_1,\dots,f_r)$. Assumption 3 is also satisfied because of Assumption 3′ and the property (4) of $\xi(\cdot)$.

Now consider the control system

$$\dot{x} = f_0(t,x) + F(t,x)u, \quad x(0) = x_0, \quad u \in U. \tag{45}$$

According to property (5) in Lemma 5.2, each $f_j(t,\cdot)$ satisfies the growth condition (42) with constant $3R$ instead of R. Therefore, Assumption 1 from section 2 is also fulfilled with S defined above in the present section.

Because of property (1) in Lemma 5.2, system (36), (37) can be viewed as an Euler discretization of (45) with step length h.

From [8] we obtain the following estimation of the Hausdorff distance between the reachable set $R(T)$ of (45) on $[0,T]$ and the reachable set R_N of (36):

$$H(R_N, R(T)) \le \frac{C_1}{N}, \tag{46}$$

where C_1 depends only on the constants listed in the formulation of the theorem (see also the first paragraph of the proof). Because of the equi-Lipschitz continuity of the functions $g \in \mathcal{G}$ we obtain

$$\sup_{g\in\mathcal{G}} \left| \inf_{x_N \in R_N} g(x_N) - \inf_{x(T)\in R(T)} g(x(T)) \right| \le \frac{C_1 L_g}{N}. \tag{47}$$

On the other hand, since Assumptions 1–4 are fulfilled by system (45), we can apply Theorem 2.1 for K (instead of N) subintervals of constant control. If $R_K(T)$ is the corresponding reachable set of (45) in the set $\mathcal{U}_K$ of admissible controls, we obtain the estimation

$$\sup_{g\in\mathcal{G}} \left| \inf_{x_K(T)\in R_K(T)} g(x_K(T)) - \inf_{x(T)\in R(T)} g(x(T)) \right| \le \frac{C_2}{K^2}. \tag{48}$$

For any $u(\cdot) \in \mathcal{U}_K$ one can discretize (45) by the Euler scheme with step h and come back to the discrete system (36), but with $u_{kM} = u_{kM+1} = \cdots = u_{(k+1)M-1}$. The accuracy with respect to the performance $\mathcal{G}$ is as in (46), which together with (47), (48) completes the proof. □

If instead of the Euler discretization scheme we apply in the last step of the proof the Euler–Cauchy scheme (or some other Runge–Kutta one) with step $\tau = T/K = Mh$, we obtain

$$x_{k+1} = x_k + 0.5\tau(f(t_k,x_k,u_k) + f(t_{k+1}, x_k + \tau f(t_k,x_k,u_k), u_k)), \quad k = 0,\dots,K-1,$$

where $f = f_0 + Fu$, $t_k = k\tau$, and u_k is the value of $u(\cdot)$ in (t_k, t_{k+1}). It is well known that

$$|x(T) - x_N| \le C_3/K^2.$$

Hence, denoting for $k = 0,\dots,K-1$

$$\mathcal{F}_M(k,y,v) = y + 0.5\tau(\mathcal{F}(kM,y,v) + \mathcal{F}((k+1)M, y + \tau\mathcal{F}(k,y,v), v)), \tag{49}$$

we obtain for the reachable set $\tilde{R}_K$ of the corresponding system (38) the estimation

$$H(\tilde{R}_K, R_K(T)) \le \frac{C_3}{K^2}.$$

Combining this with (47) and (48) we get

$$\sup_{g\in\mathcal{G}} \left| \inf_{y_K\in\tilde{R}_K} g(y_K) - \inf_{x_N\in R_N} g(x_N) \right| \le \frac{C_1 L_G}{N} + \frac{C_2 + C_3 L_G}{K^2} = C\left(\frac{1}{N} + \frac{1}{K^2}\right).$$

The right-hand side of the K-stage system (38) is here defined by (49) instead of (39), but the estimation (41) still holds. In fact, (49) is an approximation of (39), which avoids the iteration in (39).

REFERENCES

[1] Z. Artstein, *First order approximations for differential inclusions*, Set-Valued Anal., 2 (1994), pp. 2–17.

[2] J.-P. Aubin and H. Frankowska, *Set-Valued Analysis*, Birkhäuser, Boston, 1990.

[3] R. Baier, *Mengenwertige Integration und die Diskrete Approximation Erreichbarer Mengen*, Ph.D. thesis, Universität Bayreuth, Germany 1995; Bayreuther Mathematische Schriften, Heft 50.

[4] R. Baier and F. Lempio, *Computing Aumann's integrals*, in Modeling Techniques for Uncertain Systems, A. Kurzhanski and V. Veliov, eds., Progress in Systems and Control Theory 18, Birkhäuser, Boston, 1994, pp. 71–90.

[5] B. Doitchinov and V. Veliov, *Parametrisations of integrals of set-valued mappings and applications*, J. Math. Anal. Appl., 179 (1993), pp. 483–499.

[6] A. Dontchev, *Discrete approximations in optimal control*, in Nonsmooth Analysis and Geometric Methods in Optimal Control, R. R. B. Mordukhovich and H. Sussmann, eds., Proc. IMA, Springer-Verlag, New York, 1994.

[7] A. Dontchev, *An a priori estimate for discrete approximations in nonlinear optimal control*, SIAM J. Control Optim., 34 (1996), pp. 1315–1328.

[8] A. Dontchev and E. Farkhi, *Error estimates for discretized differential inclusions*, Computing, 41 (1989), pp. 349–358.

[9] A. Dontchev and W. Hager, *Euler approximation to the feasible set*, Numer. Funct. Anal. Optim., 15 (1994), pp. 245–262.

[10] A. Dontchev and F. Lempio, *Difference methods for differential inclusions: A survey*, SIAM Review, 34 (1992), pp. 263–294.

[11] G. Häckl, *Numerical approximation of reachable sets and control sets*, Random Comput. Dynamics, 1 (1992–1993), pp. 371–394.

[12] W. Hager, *Rate of convergence for discrete approximations to unconstrained control problems*, SIAM J. Numer. Anal., 13 (1976), pp. 449–472.

[13] M. Krastanov and N. Kirov, *Dynamic interactive system for analysis of linear differential inclusions*, in Modeling Techniques for Uncertain Systems, A. Kurzhanski and V. Veliov, eds., Progress in Systems and Control Theory 18, Birkhäuser, Boston, 1994, pp. 123–130.

[14] A. B. Kurzhanski and I. Valyi, *Ellipsoidal techniques for dynamic systems: Control synthesis for uncertain systems*, Dynamics and Control, 2 (1992), pp. 87–111.

[15] F. Lempio, *Difference Methods for Differential Inclusions*, Lecture Notes in Econom. and Math. Systems 378, Springer-Verlag, New York, 1992, pp. 236–273.

[16] F. Lempio, *Modified Euler methods for differential inclusions*, in Set-Valued Analysis and Differential Inclusions, A. Kurzhanski and V. Veliov, eds., Progress in Systems and Control Theory 16, Birkhäuser, Boston, 1993, pp. 131–148.

[17] B. Mordukhovich, *Discrete approximations and refined Euler-Lagrange conditions for nonconvex differential inclusions*, SIAM J. Control Optim., 3 (1995), pp. 882–915.

[18] G. Redien, *Collocation at Gauss points as a discretization in optimal control*, SIAM J. Control Optim., 2 (1979), pp. 298–306.

[19] D. Silin, *On discontinuous strategies in optimal control problems*, J. Math. Systems Estim. Control, 4 (1994), pp. 205–217.

[20] V. Veliov, *Second order discrete approximations to strongly convex differential inclusions*, Systems Control Lett., 13 (1989), pp. 263–269.

[21] V. Veliov, *Parametric and functional uncertainties in dynamical systems: Local and global relationship*, in Computer Arithmetic and Enclosure Methods, L. Atanassova and J. Herzberger, eds., North-Holland, Amsterdam, 1992.

[22] V. Veliov, *Second order discrete approximations to linear differential inclusions*, SIAM J. Numer. Anal., 29 (1992), pp. 439–451.

[23] P. Wolenski, *The exponential formula for the reachable set of Lipschitz differential inclusion*, SIAM J. Control Optim., 28 (1990), pp. 1148–1161.

SIAM J. CONTROL OPTIM.
Vol. 35, No. 5, pp. 1487–1507, September 1997

003

TRACKING FAST TRAJECTORIES ALONG A SLOW DYNAMICS: A SINGULAR PERTURBATIONS APPROACH*

ZVI ARTSTEIN† AND VLADIMIR GAITSGORY‡

Abstract. Controlled coupled slow and fast motions are examined in a singular perturbations setting. The objective is to minimize a cost functional that takes into account both the fast motion, supposing, say, tracking a fast target, and the slow dynamics. A method is offered to cope with the possibility that the fast flow has nonstationary limits. Invariant measures of the fast motion are then the controlled objects on the infinitesimal scale. Optimal amalgamation of them on the slow scale induces the variational limit, whose solutions are near optimal solutions of the perturbed system.

Key words. singular perturbations, chattering systems, tracking, invariant measures

AMS subject classifications. 49J15, 93C15, 49N10

PII. S036301299528458X

1. Introduction.

This paper examines systems where a control policy $u(t)$ determines simultaneously two moving states. One—say, $x(t)$—moves at an ordinary pace, while the second—say, $y(t)$—progresses much faster. An example that motivates our analysis is the case where $y(t)$ is supposed to track a prescribed target $\Gamma(t)$, which also progresses very fast and whose characteristics may depend on the slow moving state. The controller's goal is to minimize a cost functional that takes into account both the success in tracking and the performance of the relatively slow motion. As it is assumed that both $y(t)$ and $\Gamma(t)$ evolve considerably faster than $x(t)$, their motion could be modeled as a singular perturbation relative to the slow time scale.

Indeed, we address a singularly perturbed model, where the small parameter ϵ reflects the speed ratio of the slow and fast movements. As customary, we look for a control rule associated with the limit optimization problem as $\epsilon \to 0$, and hope that it will generate a near optimal policy for $\epsilon > 0$ small. To this end, it is useful to identify a nominal control problem associated with the limit case. This nominal problem should be such that the case $\epsilon > 0$ small can be regarded as a small perturbation.

In many control and optimal control of singular perturbations problems, the reduced-order system (where ϵ is actually set to be equal to 0) serves as an appropriate nominal problem. For a partial list of successful applications of this approach, consult Kokotovic, Khalil, and O'Reilly [13]; Kokotovic and Sannuti [15]; Chow and Kokotovic [8]; Kokotovic, O'Malley, and Sannuti [14]; Saksena, O'Reilly, and Kokotovic [20]; and Kokotovic [11]. The reduced-order approach, however, may not be adequate for the analysis of the fast tracking. Indeed, the technique assumes that the fast dynamics converges on the fast scale to an equilibrium (that may vary on the slow scale). Such a consideration may not apply to optimal tracking of a fast moving target, which may exhibit nontrivial dynamics even relative to the fast time scale.

The present paper offers a nominal limit problem which is able to analyze dynamic limits. It is based on ideas rooted in earlier works of the authors; see [1], [3], [9], [10], and references therein and also Vigodner [21], where the case of slowly varying

*Received by the editors April 17, 1995; accepted for publication (in revised form) June 3, 1996. http://www.siam.org/journals/sicon/35-5/28459.html

†Department of Theoretical Mathematics, The Weizmann Institute of Science, Rehovot 76100, Israel (mtarts@weizmann.weizmann.ac.il).

‡School of Mathematics, University of South Australia, The Levels, Pooraka, South Australia 5095, Australia (mavg@lux.levels.unisa.edu.au).

controls is considered. The essence of our approach is that the controller solves on an infinitesimal scale dynamic control problems associated with the fast motion, and integrates their solutions on the slow scale. The integrated control policy is, under appropriate conditions, a near optimal policy for the original singular perturbations problem. In this paper we do not investigate thoroughly mathematical aspects such as finding general conditions under which the procedure works. Rather, we do not hesitate to assume that the solution to the limit problem exists, and we prove that it is near optimal. In the closing section of the paper we display a broad class of tracking problems for which the solution can be exhibited.

The paper is organized as follows. Sections 2 to 9 are devoted to the general theory, while in the last two sections of the paper we address applications. The model is set in section 2. Some preliminaries on invariant measures are given in section 3. The notion of near optimal solutions is defined in section 4. Sections 5 and 6 introduce and analyze control policies operating on the infinitesimal and global time scales, respectively. These are related to the limit problem as $\epsilon \to 0$, and the stability of them with respect to the singular perturbations is checked in section 7. In sections 8 and 9 we formulate the limit problem and verify when solutions of it are indeed near optimal solutions.

2. Setting the model. In what follows, x, which represents the slow state, is an element in R^n, the n-dimensional Euclidean space. The fast state y is in R^m. The admissible controls are Borel-measurable functions into a prescribed constraint set $U \subset R^k$. The time variable is t. We use interchangeably the notations $\frac{dx}{dt}$ and $\dot{x}$ to denote derivation with respect to time.

The cost function is generated by a function

$$c(x, y, u) : R^n \times R^m \times U \to R. \tag{2.1}$$

The underlying problem is then as follows. (SP stands for singular perturbations.)

SP Problem 2.1.

$$\text{minimize} \quad \int_0^1 c(x(t), y(t), u(t))dt \tag{2.2}$$

subject to

$$\begin{aligned} \frac{dx}{dt} &= f(x, y, u), \\ \epsilon\frac{dy}{dt} &= g(x, y, u) \end{aligned} \tag{2.3}$$

with initial conditions

$$x(0) = \overline{x}, \quad y(0) = \overline{y} \tag{2.4}$$

and where the minimization is over all admissible controls $u(t) : [0, 1] \to U$. The infimal value, as well as the optimal policies, may depend on ϵ. By

$$\text{val}(\epsilon) \tag{2.5}$$

we denote the infimal value of the problem for $\epsilon > 0$ fixed.

Remark 2.2. This is an important remark. We did not include explicit dependence of the equations on the slow time t. This was done only for the sake of clarity. The

state x could include the time t, with an additional equation added, namely, $\dot{t} = 1$, to the slow part of the equation. The time t could also be a variable in the cost c, given in (2.1) and (2.2).

We collect here some technical assumptions of the model that are assumed throughout the paper.

Assumption 2.3. The functions $c(x, y, u)$, $f(x, y, u)$, and $g(x, y, u)$ are continuous on their respective domains.

Assumption 2.4. For every x fixed and an admissible control $u(t) : [0, \infty) \to U$, if there exists a solution to the equation $\dot{y} = g(x, y, u(t))$, $y(0) = y_0$, then it is unique.

The standing assumptions are not strong. This is possible as in later developments the existence of solutions with prescribed properties is assumed or checked directly rather than derived from the underlying assumptions (see the introduction). Note, however, that we assume that the functions f, g, and c are defined globally. This can be clearly relaxed by introducing partial domains. We leave out the details.

3. Invariant measures. For the convenience of the reader, we briefly recall here the notion of invariant measure. It plays a major role in our definition of a near optimal control. For background information consult Nemytskii and Stepanov [18, Chapter VI.9] or [3].

A probability measure—say, μ—on a closed subset S of R^d is a countably additive map defined on the Borel subset of S, with values in [0,1] and $\mu(S) = 1$. Let $\dot{\sigma} = L(\sigma)$ be a differential equation on R^d. Suppose that for each $\sigma_0 \in S$ there exists a unique solution $\varphi(t, \sigma_0)$ satisfying $\varphi(0, \sigma_0) = \sigma_0$ of the differential equation and $\varphi(t, \sigma_0) \in S$ for $t \geq 0$. A probability measure μ on S is invariant with respect to $\dot{\sigma} = L(\sigma)$ if

$$\mu(C) = \mu(\varphi(t, C)) \tag{3.1}$$

for all $C \subset S$ closed and all $t \geq 0$ (here $\varphi(t, C) = \{\varphi(t, c) : c \in C\}$).

The support of a measure μ, denoted supp μ, is the smallest closed set C such that $\mu(C) = 1$. It is easy to see that if $\sigma_0 \in$ supp μ, then $\varphi(t, \sigma_0) \in$ supp μ for all t; hence (3.1) holds actually for all $t \in R$.

If S is compact, then an invariant measure of $\dot{\sigma} = L(\sigma)$ exists. If there exists a unique invariant probability measure μ, then any solution $\sigma(t)$ which is bounded for $t \geq 0$ converges in distribution to μ in the following sense. If ν_T is the measure on S given by

$$\nu_T(C) = \frac{1}{T}\lambda\{t : 0 \leq t \leq T, \sigma(t) \in C\}, \tag{3.2}$$

where λ is the Lebesgue measure, then ν_T converge weakly to μ (see the following paragraph). This means in particular that the asymptotic statistics of the values of $\sigma(\cdot)$ is governed by the unique invariant measure, namely,

$$\frac{1}{T}\int_0^T h(\sigma(\tau))d\tau \to \int_{R^d} h(z)\mu(dz) \tag{3.3}$$

for all $h : R^d \to R$ bounded and continuous.

The previous claim follows directly from the definition of weak convergence. The sequence μ_k of probability measures on R^d converges weakly to μ_0 if

$$\int_{R^d} h(z)\mu_k(dz) \to \int_{R^d} h(z)\mu_0(dz) \tag{3.4}$$

for every bounded and continuous function $h : R^d \to R$. See, e.g., [5].

4. Near optimal solutions. The notion of near optimality is formally introduced in this section. In essence it does not differ from the standard considerations of singularly perturbed problems—say, for the reduced order form (see, e.g., Chow and Kokotovic [8]). The novelty in the dynamic limit form is that we allow feedback of a clock variable as follows.

DEFINITION 4.1. *Let $S \subset R^\ell$ be bounded. A differential equation $\dot\sigma = L(\sigma)$ defined for $\sigma \in S$ is called a* clock *if for every $\sigma_0 \in S$ there exists a unique solution $\sigma(t)$ satisfying $\sigma(0) = \sigma_0$ and $\sigma(t) \in S$ for $t \geq 0$. A clock is called a* proper clock *if the equation has a unique invariant measure and every solution $\sigma(t)$ converges to it in distribution (see previous section). We shall also consider a* parametrized proper clock*, namely, an equation $\dot\sigma = L(x, \sigma)$ for $\sigma \in S$, which is a proper clock for each fixed x.*

The controls $\mathbf{u}$ that we consider as candidates for solving the singular perturbation problem 2.1 have the form

$$u(x, y, \sigma, t) : R^n \times R^m \times S \times [0, 1] \to U, \tag{4.1}$$

where $\sigma = \sigma(t)$ is a solution of

$$\epsilon \frac{d\sigma}{dt} = L(x, \sigma), \quad \sigma(0) = \sigma_0, \tag{4.2}$$

with $\dot\sigma = L(x, \sigma)$ being a parametrized proper clock on a set $S \subset R^\ell$; namely, we allow feedback of the slow and fast moving states, the clock variable and the time.

Remark 4.2. Although we have stated in Remark 2.2 that the slow state x may include the variable t and thus no explicit dependence on t is needed, we choose to have t as an explicit variable in (4.1). This is done to emphasize that, even if the state equations are time-invariant, an optimal control may be time-varying.

When the control $\mathbf{u}$ of (4.1) is inserted into the system (2.4)–(2.5) and (4.2), it yields (under standard assumptions) a solution which depends of course on ϵ. We denote it by

$$(x_\epsilon(t), y_\epsilon(t), \sigma_\epsilon(t)). \tag{4.3}$$

Note that although the clock equation in (4.2) is not affected by the control, there is an indirect dependence through the x variable.

We do not demand that the solution (4.3) be the unique solution of the control equation, although in most cases it is. The reason is that our results hold without the uniqueness assumption. (We shall need uniqueness for the limit problem, though.) But, in order to simplify notations, we suppress the dependence on the triplet (4.3) when defining $c_\epsilon(\mathbf{u})$, the cost of $\mathbf{u}$, as follows:

$$c_\epsilon(\mathbf{u}) = \int_0^1 c(x_\epsilon(t), y_\epsilon(t), u(x_\epsilon(t), y_\epsilon(t), \sigma_\epsilon(t), t))dt; \tag{4.4}$$

namely, $c_\epsilon(\mathbf{u})$ is the cost when applying $\mathbf{u}$ to Problem 2.1 with ϵ given. Note, however, that when applying the control $\mathbf{u}$, the precise value of ϵ need not be available. Instead, the control employs the variable σ, which is affected by ϵ through (4.2).

DEFINITION 4.3. *The control policy $\mathbf{u}$ is near optimal if*

$$c_\epsilon(\mathbf{u}) - \mathrm{val}(\epsilon) \to 0 \tag{4.5}$$

as $\epsilon \to 0$.

A motivation to consider a control function that depends on the auxiliary clock variable σ is as follows. Consider the case where the fast variable $y(t)$ has to track a fast target, with dynamical characteristics such as periodicity, almost periodicity, recurrence, etc., which could vary with the slow dynamics. A reasonable description for such a target is $\Gamma(x, \sigma(t))$, with $\sigma(t)$ a clock variable which solves an equation of the form (4.2).

For instance, the equation $\dot{\sigma} = 1 - \sigma$ is a proper clock on $[0, 2]$. The corresponding motion of a target $\Gamma(\sigma(t))$ exhibits a convergence to the stationary point $\Gamma(1)$. The convergence becomes faster when a small ϵ multiplies the derivative. A clock that models a periodic motion is obtained by setting $S = S^1$, the unit sphere $\{(\xi, \eta) : \xi^2 + \eta^2 = 1\}$ in R^2, and letting $\sigma = (\xi, \eta)$ be parametrized by $\arg(\xi + i\eta)$; namely, σ is the angle in the polar coordinates. Then setting $\dot{\sigma} = 1$ induces a periodic motion $\Gamma(\sigma(t))$ of the target. If $\dot{\sigma} = L(x)$ with $L(x) > 0$, then the induced target $\Gamma(x, \sigma(t))$ exhibits periodicity in time, with period and location depending on the slow state. A target whose movement is generated by two incommensurable periods—say, α and β—can be modeled by a clock on $S^1 \times S^1$, with (σ_1, σ_2) being the clock variable, and the differential system $\dot{\sigma}_1 = 2\pi\alpha^{-1}$, $\dot{\sigma}_2 = 2\pi\beta^{-1}$. In these examples there is clearly a unique invariant measure on S.

This general class of problems has the form

$$\text{minimize} \int_0^1 c(x(t), y(t), u(t), \Gamma(x(t), \sigma(t)))dt \tag{4.6}$$

subject to

$$\begin{aligned} \frac{dx}{dt} &= f(x, y, u), \\ \epsilon\frac{dy}{dt} &= g(x, y, u), \\ \epsilon\frac{d\sigma}{dt} &= L(x, \sigma), \end{aligned} \tag{4.7}$$

which is a particular case of (2.2)–(2.3), namely, with (y, σ) being the fast dynamics. It is then reasonable that the near optimal controls will employ the clock variable which induces the target. A concrete example is solved in the closing section.

5. Infinitesimal control policies. In the present and the next section we identify a class of control policies, among which we shall find the near optimal solutions. In this section we study their infinitesimal properties. In the next section we examine their global structure.

DEFINITION 5.1. *Let x be fixed. (x could incorporate an instant t of the slow time; see Remark* 2.2.*) Let $\dot{\sigma} = L(x, \sigma)$ be a parametrized proper clock on $S \subset R^\ell$. The control function*

$$u(x, y, \sigma) : R^n \times R^m \times S \to U \tag{5.1}$$

is called proper *if the following hold.*

(i) *$u(x, y, \sigma)$ is continuous, and for every $(y_0, \sigma_0) \in R^m \times S$, the system*

$$\begin{aligned} \frac{dy}{d\tau} &= g(x, y, u(x, y, \sigma)), \\ \frac{d\sigma}{d\tau} &= L(x, \sigma) \end{aligned} \tag{5.2}$$

with initial conditions $y(0) = y_0$, $\sigma(0) = \sigma_0$ *has a unique solution* $(y(\tau), \sigma(\tau))$, *which is defined and bounded on* $[0, \infty)$, *and*

(ii) *There is a unique invariant probability measure of* (5.2) *on* $R^m \times S$, *and it is with compact support.*

Remark 5.2. Condition (ii) may need an explanation. Recall that the clock equation $\dot{\sigma} = L(x, \sigma)$ is assumed to have a unique invariant measure; see Definition 4.1. Thus, the asymptotic statistics of $\sigma(\tau)$, for x fixed, is determined by this invariant measure. It is therefore plausible that the trajectory $y(\tau)$, which, say, tracks $\Gamma(x, \sigma(\tau))$, will also inherit this dynamical characteristic. This is indeed the case in the available tracking examples; see, e.g., [2] for the periodic case. The example treated in the closing section is of such nature.

PROPOSITION 5.3. *Let* $u(x, y, \sigma)$ *be a proper control function, and let* ν *be the invariant probability measure on* $R^m \times S$ *associated with it. Let* $(y(\tau), \sigma(\tau))$ *be the solution of* (5.2) *with initial conditions* (y_0, σ_0). *Then the limit*

$$\lim_{T\to\infty} \frac{1}{T} \int_0^T f(x, y(\tau), u(x, y(\tau), \sigma(\tau)))d\tau \tag{5.3}$$

exists, and regardless of (y_0, σ_0), *it is equal to*

$$\int_{R^m\times S} f(x, y, u(x, y, \sigma))\nu(dy \times d\sigma). \tag{5.4}$$

Likewise, the limit

$$\lim_{T\to\infty} \frac{1}{T} \int_0^T c(x, y(\tau), u(x, y(\tau), \sigma(\tau)))d\tau \tag{5.5}$$

exists and is equal to

$$\int_{R^m\times S} c(x, y, u(x, y, \sigma))\nu(dy \times d\sigma) \tag{5.6}$$

independently of (y_0, σ_0).

Proof. This is a version of the ergodic theorem. What follows is the argument in brief (see also the background in section 3). The probability measures ν_T induced on $R^m \times S$ by $(y(\tau), \sigma(\tau)) : [0, T] \to R^m \times S$ (see (3.2)) converge weakly to ν as $T \to \infty$. Since both functions $f(x, y, u)$ and $c(x, y, u)$ are assumed continuous and since $u(x, y, \sigma)$ is assumed continuous, it follows that the integrands in (5.4) and (5.6) are continuous in (y, σ). They are also bounded on the support of ν and the range of $(y(\tau), \sigma(\tau))$. Hence the weak convergence of ν_T to ν implies the respective convergences. □

Remark 5.4. The interpretation we give to the expressions (5.3)–(5.6) is as follows. If on an infinitesimal period, where the slow state x stays put, the fast state and clock solve (5.2) on an unbounded interval, then the limit in (5.3) constitutes the right-hand side of the slow equation. We say then that (5.4) is a velocity at x, infinitesimally generated by the control. Likewise, (5.5) represents the limit of the average cost of the operation on $[0, T]$. Thus (5.6) is the cost of infinitesimally generating (5.4) at x by using the control u.

Next, we extend the definition of a proper control function and consider dependence on v. We use the same adjective, namely, proper, but no confusion should arise, since the context dictates which version is applicable.

Convention 5.5. In what follows, we refer to domains D of the form

$$D \subset R^n \times R^m \times S \times R^n.$$

We then agree that if $(x, y, \sigma, v) \in D$, then $(x, y', \sigma', v) \in D$ for all $(y', \sigma') \in R^m \times S$. This is done for convenience only; see the comment closing section 2. Then we may say that (x, v) is in D, meaning that $\{x\} \times R^m \times S \times \{v\}$ is a subset of D.

DEFINITION 5.6. *Let* $\mathbf{u}$ *be given as*

$$u(x, y, \sigma, v) : D \to U$$

with domain $D \subset R^n \times R^m \times S \times R^n$. *We say that* $\mathbf{u}$ *is a* proper control policy *if the following hold:*

(iii) $\mathbf{u}$ *is continuous on* D.

(iv) *For each* (x, v) *fixed in* D, *the function* $\mathbf{u}$ *is a proper control function.*

(v) *For each* (x, v) *fixed in* D, *the control* $\mathbf{u}$ *infinitesimally generates* v *at* x, *namely,*

$$v = \int_{R^m \times S} f(x, y, u(x, y, \sigma, v)) \nu_{x,v}(dy \times d\sigma), \tag{5.7}$$

where $\nu_{x,v}$ *is the invariant probability measure associated with* $\mathbf{u}$ *for* (x, v) *fixed.*

(vi) *For* $K \subset R^n \times R^n$ *a bounded set in the domain* D, *all the invariant probability measures* $\nu_{x,v}$ *associated with* $\mathbf{u}$ *for* $(x, v) \in K$, *are supported in a common bounded subset of* $R^m \times S$.

PROPOSITION 5.7. *Let* $\mathbf{u}$ *be a proper control policy. The cost of generating* v *at* x *by* $\mathbf{u}$, *namely,*

$$\text{cost}(\mathbf{u}, x, v) = \int_{R^m \times S} c(x, y, u(x, y, \sigma, v)) \nu_{x,v}(dy \times d\sigma), \tag{5.8}$$

is a continuous function of (x, v) *in the domain of* $\mathbf{u}$.

Proof. The continuity of the functions $\mathbf{u}$ and c implies that the integrand in (5.8) is uniformly continuous on bounded sets in $R^m \times S$. Applying this to the common support of $\nu_{x,v}$ for (x, v) in a bounded set (see condition (vi)) would imply the desired continuity, provided that $\nu_{x,v}$ is weakly continuous as a function of (x, v). The weak continuity follows from the uniqueness; see [3, Proposition 3.2, Remark 3.3]. Thus the proof is complete. ∎

6. Global control policies. A proper control policy, as introduced in the previous section, generates the velocity v at x. This provides the option that the derivative of the slow motion at x will be v. But the infinitesimal policy by itself does not suffice to induce the slow motion $x(t)$. To this end we need to combine it with a policy on the slow scale. This is achieved as follows.

DEFINITION 6.1. *Let* $\mathbf{v}$ *be a function*

$$v(x, t) : R^n \times [0, 1] \to R^n \tag{6.1}$$

continuous in (x, t) *and such that for every initial condition* $x(0) = \overline{x}$, *the solution* $x(t)$ *of the differential equation*

$$\frac{dx}{dt} = v(x, t), \quad x(0) = \overline{x} \tag{6.2}$$

is determined uniquely. (*We allow* $\mathbf{v}$ *to be defined on a subset of* $R^n \times [0,1]$.) *Let* $\mathbf{u}$ *be a proper control policy. The pair* $(\mathbf{u}, \mathbf{v})$ *is called a* proper global control policy. *We say that* $(\mathbf{u}, \mathbf{v})$ *generates the solution* $x(\cdot)$ *of* (6.2) *if for all* $t \in [0,1]$, *the pair* $(x(t), v(x(t),t))$ *is in the domain of* $\mathbf{u}$.

Remark 6.2. In the following discussion we may consider $v(x,t)$, which is defined globally, but this would be done for convenience only; see the closing comment in section 2. The explicit dependence of v on t is done for the reason mentioned in Remark 4.2.

Remark 6.3. The reasoning behind Definition 6.1 is that $\mathbf{v}$ determines the slow evolution on the slow time scale, while $\mathbf{u}$ determines the desired velocities on the infinitesimal scale. The differential equation (6.2) then becomes a chattering equation (see [1], [3]), namely,

$$\frac{dx}{dt} = \int_{R^m \times S} f(x, y, u(x, y, \sigma, v(x,t)) \nu_{x, v(x,t)}(dy \times d\sigma), \tag{6.3}$$

where $\nu_{x,v}$ is the invariant probability measure associated with $\mathbf{u}$ at (x,v). It is in this sense that the pair $(\mathbf{u}, \mathbf{v})$ generates the trajectory $x(t)$ as a solution of (6.3). Indeed, (6.3) is defined only for (x,t) such that $(x, v(x,t))$ is in the domain of $\mathbf{u}$. Note that in this domain we require that solutions of (6.3) be uniquely determined by the initial conditions.

We now identify the cost of generating a trajectory $x(t)$ by $(\mathbf{u}, \mathbf{v})$. It will, naturally, be the accumulation of the infinitesimal costs, as follows.

DEFINITION 6.4. *Let* $(\mathbf{u}, \mathbf{v})$ *be a proper global control policy which generates* $x(t)$ *on* $[0,1]$. *The cost of generating the trajectory is given by*

$$\text{cost}(\mathbf{u}, \mathbf{v}) = \int_0^1 \text{cost}(\mathbf{u}, x(t), v(x(t),t)) dt, \tag{6.4}$$

with $\text{cost}(\mathbf{u}, x, v)$ *given in* (5.8). *Notice that denoting the cost by* $\text{cost}(\mathbf{u}, \mathbf{v})$ *suppresses the dependence on the initial condition* $x(0)$, *which later is assumed to be prescribed.*

7. Stability. The proper global control policies $(\mathbf{u}, \mathbf{v})$ are obtained as combinations of a policy $\mathbf{v}$ on the slow scale and a control $\mathbf{u}$ on the fast scale. The outcome is a chattering equation (6.3). But the composed control policy, given by

$$u(x, y, \sigma, v(x,t)),$$

is of the type described in section 4 and can in particular be applied to the singular perturbation Problem 2.1. For a given $\epsilon > 0$, applying $(\mathbf{u}, \mathbf{v})$ in this manner results in the equations

$$\begin{aligned} \frac{dx}{dt} &= f(x, y, u(x, y, \sigma, v(x,t)), \\ \epsilon \frac{dy}{dt} &= g(x, y, u(x, y, \sigma, v(x,t)), \\ \epsilon \frac{d\sigma}{dt} &= L(x, \sigma), \end{aligned} \tag{7.1}$$

with initial conditions (2.4). A solution $(x_\epsilon(t), y_\epsilon(t), \sigma_\epsilon(t))$ then gives rise to the cost

$$\int_0^1 c(x_\epsilon(t), y_\epsilon(t), u(x_\epsilon(t), y_\epsilon(t), \sigma_\epsilon(t), v(x_\epsilon(t),t))) dt, \tag{7.2}$$

which we denote by $c_\epsilon(\mathbf{u},\mathbf{v})$, suppressing the dependence on the specific solution; see the discussion concerning (4.4).

The stability result that follows employs the notions of a direct integral of probability measures, and of statistical convergence. These concepts are examined in detail in [3]. We briefly recall the definitions here in the framework of the theorem. Consider a function $(y_\epsilon(t),\sigma_\epsilon(t))$ defined on $[0,1]$ with values in $R^m \times S$. With this function we associate a measure ν_ϵ defined on $R^m \times S \times [0,1]$ and given by

$$\nu_\epsilon(D) = \lambda\{t : (y_\epsilon(t),\sigma_\epsilon(t),t) \in D\}, \tag{7.3}$$

where λ is the Lebesgue measure. We say that $(y_\epsilon(t),\sigma_\epsilon(t))$ converge statistically, as $\epsilon \to 0$, to a measure ν, if ν_ϵ given in (7.3) converges weakly to ν. The motivation is that ν, which is defined on $R^m \times S \times [0,1]$, dominates the asymptotic distribution of the graph of $(y_\epsilon(t),\sigma_\epsilon(t))$.

Next consider the probability measures ν_t defined on $R^m \times S$, for $t \in [0,1]$. The direct integral of ν_t is the measure ν on $R^m \times S \times [0,1]$ determined by

$$\nu(C \times E) = \int_E \nu_t(C)dt \tag{7.4}$$

for $C \subset R^m \times S$ and $E \subset [0,1]$ measurable. On the entire Borel field of $R^m \times S \times [0,1]$ the measure ν is obtained by a standard extension.

We now state and prove the stability of applying a proper global control policy to the singular perturbations problem.

THEOREM 7.1. *Let $(\mathbf{u},\mathbf{v})$ be a proper global control policy that generates the trajectory $x(t)$, with $x(0) = \overline{x}$. Suppose also that the domain of $\mathbf{u}$ contains an open neighborhood of the graph of $(x(t),v(x(t),t))$ in $R^n \times R^n$. Let $(x_\epsilon(t),y_\epsilon(t),\sigma_\epsilon(t))$ be a solution of (7.1) for $\epsilon > 0$ fixed. Then, as $\epsilon \to 0$,*

(a) *$c_\epsilon(\mathbf{u},\mathbf{v})$ converge to $\mathrm{cost}(\mathbf{u},\mathbf{v})$;*

(b) *$x_\epsilon(t)$ converge uniformly on $[0,1]$ to $x(t)$;*

(c) *the trajectories $(y_\epsilon(t),\sigma_\epsilon(t))$ converge statistically to the direct integral of $\nu_{x(t),v(x(t),t)}$, where $\nu_{x,v}$ is the invariant probability measure associated with $\mathbf{u}$ for (x,v).*

Proof. Notice that once $(\mathbf{u},\mathbf{v})$ is inserted into the control equations of Problem 2.1, the resulting system (7.1) is a standard system of ordinary differential equations, some of them singularly perturbed. Claims (b) and (c) then follow from Theorem 2.5 in [3]. Indeed, the conditions in the latter paper are easily implied by the definition of a proper global control policy, and the chattering equation (6.3) is the chattering limit that governs the uniform limit of $x_\epsilon(t)$ and the statistical limit of the singularly perturbed part as $\epsilon \to 0$.

To verify claim (a), we use claims (b) and (c) as follows. The expression (6.4) for $\mathrm{cost}(\mathbf{u},\mathbf{v})$ can also be written as

$$\int_{R^m \times S \times [0,1]} c(x(t),y,u(x(t),y,\sigma,v(x(t),t))\nu(dy \times d\sigma \times dt), \tag{7.5}$$

with ν being the direct integral of $\nu_{x(t),v(x(t),t)}$ (see (7.4)). This is an extension of Fubini's theorem; see, e.g., [6]. The cost $c_\epsilon(\mathbf{u},\mathbf{v})$ given in (7.2) can in turn be expressed in the form

$$\int_{R^m \times S \times [0,1]} c(x_\epsilon(t),y,u(x_\epsilon(t),y,\sigma,v(x_\epsilon(t),t)))\nu_\epsilon(dy \times d\sigma \times dt), \tag{7.6}$$

where ν_ϵ is the measure obtained as the distribution of $(y_\epsilon(t), \sigma_\epsilon(t), t)$ on $R^m \times S \times [0,1]$ (see (7.3)). This expression is obtained from (7.2) by a change of notation only. Claim (b) guarantees that $x_\epsilon(t)$ converge uniformly to $x(t)$, and claim (c) implies that ν_ϵ converges weakly to ν. Hence the desired convergence of (7.6) to (7.5) as $\epsilon \to 0$ follows from the continuity of c, $\mathbf{u}$, and $\mathbf{v}$ and from their boundedness on the common supports of ν_ϵ and ν. This completes the proof. ∎

8. Infinitesimally generated velocities. As a step toward optimality considerations, in this section we broaden the notion of infinitesimal generation of velocities. The motivation is the same as the one displayed in Remark 5.4. The extension is needed for variational considerations.

DEFINITION 8.1. *Let $x \in R^n$ be fixed. Let $B \subset R^m$ be fixed. We say that the velocity vector $v \in R^n$ is* infinitesimally generated *at x using the base B if the following holds. There exist a sequence $T_k \to \infty$ and a sequence of admissible controls $u_k(\tau) : [0, T_k] \to U$ such that*

$$v = \lim_{k\to\infty} \frac{1}{T_k} \int_0^{T_k} f(x_k(\tau), y_k(\tau), u_k(\tau)) d\tau, \tag{8.1}$$

where $y_k(\tau)$ is a solution on $[0, T_k]$ of

$$\frac{dy}{d\tau} = g(x_k(\tau), y, u_k(\tau)) \tag{8.2}$$

and where $y_k(\tau) \in B$ for all k and all $\tau \in [0, T_k]$, and $\sup\{|x_k(\tau) - x| : \tau \in [0, T_k]\}$ converge to 0 as $k \to \infty$. The set of velocities that are infinitesimally generated at x from the base B is denoted by $V_B(x)$.

PROPOSITION 8.2. *For a fixed B, the set $V_B(x)$ is closed in R^n, and the mapping $x \to V_B(x)$ has a closed graph.*

Proof. The claims follow from the definition by using a diagonal argument. ∎

As mentioned, the set $V_B(x)$ represents the limit directions that the slow dynamics could move in, in a given instant, if the fast dynamics is confined to the set B. We consider a varying state $x_k(\tau)$ in the generation of v and a sequence of controls, since v is indeed the result of a limit process. Next we consider a more restrictive notion.

DEFINITION 8.3. *Let x and $B \subset R^m$ be fixed. We say that $v \in V_B(x)$ is properly generated if there exist a control $u(\tau) : [0, \infty) \to U$ and $y_0 \in B$ such that*

$$v = \lim_{T\to\infty} \frac{1}{T} \int_0^{T} f(x, y(\tau), u(\tau)) d\tau, \tag{8.3}$$

where $y(\cdot)$ is the solution of $\dot{y} = g(x, y, u(\tau))$, $y(0) = y_0$, and $y(\tau) \in B$ for $\tau \in [0, \infty)$.

Note that if (x, v) is in the domain of the proper control policy $\mathbf{u}$ and the trajectory $y(\tau)$ resulting from (5.2) is in B, then $v \in V_B(x)$, and it is properly generated.

9. Optimality. We are about to introduce two optimization problems, one on the infinitesimal (fast) scale, the other on the global (slow) time scale. A composition of their solutions would constitute a near optimal solution.

The infinitesimal problem depends on three parameters, $x \in R^n$, $B \subset R^m$, and $v \in V_B(x)$, as follows. (IG stands for infinitesimally generating.)

IG(x, v, B) Problem 9.1.

$$\text{minimize } \lim_{k\to\infty} \frac{1}{T_k} \int_0^{T_k} c(x_k(\tau), y_k(\tau), u_k(\tau)) d\tau, \tag{9.1}$$

with the minimization being over all sequences $T_k \to \infty$, admissible controls $u_k(\tau) : [0, T_k] \to U$, and functions $x_k(\tau) : [0, T_k] \to R^n$ satisfying $\sup\{|x_k(\tau) - x| : 0 \le \tau \le T_k\}$ converge to 0 as $k \to \infty$, and where $y_k(\tau)$ solve on $[0, T_k]$ the equation

$$\frac{dy}{d\tau} = g(x_k(\tau), y, u_k(\tau)) \tag{9.2}$$

and $y_k(\tau) \in B$ for $\tau \in [0, T_k]$. In addition, it is demanded that

$$v = \lim_{k\to\infty} \frac{1}{T_k} \int_0^{T_k} f(x_k(\tau), y_k(\tau), u_k(\tau))d\tau. \tag{9.3}$$

We denote by $\Phi_B(x, v)$ the infimal value in (9.1). It is clear that the IG(x, v, B) problem is simply the problem of infinitesimally generating the velocity v at x using the base B (see Definition 8.1) with a minimal averaged cost. The value $\Phi_B(x, v)$ could be equal to $+\infty$, and in fact we set $\Phi_B(x, v) = \infty$ if $v \notin V_B(x)$.

PROPOSITION 9.2. *For $B \subset R^m$ fixed, the function $\Phi_B(x, v)$ is lower semicontinuous, namely, $(x_k, v_k) \to (x_0, v_0)$ implies $\liminf \Phi_B(x_k, v_k) \ge \Phi_B(x_0, v_0)$.*

Proof. It follows from a simple diagonal argument. □

Our next step is the global problem. It depends on a parameter $B \subset R^m$ and also refers to the initial condition $x(0) = \overline{x}$ (see (2.4)) of the original singular perturbations problem, as follows. (GO stands for global optimization.)

GO(B) Problem 9.3.

$$\text{minimize} \int_0^1 \Phi_B(x(t), v(x(t), t))dt \tag{9.4}$$

subject to

$$\frac{dx}{dt} = v(x, t), \quad x(0) = \overline{x} \tag{9.5}$$

and the constraint

$$v(x, t) \in V_B(x). \tag{9.6}$$

A solution to the problem is a pair $(x(\cdot), v(\cdot, \cdot))$ such that $x(t)$ is absolutely continuous; $v(x, t)$ is continuous in x and measurable in t, defined on a subset of $R^n \times [0, 1]$; and (9.5) produces $x(t)$ as its unique solution.

Remark 9.4. The GO(B) problem is actually a standard Bolza problem of minimizing $\int_0^1 \Phi_B(x, \dot{x})dt$ subject to $x(0) = \overline{x}$ and $\dot{x} \in V_B(x)$. Indeed, we allow in our formulation $v(x, t)$ to be defined on the graph of $x(t)$ only, namely $v(x, t) = \dot{x}(t)$. We prefer the formulation with the function $v(x, t)$, for two reasons. The first one is to make transparent the fact that in the singular perturbations setting, the velocity $v(x, t)$ is infinitesimally generated. The second one is that in many problems that arise in applications, the solution is indeed given in a feedback form, generating a differential equation on a domain larger than the optimal trajectory. It is this property that enables us to derive the near optimal solutions for the fast tracking.

The mathematical characteristics of the two problems, 9.1 and 9.3, are of course of interest. An avenue to a systematic treatment of them is the concept of overtaking solutions; see, e.g., Carlson, Haurie, and Leizarowitz [7] and the recent contribution of Zaslawski [23]. As mentioned earlier, in this paper we do not pursue these problems.

Rather, we go ahead, assume that the two problems have solutions, and show that under some conditions, the composition of the solutions is near optimal for the SP Problem 2.1.

The main result is proven under a coercivity condition described in the following definition. In the formulation we refer to control functions $u_\epsilon(t)$ that, when applied to equations (2.3)–(2.4) and (4.2), produce trajectories $(x_\epsilon(t), y_\epsilon(t), \sigma_\epsilon(t))$. The resulting cost is given by

$$\text{cost } (u_\epsilon) = \int_0^1 c(x_\epsilon(t), y_\epsilon(t), u_\epsilon(t))dt. \tag{9.7}$$

Compare with (2.2). We shall write $c_\epsilon(t)$ for the integrand in (9.7). Recall that $\text{val}(\epsilon)$ is the infimal cost.

DEFINITION 9.5. *The singular perturbations problem is called* coercive *if, whenever $u_\epsilon(t)$ are control functions satisfying $\text{cost}(u_\epsilon) - \text{val}(\epsilon) \to 0$ as $\epsilon \to 0$, then for ϵ small enough (say, $0 < \epsilon \le \epsilon_0$) the trajectories $(x_\epsilon(t), y_\epsilon(t))$ are contained in one compact subset of $R^n \times R^m$ (say, $C \times B$) and $\dot{x}_\epsilon(t)$ and $c(t)$ share a common L_2-bound, namely, $\int_0^1 |\dot{x}_\epsilon(t)|^2 dt$, and $\int_0^1 |c_\epsilon(t)|^2 dt$ have a bound independent of ϵ.*

Coercivity is very common in optimal control problems; we do not elaborate on this concept here.

THEOREM 9.6. *Suppose that the singular perturbations problem is coercive, and let $C \times B$ be the compact subset of $R^n \times R^m$ provided by the coercivity. Let $(x(t), v(x(t), t))$ be a solution of the GO(B) Problem 9.3 and such that $\mathbf{v} = v(x, t)$ is defined and continuous on a neighborhood of the graph $\{(x(t), t) : 0 \le t \le 1\}$ in $R^n \times [0, 1]$. Let $\mathbf{u} = u(x, y, \sigma, v)$ be a proper control policy whose domain contains a neighborhood of the graph of $(x(\cdot), \dot{x}(\cdot))$, and furthermore assume that $\text{cost}(\mathbf{u}, x, v) = \Phi_B(x, v)$. Then the composition*

$$u(x, y, \sigma, t) = u(x, y, \sigma, v(x, t))$$

is a near optimal solution of the singular perturbations problem.

Proof. We shall denote by $c_\epsilon(\mathbf{u})$ the cost of applying the displayed composition of $\mathbf{u}$ and $\mathbf{v}$ to the problem with $\epsilon > 0$ fixed. See (4.4).

It is clear that the pair $(\mathbf{u}, \mathbf{v})$ is a proper global control policy that generates the trajectory $x(t)$. By the stability result, Theorem 7.1, the convergence $c_\epsilon(\mathbf{u}) = c_\epsilon(\mathbf{u}, \mathbf{v})$ to $\text{cost}(\mathbf{u}, \mathbf{v})$ holds. The latter, in turn, is equal (by (6.4) and the assumption of the present result) to the optimal value of the GO(B) problem. Hence we deduce that

$$c_\epsilon(\mathbf{u}) - \int_0^1 \Phi_B(x(t), v(x(t), t))dt \to 0 \tag{9.8}$$

as $\epsilon \to 0$.

Next we show that

$$\int_0^1 \Phi_B(x(t), v(x(t), t))dt - \text{val}(\epsilon) \to 0 \tag{9.9}$$

as $\epsilon \to 0$. To this end, let $u_\epsilon(t)$ be admissible control functions, generating the solutions $(x_\epsilon(t), y_\epsilon(t), \sigma_\epsilon(t))$ of (2.3)–(2.4) and (4.2), with $\epsilon > 0$ fixed, and such that $c(u_\epsilon) - \text{val}(\epsilon) \to 0$ as $\epsilon \to 0$ (namely, $u_\epsilon(\cdot)$ provides an approximately optimal solution). By the coercivity, there exists a subsequence ϵ_i, with $\epsilon_i \to 0$, such that (and we write $x_i(t)$ for $x_{\epsilon_i}(t)$, etc.)

(1) $x_i(t)$ converge uniformly on $[0,1]$, say, to $x_0(t)$,
(2) $\dot{x}_i(t) = f(x_i(t), y_i(t), u_i(t))$ converge weakly in L_2 to $\dot{x}_0(t)$,
(3) $c_i(t) = c(x_i(t), y_i(t), u_i(t))$ converge weakly in L_2, say, to $c_0(t)$.

For a fixed $h > 0$, consider a time $t \in [0, 1-h]$. The vector

$$v_h(t) = \frac{x_0(t+h) - x_0(t)}{h} \tag{9.10}$$

is the limit as $\epsilon_i \to 0$ of

$$\frac{1}{h}(x_i(t+h) - x_i(t)) = \frac{1}{h}\int_t^{t+h} f(x_i(s), y_i(s), u_i(s))ds,$$

which after the change of variables $\tau = \epsilon_i^{-1}(s-t)$ is written as

$$\frac{\epsilon_i}{h}\int_0^{h/\epsilon_i} f(x_i(\tau), y_i(\tau), u_i(\tau))d\tau. \tag{9.11}$$

Likewise, the average

$$c_h(t) = \frac{1}{h}\int_t^{t+h} c_0(s)ds$$

is the limit as $\epsilon_i \to 0$ of

$$\frac{1}{h}\int_t^{t+h} c_i(s)ds = \frac{\epsilon_i}{h}\int_0^{h/\epsilon_i} c(x_i(\tau), y_i(\tau), u_i(\tau))d\tau. \tag{9.12}$$

Since $(y_i(\tau), \sigma_i(\tau))$ on $[0, h/\epsilon_i]$ solve (9.2) with $x_i(\tau)$ the parameter, and since $x_i(\tau)$ on $[0, h/\epsilon_i]$ is near $x(t)$ if h is small, it follows that if we take a diagonal subsequence i_j associated with a sequence $h_j \to 0$, any cluster point $v(t)$ of $v_h(t)$ is in $V_B(x_0(t))$ and any cluster point $c(t)$ of $c_h(t)$ satisfies $c(t) \geq \Phi_B(x_0(t), v(t))$. This occurs by Definition 8.1 and the definition of $\Phi_B(x,v)$ in Problem 9.1.

But $v_h(t)$ converges almost everywhere to $\dot{x}_0(t)$, and $c_h(t)$ converges almost everywhere to $c_0(t)$. We therefore conclude that

$$\int_0^1 c_0(t)dt \geq \int_0^1 \Phi_B(x_0(t), \dot{x}_0(t))dt. \tag{9.13}$$

Since the left-hand side of (9.13) is the limit of $\int_0^1 c_i(t)dt$, and since $\mathbf{v}$ was the solution of the GO(B) problem, we conclude that

$$\lim \int_0^1 c_i(t)dt \geq \int_0^1 \Phi_B(x_0(t), \dot{x}_0(t))dt \geq \int_0^1 \Phi_B(x(t), v(x(t), t))dt, \tag{9.14}$$

or in other words, $\lim \operatorname{val}(\epsilon_i)$ exists, and it is bigger or equal to the right-hand side term of the displayed inequalities. But since $c_\epsilon(\mathbf{u}) \geq \operatorname{val}(\epsilon)$, it actually follows from the stability part, in particular from (9.8), that

$$\lim_{\epsilon_i \to 0} \operatorname{val}(\epsilon_i) = \int_0^1 \Phi_B(x(t), t))dt. \tag{9.15}$$

Since ϵ_i was an arbitrary subsequence for which the weak convergence holds, and since the coercivity implies weak-L_2 compactness, the limit in (9.15) holds for all $\epsilon \to 0$. This is the desired conclusion. ∎

Remark 9.7. Since $B_1 \subset B_2$ implies $\Phi_{B_2}(x,v) \le \Phi_{B_1}(x,v)$, it clearly follows that any domain containing the one given by the coercivity can be used in the calculation of the optimal proper policy. In particular, we can use $B = R^m$. The coercivity, however, is needed in the proof of the near optimality.

Remark 9.8. The proper controls $\mathbf{u}$ in the main result, Theorem 9.6, were assumed to generate a unique invariant measure $\nu_{x,v}$, with cost (given by (5.8)) equal to $\Phi_B(x,v)$. The result would hold even if the uniqueness of the invariant measure is dropped, provided that the expression (5.8) is equal to $\Phi_B(x,v)$ for every invariant measure $\nu_{x,v}$. The only change in the proof would be to refer in the stability part to Theorem 2.2 (rather than Theorem 2.5) in [3].

10. Comments. We display two examples whose formulation does not include a fast clock variable but which still cannot be solved with the order reduction method (see the introduction). We show how the reasoning of the present paper applies.

Note first that the structure of the solutions suggested by the reduced-order approach is a particular case of those offered in this paper. Indeed, equating $\epsilon = 0$ in (2.3) and providing $u(x,y)$, which makes the equation $\dot y = g(x,y,u(x,y))$ asymptotically stable around, say, $y_0(x)$, are equivalent to proper generation of the velocity $f(x,y_0(x),u(x,y_0(x)))$ at x, with a proper control that has a special type of invariant measure, namely, a measure supported at the equilibrium $y_0(x)$. So, in case our infinitesimal optimization step can be achieved with an invariant measure supported at a stable point, we get back the reduced-order procedure.

Here are the promised examples.

Example 10.1. Consider the problem with scalar variables

$$\text{minimize} \int_0^1 (x^2(t) + (y_1^2(t) + y_2^2(t) - 1)^2 + u^2(t))dt \tag{10.1}$$

subject to

$$\begin{aligned} \frac{dx}{dt} &= y_1 + y_2, \\ \epsilon \begin{pmatrix} \dot y_1 \\ \dot y_2 \end{pmatrix} &= \begin{pmatrix} 0 & 1 \\ -1 & 0 \end{pmatrix} \begin{pmatrix} y_1 \\ y_2 \end{pmatrix} + \begin{pmatrix} 0 \\ 1 \end{pmatrix} u, \\ x(0) = 0, \quad y_1(0) &= \overline{y}_1, \quad y_2(0) = \overline{y}_2. \end{aligned} \tag{10.2}$$

A direct inspection reveals that the limit of the minimal cost as $\epsilon \to 0$ is 0. A near optimal solution can be constructed by examining the fast equation

$$\frac{dy_1}{d\tau} = y_2, \quad \frac{dy_2}{d\tau} = -y_1 + u \tag{10.3}$$

and devising a feedback control $u = u(y_1, y_2)$ which stabilizes the solution of (10.3) around a periodic solution of the form $(\cos(\varphi + \tau), \sin(\varphi + \tau))$, while $u = 0$ on the limit trajectory. This can be achieved easily. Note that there is no need to employ an auxiliary clock variable; note also that setting $\epsilon = 0$ in (10.2) will not lead to a near optimal solution.

Example 10.2. In this example a nontrivial clock example is employed. Consider the system with scalar variables

$$\text{minimize} \int_0^1 (x^2(t) + y_1^2(t) + y_2^2(t) + (u_1^2(t) + u_2^2(t) - 1)^2)dt$$

subject to

$$\begin{aligned} \frac{dx}{dt} &= u_1 + u_2, \\ \epsilon\frac{dy_1}{dt} &= u_1, \\ \epsilon\frac{dy_2}{dt} &= u_2, \\ x(0) = 0, \quad y_1(0) &= \overline{y}_1, \quad y_2(0) = \overline{y}_2. \end{aligned} \tag{10.4}$$

A direct inspection reveals that the limit of the minimal value as $\epsilon \to 0$ is 0. A near optimal solution can be constructed using the clock equation $\epsilon\dot{\sigma} = 1$, with $\sigma \in S^1$ (namely, $\sigma = \arg(\xi + i\eta)$ on the unit circle in R^2). Then

$$u_1 = -y_1 + \sin\sigma, \quad u_2 = -y_2 + \cos\sigma \tag{10.5}$$

is a near optimal solution.

11. A linear-quadratic tracking. The technique offered in the previous sections is applied in this section to a system with linear state equations and a quadratic cost criterion. The target is assumed to be periodic. Note that the corresponding regulator problem was analyzed in the literature; see Kokotovic and Yackel [16], O'Malley [19], and Chow and Kokotovic [8].

In what follows, a prime over a matrix—say, A'—denotes the transposed matrix; for vectors y we use y' to denote the row version. For a matrix Q symmetric and positive semidefinite, we denote by $\|x\|_Q^2$ the quadratic form $x'Qx$.

The target to be tracked is given by

$$\Gamma(x, \sigma) : R^n \times R \to R^m, \tag{11.1}$$

assuming that $\Gamma(x, \sigma)$ is continuous and periodic in σ, with period $T(x) > 0$ (not necessarily the minimal one) continuous in x.

The tracking problem is as follows. (SP-LQPT stands for singularly perturbed linear quadratic periodic tracking.)

SP-LQPT Problem 11.1.

$$\text{minimize} \int_0^1 (\|x(t)\|_Q^2 + \|y(t) - \Gamma(x(t), \sigma(t))\|_W^2 + \|u(t)\|_R^2)dt \tag{11.2}$$

subject to

$$\begin{aligned} \frac{dx}{dt} &= Fx + Gy + Hu, \\ \epsilon\frac{dy}{dt} &= Cx + Ay + Bu, \\ \epsilon\frac{d\sigma}{dt} &= 1, \end{aligned} \tag{11.3}$$

with initial conditions

$$x(0) = \overline{x}, \quad y(0) = \overline{y}, \quad \sigma(0) = \overline{\sigma}, \tag{11.4}$$

and where the matrices appearing in (11.2)–(11.3) are constant and have the appropriate dimensions. Compare with (4.6)–(4.7).

The following assumption is standard in the linear-quadratic trait; it is employed throughout.

Assumption 11.2. The matrices Q and W are symmetric and positive semidefinite; the matrix R is symmetric and positive definite. The pair (A, B) is controllable and (A, W) is observable.

Remark 11.3. The form in which the SP-LQPT problem is presented is with the fast time σ being in R and with our stating explicitly that $\Gamma(x, \cdot)$ is $T(x)$-periodic. This deviates from the clock description of Definition 4.1. The latter would consider σ in S^1. We find (11.3) a bit more convenient in the purely periodic case. It is clear that the two forms are equivalent.

The analysis of the SP-LQPT problem will follow several steps.

Note first that the positive definiteness of R implies that the problem (11.2)–(11.3) is coercive (see Definition 9.5). Hence the base set appearing in Problems 9.1 and 9.3 can be taken as any large enough bounded set in R^n. A reference to it is therefore suppressed in the formulas that follow.

Step I. Determination of the infinitesimally generated velocities.

For each $x \in R^n$ we have to determine the set $V(x)$ of velocities that can be infinitesimally generated at x; see Definition 8.1. To this end, consider the $m \times (m+k)$ matrix $[A, B]$, operating on pairs (y, u) in $R^m \times R^k$ and with values in R^m. The controllability of (A, B) implies that $[A, B]$ has a full range. Therefore the $m \times m$ matrix

$$M = [A, B] \begin{bmatrix} A' \\ B' \end{bmatrix} \tag{11.5}$$

is invertible. Denote by $\ker[A, B]$ the kernel of $[A, B]$, namely, the family of pairs (y, u) such that $Ay + Bu = 0$. Finally, let

$$V = \{Gy + Hu : (y, u) \in \ker[A, B]\}. \tag{11.6}$$

It is clear that V is a linear subspace of R^n.

PROPOSITION 11.4. *The infinitesimally generated velocities at x form a translation of a linear space, given by*

$$V(x) = \left(F - [G, H] \begin{bmatrix} A' \\ B' \end{bmatrix} M^{-1} C \right) x + V \tag{11.7}$$

(*with M and V given in* (11.5)–(11.6)).

Proof. An infinitesimally generated velocity has the form

$$\lim_{k \to \infty} \frac{1}{T_k} \int_0^{T_k} (Fx_k(\tau) + Gy_k(\tau) + Hu_k(\tau)) d\tau, \tag{11.8}$$

with constraints as described in Definition 8.1 and in particular when $(y_k(\cdot), u_k(\cdot))$ satisfy

$$\frac{dy_k}{d\tau}(\tau) = Ay_k(\tau) + Bu_k(\tau) + Cx_k(\tau) \tag{11.9}$$

and $y_k(\tau)$ are uniformly bounded. Taking averages on $[0, T_k]$ and using the fact that $y_k(\tau)$ are bounded (and $x_k(\tau) \to x$) imply that the limit in (11.8) has the form

$$Fx + Gy + Hu \tag{11.10}$$

under the constraint

$$Ay + Bu + Cx = 0, \tag{11.11}$$

and the controllability clearly implies that any element obtained in this way is a limit of the form (11.8) (in fact with $x_k(\tau) = x$). The passage from the formulation (11.10)–(11.11) to the representation (11.7) follows simple linear algebra considerations. ∎

Step II. An auxiliary problem.

The periodicity in σ of the target $\Gamma(x, \sigma)$ suggests that the infinitesimal generation problem (i.e., IG(x, v) problem 9.1) could be solved with a proper control policy that induces an invariant measure supported on a periodic orbit. We shall see later that this is indeed the case. At present we formulate a variational problem that determines, as we shall see, the limit trajectory.

In what follows, $\Gamma(\tau)$ is T-periodic, with $T > 0$, $c \in R^m$, and $q \in R^n$ (with a constraint on q displayed later). The rest of the data are taken from Problem 11.1. (LQPBVT stands for linear-quadratic periodic boundary value tracking.)

LQPBVT$(\Gamma(\cdot), c, q)$ *Problem* 11.5.

$$\text{minimize } \frac{1}{T}\int_0^T (\|y(\tau) - \Gamma(\tau)\|_W^2 + \|u(\tau)\|_R^2)d\tau \tag{11.12}$$

subject to

$$\begin{aligned} \frac{dy}{d\tau} &= Ay + Bu + c, \\ \frac{dz}{d\tau} &= Gy + Hu, \end{aligned} \tag{11.13}$$

with the boundary condition

$$\begin{aligned} y(T) &= y(0), \\ z(T) &= z(0) + Tq, \end{aligned} \tag{11.14}$$

where q belongs to a translation of a subspace, namely, to

$$-[G, H]\begin{bmatrix} A' \\ B' \end{bmatrix} M^{-1}c + V, \tag{11.15}$$

with V the subspace given in (11.6). Clearly, a solution is a pair $(y(\tau), u(\tau))$ defined on $[0, T]$.

In what follows, we consider the varying data $\Gamma(\cdot)$, c, and q, where the period T may vary as well. When we say that $\Gamma_j(\cdot)$ converge to $\Gamma_0(\cdot)$, we mean that the periods converge and the periodic continuations converge uniformly on compact intervals.

PROPOSITION 11.6. *The problem LQPBVT* (Γ, c, q) *has a unique optimal solution* $(y^*(\tau), u^*(\tau))$. *It satisfies* $u^*(0) = u^*(T)$, *and its periodic continuation is uniformly continuous on compact intervals with respect to the parameters* $(\Gamma(\cdot), c, q)$.

Proof. First we note that a feasible pair $(y(\tau), u(\tau))$, namely, one that satisfies the constraints, exists. In fact, there is a constant pair (y, u) which is feasible. Indeed, if

$$q = \frac{1}{T}\left(-[G, H]\begin{bmatrix} A' \\ B' \end{bmatrix} M^{-1}c + Gy_0 + Hu_0\right) \tag{11.16}$$

with $(y_0, u_0) \in \ker[A, B]$, take

$$(y, u) = -\begin{bmatrix} A' \\ B' \end{bmatrix} M^{-1}c + (y_0, u_0) \tag{11.17}$$

(see also Remark 11.7). The constant feasible control may not be an optimal one. Existence and uniqueness of an optimal control follow standard arguments; see, for instance, Lee and Markus [17, Chapters 2 and 3].

Next we verify the claim that the optimal control $u^*(\tau)$ is continuous and periodic. Applying to (11.12)–(11.13) the standard necessary and sufficient conditions for extrema in linear-quadratic systems (see, e.g., Lee and Markus [17, pp. 180, 191]), we get that the control $u^*(\tau)$ is of the form

$$u^*(\tau) = R^{-1}[B, H] \begin{bmatrix} \eta^*(\tau) \\ \zeta^*(\tau) \end{bmatrix}, \tag{11.18}$$

where $(y^*(\tau), z^*(\tau), \eta^*(\tau), \zeta^*(\tau))$ satisfy the equations

$$\frac{d}{d\tau} \begin{pmatrix} y \\ z \\ \eta \\ \zeta \end{pmatrix} = \begin{pmatrix} A & 0 & BR^{-1}B' & BR^{-1}H' \\ G & 0 & HR^{-1}B' & HR^{-1}H' \\ W & 0 & -A' & -G' \\ 0 & 0 & 0 & 0 \end{pmatrix} \begin{pmatrix} y \\ z \\ \eta \\ \zeta \end{pmatrix} + \begin{pmatrix} c \\ 0 \\ -W\Gamma(\tau) \\ 0 \end{pmatrix}. \tag{11.19}$$

We see right away that $u^*(\tau)$ must be continuous. Since the dual variable $\zeta^*(\tau)$ is constant, it follows that $\eta^*(\tau)$ must be periodic (by the periodicity of $y(\tau)$ and $\Gamma(\tau)$ and by the transversality conditions); hence from (11.18) it follows that $u^*(\tau)$ is indeed periodic. (Another argument for the periodicity is that, by solving the problem on $[0, 2T]$, we should get on $[T, 2T]$ a replica of the solution on $[0, T]$; otherwise the uniqueness is violated.) (Note that (11.18)–(11.19) can be used to compute the optimal trajectory and control.)

Finally we note that indeed $(y^*(\tau), u^*(\tau))$ depends continuously on the data. This follows easily from continuous dependence arguments and the uniqueness. This completes the proof of the proposition. □

Remark 11.7. We showed that if q satisfies the constraint (11.15), then there is a constant feasible pair (y, u). Note that satisfying (11.15) is a necessary condition for the existence of any feasible control $(y(\tau), u(\tau))$. This follows from the considerations of Proposition 11.4.

Step III. The infinitesimal generation problem.

Fix $x \in R^n$ and $v \in V(x)$ (see Proposition 11.4). We consider the IG(x, v) Problem 9.1. It is shown in this step that a solution is essentially provided by applying the solution of Problem 11.5 periodically (and where q and v are related by $v = Fx+q$; see (11.7) and (11.15)), and $c = Cx$.

The linearity, which is reflected by the separation of x and (y, u), and the required boundedness of $y(t)$ imply that the infinitesimal generation problems with data (x, v) take the form as follows. (Compare with (9.1)–(9.3); note that here proper generation, in particular x fixed, suffices.)

IG(x, v) *Problem* 11.8.

$$\text{minimize } \lim_{T\to\infty} \frac{1}{T} \int_0^T (\|x\|_Q^2 + \|y(\tau) - \Gamma(x, \tau)\|_W^2 + \|u(\tau)\|_R^2) d\tau \tag{11.20}$$

subject to

$$\frac{dy}{d\tau} = Ay + Bu + Cx \tag{11.21}$$

and

$$v = \lim_{T\to\infty} \frac{1}{T} \int_0^T (Fx + Gy(\tau) + Hu(\tau)) d\tau. \tag{11.22}$$

Associated with Problem 11.8, consider the periodic-boundary value problem LQPBVT $(\Gamma(\cdot), c, q)$ Problem 11.5 with $\Gamma(t) = \Gamma(x,t)$, $c = Cx$, and $q = v - Fx$. For the latter problem, we have determined an optimal pair $(y^*(\tau), u^*(\tau))$, which we now denote by

$$(y^*(\tau, x, v), u^*(\tau, x, v)), \tag{11.23}$$

and furthermore, the periodicity allows us to extend the functions periodically to the entire line.

PROPOSITION 11.9. *$(y^*(\tau,x,v), u^*(\tau,x,v))$ is an optimal solution of the IG(x,v) Problem* 11.8*, and consequently the minimal value of the problem is given by*

$$\Phi(x,v) = \frac{1}{T(x)} \int_0^{T(x)} (\|x\|_Q^2 + \|y^*(\tau,x,v) - \Gamma(x,\tau)\|_W^2 + \|u^*(\tau,x,v\|_R^2) d\tau. \tag{11.24}$$

Proof. Let $(y(\tau), y(\tau))$ be any pair which satisfies the constraint (11.22). Take $T = kT(x)$ with k large enough, such that the average

$$\frac{1}{T} \int_0^T (Fx + Gy(\tau) + Hu(\tau)) d\tau \tag{11.25}$$

is very close to v. The controllability of (A, B) implies that we can make $u(T) = u(0)$ and $y(T) = y(0)$ without changing much the value in (11.25). Averaging of the k translation $u_j(\tau) = u(\tau + jT(x))$, $y_j(\tau) = y(\tau + jT(x))$, with $j = 0, \ldots, k-1$ (and interpreting $(u(\cdot), y(\cdot))$ as periodic, with $kT(x)$ being the period) induces a $T(x)$-periodic pair, with cost smaller than or equal to the cost associated with $(u(\cdot), y(\cdot))$; this by the convexity. Still, the cost of the generated periodic pair is greater than or equal to that of $(u^*(\tau), y^*(\tau))$, the latter being the optimal periodic one. ∎

Step IV. Stabilization and proper control policies.

We already know that the periodic pair $(y^*(\tau,x,v), u^*(\tau,x,v))$ forms an optimal trajectory for the generation of v at x. But we wish to construct a control $\mathbf{u} = u(x,y,\tau,v)$ which would be proper (see Definition 5.6, but note that we work with periodic functions on R (see Remark 11.3)) and for (x,v) fixed would solve the infinitesimal generation problem. What we do is simply find a feedback control that, for (x,v) fixed, stabilizes the linear equation around $y^*(\tau,x,v)$ as follows.

Let K be the unique positive definite symmetric solution of the Riccati equation

$$-KA - A'K + KBR^{-1}B'K - W = 0. \tag{11.26}$$

Then it is well known that the controllability of (A,B) and the observability of (A,W) imply that such a K exists and is unique, and furthermore, it is known that $G = A - BR^{-1}B'K$ is a stable matrix; see, e.g., Athans and Falb [4, p. 773].

PROPOSITION 11.10. *Consider the control*

$$u(x,y,\tau,v) = -R^{-1}B'K(y - y^*(\tau,x,v)) + u^*(\tau,x,v). \tag{11.27}$$

Then it is a periodic control, continuous in all its variables, and for each (x,v) fixed, the solutions of

$$\frac{dy}{d\tau} = Ay + Bu + Cx$$

converge uniformly to $y^*(\tau,x,v)$.

Proof. The periodicity and continuity follow from the corresponding properties of $y^*(\tau)$ and $u^*(\tau)$, namely, Proposition 11.6. The stabilization is standard. ∎

COROLLARY 11.11. *For* $\mathbf{u}$ *given in* (11.27), $\mathrm{cost}(\mathbf{u}, x, v) = \Phi(x, v)$, *where* $\Phi(x, v)$ *is the infimal cost of generating* v *at* x.

Proof. It follows directly from the stabilization statement of Proposition 11.10 and the form (11.24) for $\Phi(x, v)$. ∎

Step V. The global problem.

Once the cost $\Phi(x, v)$ of optimally generating the velocity v at x is determined (see (11.24)), the global problem $GO(R^n)$ can be formed; see Problem 9.3. In the tracking case we study, it can be presented in the following form:

$$\text{minimize} \int_0^1 (\|x(t)\|_Q^2 + \Psi(x(t), w(t)))dt \tag{11.28}$$

subject to

$$\begin{aligned} \frac{dx}{dt} &= \left(F - [G, H]\begin{bmatrix} A' \\ B' \end{bmatrix} M^{-1}C\right)x + w, \\ x(0) &= \overline{x}, \\ w &\in V, \end{aligned} \tag{11.29}$$

with V given in (11.6) and the vectors v and w related by

$$v = w - [G, H]\begin{bmatrix} A' \\ B' \end{bmatrix} M^{-1}Cx \tag{11.30}$$

and where $\Psi(x, w)$ is related to $\Phi(x, v)$ by $\Phi(x, v) = \Psi(x, w) + \|x\|_Q^2$, namely, (by (11.24)) $\Psi(x, w)$ is given by

$$\Psi(x, w) = \frac{1}{T(x)} \int_0^{T(x)} (\|y^*(\tau, x, v) - \Gamma(x, \tau)\|_W^2 + \|u^*(\tau, x, v)\|_R^2)d\tau. \tag{11.31}$$

Note that the state equation for the global optimization is linear in x, with coefficients easily calculable from the data. The cost $\Psi(x, w)$ is not separable in general. (When $C = 0$ and Γ is independent of x, then $\Psi(x, w)$ is independent of x.)

PROPOSITION 11.12. *The cost* $\Psi(x, w)$ *is continuous in* (x, w), *it tends to* $+\infty$ *as* $|w| \to \infty$, *and it is a convex function of the variable* w.

Proof. The continuity follows from the relations (11.31) and (11.30) together with the continuous dependence statement in Proposition 11.6. The growth to $+\infty$ follows directly from the positive definiteness of R and the boundary condition (11.14). The convexity in the variable w follows as pairs $(y^*(\tau, x, v_i), u^*(\tau, x, v_i))$ for $i = 1, 2$ can be averaged, producing a feasible pair for the averaged v_i (by the linearity) with cost greater than or equal to the infimal one. The convexity of the quadratic cost concludes the argument. ∎

Step VI. The near optimal solution.

We conclude by writing down explicitly the near optimal solution to the SP-LQPT Problem 11.1. The construction (which includes an existence statement) is as follows:

For each $w \in V$ the cost $\Psi(x, w)$ is determined by (11.31), with $y^*(\tau, x, v)$ and $u^*(\tau, x, v)$ being the solution of the associated periodic boundary value, linear-quadratic tracking. Once a solution $w(x, t)$ to the global problem (11.28)–(11.29) is found, we write

$$v(x, t) = w(x, t) - [G, H]\begin{bmatrix} A' \\ B' \end{bmatrix} M^{-1}Cx,$$

and then a near optimal solution to the original singular perturbations tracking is given by

$$u(x,y,\sigma,t) = -R^{-1}B'K(y - y^*(\sigma, x, v(x,t))) + u^*(\sigma, x, v(x,t)). \tag{11.32}$$

The near optimality of this feedback control is guaranteed by the main result, Theorem 9.6. Note that indeed the near optimal solution does not employ the parameter ϵ. It is a feedback, though, of the fast time σ. Note that the near optimal solution (11.32) for the linear-quadratic tracking example is explicit in terms of solutions of the explicitly given two problems on finite intervals.

REFERENCES

[1] Z. Artstein, *Chattering variational limits of control systems*, Forum Math., 5 (1993), pp. 369–403.

[2] Z. Artstein and A. Leizarowitz, *Tracking periodic signals with the overtaking criterion*, IEEE Trans. Automat. Control, AC-30 (1985), pp. 1123–1126.

[3] Z. Artstein and A. Vigodner, *Singularly perturbed ordinary differential equations with dynamic limits*, Proc. Royal Soc. Edinburgh Sect. A, 126 (1996), pp. 541–569.

[4] M. Athans and P. L. Falb, *Optimal Control*, McGraw-Hill, New York, 1966.

[5] P. Billingsley, *Convergence of Probability Measures*, Wiley, New York, 1968.

[6] N. Bourbaki, *Integration*, Herman, Paris, 1959, chapter 6.

[7] D. A. Carlson, A. B. Haurie, and A. Leizarowitz, *Infinite Horizon Optimal Control*, Springer-Verlag, Berlin, 1987.

[8] J. H. Chow and P. V. Kokotovic, *Decomposition of near-optimum regulators for systems with slow and fast modes*, IEEE Trans. Automat. Control, AC-21 (1976), pp. 701–705; reprinted in [12].

[9] V. Gaitsgory, *Suboptimization of singularly perturbed control systems*, SIAM J. Control Optim., 30 (1992), pp. 1228–1249.

[10] V. Gaitsgory, *Suboptimal control of singularly perturbed systems and periodic optimization*, IEEE Trans. Automat. Control, 38 (1993), pp. 888–903.

[11] P. V. Kokotovic, *Applications of singular perturbations techniques to control problems*, SIAM Rev., 26 (1984), pp. 501–550.

[12] P. V. Kokotovic and H. K. Khalil, *Singular Perturbations in Systems and Control*, IEEE Selected Reprint Series, IEEE Press, New York, 1986.

[13] P. V. Kokotovic, H. K. Khalil, and J. O'Reilly, *Singular Perturbations in Control Analysis and Design*, Academic Press, New York, 1986.

[14] P. V. Kokotovic, R. E. O'Malley, and P. Sannuti, *Singular perturbations and order reduction in control theory—An overview*, Automatica, 12 (1976), pp. 123–132; reprinted in [12].

[15] P. V. Kokotovic and P. Sannuti, *Singular perturbation method for reducing the model order in optimal control design*, IEEE Trans. Automat. Control, AC-13 (1968), 377–384; reprinted in [12].

[16] P. V. Kokotovic and R. A. Yackel, *Singular perturbations of linear regulators: Basic theorems*, IEEE Trans. Automat. Control, AC-17 (1972), pp. 29–37; reprinted in [12].

[17] E. B. Lee and L. Markus, *Foundation of Optimal Control Theory*, Wiley, New York, 1967.

[18] V. V. Nemytskii and V. V. Stepanov, *Qualitative Theory of Differential Equations*, Princeton University Press, Princeton, NJ, 1960.

[19] R. E. O'Malley, *The singularly perturbed linear state regulator problem*, SIAM J. Control, 10 (1972), pp. 399–413; reprinted in [12].

[20] V. R. Saksena, J. O'Reilly, and P. V. Kokotovic, *Singular perturbations and time-scale methods in control theory: Survey* 1976–1983, Automatica, 20 (1984), pp. 273–293; reprinted in [12].

[21] A. Vigodner, *Limits of singularly perturbed control problems with statistical dynamics of fast motions*, SIAM J. Control Optim., 35 (1997), pp. 1–28.

[22] T. Yoshizawa, *Stability Theory and the Existence of Periodic Solutions and Almost Periodic Solutions*, Springer-Verlag, New York, 1975.

[23] A. Zaslavski, *Optimal programs on infinite horizon* 1, SIAM J. Control Optim., 33 (1995), pp. 1643–1660.

SIAM J. CONTROL OPTIM.
Vol. 35, No. 5, pp. 1508–1523, September 1997

004

A HOMEOMORPHIC CHARACTERIZATION OF MINIMAL SPECTRAL FACTORS*

AUGUSTO FERRANTE†

Abstract. In this paper a new characterization of the class of all minimal square spectral factors of a given rational spectral density is presented. This characterization, which is established without assumptions on poles and zeroes of the spectral density, extends a result presented in [A. Ferrante, *IEEE Trans. Automat. Control*, 39 (1994), pp. 2122–2126]. The characterization consists of two bijective maps which relate the set of minimal square spectral factors to a set of invariant subspaces of a certain matrix and to a set of symmetric solutions of an algebraic Riccati equation (ARE). In the second part of the paper it is proven that these two maps are homeomorphisms. This result extends and applies to spectral factorization theory recent results of H. Wimmer [*Integral Equations Operator Theory*, 21 (1995), pp. 362–375], where it is proven that the well-known relation between solutions of ARE and invariant subspaces of a certain matrix is, in fact, a homeomorphism.

Key words. spectral factorization problem, characterization of minimal spectral factors, algebraic Riccati equation, continuity

AMS subject classifications. 93C45, 93E10

PII. S0363012995289336

1. Introduction.

The spectral factorization problem is a cornerstone of many areas of systems and control theory, circuit theory, and prediction theory. Indeed, this problem is crucial in both linear quadratic optimal control and optimal filtering [16], [10]. In recent years, we have witnessed an increasing interest in *acausal* spectral factors, i.e., spectral factors whose poles and zeroes need not be in the left half complex plane [14], [15], [13], [4], [2], [3], [6], [7], [11]. These spectral factors are employed in problems of estimation in which the available information is not organized in a causal structure. For example, in [12, sect. 8] a noncausal estimation problem is considered, and the solution is given (in the spectral domain) in terms of an acausal spectral factor.

Also, great effort has been made to understand the relation between solutions of the spectral factorization problem and solutions of the ARE: since the first works in the late 1960s, the spectral factorization problem has been related to the Kalman–Popov–Yacubovich positive-real lemma, and this, in turn, to linear matrix inequality (see, for example, [16]) and to the ARE. In [14] and [15], the set of minimal spectral factors with given structure of poles or zeroes was characterized in terms of solutions of a homogeneous ARE. In [4] and [2], a characterization of the set of all minimal square spectral factors was provided for a class (a certain condition on sets of poles and zeroes of the spectral density was assumed) of multivariate rational spectral densities, in terms of a homogeneous ARE of double dimension with respect to that of [14] and [15]. In [6], a different characterization of all minimal spectral factors was given, without any assumption on the zeroes and poles of the spectral density. In the first part of this paper we obtain a characterization alternative, but equivalent, to that of [6] extending the results presented in [4] and [2] to an arbitrary rational coercive spectral density. More precisely, we show that, given a rational coercive spectral

*Received by the editors July 21, 1995; accepted for publication (in revised form) June 5, 1996.
http://www.siam.org/journals/sicon/35-5/28933.html
†Dipartimento di Ingegneria Elettrica, Gestionale e Meccanica, Università di Udine, via delle Scienze 208, 33100 Udine, Italy (ferrante@picolit.diegm.uniud.it).

density $\Phi(s)$, there is a one-to-one correspondence between the set of its minimal spectral factors and a class of symmetric solutions of an homogeneous ARE. This class, in turn, is in one-to-one correspondence with a set of invariant subspaces of a certain matrix Z which contains all the information about poles and zeroes of the spectral density. This matrix can be computed directly from the spectral density $\Phi(s)$. The correspondence between the set of invariant subspaces of Z and the set of solutions of the ARE has been extensively studied, and it is given by a standard procedure; see [19], [17], [20], and references therein.

In the second part of this paper we address the issue of the continuity of the maps which relate invariant subspaces of the above-mentioned matrix Z, solutions of the ARE, and minimal spectral factors. In the cases in which the ARE—and the related spectral factorization problem—have a finite number of solutions, the map which assigns a minimal spectral factor to a solution of the ARE is clearly continuous when we endow the domain and the image of the map with any reasonable topology. The same is true for the map relating invariant subspaces of the matrix Z and solutions of the ARE. However, when the ARE—and the related spectral factorization problem—have a continuum of solutions, the issue of continuity becomes more interesting. Indeed, in recent literature we find a relevant amount of work studying topological properties of solutions of the ARE; see, for instance, [8], [18], [19]. In particular, in [19] the problem of the continuity of the map which relates invariant subspaces of Z and solutions of the ARE has been investigated. In that paper it was shown that the function which maps an invariant subspace into a solution of the ARE is a homeomorphism when we endow the set of subspaces of $\mathbb{R}^n$ with the topology induced by the *gap metric*. In this paper we show that also the map which relates solutions of the ARE to minimal spectral factors is a homeomorphism when we endow the set of spectral factors of $\Phi(s)$ with the topology induced by the $\mathcal{L}^\infty$ norm. By composition, we have that also the map relating invariant subspaces of Z and minimal spectral factors is a homeomorphism.

Moreover we show that the map which assigns a minimal spectral factor to a solution of the ARE, and its inverse, are *Lipschitz continuous*, and we provide an upper-bound for the Lipschitz constant. This fact may be interesting in numerical solutions of the spectral factorization problem where a minimal spectral factor is computed starting from a numerical approximation of a fixed solution of the ARE. The Lipschitz continuity is interesting also in the case when the ARE has a finite number of solutions because it provides a bound on the distance of two minimal spectral factors given the distance of the corresponding solutions of the ARE.

2. Mathematical preliminaries and definitions. In this paper we deal with multivariate *real rational coercive spectral densities*, i.e., $m \times m$ matrix-valued rational functions $\Phi(s)$ with real coefficients and such that (s.t.) the following properties hold:

(i) $\Phi(s) = \Phi(-s)^T$.

(ii) $\Phi(s)$ is analytic on an open strip including the imaginary axis.

(iii) $\exists\ c \in \mathbb{R}_+$ s.t. $\Phi(i\omega) > cI > 0 \quad \forall \omega \in \mathbb{R}$.

Note that property (iii) implies that $R := \Phi(i\infty)$ is a symmetric positive definite matrix.

A matrix function $W(s)$ of dimensions $(m \times p)$ is called a *spectral factor* of $\Phi(s)$ if

$$W(s)W(-s)^T = \Phi(s). \tag{2.1}$$

Condition (iii) implies that $p \geq m$; if $p = m$, we have a *square* spectral factor. $W(s)$ is said to be *minimal* if it has least possible McMillan degree. In this paper we are

interested in minimal square spectral factors. Therefore in the following we shall say that a minimal square spectral factor is a solution of the spectral factorization problem for $\Phi(s)$. Clearly if $W(s)$ is a minimal spectral factors and O is a constant orthogonal matrix, then also $W_O(s) := W(s)O$ is a minimal spectral factor. However, $W(s)$ and $W_O(s)$, considered as transfer functions of linear systems, correspond to the same dynamics, and they differ only for a change of basis on the input space. For this reason we shall identify spectral factors which differ by multiplication on the right by a constant orthogonal matrix. Hence, without loss of generality, we shall suppose that $W(i\infty) = \Phi(i\infty)^{1/2} = R^{1/2}$.

Given a minimal realization

$$W(s) = R^{1/2} + C(sI - A)^{-1}B \tag{2.2}$$

of a spectral factor, we define the *pole structure* of $W(s)$ to be the Jordan structure of the *state matrix* A. We define the *zero structure* of $W(s)$ to be the Jordan structure of the *zero matrix* $\Gamma := A - BR^{-1/2}C$. It is immediate to see that the zero matrix Γ is the state matrix of a minimal realization of $W(s)^{-1}$. The eigenvalues of A and Γ will be called *poles* and *zeroes* of $W(s)$, respectively. The pole and zero structures of $W(s)$ are well defined since the Jordan forms of A and Γ do not depend on the realization (2.2).

It is well known (see, for instance, [10]) that, given a spectral density, there exists a minimal spectral factor $\overline{W}_-(s)$ which is *antistable* and *minimum phase*; i.e., all the poles of $\overline{W}_-(s)$ are in the open right half plane, and all the zeroes of $\overline{W}_-(s)$ are on the open left half plane. This spectral factor, which may be computed from $\Phi(s)$ via various algorithms [1], [15], will play a crucial role in this paper since it will be taken as the reference spectral factor.

3. Characterization of minimal square spectral factors. Let $\Phi(s)$ be a rational coercive spectral density and

$$\overline{W}_-(s) = R^{1/2} + C(sI + A^T)^{-1}B \tag{3.1}$$

be a minimal realization of its minimal spectral factor $\overline{W}_-(s)$. Define

$$\Gamma := -A^T - BR^{-1/2}C \tag{3.2}$$

to be the zero matrix of $\overline{W}_-(s)$. In what follows, A, B, and C will be matrices of dimensions $n \times n$, $n \times m$, and $m \times n$, respectively. Clearly Γ is an $n \times n$ matrix. Notice that, since both A and Γ are stability matrices, their spectra $\sigma(A)$ and $\sigma(\Gamma)$ are unmixed; i.e.,

$$\sigma(A) \cap \sigma(-A) = \emptyset, \tag{3.3}$$

$$\sigma(\Gamma) \cap \sigma(-\Gamma) = \emptyset. \tag{3.4}$$

Let $\mathcal{S}_A$ and $\mathcal{S}_\Gamma$ be the sets of invariant subspaces of A^T and Γ^T, respectively, and let $\mathcal{S}$ be the set defined by

$$\mathcal{S} := \{(S_A, S_\Gamma) : \quad S_A \in \mathcal{S}_A, \quad S_\Gamma \in \mathcal{S}_\Gamma\}. \tag{3.5}$$

$\mathcal{S}$ is the set of pairs of subspaces, one invariant for A^T and the other for Γ^T.

The following lemma gives a first characterization of the set of minimal square spectral factors.

LEMMA 3.1. *There exists a one-to-one correspondence between the set $\mathcal{W}$ of the minimal square spectral factors of $\Phi(s)$ and the set $\mathcal{S}$ defined above.*

Proof. Let $S := (S_A, S_\Gamma)$ be an element of $\mathcal{S}$. The first step of the proof is to compute a spectral factor $W(s) \in \mathcal{W}$ corresponding to such S.

Consider the following homogeneous ARE:

$$AQ + QA^T + QBB^TQ = 0. \tag{3.6}$$

Since A has unmixed spectrum, there is a one-to-one correspondence between symmetric solutions of (3.6) and invariant subspaces of A^T: this correspondence is given by the map which assigns to each solution $\bar{Q}$ the A^T-invariant subspace $\bar{S} := \ker \bar{Q}$; see, for instance, [14], [15], or [17]. Then let Q_{S_A} be the (unique) symmetric solution of (3.6) whose kernel is S_A, and let

$$H := C - R^{1/2}B^TQ_{S_A}. \tag{3.7}$$

Now let P_{S_Γ} be the (unique) symmetric solution of the ARE

$$\Gamma P + P\Gamma^T + PH^TR^{-1}HP = 0, \tag{3.8}$$

whose kernel is S_Γ. Again the existence and uniqueness of such solution is guaranteed since Γ has unmixed spectrum [14].

Now define the all-pass functions

$$K(s) := I - B^T(sI - A)^{-1}Q_{S_A}B \tag{3.9}$$

and

$$Q(s) := I - R^{-1/2}H(sI - \Gamma)^{-1}P_{S_\Gamma}H^TR^{-1/2}. \tag{3.10}$$

Computing the product $\overline{W}_-(s)K(s)Q(s)$ we get

$$W_S(s) := \overline{W}_-(s)K(s)Q(s) \tag{3.11}$$

$$= [R^{1/2} + C(sI + A^T)^{-1}B][I - B^T(sI - A)^{-1}Q_{S_A}B]Q(s) \tag{3.12}$$

$$= [R^{1/2} + (C - R^{1/2}B^TQ_{S_A})(sI + A^T + BB^TQ_{S_A})^{-1}B]$$

$$\times [I - R^{-1/2}H(sI - \Gamma)^{-1}P_{S_\Gamma}H^TR^{-1/2}] \tag{3.13}$$

$$= R^{1/2} + (C - R^{1/2}B^TQ_{S_A})(sI + A^T + BB^TQ_{S_A})^{-1}(B - P_{S_\Gamma}C^TR^{-1/2}), \tag{3.14}$$

and hence $W_S(s)$ is clearly a minimal spectral factor.

Conversely, given any minimal spectral factor $W_1(s)$, the all-pass function defined by $U_1(s) := \overline{W}_-(s)^{-1}W_1(s)$ is an inner function that admits the inner factorization $U_1(s) = K(s)Q(s)$, where $K(s)$ is such that $\overline{W}_-(s)$ and $W_{1,-}(s) := \overline{W}_-(s)K(s)$ have the same zero structure and $Q(s)$ is such that $W_{1,-}(s)$ and $W_1(s) = W_{1,-}(s)Q(s)$ have the same pole structure [4, Lemma 3.6]. From this fact, using Theorems 3.1 and 3.2 in [14], we can conclude that $K(s)$ has a realization of the form (3.9) and $Q(s)$ has a realization of the form (3.10) for a certain couple Q_{S_A} and P_{S_Γ} of solutions of (3.6) and (3.8), respectively. To the pair (Q_{S_A}, P_{S_Γ}), there corresponds the couple of subspaces $S_A := \ker Q_{S_A}$ and $S_\Gamma := \ker P_{S_\Gamma}$, invariant for A^T and Γ^T, respectively.

It remains to show that the map assigning a minimal spectral factor to each $S \in \mathcal{S}$ is injective. From equation (3.14) we see that the state matrix of a minimal

realization of $W_S(s)$ is given by $F := -A^T - BB^T Q_{S_A}$, where Q_{S_A} is the unique solution of equation (3.6) s.t. its kernel is S_A. Since the subspace S_A is A^T-invariant, we have that, in a basis where the first vectors form a basis for S_A and the remaining vectors form a basis for $(S_A)^\perp$, A^T has the structure $A^T = \left[\begin{smallmatrix} A_{11} & A_{12} \\ 0 & A_{22} \end{smallmatrix}\right]$. In this basis, the matrix Q_{S_A} has the structure $Q_{S_A} = \left[\begin{smallmatrix} 0 & 0 \\ 0 & Q_R \end{smallmatrix}\right]$, where Q_R is a nonsingular matrix. In fact Q_{S_A} is a symmetric matrix whose kernel is S_A. Moreover, denoting by M the lower right block of the matrix BB^T partitioned conforming with A^T, Q_R is the unique nonsingular solution of the following reduced-order ARE:

$$A_{22}^T X + X A_{22} + X M X = 0. \tag{3.15}$$

Hence the state matrix F of $W_S(s)$ can be written as

$$F = -A^T - BB^T Q_{S_A} = \begin{bmatrix} -A_{11} & -A_{12} \\ 0 & -A_{22} \end{bmatrix} - \begin{bmatrix} 0 & * \\ 0 & MQ_R \end{bmatrix} \tag{3.16}$$

$$= \begin{bmatrix} -A_{11} & ** \\ 0 & -A_{22} - MQ_R \end{bmatrix} = \begin{bmatrix} -A_{11} & ** \\ 0 & Q_R^{-1} A_{22}^T Q_R \end{bmatrix}, \tag{3.17}$$

where the values of the blocks $*$ and $**$ have no influence on the argument. Since A has the structure

$$A = \begin{bmatrix} A_{11}^T & 0 \\ A_{12}^T & A_{22}^T \end{bmatrix},$$

we can conclude that the restriction of F to S_A coincides with the restriction of $-A^T$ to S_A, and the map induced by F in the quotient space $\mathbb{R}^n/S_A$ is similar to the map induced by A in the same quotient space. Since A has unmixed spectrum, this implies that to different A^T-invariant subspaces there correspond spectral factors with different pole structure which are therefore necessarily different.

A similar argument shows that to different Γ^T-invariant subspaces there correspond different spectral factors. □

The following corollary characterizes the state and zero dynamics of the spectral factor $W_S(s)$ corresponding to the pair $S = (S_A,\ S_\Gamma)$.

COROLLARY 3.2. *The minimal spectral factor $W_S(s)$ has a minimal realization where the state and the zero matrices F and Λ are s.t.*

1. *the restriction of F to S_A coincides with the restriction of $-A^T$ to S_A.*
2. *the map induced by F in the quotient space $\mathbb{R}^n/S_A$ is similar to the map induced by A in the same quotient space.*
3. *the restriction of Λ to S_Γ coincides with the restriction of Γ to S_Γ.*
4. *the map induced by Λ in the quotient space $\mathbb{R}^n/S_\Gamma$ is similar to the map induced by $-\Gamma^T$ in the same quotient space.*

Proof. The first two points of the corollary are proven by equation (3.16). The proof of points (3) and (4) is similar. □

Remark 3.3. We notice that, in Lemma 3.1, the choice of $\overline{W}_-(s)$ as the reference spectral factor has been made just for the sake of simplicity. In fact, from the proof of that lemma, we see that its validity is independent on this assumption, and it remains true starting from any minimal spectral factor $W(s)$ such that the corresponding state and zero matrices A and Γ satisfies conditions (3.3). For example, a spectral factor (different from $\overline{W}_-(s)$) which satisfies condition (3.3) is $W_+(s)$, the unique minimal stable spectral factor whose inverse is antistable.

We now give an explicit and direct characterization of the set $\mathcal{W}$ in term of a set of solutions of a unique ARE of dimension $2n$. To this aim we define the two matrices L and Z in the following way:

$$L := [-B^T \mid R^{-1/2}C], \qquad Z := \begin{bmatrix} A & 0 \\ 0 & \Gamma \end{bmatrix}. \tag{3.18}$$

Notice that, since both A and Γ are stability matrices, also the spectrum $\sigma(Z)$ of Z is unmixed, i.e.,

$$\sigma(Z) \cap \sigma(-Z) = \emptyset. \tag{3.19}$$

For compactness of notation we shall represent each element $S = (S_A, S_\Gamma) \in \mathcal{S}$ as a subspace of $\mathbb{R}^{2n}$. More precisely,

$$\mathcal{S} := \left\{ S = \left\{ z = \begin{bmatrix} x \\ 0 \end{bmatrix} + \begin{bmatrix} 0 \\ y \end{bmatrix} : \quad x \in S_A,\ y \in S_\Gamma \right\} : \quad S_A \in \mathcal{S}_A, \quad S_\Gamma \in \mathcal{S}_\Gamma \right\}.$$

Clearly each $S \in \mathcal{S}$ is an invariant subspace of Z. The converse is true if and only if A and Γ have disjoint spectra:

$$\sigma(A) \cap \sigma(\Gamma) = \emptyset. \tag{3.20}$$

Consider the following ARE whose solutions are in one-to-one correspondence with the set of invariant subspaces of Z^T:

$$Z\Delta + \Delta Z^T + \Delta L^T L \Delta = 0. \tag{3.21}$$

In view of such correspondence, with each $S \in \mathcal{S}$ we can associate a symmetric solution of equation (3.21): we shall denote by Δ_S such solution. Let $\mathcal{D}_\mathcal{S}$ be the set of solution of (3.21) corresponding to $\mathcal{S}$:

$$\mathcal{D}_\mathcal{S} := \{\Delta_S : \quad S \in \mathcal{S}\}. \tag{3.22}$$

We observe that, in the case when A and Γ have disjoint spectra, $\mathcal{D}_\mathcal{S}$ is the set of all symmetric solutions of (3.21). Also, notice that, given the subspace $S \in \mathcal{S}$, the solution Δ_S may be computed by employing a standard procedure.

In the following proposition we provide an explicit characterization of the set $\mathcal{W}$ in terms of the elements of $\mathcal{D}_\mathcal{S}$.

PROPOSITION 3.4. *There exists a one-to-one correspondence between the set $\mathcal{W}$ of the minimal square spectral factors of $\Phi(s)$ and the set $\mathcal{D}_\mathcal{S}$ above defined. More precisely, we have*

$$\mathcal{W} = \{\overline{W}_-(s)[I - L(sI - Z)^{-1}\Delta L^T] \ : \quad \Delta \in \mathcal{D}_\mathcal{S}\ \}. \tag{3.23}$$

Proof. Let $S = \left\{ \left[\begin{smallmatrix} x \\ 0 \end{smallmatrix} \right] + \left[\begin{smallmatrix} 0 \\ y \end{smallmatrix} \right] : x \in S_A,\ y \in S_\Gamma \right\} \in \mathcal{S}$, and denote by Q_{S_A}, P_{S_Γ}, and Δ_S the solutions of (3.6), (3.8), and (3.21) corresponding to S_A, S_Γ, and S, respectively. Moreover, set

$$\Delta := \begin{bmatrix} Q_{S_A}P_{S_\Gamma}Q_{S_A} + Q_{S_A} & Q_{S_A}P_{S_\Gamma} \\ P_{S_\Gamma}Q_{S_A} & P_{S_\Gamma} \end{bmatrix}. \tag{3.24}$$

We now show that

$$\Delta_S = \Delta. \tag{3.25}$$

To this aim, it is sufficient to prove that the matrix Δ given by (3.24) solves equation (3.21) and that $\ker \Delta = S$. The first of these two facts may be checked by plugging the left-hand side of (3.24) into equation (3.21) and analyzing the resulting equation block by block. The upper left block is given by

$$\begin{aligned}
&AQ_{S_A} + Q_{S_A}A^T + Q_{S_A}BB^TQ_{S_A} \\
&+ (Q_{S_A}BB^TQ_{S_A} + AQ_{S_A} - Q_{S_A}BR^{-1/2}C)P_{S_\Gamma}Q_{S_A} \\
&+ Q_{S_A}P_{S_\Gamma}(Q_{S_A}BB^TQ_{S_A} + Q_{S_A}A^T - C^TR^{-1/2}B^TQ_{S_A}) \\
&+ Q_{S_A}P_{S_\Gamma}(Q_{S_A}BB^TQ_{S_A} - C^TR^{-1/2}B^TQ_{S_A} - Q_{S_A}BR^{-1/2}C + C^TR^{-1}C)P_{S_\Gamma}Q_{S_A}.
\end{aligned} \tag{3.26}$$

The sum of the first three terms vanishes in view of equation (3.6). Again, in view of equation (3.6) and using the definition of Γ, the terms in the first parenthesis reduce to $Q_{S_A}\Gamma$. Symmetrically the second parenthesis reduces to $\Gamma^TQ_{S_A}$. Finally, it is easy to see that the terms in the last parenthesis add up to $HR^{-1}H$, where H is defined in (3.7). The upper left block is therefore $Q_{S_A}(\Gamma P_{S_\Gamma} + P_{S_\Gamma}\Gamma^T + P_{S_\Gamma}C_1R^{-1}C_1P_{S_\Gamma})Q_{S_A}$, and this clearly vanishes since P_{S_Γ} solves (3.8). Similarly one can check that the other blocks are equal to zero. Therefore Δ is a symmetric solution of (3.21).

The fact that $\ker \Delta = S$ follows from the identity

$$\begin{bmatrix} Q_{S_A}P_{S_\Gamma}Q_{S_A} + Q_{S_A} & Q_{S_A}P_{S_\Gamma} \\ P_{S_\Gamma}Q_{S_A} & P_{S_\Gamma} \end{bmatrix} = \begin{bmatrix} I & Q_{S_A} \\ 0 & I \end{bmatrix} \begin{bmatrix} Q_{S_A} & 0 \\ P_{S_\Gamma}Q_{S_A} & P_{S_\Gamma} \end{bmatrix}, \tag{3.27}$$

which implies

$$\begin{aligned}
\ker \Delta &= \ker \begin{bmatrix} Q_{S_A} & 0 \\ P_{S_\Gamma}Q_{S_A} & P_{S_\Gamma} \end{bmatrix} \\
&= \left\{ \begin{bmatrix} x \\ y \end{bmatrix} \ : \ Q_{S_A}x = 0, \ \ P_{S_\Gamma}Q_{S_A}x + P_{S_\Gamma}y = 0 \right\} \qquad (3.28) \\
&= \left\{ \begin{bmatrix} x \\ y \end{bmatrix} \ : \ x \in \ker Q_{S_A}, \ \ y \in \ker P_{S_\Gamma} \right\} = S.
\end{aligned}$$

In this way equation (3.25) remains proven.

Now we show that

$$I - L(sI - Z)^{-1}\Delta_S L^T = K(s)Q(s), \tag{3.29}$$

where $K(s)$ and $Q(s)$ are defined by (3.9) and (3.10). Taking into account that

$$\begin{aligned}
Q_{S_A}BR^{-1/2}H &= Q_{S_A}BR^{-1/2}C - Q_{S_A}BB^TQ_{S_A} \\
&= -Q_{S_A}A^T - Q_{S_A}\Gamma - Q_{S_A}BB^TQ_{S_A} \qquad (3.30) \\
&= AQ_{S_A} - Q_{S_A}\Gamma = Q_{S_A}(sI - \Gamma) - (sI - A)Q_{S_A}, \qquad (3.31)
\end{aligned}$$

where the first equality follows from the definition (3.7) of H, the second from the definition of Γ, and the third from (3.6), we get

$$K(s)Q(s) = B^T(sI - A)^{-1}Q_{S_A}(P_{S_\Gamma}H^TR^{-1/2} - B) - (R^{-1/2}H + B^TQ_{S_A})(sI - \Gamma)^{-1}P_{S_\Gamma}H^TR^{-1/2} + I. \tag{3.32}$$

From this equation, using again the definition (3.7) of H we immediately get equation (3.29).

The previous argument, together with equation (3.11) and Lemma 3.1, proves equation (3.23). □

Remark 3.5. Proposition 3.4 extends a result presented in [4], where the assumption (3.20) was made, to arbitrary coercive spectral densities. In particular, as was pointed out in [2], if condition (3.20) fails, then there exist solutions of equation (3.21) corresponding to nonminimal spectral factors: Proposition 3.4 characterizes the set $\mathcal{D}_\mathcal{S}$ of the solutions of (3.21) corresponding to minimal spectral factors.

Remark 3.6. In [6] a characterization of minimal spectral factors in terms of the solutions of a pair of n-dimensional ARE was presented. Proposition 3.4 gives a characterization of minimal spectral factors *at once* in terms of the solutions of a unique $2n$-dimensional ARE. The advantage of such characterization is that of representing the class of minimal spectral factors in terms of a unique parameter $\Delta \in \mathcal{D}_\mathcal{S}$, as shown by equation (3.23). In this way, the comparison of different spectral factors is more direct as is shown, for example, in the next section, where the difference of minimal spectral factors is considered. Also, such a representation of the class $\mathcal{W}$ in terms of a unique parameter Δ seems to be very natural when one minimal spectral factor has to be selected in the class $\mathcal{W}$ by minimizing a certain cost function. Finally, from the numerical point of view, the $2n$-dimensional ARE (3.21) is not more complex than a pair of n-dimensional AREs. In fact, since the matrix Z is block-diagonal, equation (3.21) may be easily decoupled.

Remark 3.7. Let Δ be an element of $\mathcal{D}_\mathcal{S}$ partitioned into four $n \times n$ blocks:

$$\Delta = \begin{bmatrix} \Delta_{11} & \Delta_{12} \\ \Delta_{21} & \Delta_{22} \end{bmatrix}. \tag{3.33}$$

As we have shown in the proof of the previous proposition, Δ has the form (3.24). Hence we have[1]

$$\begin{cases} P_{S_\Gamma} = \Delta_{22}, \\ P_{S_A} = \Delta_{11} - \Delta_{12}\Delta_{22}^\sharp\Delta_{21}. \end{cases} \tag{3.34}$$

This observation and equations (3.11) provide a minimal realization of the spectral factor $W_\Delta(s)$ directly in terms of A, B, C, R, and the solution Δ.

[1] In this paper, if Δ is a matrix, $\Delta^\sharp$ will denote the Moore–Penrose pseudoinverse of Δ, i.e., the unique matrix s.t.
(1) $\Delta\Delta^\sharp\Delta = \Delta$,
(2) $\Delta^\sharp\Delta\Delta^\sharp = \Delta^\sharp$,
(3) $(\Delta\Delta^\sharp)^T = \Delta\Delta^\sharp$,
(4) $(\Delta^\sharp\Delta)^T = \Delta^\sharp\Delta$.

4. Continuity. In this section we investigate the properties of continuity of the function Θ, which associates a solution $\Delta_S \in \mathcal{D}_{\mathcal{S}}$ with the Z^T-invariant subspace $S \in \mathcal{S}$, and of the function Ξ, which associates a minimal spectral factor $W_\Delta(s) \in \mathcal{W}$ to $\Delta \in \mathcal{D}_{\mathcal{S}}$. To this aim, we specify a metric for each of the three sets $\mathcal{S}$, $\mathcal{D}_{\mathcal{S}}$, and $\mathcal{W}$. We endow the vector space $\mathbb{C}^n$ ($n \in \mathbb{N}$) with the usual Euclidean norm; i.e., for $x \in \mathbb{C}^n$, $\|x\|_e := (x^*x)^{1/2}$ (with the symbol x^* we denote the conjugate transpose of x). We define, as usual, $\|Y\| := \max\{\|Yx\|_e : \ x \in \mathbb{C}^n, \|x\|_e = 1\}$ to be the norm of a matrix $Y \in \mathbb{C}^{m\times n}$. It is well known that $\|Y\|$ equals the largest singular value of the matrix Y, i.e., the square root of the largest eigenvalue of Y^*Y. Since $\mathcal{D}_{\mathcal{S}} \subset \mathbb{C}^{n\times n}$, the norm $\|\cdot\|$ defined above induces a metric in the set $\mathcal{D}_{\mathcal{S}}$.

Since $\Phi(s)$ is a spectral density, it is analytic on an open strip including the imaginary axis. Then any minimal spectral factor $W(s) \in \mathcal{W}$ is analytic in the same strip, and this implies that the matrix norm of $W(i\omega)$ is a bounded function of $\omega \in \mathbb{R}$, or, more compactly, $W(s) \in \mathcal{L}^\infty_{m\times m}(\mathbb{I})$. Hence, we endow the set $\mathcal{W}$ with the metric induced by the $\mathcal{L}^\infty_{m\times m}(\mathbb{I})$ norm; i.e., if $W(s) \in \mathcal{W}$, $\|W(s)\|_\infty := \sup_{\omega\in\mathbb{R}} \|W(i\omega)\|$, where by $\|W(i\omega)\|$ we intend the norm of the matrix $W(i\omega)$ as defined above. Finally we endow the set $\mathcal{S}$ with the *gap metric*. This is defined as follows: Let $S_1, S_2 \in \mathcal{S}$, and define

$$
\begin{array}{rccl}
d: & \mathcal{S}\times\mathcal{S} & \longrightarrow & \mathbb{R}_+, \\
 & S_1 \times S_2 & \mapsto & d(S_1,S_2) := ||P_{S_1} - P_{S_2}||,
\end{array}
\tag{4.1}
$$

where P_{S_i}, $i = 1, 2$, is the (linear) operator of orthogonal projection onto the space S_i. (This function d defines a metric in the set of subspaces of $\mathbb{R}^{2n}$). Next we recall some well-known facts. Let $\mathcal{L}^2_m[a, b]$ be the space of m-dimensional vector-valued functions $v(t)$ such that $\int_a^b v^T(t)v(t)dt < +\infty$. The number $\int_a^b v^T(t)v(t)dt$ is the $\mathcal{L}^2_m[a, b]$-norm of $v(t)$, and it is denoted by $\|v\|_2$. As a standard application of Parseval equality and the Bochner theorem (see, e.g., [5]) we have that if

$$
\begin{cases}
\dot{x}(t) = Ax(t) + Bu(t), \\
y(t) = Cx(t) + Du(t)
\end{cases}
\tag{4.2}
$$

is a linear system where $A \in \mathbb{R}^{n\times n}$ is a stability matrix, $B \in \mathbb{R}^{n\times m}$, and $C \in \mathbb{R}^{p\times n}$, then the transfer function $G(s) = C(sI - A)^{-1}B + D$ is in $\mathcal{L}^\infty_{p\times m}(\mathbb{I})$ and its norm is given by

$$
\|G(s)\|_\infty = \sup\{\|y_u(t)\|_2 : \ u \in \mathcal{L}^2_m[0,\infty], \ \|u(t)\|_2 = 1\},
\tag{4.3}
$$

where $y_u(t)$ is the forced response of system (4.2) to the input $u(t)$. The function $G(s)$ is in fact in $H^\infty_{p\times m}$ [5], and (4.3) is also the H^∞ norm of $G(s)$.

To prove our result on the continuity properties of Θ, Ξ, we need some preliminary results. First, we recall from [9] that, given two matrices $A \in \mathbb{R}^{n\times n}$ and $B \in \mathbb{R}^{n\times m}$ and denoting by $\lambda_m(A)$ the smallest singular value of A, the following inequality holds:

$$
\|AB\| \geq \lambda_m(A)\|B\|.
\tag{4.4}
$$

The next two lemmas collect some results that will be useful in what follows.

LEMMA 4.1. *The pair* (L, Z) *defined by equation* (3.18) *is observable.*

Proof. Since equation (3.1) is a minimal realization of a minimal spectral factor, the spectral density $\Phi(s) = \overline{W}_-(s)\overline{W}_-^T(-s)$ has McMillan degree $2n$ (where n is the dimension of the square matrix A). Using the realization (3.1) we get

$$\begin{aligned}\Phi(s) &= \overline{W}_-(s)\overline{W}_-^T(-s)\\ &= R + C(sI + A^T)^{-1}BR^{1/2}\\ &\quad - R^{1/2}B^T(sI - A)^{-1}C^T - C(sI + A^T)^{-1}BB^T(sI - A)^{-1}C^T.\end{aligned} \tag{4.5}$$

We now define N to be the unique solution of the following Lyapunov equation:

$$BB^T = A^TN + NA. \tag{4.6}$$

The existence and uniqueness of such N are due to the stability of A. Equation (4.6) yields immediately $BB^T = (sI + A^T)N - N(sI - A)$. Plugging this expression of BB^T in equation (4.5) we get

$$\begin{aligned}\Phi(s) = &\begin{bmatrix} -CN - R^{1/2}B^T & C \end{bmatrix} \begin{bmatrix} sI - A & 0 \\ 0 & sI + A^T \end{bmatrix}^{-1}\\ &\times \begin{bmatrix} C^T \\ BR^{1/2} + NC^T \end{bmatrix} + R.\end{aligned} \tag{4.7}$$

Since $\Phi(s)$ has McMillan degree $2n$, its realization (4.7) is minimal and hence observable. Therefore, using the celebrated Popov–Bielevich–Hautus test, we have

$$\text{Rank} \begin{bmatrix} \lambda I - A & 0 \\ 0 & \lambda I + A^T \\ -CN - R^{1/2}B^T & C \end{bmatrix} = 2n \qquad \forall \lambda \in \mathbb{C}. \tag{4.8}$$

This immediately implies

$$2n = \text{Rank} \begin{bmatrix} \lambda I - A & 0 \\ 0 & \lambda I + A^T \\ -CN - R^{1/2}B^T & C \end{bmatrix} \tag{4.9}$$

$$\begin{aligned}= \text{Rank} \Bigg(&\begin{bmatrix} I & 0 & 0 \\ -N & I & 0 \\ 0 & 0 & I \end{bmatrix} \begin{bmatrix} I & 0 & 0 \\ 0 & I & BR^{-1/2} \\ 0 & 0 & R^{-1/2} \end{bmatrix}\\ &\times \begin{bmatrix} \lambda I - A & 0 \\ 0 & \lambda I + A^T \\ -CN - R^{1/2}B^T & C \end{bmatrix} \begin{bmatrix} I & 0 \\ N & I \end{bmatrix} \Bigg)\end{aligned} \tag{4.10}$$

$$= \text{Rank} \begin{bmatrix} \lambda I - Z \\ L \end{bmatrix} \qquad \forall \lambda \in \mathbb{C}. \tag{4.11}$$

This concludes the proof. ∎

Before introducing the next lemma we give a definition and recall some well-known results. Given a pair of matrices $Z \in \mathbb{R}^{n\times n}$ and $G \in \mathbb{R}^{n\times m}$, we define $\mathcal{R}_{[0.T]}^{(Z,G)}$ as the following reachability operator:

$$\begin{aligned}\mathcal{R}_{[0.T]}^{(Z,G)}: \quad \mathcal{L}_m^2[0,T] &\longrightarrow \mathbb{R}^n,\\ u(\cdot) &\mapsto \mathcal{R}_{[0.T]}^{(Z,G)}(u) := \int_0^T e^{Zt}Gu(t)dt.\end{aligned} \tag{4.12}$$

For simplicity of notation, from now on, we shall drop the (Z, G) and simply write $\mathcal{R}_{[0,T]}$. We define the norm $\|\mathcal{R}_{[0,T]}\|_o$ as the operator norm induced on $\mathcal{R}_{[0,T]}$ by the usual norms $\|\cdot\|_2$ and $\|\cdot\|_e$ on $\mathcal{L}^2_m[0,T]$ and $\mathbb{R}^n$, respectively. Precisely,

$$\|\mathcal{R}_{[0,T]}\|_o = \sup\{\|\mathcal{R}_{[0,T]}(u)\|_e : \ u \in \mathcal{L}^2_m[0,T], \ \|u(t)\|_2 = 1\}. \tag{4.13}$$

It is easy to see that

$$\|\mathcal{R}_{[0,T]}\|_o = \left\| \int_0^T e^{Zt} G G^T e^{Z^T t} dt \right\|^{1/2}. \tag{4.14}$$

In fact $\mathcal{R}_{[0,T]}$ is a linear operator on Hilbert spaces, and hence the well-known relation

$$\|\mathcal{R}_{[0,T]}\|_o = \|\mathcal{R}^*_{[0,T]} \mathcal{R}_{[0,T]}\|_o^{1/2}, \tag{4.15}$$

where $\mathcal{R}^*_{[0,T]}$ is the adjoint of $\mathcal{R}_{[0,T]}$, holds.

LEMMA 4.2. *Let $Z \in \mathbb{R}^{n\times n}$ be a matrix with no eigenvalues on the imaginary axis, $G \in \mathbb{R}^{n\times m}$, and $\mathcal{R}_{[0,1]}$ be the corresponding reachability operator defined in* (4.12). *Then, there exist a real constant $k > 0$ independent of G and a function $\bar{u} \in \mathcal{L}^2_m[0,1]$ s.t. $\|\bar{u}\|_2 = 1$ and*

$$\|\mathcal{R}_{[0,1]}\|_o \geq \|\mathcal{R}_{[0,1]}(\bar{u})\|_e \geq k\|G\|. \tag{4.16}$$

Proof. By definition we have $\|G\| = \max_{\|u\|_e=1} \|Gu\|_e$. Then, denoting by u_m the vector $u_m := \arg\max_{\|u\|_e=1} \ \|Gu\|_e$ and by g the vector $g := Gu_m$ we have

$$\|G\| = \|g\|_e. \tag{4.17}$$

Set $\bar{u}(t) := u_m$, $t \in [0,1]$. Clearly, $\bar{u} \in \mathcal{L}^2_m[0,1]$ and $\|\bar{u}\|_2 = 1$.

Since Z has no eigenvalues on the imaginary axis, it is nonsingular, and hence we can write

$$\|\mathcal{R}_{[0,1]}\|_o \geq \|\mathcal{R}_{[0,1]}(\bar{u})\|_e = \left\| \int_0^1 e^{Zt} G u_m dt \right\|_e \tag{4.18}$$

$$= \left\| \int_0^1 e^{Zt} dt g \right\|_e = \|Z^{-1}(e^Z - I)g\|_e \tag{4.19}$$

$$\geq \lambda_m(Z^{-1}(e^Z - I))\|g\|_e \tag{4.20}$$

$$= k\|G\|, \tag{4.21}$$

where $k := \lambda_m(Z^{-1}(e^Z - I))$ is the smallest singular value of $Z^{-1}(e^Z - I)$ and (4.20) is obtained in view of equation (4.4). We now observe that Z has no eigenvalues on the imaginary axis, and hence $e^Z - I$ is nonsingular. Then, so is $Z^{-1}(e^Z - I)$, and therefore k, the smallest singular value of $Z^{-1}(e^Z - I)$, is strictly positive. ∎

The following theorem, which is our main result, establishes the continuity of functions Θ and Ξ and their inverses.

THEOREM 4.3. *The maps*

$$\begin{array}{rccc} \Theta : & \mathcal{S} & \longrightarrow & \mathcal{D}_{\mathcal{S}}, \\ & S & \mapsto & \Delta_S \end{array} \tag{4.22}$$

and

$$\begin{array}{rccc} \Xi: & \mathcal{D}_{\mathcal{S}} & \longrightarrow & \mathcal{W}, \\ & \Delta & \mapsto & W_{\Delta}(s) \end{array} \tag{4.23}$$

are homeomorphisms.

Proof. The proof for the map Θ can be obtained with trivial modifications from [19, Theorem 4.3].

We now prove the continuity of the map Ξ. Let $\mu := \|\overline{W}_-(s)\|_\infty$ and $\nu := \|L(sI - Z)^{-1}\|_\infty$. Since A and Z are stability matrices, they have no eigenvalues on the imaginary axis. Hence μ and ν are finite nonnegative real numbers. Now let $\Delta, \Delta_1 \in \mathcal{D}_{\mathcal{S}}$. In view of Proposition 3.4, we have

$$\|\Xi(\Delta) - \Xi(\Delta_1)\|_\infty = \|\overline{W}_-(s)[L(sI - Z)^{-1}(\Delta - \Delta_1)L^T]\|_\infty. \tag{4.24}$$

Now recall that if $U(s) \in \mathcal{L}^\infty_{p\times q}$, $V(s) \in \mathcal{L}^\infty_{q\times r}$, then $\|U(s)V(s)\|_\infty \le \|U(s)\|_\infty \|V(s)\|_\infty$. Then from equation (4.24) it follows that

$$\|\Xi(\Delta) - \Xi(\Delta_1)\|_\infty \le \mu\nu\|L\|\|\Delta - \Delta_1\|, \tag{4.25}$$

and this immediately yields

$$\|\Xi(\Delta) - \Xi(\Delta_1)\|_\infty \xrightarrow{\|\Delta-\Delta_1\|\to 0} 0. \tag{4.26}$$

It remains to prove continuity of the map Ξ^{-1}. Assume that $W(s)$ and $W_1(s)$ are two minimal spectral factors and Δ and Δ_1 are the corresponding solutions of equation (3.21).

Since the spectral density $\Phi(s)$ is, by assumption, coercive, the smallest singular value $\lambda_m(\overline{W}_-(i\omega))$ of $\overline{W}_-(i\omega)$, is, as a function of ω, bounded away from zero, i.e.,

$$\inf_\omega \lambda_m(\overline{W}_-(i\omega)) = c > 0. \tag{4.27}$$

Thus, in view of equation (4.4), we have

$$\|\Xi(\Delta) - \Xi(\Delta_1)\|_\infty \tag{4.28}$$

$$= \|\overline{W}_-(s)L(sI - Z)^{-1}(\Delta - \Delta_1)L^T\|_\infty \tag{4.29}$$

$$= \sup_\omega \|\overline{W}_-(i\omega)L(i\omega I - Z)^{-1}(\Delta - \Delta_1)L^T\| \tag{4.30}$$

$$\ge \sup_\omega [\lambda_m(\overline{W}_-(i\omega))\|L(i\omega I - Z)^{-1}(\Delta - \Delta_1)L^T\|] \tag{4.31}$$

$$\ge \inf_\omega [\lambda_m(\overline{W}_-(i\omega))] \sup_\omega \|L(i\omega I - Z)^{-1}(\Delta - \Delta_1)L^T\| \tag{4.32}$$

$$= c\|L(sI - Z)^{-1}(\Delta - \Delta_1)L^T\|_\infty. \tag{4.33}$$

Since $c > 0$, the comparison of equations (4.33) and (4.28) proves that

$$\|L(sI - Z)^{-1}(\Delta - \Delta_1)L^T\|_\infty \xrightarrow{\|\Xi(\Delta)-\Xi(\Delta_1)\|_\infty\to 0} 0. \tag{4.34}$$

We now use Lemmas 4.2 and 4.1 to prove that if $\|L(sI-Z)^{-1}(\Delta-\Delta_1)L^T\|_\infty \to 0$, then also

$$\|(\Delta - \Delta_1)L^T\| \to 0. \tag{4.35}$$

Define $G := (\Delta - \Delta_1)L^T$. Since Z is a stability matrix, as we have recalled in (4.3), the norm $\|L(sI - Z)^{-1}G\|_\infty$ is given by

$$\|L(sI - Z)^{-1}G\|_\infty = \sup_{\substack{u \in \mathcal{L}^2_m[0,+\infty) \\ \|u\|_2 = 1}} \|y_u(\cdot)\|_2, \tag{4.36}$$

where $y_u(\cdot)$ is defined by

$$y_u(t) := \int_0^t Le^{Z(t-\sigma)}Gu(\sigma)d\sigma. \tag{4.37}$$

From (4.36) we easily get

$$\|L(sI - Z)^{-1}G\|_\infty \geq \sup_{\substack{u \in \mathcal{L}^2_m[0,1] \\ \|u\|_2 = 1}} \|y_u(\cdot)\|_2. \tag{4.38}$$

We observe that if $u \in \mathcal{L}^2_m[0,1]$, then $y_u(t)$ is, for $t \geq 1$, the free response the system (G, Z, L) corresponding to the initial state $x(1) = \mathcal{R}_{[0,1]}(u)$, where $\mathcal{R}_{[0,1]}$ is the operator defined in (4.12). Also, it is obvious that $\int_0^{+\infty} \|y(t)\|_e^2 dt \geq \int_1^{+\infty} \|y(t)\|_e^2 dt$, and hence we can write

$$\sup_{\substack{u \in \mathcal{L}^2_m[0,1] \\ \|u\|_2 = 1}} \|y_u(\cdot)\|_2 \geq \sup_{\substack{u \in \mathcal{L}^2_m[0,1] \\ \|u\|_2 = 1}} \left(\int_1^{\infty} \|y_u(t)\|_e^2 dt \right)^{1/2} \tag{4.39}$$

$$= \sup_{\substack{u \in \mathcal{L}^2_m[0,1] \\ \|u\|_2 = 1}} \left(\int_1^{+\infty} \|Le^{Z(t-1)}\mathcal{R}_{[0,1]}(u)\|_e^2 dt \right)^{1/2} \tag{4.40}$$

$$= \sup_{\substack{u \in \mathcal{L}^2_m[0,1] \\ \|u\|_2 = 1}} \left(\int_0^{+\infty} \|Le^{Zt}\mathcal{R}_{[0,1]}(u)\|_e^2 dt \right)^{1/2}, \tag{4.41}$$

where the derivation of equation (4.41) is immediate.

Since Z is a stability matrix, it has no eigenvalues on the imaginary axis. Thus, we can employ Lemma 4.2, which proves the existence of an $\mathcal{L}^2_m[0,1]$ function $\bar{u}$ with unitary norm and of a positive constant k independent of G s.t., defining $\bar{x} := \mathcal{R}_{[0,1]}(\bar{u})$, we have

$$\|\bar{x}\|_e \geq k\|G\|. \tag{4.42}$$

From the latter equation and equations (4.36), (4.38), and (4.39) we get

$$\|L(sI - Z)^{-1}G\|_\infty \geq \left(\int_0^{+\infty} \|Le^{Zt}\bar{x}\|_e^2 dt \right)^{1/2} = \left(\bar{x}^T \mathcal{O} \bar{x}\right)^{1/2}, \tag{4.43}$$

where $\mathcal{O}$ is the observability Gramian of the pair (Z, L). Since, as we proved in Lemma 4.1, the pair (Z, L) is observable, the matrix $\mathcal{O}$ is positive definite and hence it may be factored as $\mathcal{O} = V^T V$, where V is nonsingular. Thus, the smallest singular value $\lambda_m(V)$ of V is strictly positive. Therefore, employing equation (4.4), we get

$$\|L(sI - Z)^{-1}G\|_\infty \geq \|V\bar{x}\| \geq \lambda_m(V)\|\bar{x}\|_e \geq k_1\|G\|, \tag{4.44}$$

where $k_1 := k\lambda_m(V)$ is strictly positive. We can then conclude that if $\|L(sI - Z)^{-1}(\Delta - \Delta_1)L^T\|_\infty \to 0$, then also $\|(\Delta - \Delta_1)L^T\| = \|G\| \to 0$. This clearly implies that

$$\|\Delta L^T L\Delta - \Delta_1 L^T L\Delta_1\| \to 0. \tag{4.45}$$

In fact, defining $X := \Delta L^T$ and $X_1 := \Delta_1 L^T$, we have

$$\|\Delta L^T - \Delta_1 L^T\| = \|X - X_1\| \tag{4.46}$$

$$\geq \frac{1}{2\|X_m\|}[\|X - X_1\|\|X_1^T\| + \|X\|\|X^T - X_1^T\|] \tag{4.47}$$

$$\geq \frac{1}{2\|X_m\|}\|X_1X_1^T - XX^T\| \tag{4.48}$$

$$= k_2\|\Delta L^T L\Delta - \Delta_1 L^T L\Delta_1\|, \tag{4.49}$$

with $k_2 := \frac{1}{2\|X_m\|} > 0$ and $X_m := \Delta_m L^T$, where Δ_m is the solution of (3.21) which maximizes the norm $\|\Delta_m L^T\|$. (Such solution exists since the set of solutions of (3.21) is compact.)

Since both Δ and Δ_1 solve (3.21), equations (4.46) to (4.49) may be rewritten as

$$\|\Delta L^T - \Delta_1 L^T\| \geq k_2\|Z(\Delta - \Delta_1) + (\Delta - \Delta_1)Z^T\|, \tag{4.50}$$

where k_2 is strictly positive. This clearly implies that if $\|(\Delta - \Delta_1)L^T\| \to 0$ then also

$$\|Z(\Delta - \Delta_1) + (\Delta - \Delta_1)Z^T\| \to 0. \tag{4.51}$$

Finally we have to prove that (4.51) implies

$$\|\Delta - \Delta_1\| \to 0. \tag{4.52}$$

To see this define $Y := \Delta - \Delta_1$ and $W := Z(\Delta - \Delta_1) + (\Delta - \Delta_1)Z^T$. Since Z is a stability matrix, the Lyapunov equation

$$ZY + YZ^T = W \tag{4.53}$$

has the unique solution

$$Y = \Delta - \Delta_1 = \int_0^\infty e^{Zt}We^{Z^Tt}dt, \tag{4.54}$$

and then we have

$$\|\Delta - \Delta_1\| = \left\|\int_0^\infty e^{Zt}We^{Z^Tt}dt\right\| \tag{4.55}$$

$$\leq \int_0^\infty \|e^{Zt}We^{Z^Tt}\|dt \tag{4.56}$$

$$\leq \int_0^\infty \|e^{Zt}\|^2dt\|W\| \tag{4.57}$$

$$= k_3\|W\|, \tag{4.58}$$

where $k_3 := \int_0^\infty \|e^{Zt}\|^2 dt$ is finite since Z is a stability matrix. This proves formula (4.52) and concludes the proof. □

COROLLARY 4.4. *The map Ξ and its inverse are Lipschitz continuous.*

Proof. From the first part of the proof of Theorem 4.3, it immediately follows that if Δ and Δ_1 are elements of $\mathcal{D}_\mathcal{S}$, then

$$\|\Xi(\Delta) - \Xi(\Delta_1)\|_\infty \leq K\|\Delta - \Delta_1\|, \tag{4.59}$$

where K is given by

$$K = \|\overline{W}_-(s)\|_\infty \|L\| \cdot \|L(sI - Z)^{-1}\|_\infty. \tag{4.60}$$

On the other hand, comparing inequalities (4.28)–(4.33), (4.44), (4.50), and (4.55)–(4.58) we get

$$\|\Delta - \Delta_1\| \leq \bar{K}\|\Xi(\Delta) - \Xi(\Delta_1)\|_\infty, \tag{4.61}$$

where $\bar{K} := \frac{k_3}{ck_1k_2}$ is finite since k_3 is finite and c, k_1, and k_2 are strictly positive. (See the definitions of c, k_1, k_2, and k_3, in (4.33), (4.44), (4.50), and (4.58), respectively.) □

Remark 4.5. It is worth noting that equations (4.59) and (4.60) give

$$\frac{\|W_\Delta(s) - W_{\Delta_1}(s)\|_\infty}{\|\overline{W}_-(s)\|_\infty} \leq k\|\Delta - \Delta_1\|, \tag{4.62}$$

where $k = \|L\| \cdot \|L(sI - Z)^{-1}\|_\infty$. We observe that for all spectral factors $W(s)$ of $\Phi(s)$ we have $\|\overline{W}_-(s)\|_\infty = \|W(s)\|_\infty = \|\Phi(s)\|_\infty^{1/2}$. Then the denominator of the left-hand side of (4.62) is a constant depending only on $\Phi(s)$, and thus the left-hand side of (4.62) has the meaning of relative error on determination of the spectral factor given an error $\|\Delta - \Delta_1\|$ on the solution of the ARE (3.21). From equation (4.62) we see that given a spectral density corresponding to a *small* k, if Δ_1 is *close* to Δ, then the relative error on $W_{\Delta_1}(s)$ (with respect to $W_\Delta(s)$) is *small*. On the contrary, for spectral density corresponding to a *large* k, this is not guaranteed.

The value of k depends on L and Z, which depend only on the data of the problem. Roughly speaking, once the matrix L has been fixed, the value of k depends on the stability margin of Z and hence on the stability margin of A and Γ. Thus for spectral factors with zeros and poles *far* from the imaginary axis, the relative error on the spectral factor is not much larger of the error on the solution of the ARE (3.21).

REFERENCES

[1] P. FAURRE, M. CLERGET, AND F. GERMAIN, *Opérateurs Rationnels Positifs*, Dunod, Paris, 1979.

[2] A. FERRANTE, *A parametrization of minimal stochastic realizations*, IEEE Trans. Automat. Control, 39 (1994), pp. 2122–2126.

[3] A. FERRANTE, *A parametrization of the minimal square spectral factors of a nonrational spectral density*, J. Math. Systems Estim. Control, 4 (1994), pp. 489–492. Summary. Full paper in publication and available via ftp from the publisher.

[4] A. FERRANTE, G. MICHALETZKY, AND M. PAVON, *Parametrization of all minimal square spectral factors*, Systems Control Lett., 21 (1993), pp. 249–254.

[5] B. FRANCIS, *A Course in H_∞ Control Theory*, Springer-Verlag, New York, 1987.

[6] P. A. FUHRMANN, *On the characterization and parametrization of minimal spectral factors*, J. Math. Systems Estim. Control, (1995). Accepted for publication and available via ftp from the publisher.

[7] P. A. Fuhrmann and R. Ober, *State space formulas for coprime factorizations*, in Contributions to Operator Theory and Its Applications, Oper. Theory Adv. Appl. 62, T. Furuta, I. Gohberg, and T. Nakazi, eds., Birkhäuser, Basel, 1993, pp. 39–75.

[8] P. Gahinet and A. Laub, *Computable bounds for the sensitivity of the algebraic Riccati equation*, SIAM J. Control Optim., 28 (1990), pp. 1461–1480.

[9] R. Horn and C. Johnson, *Matrix Analysis*, Cambridge University Press, Cambridge, UK, 1985.

[10] A. Lindquist and G. Picci, *On the stochastic realization problem*, SIAM J. Control Optim., 17 (1979), pp. 365–389.

[11] A. Lindquist and G. Picci, *Realization theory for multivariate stationary Gaussian processes*, SIAM J. Control Optim., 23 (1985), pp. 809–857.

[12] A. Lindquist and G. Picci, *A geometric approach to modelling and estimation of linear stochastic systems*, J. Math. Systems Estim. Control, 1 (1991), pp. 241–333.

[13] M. Pavon, *On the parametrization of nonsquare spectral factors*, in Systems and Networks: Mathematical Theory and Application—Proceedings of the Int. Symp. MTNS '93, U. Helmke, R. Mennicken, and J. Saurer, eds., vol. II, Regensburg, Germany, 2–6 August 1993, Akademie Verlag, Berlin, pp. 413–416.

[14] G. Picci and S. Pinzoni, *Acausal models of stationary processes*, in Proc. First European Control Conference, 1991, pp. 613–616.

[15] G. Picci and S. Pinzoni, *Acausal models and balanced realizations of stationary processes*, Linear Algebra Appl., 205-206 (1994), pp. 997–1043.

[16] J. C. Willems, *Least squares stationary optimal control and the algebraic Riccati equation*, IEEE Trans. Automat. Control, AC-16 (1971), pp. 621–634.

[17] H. Wimmer, *Decomposition and parametrization of semidefinite solutions of the continuous-time algebraic Riccati equation*, SIAM J. Control Optim., 32 (1994), pp. 995–1007.

[18] H. Wimmer, *A Galois correspondence between sets of semidefinite solutions of continuous-time algebraic Riccati equations*, Linear Algebra Appl., 205-206 (1994), pp. 1253–1270.

[19] H. Wimmer, *Isolated semidefinite solutions of the continuous-time algebraic Riccati equation*, Integral Equations Operator Theory, 21 (1995), pp. 362–375.

[20] H. Wimmer, *Lattice properties of sets of semidefinite solutions of continuous-time algebraic Riccati equations*, Automatica J. IFAC, 31 (1995), pp. 173–182.

SIAM J. CONTROL OPTIM.
Vol. 35, No. 5, pp. 1524–1543, September 1997

005

AUGMENTED LAGRANGIAN TECHNIQUES FOR ELLIPTIC STATE CONSTRAINED OPTIMAL CONTROL PROBLEMS*

MAÏTINE BERGOUNIOUX† AND KARL KUNISCH‡

Abstract. We propose augmented Lagrangian methods to solve state and control constrained optimal control problems. The approach is based on the Lagrangian formulation of nonsmooth convex optimization in Hilbert spaces developed in [K. Ito and K. Kunisch, *Augmented Lagrangian Methods for Nonsmooth Convex Optimization in Hilbert Spaces*, preprint, 1994]. We investigate a linear optimal control problem with a boundary control function as in [M. Bergounioux, *Numer. Funct. Anal. Optim.*, 14 (1993), pp. 515–543]. Both the equation and the constraints are augmented. The proposed methods are general and can be adapted to a much wider class of problems.

Key words. state and control constrained optimal control problems, augmented Lagrangian, elliptic equations

AMS subject classifications. 49J20, 49M29

PII. S036301299529330X

1. Setting of the problem. Let Ω be an open, bounded subset of $\mathbb{R}^n$, $n \leq 3$, with a smooth boundary Γ. We consider the following optimal control problem:

$$(\mathcal{P}) \qquad \min \quad J(y,u) = \frac{1}{2}\int_\Omega (y - z_d)^2 \, dx + \frac{\alpha}{2}\int_\Gamma (u - u_d)^2 \, d\sigma,$$

$$Ay = f \text{ in } \Omega, \quad y = u \quad \text{on } \Gamma, \tag{1.1}$$

$$\Lambda_1 y \in K, \ u \in U, \tag{1.2}$$

where

- $f,\ z_d \in L^2(\Omega),\ u,\ u_d \in L^2(\Gamma)$, and either $\alpha > 0$ or U_{ad} is bounded in $L^2(\Gamma)$;
- A is an elliptic operator defined by

$$\begin{cases} Ay = -\displaystyle\sum_{i,j=1}^{n} \partial_{x_i}(a_{ij}(x)\partial_{x_j} y) + a_0(x)y \ \text{ with} \\ a_{ij}, a_0 \in \mathcal{C}^2(\bar{\Omega}) \text{ for } i,j = 1,\dots,n, \ \inf\ \{a_0(x) \mid x \in \bar{\Omega}\} > 0, \\ \displaystyle\sum_{ij=1}^{n} a_{ij}(x)\xi_i\xi_j \geq \delta \sum_{i=1}^{n} \xi_i^2 \ \forall x \in \bar{\Omega}, \forall \xi \in \mathbb{R}^n, \delta > 0; \end{cases} \tag{1.3}$$

- L is a Hilbert space (with dual L' identified with L) and $\Lambda_1 \in \mathcal{L}(W, L)$ (W is defined just below).
- K and U are nonempty, closed, convex subsets of L and $L^2(\Gamma)$, respectively.

*Received by the editors October 18, 1995; accepted for publication (in revised form) June 13, 1996. This research was supported in part by EEC HCM contract CHRX-CT94-0471.
http://www.siam.org/journals/sicon/35-5/29330.html

†UMR-CNRS 6628, Université d'Orléans, U.F.R. Sciences, B.P. 6759, F-45067 Orléans Cedex 2, France (maitine@univ-orleans.fr).

‡Fachbereich Mathematik, Technische Universität Berlin, Strasse des 17 Juni 136, D-10623 Berlin, Germany (Kunisch@math.tu-berlin.de). The research of this author was supported in part by the Christian Doppler Laboratory for Parameter Identification and Inverse Problems.

System (1.1) is well-posed: for every $(u, f) \in L^2(\Gamma) \times L^2(\Omega)$ there exists a unique solution $y = \mathcal{T}(u, f)$ in W, where

$$W = \{\ y \in L^2(\Omega) \mid Ay \in L^2(\Omega)\ ,\ y_{|\Gamma} \in L^2(\Gamma)\}.$$

Moreover $\mathcal{T}$ is continuous from $L^2(\Gamma) \times L^2(\Omega)$ to W, when W is endowed with the graph norm:

$$|y|_W^2 = |y|_\Omega^2 + |Ay|_\Omega^2 + |y_{|\Gamma}|_\Gamma^2.$$

For more details, one may refer to Lions and Magenes [7, Vol. 1, Chap. 2]. From now on, when H is a Hilbert space, we denote by $(\ ,\ )_H$ (resp., $(\ ,\ )_\Omega$ and $(\ ,\ )_\Gamma$) the H (resp., $L^2(\Omega)$ and $L^2(\Gamma)$) inner products and by $|\ |_H$, $|\ |_\Omega$, $|\ |_\Gamma$ the H, $L^2(\Omega)$, and $L^2(\Gamma)$ norms, respectively.

We assume that the feasible domain

$$\mathcal{D}\ =\ \{\ (y,u) \in W \times L^2(\Gamma) \mid Ay\ =\ f\ \text{ in } \Omega,\ y = u\ \text{ on } \Gamma,\ (\Lambda_1 y, u) \in K \times U\ \}$$

is nonempty. It is easy to see that problem $(\mathcal{P})$ has a unique solution $(\bar{y}, \bar{u})$ since the functional J is strictly convex and coercive and $\mathcal{D}$ is convex, closed, and nonempty. Our main purpose is to retrieve optimality conditions for such a problem, with a new "penalization" method and to use them as a basis for numerical algorithms. Indeed, this has been done via a penalization of the state equation only in Bergounioux [1], where the existence of Lagrange multipliers for the state equation has been proved under appropriate qualification conditions. Here we use a different point of view, since we also use a penalization of the nonsmooth constraints $\Lambda_1 y \in K,\ u \in U$ with an augmented Lagrangian method as in Ito and Kunisch [6]. Optimality systems have been derived by other authors before. We mention, for instance, the work of Bonnans and Casas [3, 4] and the references given in [1]. In contrast to our work, these contributions are not based on augmented Lagrangian formulations, and they do not analyze algorithmic aspects. Similarly, the algorithm we present in section 4 is based on the augmentation of both the state equation and *the state and control constraints*. The main contribution of this research is the elimination of these latter constraints from the set of explicit constraints by augmentation. Commonly augmented Lagrangian algorithms are based on the augmentation of the state equation only. This is the case for instance for all the methods described in [5].

2. Augmented Lagrangian formulation. In this section we use the framework of [6]. We denote

$$X = W \times L^2(\Gamma),\quad H = L \times L^2(\Gamma),\quad D = K \times U.$$

Let Λ be defined from X to H by $\Lambda(y,u) = (\Lambda_1 y, u)$ and φ be the characteristic function of the convex set D. Then, following [6], we define (for any $c > 0$) the function $\varphi_c : H \times H \to \mathbb{R}$ by

$$\varphi_c(x,\lambda) = \inf_{\xi \in H} \left\{\ \varphi(x-\xi) + (\lambda,\xi)_H + \frac{c}{2}|\xi|_H^2\ \right\}, \tag{2.1}$$

where $x = (y,u)$.

Here $(\ ,\ )_H$ is given by $(\lambda,\xi)_H = (\lambda_1,\xi_1)_\Omega + (\lambda_2,\xi_2)_\Gamma$, with $\lambda = (\lambda_1,\lambda_2)$ and $\xi = (\xi_1,\xi_2)$.

Let us give some properties of the function φ_c. (For the proof we refer to [6].)

PROPOSITION 2.1. *For all $x \in H,\ \lambda \in H$, the infimum in* (2.1) *is attained at a unique point $\xi_c(x,\lambda)$. φ_c is convex and Lipschitz-continuously Fréchet-differentiable in x, and*

$$\varphi_c'(x,\lambda) = \lambda + c\ \xi_c(x,\lambda)\ . \tag{2.2}$$

Moreover $\lim_{c\to+\infty} \varphi_c(x,\lambda) = \varphi(x)$. □

Now we compute φ_c for our case.

PROPOSITION 2.2. *For all $x = (y,u) \in H$ and $\lambda = (\lambda_1,\lambda_2) \in H$*

$$\begin{aligned}\varphi_c(x,\lambda) &= \frac{c}{2}\left|x - P_D\left(x+\frac{\lambda}{c}\right)\right|_H^2 + \left(\lambda, x - P_D\left(x+\frac{\lambda}{c}\right)\right)_H \\ &= \frac{c}{2}\left|y - P_K\left(y+\frac{\lambda_1}{c}\right)\right|_L^2 + \left(\lambda_1, y - P_K\left(y+\frac{\lambda_1}{c}\right)\right)_L \\ &\quad + \frac{c}{2}\left|u - P_U\left(u+\frac{\lambda_2}{c}\right)\right|_\Gamma^2 + \left(\lambda_2, u - P_U\left(u+\frac{\lambda_2}{c}\right)\right)_\Gamma,\end{aligned} \tag{2.3}$$

$$\varphi_c'(x,\lambda) = c\left(y + \frac{\lambda_1}{c} - P_K\left(y+\frac{\lambda_1}{c}\right), u + \frac{\lambda_2}{c} - P_U\left(u+\frac{\lambda_2}{c}\right)\right), \tag{2.4}$$

where P_K (resp., P_U, P_D) is the L (resp., $L^2(\Gamma)$, H) projection on K (resp., on U, D).

Proof. Setting $\tilde\xi = x - \xi$ in (2.1) we obtain the equivalent representation

$$\varphi_c(x,\lambda) = \inf_{\tilde\xi\in H}\left\{\ \varphi(\tilde\xi) + \left(\lambda, x-\tilde\xi\right)_H + \frac{c}{2}\ |x-\tilde\xi|_H^2\ \right\}.$$

As $(\lambda, x-\tilde\xi)_H + \frac{c}{2}\ |x-\tilde\xi|_H^2 = \frac{c}{2}\ |\tilde\xi - (x+\frac{\lambda}{c})|_H^2 - \frac{1}{2c}|\lambda|_H^2$ and $\varphi(\tilde\xi) = +\infty$, if $\tilde\xi \notin D$, $\varphi(\tilde\xi) = 0$ else, it follows that

$$\varphi_c(x,\lambda) = \frac{c}{2}\left[\inf_{\tilde\xi\in D}\left|\tilde\xi - \left(x+\frac{\lambda}{c}\right)\right|_H^2\right] - \frac{1}{2c}|\lambda|_H^2. \tag{2.5}$$

The infimum is attained at $\tilde\xi = P_D(x+\frac{\lambda}{c})$. We define ξ_c so that $\tilde\xi = x - \xi_c$; that is,

$$\xi_c = \left[y - P_K\left(y+\frac{\lambda_1}{c}\right), u - P_U\left(u+\frac{\lambda_2}{c}\right)\right]$$

and

$$\begin{aligned}\varphi_c(x,\lambda) &= \frac{c}{2}\ \left|P_D\left(x+\frac{\lambda}{c}\right) - \left(x+\frac{\lambda}{c}\right)\right|_H^2 - \frac{1}{2c}\ |\lambda|_H^2 \\ &= \frac{c}{2}\ \left|x - P_D\left(x+\frac{\lambda}{c}\right)\right|_H^2 + \left(\lambda, x - P_D\left(x+\frac{\lambda}{c}\right)\right)_H.\end{aligned}$$

Now we compute φ_c' with formula (2.2) of Proposition 2.1, $\varphi_c'(x,\lambda) = \lambda + c\ \xi_c(x,\lambda)$, and the desired result follows. □

Next we consider the "augmented" problem:

$$(\mathcal{P}_{c,\lambda}) \qquad \min\ \ \{\ F_{c,\lambda}(y,u)\ |\ Ay = f \text{ in } \Omega\ , y = u \text{ on } \Gamma\ \},$$

where $F_{c,\lambda}(y,u) = J(y,u) + \varphi_c(\Lambda(y,u),\lambda)$ is the augmented Lagrangian function of $(\mathcal{P})$ associated with the constraint $\Lambda(y,u) \in D$. We have first an asymptotic result.

THEOREM 2.1. *For all $\lambda \in H$ and $c > 0$, problem $(\mathcal{P}_{c,\lambda})$ has a unique solution $(y_{c,\lambda}, u_{c,\lambda})$.*

Moreover for every fixed $\lambda \in H$

$$\lim_{c\to+\infty} y_{c,\lambda} = \bar{y} \text{ strongly in } W \text{ and } \lim_{c\to+\infty} u_{c,\lambda} = \bar{u} \text{ strongly in } L^2(\Gamma).$$

Proof. Let $\lambda \in H$ be fixed. For convenience, we shall omit the subscript λ and write x_c for $x_{c,\lambda}$. Existence and uniqueness of a solution (y_c,u_c) to $(\mathcal{P}_{c,\lambda})$ follows easily since the feasible domain is nonempty, closed, and convex and $F_{c,\lambda}$ is strictly convex and coercive. We set $x_c = (\Lambda_1 y_c, u_c) \in H$. To prove convergence of (y_c,u_c) to the solution $(\bar{y},\bar{u})$ of $(\mathcal{P})$ we first argue that $\{(y_c,u_c)\}_{c\geq c_0}$ is bounded, where $c_0 > 0$ is arbitrary and fixed. Since $(\bar{y},\bar{u})$ is feasible for $(\mathcal{P}_{c,\lambda})$, we have

$$F_{c,\lambda}(y_c,u_c) \leq F_{c,\lambda}(\bar{y},\bar{u}) \text{ for all } c > 0.$$

Observe from the definition of φ_c in (2.1) that $\varphi_c(\Lambda(\bar{y},\bar{u}),\lambda) = 0$ for all c. Hence, using (2.5), we obtain

$$J(y_c,u_c) - \frac{1}{2c}|\lambda|_H^2 \leq J(y_c,u_c) + \varphi_c(\Lambda(y_c,u_c),\lambda) \leq J(\bar{y},\bar{u}) \quad \text{for all } c > 0. \tag{2.6}$$

It follows that $\{(y_c,u_c)\}_{c\geq c_0}$ is bounded in $L^2(\Omega)\times L^2(\Gamma)$. Since $Ay_c = f$ for all $c > 0$, the set $\{(y_c,u_c)\}_{c\geq c_0}$ is bounded in X as well. Hence there exists $(\tilde{y},\tilde{u}) \in X$ such that a subsequence of $\{(y_c,u_c)\}_{c>0}$, denoted by the same symbol, converges weakly in X to $(\tilde{y},\tilde{u})$. Wellposedness of (1.1) in W implies that $A\tilde{y} = f$ in Ω and $\tilde{y} = \tilde{u}$ on Γ. Due to Proposition 2.2 and (2.6)

$$\left|x_c - P_D\left(x_c + \frac{\lambda}{c}\right)\right|_H^2 + \frac{2}{c}\left(\lambda, x_c - P_D\left(x_c + \frac{\lambda}{c}\right)\right)_H \leq \frac{1}{c}J(\bar{y},\bar{u})$$

and consequently

$$\left|x_c + \frac{\lambda}{c} - P_D\left(x_c + \frac{\lambda}{c}\right)\right|_H^2 \leq \frac{2J(\bar{y},\bar{u})}{c} + \frac{|\lambda|^2}{c^2} \quad \text{for all } c > 0. \tag{2.7}$$

Thus $(x_c + \frac{\lambda}{c} - P_D(x_c + \frac{\lambda}{c}))$ converges strongly to 0 in H. As (y_c,u_c) converges weakly to $(\tilde{y},\tilde{u})$ in X and Λ is linear continuous, $(x_c + \frac{\lambda}{c})$ converges weakly to $\tilde{x} = (\Lambda_1\tilde{y},\tilde{u})$ in H. This yields that $P_D(x_c + \frac{\lambda}{c})$ converges weakly to $\tilde{x}$ as well. Since D is closed in H and convex, it is also weakly closed and $\tilde{x} \in D = K \times U$. Thus $(\tilde{y},\tilde{u})$ is a feasible pair for $(\mathcal{P})$.

Let us prove the strong convergence of (y_c,u_c) to $(\tilde{y},\tilde{u})$ in X. First we note that due to Proposition 2.1 $\lim_{c\to+\infty}\varphi_c(\tilde{x},\lambda) = \varphi(\tilde{x}) = 0$. As $(\tilde{y},\tilde{u})$ is a feasible pair for $(\mathcal{P})$, it is also a feasible pair for $(\mathcal{P}_{c,\lambda})$ for any $c > 0$ and we have

$$J(y_c,u_c) + \varphi_c(\Lambda(y_c,u_c),\lambda) \leq J(\tilde{y},\tilde{u}) + \varphi_c(\Lambda(\tilde{y},\tilde{u}),\lambda) \quad \text{for all } c > 0.$$

Relation (2.3) implies that $\varphi_c(\Lambda(y_c,u_c),\lambda) \geq (\lambda, x_c - P_D(x_c + \frac{\lambda}{c}))_H$, and consequently

$$J(y_c,u_c) + \left(\lambda, x_c - P_D\left(x_c + \frac{\lambda}{c}\right)\right)_H \leq J(\tilde{y},\tilde{u}) + \varphi_c(\Lambda(\tilde{y},\tilde{u}),\lambda) \text{ for all } c > 0. \tag{2.8}$$

We take the limits inferior in this relation. With the strong convergence of $(x_c - P_D(x_c + \frac{\lambda}{c}))$ to 0 in H, we obtain

$$0 \leq J(\tilde{y},\tilde{u}) \leq \liminf_{c\to+\infty} J(y_c,u_c) \leq J(\tilde{y},\tilde{u}) + \lim_{c\to+\infty} \varphi_c(\Lambda(\tilde{y},\tilde{u}),\lambda) = J(\tilde{y},\tilde{u}).$$

Finally

$$\lim_{c\to+\infty} J(y_c,u_c) = J(\tilde{y},\tilde{u}). \tag{2.9}$$

This implies that (y_c,u_c) converges strongly to $(\tilde{y},\tilde{u})$ in $L^2(\Omega)\times L^2(\Gamma)$. Moreover $Ay_c = f = A\tilde{y}$, and therefore (y_c,u_c) converges to $(\tilde{y},\tilde{u})$ strongly in X.

It remains to prove that $(\tilde{y},\tilde{u}) = (\bar{y},\bar{u})$. We use relation (2.8) with $(\bar{y},\bar{u})$ as a feasible pair for $(\mathcal{P})$ instead of $(\tilde{y},\tilde{u})$ and obtain

$$J(y_c,u_c) + \left(\lambda, x_c - P_D\left(x_c + \frac{\lambda}{c}\right)\right)_H \leq J(\bar{y},\bar{u}) + \varphi_c(\Lambda(\bar{y},\bar{u}),\lambda) \quad \text{for all } c>0.$$

Taking the limit as c tends to $+\infty$ we have $0 \leq J(\tilde{y},\tilde{u}) \leq J(\bar{y},\bar{u})(\leq J(\tilde{y},\tilde{u}))$. As $(\bar{y},\bar{u})$ is the unique solution of $(\mathcal{P})$ we get the result. □

The following section will be devoted to deriving optimality conditions. We first consider the augmented problem $\mathcal{P}_{c,\lambda}$, and we shall then pass to the limit with respect to c.

3. Optimality conditions.

3.1. Penalized optimality conditions. We first write the necessary optimality conditions for problem $(\mathcal{P}_{c,\lambda})$. This problem can be expressed as

$$\min \{ F_{c,\lambda}(y,u) \mid e(y,u) = 0\},$$

where e is defined by

$$\begin{array}{rccc} e: & W\times L^2(\Gamma) & \to & L^2(\Omega)\times L^2(\Gamma), \\ & (y,u) & \mapsto & (Ay-f, y_{|\Gamma}-u). \end{array}$$

As the Fréchet derivative $e'(y_c,u_c)$ of e at (y_c,u_c) given by

$$\begin{array}{rccc} e'(y_c,u_c): & W\times L^2(\Gamma) & \to & L^2(\Omega)\times L^2(\Gamma), \\ & (y,u) & \mapsto & (Ay, y_{|\Gamma}-u) \end{array}$$

is surjective, we may apply the general theory of Lagrange multipliers: There exist $q_c \in L^2(\Omega)$ and $r_c \in L^2(\Gamma)$ such that the (generalized) Lagrange functional

$$\mathcal{L}_{c,\lambda}(y,u,q,r) = J(y,u) + \varphi_c(\Lambda(y,u),\lambda) + (q,Ay-f)_\Omega + (r,y-u)_\Gamma \tag{3.1}$$

satisfies the optimality condition

$$\nabla_{(y,u)}\mathcal{L}_{c,\lambda}(y_c,u_c,q_c,r_c) = 0. \tag{3.2}$$

Let us detail the above relation: we may decouple and obtain

$$(y_c - z_d, y)_\Omega + (q_c, Ay)_\Omega + (r_c,y)_\Gamma + (\mu_{1,c},\Lambda_1 y)_L = 0 \quad \text{for all } y\in W,$$

$$\alpha\ (u_c-u_d,u)_\Gamma - (r_c,u)_\Gamma + (\mu_{2,c},u)_\Gamma = 0 \quad \text{for all } u\in L^2(\Gamma),$$

where $\mu_{1,c} = \nabla_y \varphi_c(\Lambda(y_c, u_c), \lambda) = c[\Lambda_1 y_c + \frac{\lambda_1}{c} - P_K(\Lambda_1 y_c + \frac{\lambda_1}{c})] \in L$ and $\mu_{2,c} = \nabla_u \varphi_c(\Lambda(y_c, u_c), \lambda) = c[u_c + \frac{\lambda_2}{c} - P_U(u_c + \frac{\lambda_2}{c})] \in L^2(\Gamma)$.

We summarize these calculations in the following theorem.

THEOREM 3.1. *Let λ be fixed in H and (y_c, u_c) be the optimal solution of $(\mathcal{P}_{c,\lambda})$. Then there exist $(\mu_{1,c}, \mu_{2,c}) \in H$ and $(q_c, r_c) \in L^2(\Omega) \times L^2(\Gamma)$ such that*

$$Ay_c = f \text{ in } \Omega, \quad y_c = u_c \text{ on } \Gamma,$$

$$(y_c - z_d, y)_\Omega + (q_c, Ay)_\Omega + (r_c, y)_\Gamma + (\mu_{1,c}, \Lambda_1 y)_L = 0 \quad \text{for all } y \in W, \tag{3.3}$$

$$\alpha\ (u_c - u_d) - r_c + \mu_{2,c} = 0, \tag{3.4}$$

where

$$\begin{aligned} \mu_{1,c} &= c\ \left[\Lambda_1 y_c + \frac{\lambda_1}{c} - P_K\left(\Lambda_1 y_c + \frac{\lambda_1}{c}\right)\right] \in L, \\ \mu_{2,c} &= c\ \left[u_c + \frac{\lambda_2}{c} - P_U\left(u_c + \frac{\lambda_2}{c}\right)\right] \in L^2(\Gamma). \end{aligned} \tag{3.5}$$

3.2. Passage to the limit. The approximate optimality systems of Theorem 3.1 were obtained without assumption beyond those that are required to ascertain existence of a solution to $(\mathcal{P})$. To obtain an optimality system for $(\mathcal{P})$ itself we pass to the limit as c tends to $+\infty$ in (3.3)–(3.5). This requires a priori estimates for q_c and r_c which depend upon qualification conditions.

More precisely, let $V = V_1 \times V_2$ be a dense separable Banach subspace of $L^2(\Omega) \times L^2(\Gamma)$. We introduce the following assumption:

$(\mathcal{H})$ *There exists a bounded (in $L^2(\Omega) \times L^2(\Gamma)$)* subset $\mathcal{M}$ of X *such that* $\Lambda(\mathcal{M}) \subset K \times U$ *and* $0 \in \mathrm{Int}_V(\mathcal{V}(\mathcal{M}))$, *where* Int_V *denotes the interior with respect to the V-topology and* $\mathcal{V}(y, u) = (Ay - f, y|_\Gamma - u)$.

We note that $(\mathcal{H})$ is equivalent to the following:

There exists an $L^2(\Omega) \times L^2(\Gamma)$-bounded subset $\mathcal{M} \subset X$ and $\rho > 0$ such that for all $\xi = (\xi_1, \xi_2) \in B_V(0, 1)$, there exists $(y_\xi, u_\xi) \in \mathcal{M}$ satisfying $(\Lambda_1 y_\xi, u_\xi) \in K \times U$ and $Ay_\xi = f - \rho\xi_1$ in Ω, $y_\xi = u_\xi - \rho\xi_2$ on Γ.

Here $B_V(0, 1)$ denotes the unit ball in V. For $f \equiv 0$ and $V = L^2(\Omega) \times L^2(\Gamma)$ condition $(\mathcal{H})$ is satisfied, for example, if $0 \in \mathrm{int}_{L \times L^2(\Gamma)}(K \times U)$.

Under this hypothesis we can pass to the limit in the previous optimality conditions to obtain the main result of this section.

THEOREM 3.2. *Let $(\bar{y}, \bar{u})$ be the optimal solution of $(\mathcal{P})$, and assume that $(\mathcal{H})$ holds. Then there exist $(\bar{q}, \bar{r}) \in V_1' \times V_2'$ and $(\bar{\mu}_1, \bar{\mu}_2) \in L_1' \times V_2'$ such that*

$$\begin{aligned} &A\bar{y} = f \text{ in } \Omega, \quad \bar{y} = \bar{u} \text{ on } \Gamma, \\ &(\bar{y} - z_d, y)_\Omega + \langle \bar{q}, Ay \rangle_{V_1', V_1} \\ &+ \langle \bar{r}, y \rangle_{V_2', V_2} + \langle \bar{\mu}_1, \Lambda_1 y \rangle_{L_1', L_1} = 0 \quad \text{for all } y \in W_{1,2}, \end{aligned} \tag{3.6}$$

$$\alpha\ (\bar u - u_d) - \bar r + \bar\mu_2 = 0 \ \textit{in}\ V_2', \tag{3.7}$$

$$\begin{aligned} &\langle \bar\mu_1, \Lambda_1(y-\bar y)\rangle_{L_1',L_1} \le 0 \quad \textit{for all}\ y \in \{\ \bar y + W_{1,2}\}\ \textit{such that}\ \Lambda_1 y \in K,\\ &\langle \bar\mu_2, u-\bar u\rangle_{V_2',V_2} \le 0 \quad \textit{for all}\ u \in U \cap \{\bar u + V_2\}, \end{aligned} \tag{3.8}$$

where $W_{1,2} = \{\ y \in L^2(\Omega) \mid Ay \in V_1,\ y|_\Gamma \in V_2\ \}$ *endowed with the norm*

$$|y|^2_{W_{1,2}} = |y|^2_\Omega + |Ay|^2_{V_1} + |y_{|\Gamma}|^2_{V_2},$$

$L_1 = \Lambda_1(W_{1,2})$ *endowed with the graph norm, and* $\langle\ \cdot\ ,\ \cdot\ \rangle_{V',V}$ *denotes the duality product between* V *and* V'.

Proof. Throughout the proof we assume that $(\lambda_1,\lambda_2) \in L_1 \times V_2$. We first remark that (3.5) implies

$$\begin{aligned} &(\mu_{1,c}, z - \Lambda_1 y_c)_L + \frac{|\mu_{1,c}|^2_L}{c} \le \frac{1}{c}\,(\mu_{1,c},\lambda_1)_L \quad \text{for all } z \in K,\\ &(\mu_{2,c}, u - u_c)_\Gamma + \frac{|\mu_{2,c}|^2}{c} \le \frac{1}{c}\,(\mu_{2,c},\lambda_2)_\Gamma \quad \text{for all } u \in U. \end{aligned} \tag{3.9}$$

We just prove the first inequality (the second one may be proved quite similarly).

The projection $P_K(\Lambda_1 y_c + \frac{\lambda_1}{c})$ is characterized by

$$\left(z - P_K\left(\Lambda_1 y_c + \frac{\lambda_1}{c}\right), \Lambda_1 y_c + \frac{\lambda_1}{c} - P_K\left(\Lambda_1 y_c + \frac{\lambda_1}{c}\right)\right)_L \le 0 \quad \text{for all } z \in K,$$

and with (3.5) this yields

$$(\mu_{1,c}, z - \Lambda_1 y_c)_L + \frac{|\mu_{1,c}|^2}{c} \le \frac{1}{c}\,(\mu_{1,c},\lambda_1)_L \quad \text{for all } z \in K.$$

Thus the first inequality in (3.9) is verified. We next note that (3.5) may be written as

$$\frac{\mu_c}{c} = x_c + \frac{\lambda}{c} - P_D\left(x_c + \frac{\lambda}{c}\right),$$

where $x_c = \Lambda(y_c,u_c)$. We have seen in the proof of Theorem 2.1 that $x_c - P_D(x_c + \frac{\lambda}{c})$ converges strongly to 0 in H. Therefore $\frac{\mu_c}{c}$ converges strongly to 0 in H as well, and there exists $c_o > 0$ and M such that

$$\left(\frac{\mu_{1,c}}{c},\lambda_1\right)_L + \left(\frac{\mu_{2,c}}{c},\lambda_2\right)_\Gamma \le M \quad \text{for all } c \ge c_0. \tag{3.10}$$

Now we may obtain estimates on q_c and r_c. Let ξ be in $B_V(0,1)$ and (y_ξ,u_ξ) be the associated pair given by $(\mathcal{H})$. We add relations (3.3) and (3.4) used with the pair $(y_\xi - y_c, u_\xi - u_c)$ to obtain

$$\begin{aligned} &(y_c - z_d, y_\xi - y_c)_\Omega + (q_c, A(y_\xi - y_c))_\Omega + (r_c, y_\xi - y_c)_\Gamma + (\mu_{1,c}, \Lambda_1(y_\xi - y_c))_L\\ &\quad +\alpha\ (u_c - u_d, u_\xi - u_c)_\Gamma + (\mu_{2,c}, u_\xi - u_c)_\Gamma - (r_c, u_\xi - u_c)_\Gamma = 0 \end{aligned}$$

and consequently

$$\begin{aligned} (q_c,\rho\xi_1)_\Omega + (r_c,\rho\xi_2)_\Gamma \ &= \ (y_c - z_d, y_\xi - y_c)_\Omega + \alpha\ (u_c - u_d, u_\xi - u_c)_\Gamma\\ &\quad + (\mu_{1,c}, \Lambda_1(y_\xi - y_c))_L + (\mu_{2,c}, u_\xi - u_c)_\Gamma\,. \end{aligned} \tag{3.11}$$

Furthermore, relations (3.9) and (3.10) imply that

$$(\mu_{1,c}, \Lambda_1(y_\xi - y_c))_L + (\mu_{2,c}, u_\xi - u_c)_\Gamma \le \left(\frac{\mu_{1,c}}{c}, \lambda_1\right)_L + \left(\frac{\mu_{2,c}}{c}, \lambda_2\right)_\Gamma \le M \text{ for } c \ge c_0.$$

The convergence properties of Theorem 2.1 and the boundedness assumption on $\mathcal{M}$ in $L^2(\Omega) \times L^2(\Gamma)$ imply that $(y_c - z_d, y_\xi - y_c)_\Omega + \alpha\ (u_c - u_d, u_\xi - u_c)_\Gamma$ is uniformly bounded with respect to $c \ge c_0$. So we obtain with (3.11) the existence of $k > 0$ such that

$$\langle q_c, \xi_1\rangle_{V_1', V_1} + \langle r_c, \xi_2\rangle_{V_2', V_2} \le \frac{k}{\rho} \quad \text{for all } \xi = (\xi_1, \xi_2) \in B_V(0,1) \quad \text{and } c \ge c_0.$$

Therefore $\{q_c\}_{c \ge c_0}$ is bounded in V_1', and a subsequence, again denoted by q_c, converges weakly * to some $\bar{q}$ in V_1'. Similarly r_c is bounded in V_2' and converges weakly * to some $\bar{r}$ in V_2'. As we have chosen $V_1 \subset L^2(\Omega)$ and $V_2 \subset L^2(\Gamma)$ we may apply to (3.3) "smooth" test functions in $W_{1,2}$. Let us consider the Gelfand triple

$$L_1 \subset \overline{\Lambda_1(W_{1,2})} \subset L_1', \tag{3.12}$$

where $\overline{\Lambda_1(W_{1,2})}$ is considered as a subset of L and $\overline{\Lambda_1(W_{1,2})}$ denotes the closure of $\Lambda_1(W_{1,2})$ in L. Further let $\mu_{1,c}^P$ denote the projection of $\mu_{1,c}$ in L onto $\overline{\Lambda_1(W_{1,2})}$. It follows that

$$\left\langle \mu_{1,c}^P, \Lambda_1 y\right\rangle_{L_1', L_1} = -(y_c - z_d, y)_\Omega - \langle q_c, Ay\rangle_{V_1', V_1} - \langle r_c, y\rangle_{V_2', V_2} \quad \text{for all } y \in W_{1,2}.$$

It follows that, $\mu_{1,c}^P$ is bounded in L_1'. Moreover the separability of V_1 and V_2 implies the separability of L_1 (see Lemma 3.1 below). So a subsequence of $\mu_{1,c}^P$ converges weakly * to $\bar{\mu}_1$ in L_1'. Taking the limit in the above equality gives

$$(\bar{y} - z_d, y)_\Omega + \langle \bar{\mu}_1, \Lambda_1 y\rangle_{L_1', L_1} + \langle \bar{q}, Ay\rangle_{V_1', V_1} + \langle \bar{r}, y\rangle_{V_2', V_2} = 0 \quad \text{for all } y \in W_{1,2}.$$

Similarly $\mu_{2,c} = r_c - \alpha(u_c - u_d)$ converges weakly to $\bar{\mu}_2 = \bar{r} - \alpha(\bar{u} - u_d)$ in V_2'. Thus (3.6) and (3.7) are verified. It remains to show (3.8). Let $y \in \bar{y} + W_{1,2}$ and $u \in \bar{u} + V_2$ be such that $\Lambda_1 y \in K$ and $u \in U$. Then we add (3.3) with $y - y_c = (y - \bar{y}) + (\bar{y} - y_c) \in W$, and the relation that results from taking the inner product of (3.4) with $u - u_c \in L^2(\Gamma)$:

$$\begin{aligned}(y_c - z_d, y - y_c)_\Omega + \alpha\,(u_c - u_d, u - u_c)_\Gamma + (q_c, A(y - y_c))_\Omega \\ +\ (r_c, (y - u) - (y_c - u_c))_\Gamma = -(\mu_{1,c}, \Lambda_1(y - y_c))_L - (\mu_{2,c}, u - u_c)_\Gamma.\end{aligned}$$

As $Ay_c = f = A\bar{y}$ in Ω and $y_c = u_c$ on Γ, we get

$$\begin{aligned}(y_c - z_d, y - \bar{y})_\Omega + \alpha\ (u_c - u_d, u - \bar{u})_\Gamma + (q_c, A(y - \bar{y}))_\Omega \\ +\ (r_c, y - u)_\Gamma = -(\mu_{1,c}, \Lambda_1(y - y_c))_L - (\mu_{2,c}, u - u_c)_\Gamma \\ -\ (y_c - z_d, \bar{y} - y_c)_\Omega - \alpha\ (u_c - u_d, \bar{u} - u_c)_\Gamma.\end{aligned} \tag{3.13}$$

Moreover relation (3.9) implies

$$\begin{aligned}-(\mu_{1,c}, \Lambda_1(y - y_c))_L \ge -\frac{1}{c}(\mu_{1,c}, \lambda_1)_\Omega \quad \text{for all } y \in W \text{ such that } \Lambda_1 y \in K, \\ -(\mu_{2,c}, u - u_c)_\Gamma \ge -\frac{1}{c}(\mu_{2,c}, \lambda_2)_\Gamma \quad \text{for all } u \in U.\end{aligned} \tag{3.14}$$

Thus (3.13) becomes

$$
\begin{aligned}
&(y_c - z_d, y - \bar{y})_\Omega + \alpha\ (u_c - u_d, u - \bar{u})_\Gamma \\
&\quad + \ (q_c, A(y - \bar{y}))_\Omega + (r_c, y - u)_\Gamma \\
&\geq \ -\frac{1}{c}\left[(\mu_{1,c}, \lambda_1)_\Omega + (\mu_{2,c}, \lambda_2)_\Gamma\right] \\
&- \ (y_c - z_d, \bar{y} - y_c)_\Omega - \alpha\ (u_c - u_d, \bar{u} - u_c)_\Gamma\,.
\end{aligned} \tag{3.15}
$$

Let us denote by σ_c the term on the right-hand side of (3.15). Since by assumption $(\lambda_1, \lambda_2) \in L_1 \times V_2$, it is easy to see that $\lim_{c\to+\infty} \sigma_c = 0$.

Next we set successively $u = \bar{u}$ and $y = \bar{y}$. First, we choose $u = \bar{u}$ so that inequality (3.15) becomes

$$(y_c - z_d, y - \bar{y})_\Omega + (q_c, A(y - \bar{y}))_\Omega + (r_c, y - \bar{u})_\Gamma \geq \sigma_c,$$

and consequently $(y_c - z_d, y - \bar{y})_\Omega + \langle q_c, A(y - \bar{y})\rangle_{V_1',V_1} + \langle r_c, y - \bar{y}\rangle_{V_2',V_2} \geq \sigma_c$. We may now pass to the limit in the previous expression to get

$$(\bar{y} - z_d, y - \bar{y})_\Omega + \langle \bar{q}, A(y - \bar{y})\rangle_{V_1',V_1} + \langle \bar{r}, y - \bar{y}\rangle_{V_2',V_2} \geq 0.$$

With (3.6) we finally have

$$\langle \bar{\mu}_1, y - \bar{y}\rangle_{W_{1,2}',W_{1,2}} \leq 0 \quad \text{for all } y \in \{\ \bar{y} + W_{1,2}\} \text{ such that } \Lambda_1 y \in K.$$

Now we choose $y = \bar{y}$ and inequality (3.15) gives

$$
\begin{aligned}
\alpha\ (u_c - u_d, u - \bar{u})_\Gamma + (r_c, \bar{y} - u)_\Gamma &\geq \sigma_c, \\
\alpha\ (u_c - u_d, u - \bar{u})_\Gamma - \langle r_c, u - \bar{u}\rangle_{V_2',V_2} &\geq \sigma_c.
\end{aligned}
$$

Once again, we may pass to the limit, and we obtain

$$\alpha\ (\bar{u} - u_d, u - \bar{u})_\Gamma - \langle \bar{r}, u - \bar{u}\rangle_{V_2',V_2} \geq 0.$$

Together with (3.7) this implies the second inequality in (3.8), and the proof is finished as soon as the following lemma is proved.

LEMMA 3.1. *L_1 is separable.*

Proof. As $L_1 = \Lambda_1(W_{1,2})$ with Λ_1 continuous, it is sufficient to prove that $W_{1,2}$ is separable. Let D_1 (resp., D_2) be dense countable subsets of V_1 (resp., V_2). Then the subset $D = \{y \in L^2(\Omega) \mid Ay \in D_1,\ y_{|\Gamma} \in D_2\}$ is a countable subset of $W_{1,2}$ (since $\mathcal{T}$ defined in section 1 is a bijection from $D_1 \times D_2$ onto D). Moreover, the linear operator $\mathcal{T}$ is continuous from $V_1 \times V_2$ to $W_{1,2}$. We may therefore assert that D is dense because of the properties of V_i and the continuity of $\mathcal{T}$. ∎

REMARK 3.1. *Let us still denote by Λ_1 the restriction of Λ_1 to $W_{1,2}$ (i.e., from $W_{1,2}$ to L_1). Then the adjoint operator Λ_1^* is defined from L_1' to $W_{1,2}'$ and satisfies*

$$\langle \mu, \Lambda_1 y\rangle_{L_1',L_1} = \langle \Lambda_1^* \mu, y\rangle_{W_{1,2}',W_{1,2}} \ \textit{for all } (\mu, y) \in L_1' \times W_{1,2}.$$

Then relation (3.6) *and the first part of relation* (3.8) *may be written as*

$$(\bar{y} - z_d, y)_\Omega + \langle \bar{q}, Ay\rangle_{V_1',V_1} + \langle \bar{r}, y\rangle_{V_2',V_2} + \langle \bar{\nu}_1, y\rangle_{W_{1,2}',W_{1,2}} = 0, \ \textit{for all } y \in W_{1,2}$$

$$\textit{and} \quad \langle \bar{\nu}_1, y - \bar{y}\rangle_{W_{1,2}',W_{1,2}} \leq 0 \quad \textit{for all } y \in \{\ \bar{y} + W_{1,2}\} \textit{ such that } \Lambda_1 y \in K,$$

where $\bar{\nu}_1 = \Lambda_1^* \bar{\mu}_1 \in W_{1,2}'$.

3.3. Example. To illustrate the previous abstract result, we give an example for a particular choice of spaces V and L. Here we set $V_1 = L^2(\Omega)$ and $V_2 = L^2(\Gamma)$ so that $V_1' = L^2(\Omega)$, $V_2' = L^2(\Gamma)$, $W_{1,2} = W$.

The previous theorem gives the following optimality system:

$$\begin{gathered}
\bar q \in L^2(\Omega),\ \bar r \in L^2(\Gamma), \quad \text{and } \bar\mu_1 \in L_1',\ \bar\mu_2 \in L^2(\Gamma),\\
A\bar y = f \quad \text{in } \Omega, \quad \bar y = \bar u \quad \text{on } \Gamma,\\
(\bar y - z_d, y)_\Omega + (\bar q, Ay)_\Omega + (\bar r, y)_\Gamma + \langle \bar\mu_1, \Lambda_1 y\rangle_{L_1', L_1} = 0 \quad \text{for all } y \in W,\\
\alpha\ (\bar u - u_d) = \bar r - \bar\mu_2 \in L^2(\Gamma),\\
\langle \bar\mu_1, \Lambda_1(y - \bar y)\rangle_{L_1', L_1} \le 0 \quad \text{for all } y \in W \text{ such that } \Lambda_1 y \in K,\\
(\bar\mu_2, u - \bar u)_\Gamma \le 0 \quad \text{for all } u \in U.
\end{gathered} \tag{3.16}$$

If in addition L is finite dimensional, we may identify the spaces L_1, $\overline{\Lambda_1(W)}$, and L_1' of the Gelfand triple in (3.12). In this very case the optimality system becomes

$$\bar q \in L^2(\Omega),\ \bar r \in L^2(\Gamma), \text{ and } \bar\mu_1 \in \Lambda_1(W),\ \bar\mu_2 \in L^2(\Gamma), \tag{3.17}$$

$$A\bar y = f \quad \text{in } \Omega,\ \bar y = \bar u \quad \text{on } \Gamma, \tag{3.18}$$

$$(\bar y - z_d, y)_\Omega + (\bar q, Ay)_\Omega + (\bar r, y)_\Gamma + (\bar\mu_1, \Lambda_1 y)_L = 0 \quad \text{for all } \ y \in W, \tag{3.19}$$

$$\alpha\ (\bar u - u_d) = \bar r - \bar\mu_2 \in L^2(\Gamma), \tag{3.20}$$

$$(\bar\mu_1, \Lambda_1(y - \bar y))_L \le 0 \quad \text{for all } y \text{ such that } \Lambda_1 y \in K, \tag{3.21}$$

$$(\bar\mu_2, u - \bar u)_\Gamma \le 0 \quad \text{for all } u \in U. \tag{3.22}$$

As a specific example, L can be chosen as the set of linear finite elements with respect to a triangulation of Ω and Λ_1: $W \to L$ can be the L^2-projection. (H^1-projection is not admitted since the elements of W are not in general H^1-smooth.)

REMARK 3.2. *Let us still consider the case with $V_1 = L^2(\Omega)$, $V_2 = L^2(\Gamma)$, and L finite dimensional and assume that $\Lambda_1^T(L) \subset L^2(\Omega)$, where $\Lambda_1^T : L \to W'$ denotes the transpose of Λ_1. In this case we may introduce $\bar p \in H^2(\Omega) \cap H_o^1(\Omega)$ as the solution of*

$$A^*\bar p = -(\bar y - z_d + \Lambda_1^T \bar\mu_1) \text{ in } \Omega,\ \bar p = 0 \ \text{ on } \ \Gamma, \tag{3.23}$$

where A^* is the adjoint operator of A. Then with Green's formula relation (3.19) *becomes*

$$(\bar q - \bar p, Ay)_\Omega + \left(\bar r - \frac{\partial \bar p}{\partial \nu_{A^*}}, y\right)_\Gamma = 0 \quad \text{for all } y \in W. \tag{3.24}$$

For all $z \in L^2(\Omega)$ there exists $y \in H^2(\Omega) \cap H_o^1(\Omega) \subset W$ such that $Ay = z$ in Ω. So (3.24) *implies $(\bar q - \bar p, z)_\Omega = 0$ for all $z \in L^2(\Omega)$, that is $\bar q = \bar p$. Then* (3.24) *gives*

$$\bar r = \frac{\partial \bar p}{\partial \nu_{A^*}}.$$

Thus we see that relations (3.19), (3.20) *are equivalent to*

$$\begin{gathered}
A^*\bar p + \bar y - z_d + \Lambda_1^T \bar\mu_1 = 0, \quad \bar p \in H^2(\Omega) \cap H_o^1(\Omega),\\
\alpha\ (\bar u - u_d) - \frac{\partial \bar p}{\partial \nu_{A^*}} + \bar\mu_2 = 0.
\end{gathered} \tag{3.25}$$

A specific case in which Λ_1 satisfies the assumption $\Lambda_1^T(L) \subset L^2(\Omega)$ is given if L is a finite-dimensional subspace of $L^2(\Omega)$ and Λ_1 is the L^2-orthogonal projection onto L.

4. Lagrangian algorithms. In this section we turn to the numerical realization of the constrained optimal control problem $(\mathcal{P})$. We shall combine the techniques from [1] and [6] augmenting the state equation *as well as the constraints* characterizing the feasible set D to obtain well-performing algorithms.

First we recall an augmented Lagrangian algorithm based on the penalization of the state equation (see [5], [1], and the references therein).

ALGORITHM $\mathcal{A}_o$.

- Step 1. Initialization: Set $n = 0$, and choose $\gamma > 0$, $q_o \in L^2(\Omega)$, $r_o \in L^2(\Gamma)$.
- Step 2. Compute

$$(y_n, u_n) = \text{ Arg min } \{ L_\gamma(y, v, q_n, r_n) \mid \Lambda(y, u) \in K \times U\},$$

where

$$L_\gamma(y, u, q, r) = J(y, u) + (q, Ay)_\Omega + (r, y - u)_\Gamma + \frac{\gamma}{2}|Ay - f|^2_\Omega + \frac{\gamma}{2}|y - u|^2_\Gamma$$

is the augmented Lagrangian with respect to the *state equation* constraint.

- Step 3. Set

$q_{n+1} = q_n + \rho_1\ (Ay_n - f), \quad$ where $\rho_1 \in (0, 2\gamma]$,
$r_{n+1} = r_n + \rho_2\ (y_{n|\Gamma} - u_n), \quad$ where $\rho_2 \in (0, 2\gamma]$.

The analysis of this algorithm is rather standard; see [1, 5] and the references there.

THEOREM 4.1. *Let $(\bar y, \bar u)$ be the solution to $(\mathcal{P})$, and suppose that $(\mathcal{H})$ holds with $V = L^2(\Omega) \times L^2(\Gamma)$. Then the iterates of Algorithm $\mathcal{A}_o$ satisfy*

$$|y_n - \bar y|^2_\Omega + \alpha|u_n - \bar u|^2_\Gamma + \frac{1}{2\rho_1}|q_{n+1} - \bar q|^2_\Omega + \frac{1}{2\rho_2}|r_{n+1} - \bar r|^2_\Gamma$$

$$+\ (\gamma - \frac{\rho_1}{2})|Ay_n - f|^2_\Omega + (\gamma - \frac{\rho_2}{2})|y_n - u_n|^2_\Gamma \le \frac{1}{2\rho_1}|q_n - \bar q|^2_\Omega + \frac{1}{2\rho_2}|r_n - \bar r|^2_\Gamma \tag{4.1}$$

for all $n = 0, 1, 2, \ldots$. With ρ_1 and ρ_2 given as in Step 3, this implies

$$\sum_{n=0}^{\infty} |y_n - \bar y|^2_\Omega + \alpha \sum_{n=0}^{\infty} |u_n - \bar u|^2_\Gamma + \Big(\gamma - \frac{\rho_1}{2}\Big) \sum_{n=0}^{\infty} |Ay_n - f|^2_\Omega$$

$$+\ \Big(\gamma - \frac{\rho_2}{2}\Big) \sum_{n=0}^{\infty} |y_n - u_n|^2_\Gamma \le \frac{1}{2\rho_1}|q_0 - \bar q|^2_\Omega + \frac{1}{2\rho_2}|r_0 - \bar r|^2_\Gamma \tag{4.2}$$

and in particular strong convergence of $(y_n, u_n) \to (\bar y, \bar u)$ in $L^2(\Omega) \times L^2(\Gamma)$ and boundedness of $\{(q_n, r_n)\}$. If, moreover, $\rho_1 < 2\gamma$ and $\rho_2 < 2\gamma$, then $(y_n, u_n) \to (\bar y, \bar u)$ in X, and every weak limit $(\tilde q, \tilde r)$ of (q_n, r_n) has the property that $(\bar y, \bar u, \tilde q, \tilde r)$ satisfies, for all $\Lambda(y, u) \in K \times U$,

$$\Big(J'(\bar y, \bar u), (y, u) - (\bar y, \bar u)\Big)_{\Omega\times\Gamma} + \Big(\bar q, A(y - \bar y)\Big)_\Omega + \Big(\bar r, y - \bar y - (u - \bar u)\Big)_\Gamma \ge 0$$

Proof. We refer to [2]. □

The main remaining problem is the resolution of the auxiliary problem of Step 2 in Algorithm $\mathcal{A}_o$, which can be written as

$$(y_n, u_n) = \text{Arg min } \{ L_\gamma(y, u) \mid \Lambda(y, u) \in D\}.$$

To simplify the notation we omit to indicate the dependence of L_γ on q and r. During Step 2 these functions are fixed. We are going to use the following algorithm and a splitting variant to solve the auxiliary problem.

ALGORITHM $\mathcal{A}_1$.

- Step 1. Initialization: Choose $\lambda^o \in H$ and $c > 0$.
- Step 2. Compute

$$(y^j, u^j) = \text{Arg min } \{ \ L_\gamma(y,u) + \varphi_c(\Lambda(y,u), \lambda^j) \mid \Lambda(y,u) \in X\},$$

  where φ_c has been defined in the previous section.
- Step 3. Set (see 2.4)

$$\begin{aligned}\lambda^{j+1} &= \varphi_c'(\Lambda(y^j,u^j),\lambda^j)\\ &= c\left(\Lambda_1 y^j + \frac{\lambda_1^j}{c} - P_K\left(\Lambda_1 y^j + \frac{\lambda_1^j}{c}\right), u^j + \frac{\lambda_2^j}{c} - P_U\left(u^j + \frac{\lambda_2^j}{c}\right)\right).\end{aligned}$$

The convergence of this algorithm under the assumption that L is finite dimensional follows from the result in [6]. The assumption on finite dimensionality of L entails that the duality pairing between L_1 and L_1' in (3.16) can be replaced by the inner product in L (see (3.19)), which is necessary for the convergence proof. We now write the version in which Algorithm $\mathcal{A}_1$ appears as an inner loop in algorithm $\mathcal{A}_o$.

ALGORITHM $\mathcal{A}$.

- Step 1. Initialization: Set $n = 0$, and choose $\gamma > 0,\ c > 0$.
  Choose $(q_o, r_o) \in L^2(\Omega) \times L^2(\Gamma)$ and $\lambda_o = (\lambda_{o1}, \lambda_{o2}) \in L \times L^2(\Gamma)$.
- Step 2. Choose $k_n \in \mathbb{N}$, set $\lambda_n^o = \lambda_n$, and for $j = 0, \dots, k_n$

$$(y_n^j, u_n^j) = \ \text{Arg min } \{L_\gamma(y,u,q_n,r_n) + \varphi_c(\Lambda(y,u),\lambda_n^j) \mid (y,u) \in W \times L^2(\Gamma)\},$$

$$\lambda_n^{j+1} = (\lambda_{n,1}^{j+1}, \lambda_{n,2}^{j+1}) \quad \text{with} \quad \begin{cases} \lambda_{n,1}^{j+1} = c\left[\Lambda_1 y_n^j + \dfrac{\lambda_{n,1}^j}{c} - P_K\left(\Lambda_1 y_n^j + \dfrac{\lambda_{n,1}^j}{c}\right)\right], \\ \lambda_{n,2}^{j+1} = c\left[u_n^j + \dfrac{\lambda_{n,2}^j}{c} - P_U\left(u_n^j + \dfrac{\lambda_{n,2}^j}{c}\right)\right]. \end{cases}$$

  End of the inner loop: $\lambda_{n+1} = \lambda_n^{k_n+1},\ y_n = y_n^{k_n},\ u_n = u_n^{k_n}$.
- Step 3. $q_{n+1} = q_n + \frac{\rho_1}{k_n+1}(\sum_{j=0}^{k_n} A y_n^j - f)$, where $\rho_1 \in (0, 2\gamma]$,
  $r_{n+1} = r_n + \frac{\rho_2}{k_n+1}(\sum_{j=0}^{k_n}(y_n^j|_\Gamma - u_n^j))$, $\rho_2 \in (0, 2\gamma]$.

THEOREM 4.2. *Let $(\bar y, \bar u)$ be the solution to $(\mathcal{P})$, and suppose that $(\mathcal{H})$ holds with $V = L^2(\Omega) \times L^2(\Gamma)$ and that L is finite dimensional. Let $(\bar q, \bar r, \bar\mu) \in L^2(\Omega) \times L^2(\Gamma) \times L \times L^2(\Gamma)$ be an associated Lagrange multiplier. Then the iterates of Algorithm $\mathcal{A}$ satisfy*

$$\begin{aligned}&|y_n - \bar y|_\Omega^2 + \alpha|u_n - \bar u|_\Gamma^2 + \frac{k_n+1}{2\rho_1}|q_{n+1} - \bar q|_\Omega^2 + \frac{k_n+1}{2\rho_2}|r_{n+1} - \bar r|_\Gamma^2 \\ &+ \left(\gamma - \frac{\rho_1}{2}\right)|Ay_n - f|_\Omega^2 + \left(\gamma - \frac{\rho_2}{2}\right)|u_n - y_n|_\Gamma^2 + \frac{1}{2c}|\lambda_{n+1} - \bar\mu|_{L\times L^2(\Gamma)}^2 \\ &\le \frac{k_n+1}{2\rho_1}|q_n - \bar q|_\Omega^2 + \frac{k_n+1}{2\rho_2}|r_n - \bar r|_\Gamma^2 + \frac{1}{2c}|\lambda_n - \bar\mu|_{L\times L^2(\Gamma)}^2\end{aligned} \tag{4.3}$$

for all $n = 0, 1, 2, \ldots$. *If* k_n *is nonincreasing, this implies*

$$\begin{aligned} &\sum_{n=0}^{\infty} |y_n - \bar{y}|_\Omega^2 + \alpha \sum_{n=0}^{\infty} |u_n - \bar{u}|_\Gamma^2 \\ &+ \left(\gamma - \frac{\rho_1}{2}\right) \sum_{n=0}^{\infty} |Ay_n - f|_\Omega^2 + \left(\gamma - \frac{\rho_2}{2}\right) \sum_{n=0}^{\infty} |u_n - y_n|_\Gamma^2 \\ &\leq \frac{k_0+1}{2\rho_1} |q_0 - \bar{q}|_\Omega^2 + \frac{k_0+1}{2\rho_2} |r_0 - \bar{r}|_\Gamma^2 + \frac{1}{2c} |\lambda_0 - \bar{\mu}|_{L\times L^2(\Gamma)}^2 \end{aligned} \tag{4.4}$$

and in particular strong convergence of $(y_n, u_n) \to (\bar{y}, \bar{u})$ *in* $L^2(\Omega) \times L^2(\Gamma)$ *and boundedness of* $\{(q_n, r_n, \lambda_n)\}$. *If, moreover,* $\rho_1 < 2\gamma$ *and* $\rho_2 < 2\gamma$, *then* $(y_n, u_n) \to (\bar{y}, \bar{u})$ *in* X *and every weak limit* $(\tilde{q}, \tilde{r}, \tilde{\lambda})$ *of* $\{(q_n, r_n, \lambda_n)\}$ *has the property that* $(\bar{y}, \bar{u}, \tilde{q}, \tilde{r}, \tilde{\lambda})$ *satisfies* (3.19), (3.20).

Proof. See [2]. □

REMARK 4.1. *The resolution of the unconstrained minimization problem occurring in algorithm* $\mathcal{A}$ *is equivalent to the resolution of*

$$\nabla_{(y,u)} L_\gamma(y_n, u_n, q_n, r_n) + \varphi_c'(\Lambda(y_n, u_n), \lambda_n^j) = 0;$$

that is,

$$\begin{aligned} &\nabla_y \; L_\gamma(y_n, u_n, q_n, r_n) + c \left[\Lambda_1 y_n + \frac{\lambda_{n,1}^j}{c} - P_K\left(\Lambda_1 y_n + \frac{\lambda_{n,1}^j}{c}\right)\right] = 0, \\ &\nabla_u \; L_\gamma(y_n, u_n, q_n, r_n) + c \left[u_n + \frac{\lambda_{n,2}^j}{c} - P_U\left(u_n + \frac{\lambda_{n,2}^j}{c}\right)\right] = 0. \end{aligned} \tag{4.5}$$

This can be done with a Newton or a descent method for instance.

Our final goal is the analysis of Gauss–Seidel splitting techniques to solve the auxiliary problems. The splitting avoids the minimization of the auxiliary problem with respect to y and u simultaneously. The resulting algorithm is as follows.

ALGORITHM $\mathcal{A}_o^{GS}$.

- Step 1. Initialization: Set $n = 0$; choose $\gamma > 0$, $q_o \in L^2(\Omega)$, $r_o \in L^2(\Gamma)$, $u_{-1} \in U$.
- Step 2.

$$\begin{aligned} y_n &= \text{ Arg min } \{ L_\gamma(y, u_{n-1}, q_n, r_n) \mid \Lambda_1 y \in K\}, \\ u_n &= \text{ Arg min } \{L_\gamma(y_n, u, q_n, r_n) \mid u \in U\}. \end{aligned}$$

- Step 3.

$$\begin{aligned} q_{n+1} &= q_n + \rho_1 \, (Ay_n - f), \quad \text{where } \rho_1 \in (0, 2\gamma], \\ r_{n+1} &= r_n + \rho_2 \, (y_{n|\Gamma} - u_n), \quad \text{where } \rho_2 \in (0, 2\gamma]. \end{aligned}$$

Once again, we may use Algorithm $\mathcal{A}_1$ to solve the first subproblem of Step 2. The second one is easily solved directly; see Remark 4.2 below. For convenience we shall henceforth delete the index 1 in the notation of the state component of the multiplier.

ALGORITHM $\mathcal{A}^{GS}$.

- Step 1. Initialization: Set $n = 0$ and choose $\gamma > 0$, $c > 0$.
  Choose $(q_o, r_o) \in L \times L^2(\Gamma)$, $\lambda_o \in L^2(\Omega)$ and $u_{-1} \in L^2(\Gamma)$.
- Step 2. Choose $k_n \in \mathbb{N}$; set $\lambda_n^o = \lambda_n, u_n^{-1} = u_{n-1}$, and for $j = 0, \ldots, k_n$

$$y_n^j = \text{ Arg min } \{L_\gamma(y, u_n^{j-1}, q_n, r_n) + \varphi_c(\Lambda(y, u_n^{j-1}), (\lambda_n^j, 0)) \mid y \in W\},$$

$$\lambda_n^{j+1} = c\left[\Lambda_1 y_n^j + \frac{\lambda_n^j}{c} - P_K\left(\Lambda_1 y_n^j + \frac{\lambda_n^j}{c}\right)\right],$$

$$u_n^j = \text{ Arg min } \{\ L_\gamma(y_n^j, u, q_n, r_n) \mid u \in U\}.$$

End of the inner loop: $\lambda_{n+1} = \lambda_n^{k_n+1},\ y_n = y_n^{k_n}\ , u_n = u_n^{k_n}$.

- Step 3.

$$q_{n+1} = q_n + \frac{\rho_1}{k_n+1}\sum_{j=0}^{k_n}(Ay_n^j - f), \qquad \text{where } \rho_1 \in (0, 2\gamma],$$

$$r_{n+1} = r_n + \frac{\rho_2}{k_n+1}\sum_{j=0}^{k_n}(y_{n|\Gamma}^j - u_n^j), \qquad \text{where } \rho_2 \in (0, \gamma].$$

REMARK 4.2. *We may solve the first unconstrained minimization problem occurring in the previous algorithm, as was mentioned in Remark* 4.1. *The second minimization problem is indeed equivalent to*

$$u_n^j = \text{ Arg min } \left\{ \left| u - \frac{\alpha u_d + r_n + \gamma y_n^j}{\alpha+\gamma}\right|_\Gamma \ : \ u \in U \right\};$$

that is, u_n^j *is the* $L^2(\Gamma)$*-projection of* $\frac{\alpha u_d + r_n + \gamma y_n^j}{\alpha+\gamma}$ *on* U.

We end this section with a convergence analysis for Algorithm $\mathcal{A}^{GS}$.

THEOREM 4.3. *Let* $(\bar y, \bar u)$ *be the solution to* $(\mathcal{P})$*, and suppose that* $(\mathcal{H})$ *holds with* $V = L^2(\Omega) \times L^2(\Gamma)$ *and that* L *is finite dimensional. Let* $(\bar q, \bar r, \bar\mu) \in L^2(\Omega) \times L^2(\Gamma) \times L \times L^2(\Gamma)$ *be a Lagrange multiplier associated with the state equation and the state constraint.*

Then the iterates (y_n, u_n, q_n, r_n) *of Algorithm* $\mathcal{A}^{GS}$ *satisfy*

$$\begin{aligned} &|y_n - \bar y|_\Omega^2 + \left(\alpha + \frac{\gamma}{2}\right)|u_n - \bar u|_\Gamma^2 + \frac{k_n+1}{2\rho_1}|q_{n+1} - \bar q|_\Omega^2 + \frac{k_n+1}{2\rho_2}|r_{n+1} - \bar r|_\Gamma^2 \\ &\quad + \left(\gamma - \frac{\rho_1}{2}\right)|Ay_n - f|_\Omega^2 + \frac{\gamma - \rho_2}{2}|u_n - y_n|_\Gamma^2 + \frac{1}{2c}|\lambda_{n+1} - \bar\mu|_L^2 \\ &\qquad \le \frac{k_n+1}{2\rho_1}|q_n - \bar q|_\Omega^2 + \frac{k_n+1}{2\rho_2}|r_n - \bar r|_\Gamma^2 \\ &\quad + \frac{1}{2c}|\lambda_n - \bar\mu|_L^2 + \frac{\gamma - \rho_2}{2}|u_{n-1} - y_{n-1}|_\Gamma^2 + \frac{\gamma}{2}|u_{n-1} - \bar u|_\Gamma^2 \end{aligned} \tag{4.6}$$

for all $n = 1, 2, \ldots$. *If* k_n *is nonincreasing, this implies*

$$\begin{aligned} &\sum_{n=1}^{\infty}\left(|y_n - \bar y|_\Omega^2 + \alpha|u_n - \bar u|_\Gamma^2 + \left(\gamma - \frac{\rho_1}{2}\right)|Ay_n - f|_\Omega^2 + \frac{\gamma}{2}|u_n - y_n|_\Gamma^2\right) \\ \le\ &\frac{k_1+1}{2\rho_1}|q_1 - \bar q|_\Omega^2 + \frac{k_1+1}{2\rho_2}|r_1 - \bar r|_\Gamma^2 + \frac{1}{2c}|\lambda_1 - \bar\mu|_L^2 + \frac{\gamma - \rho_2}{2}|y_o - u_o|_\Gamma^2 + \frac{\gamma}{2}|u_o - \bar u|_\Gamma^2. \end{aligned}$$

Proof. We use the optimality conditions issued from Step 2 of Algorithm $\mathcal{A}^{GS}$. The iterates (y_n^j, u_n^j) of Step 2 satisfy, for $j = 0, \ldots, k_n$ and for all $y \in W$,

$$
\begin{aligned}
&\big(J_y'(y_n^j,u_n^{j-1}),y)\big)_\Omega + \left(q_n+\frac{\rho_1}{k_n+1}(Ay_n^j-f),Ay\right)_\Omega\\
&+\left(\gamma-\frac{\rho_1}{k_n+1}\right)(Ay_n^j-f,Ay)_\Omega+\left(r_n+\frac{\rho_2}{k_n+1}(y_n^j-u_n^{j-1}),y\right)_\Gamma\\
&+\left(\gamma-\frac{\rho_2}{k_n+1}\right)(y_n^j-u_n^{j-1},y)_\Gamma+\big(\varphi_{c,1}'(\Lambda_1 y_n^j,\lambda_n^j),\Lambda_1 y)\big)_L=0,
\end{aligned}
\tag{4.7}
$$

and for all $u\in U_{ad}$

$$
\begin{aligned}
&\Big(J_u'(y_n^j,u_n^j),u-u_n^j\Big)_\Gamma-\left(r_n+\frac{\rho_2}{k_n+1}(y_n^j-u_n^j),u-u_n^j\right)_\Gamma\\
&-\left(\gamma-\frac{\rho_2}{k_n+1}\right)(y_n^j-u_n^j,u-u_n^j)_\Gamma\ge 0.
\end{aligned}
\tag{4.8}
$$

From (3.19) and (3.20) it follows that

$$
\begin{aligned}
&\Big(J'(\bar y,\bar u),(y,u-\bar u)\Big)_{\Omega\times\Gamma}\\
&+(\bar q,Ay)_\Omega+(\bar r,y-(u-\bar u))_\Gamma+\Big(\bar\mu,\Lambda_1 y\Big)_{L\times L^2(\Gamma)}\ge 0
\end{aligned}
\tag{4.9}
$$

for all $(y,u)\in W\times U_{ad}$. From [6] it is known that

$$
\begin{aligned}
&\Big(\varphi_c'(\Lambda(y_n^j,u_n^j),\lambda_n^j)-\varphi_c'(\Lambda(\bar y,\bar u),\bar\mu),\Lambda(y_n^j,u_n^j)-\Lambda(\bar y,\bar u)\Big)\\
&\qquad\ge\frac{1}{2c}|\lambda_n^{j+1}-\bar\mu|^2-\frac{1}{2c}|\lambda_n^j-\bar\mu|^2
\end{aligned}
\tag{4.10}
$$

for $j=0,1,\dots,k_n$. Combining (4.7)–(4.9) and (4.10) and setting

$$
q_n^j=q_n+\frac{\rho_1}{k_n+1}\sum_{i=0}^{j}(Ay_n^i-f)\ \text{and}\ r_n^j=r_n+\frac{\rho_2}{k_n+1}\sum_{i=0}^{j}(y_n^i-u_n^i)
$$

for $j=0,\dots,k_n$ and $q_n^{-1}:=q_n,\ r_n^{-1}:=r_n$ imply

$$
\begin{aligned}
&|y_n^j-\bar y|_\Omega^2+\alpha|u_n^j-\bar u|_\Gamma^2+\frac{k_n+1}{2\rho_1}|q_n^j-\bar q|_\Omega^2-\frac{k_n+1}{2\rho_1}|q_n^{j-1}-\bar q|_\Omega^2\\
&+\left(\gamma-\frac{\rho_1}{2(k_n+1)}\right)|Ay_n^j-f|_\Omega^2+\frac{k_n+1}{2\rho_2}|r_n^j-\bar r|_\Gamma^2-\frac{k_n+1}{2\rho_2}|r_n^{j-1}-\bar r|_\Gamma^2\\
&-\frac{\rho_2}{k_n+1}\sum_{i=0}^{j-1}(y_n^i-u_n^i,y_n^j-u_n^j)_\Gamma-\frac{\rho_1}{k_n+1}\sum_{i=0}^{j-1}(Ay_n^i-f,Ay_n^j-f)_\Omega\\
&-\left(\gamma-\frac{\rho_2}{2(k_n+1)}\right)|y_n^j-u_n^j|_\Gamma^2+\frac{1}{2c}(|\lambda_n^{j+1}-\bar\mu|_L^2-|\lambda_n^j-\bar\mu|_L^2)\\
&+\gamma\Big(u_n^j-u_n^{j-1},y_n^j-\bar y\Big)_\Gamma\le 0
\end{aligned}
$$

for $n,j=0,1,\dots,k_n$. Summing the above inequality over j and using the fact that

$$
\sum_{j=1}^{k_n}\sum_{i=0}^{j-1}(a_i,a_j)_H\le\frac{k_n}{2}\sum_{j=0}^{k_n}|a_j|^2
$$

we arrive at

$$\sum_{j=0}^{k_n}(|y_n^j-\bar y|_\Omega^2+\alpha|u_n^j-\bar u|_\Gamma^2)+\frac{k_n+1}{2\rho_1}|q_n^{k_n}-\bar q|_\Omega^2+\frac{k_n+1}{2\rho_2}|r_n^{k_n}-\bar r|_\Gamma^2$$
$$+\frac{1}{2c}|\lambda_{n+1}-\bar\mu|_L^2+\Big(\gamma-\frac{\rho_1}{2}\Big)\sum_{j=0}^{k_n}|Ay_n^j-f|_\Omega^2+\Big(\gamma-\frac{\rho_2}{2}\Big)\sum_{j=0}^{k_n}|y_n^j-u_n^j|_\Gamma^2$$
$$+\gamma\sum_{j=0}^{k_n}\Big(u_n^j-u_n^{j-1},y_n^j-\bar y\Big)_\Gamma\le\frac{k_n+1}{2\rho_1}|q_n-\bar q|_\Omega^2+\frac{k_n+1}{2\rho_2}|r_n-\bar r|_\Gamma^2+\frac{1}{2c}|\lambda_n^0-\bar\mu|_L^2.$$

The estimation of $(u_n^j-u_n^{j-1},y_n^j-\bar y)_\Gamma$ is standard (see [5]): we obtain, for $j=1,2,\dots$ and $n=0,1,\dots$

$$\gamma\Big(u_n^{j-1}-u_n^j,\bar y-y_n^j\Big)_\Gamma\ge\alpha|u_n^j-u_n^{j-1}|_\Gamma^2-\frac{\gamma}{2}(|y_n^{j-1}-u_n^{j-1}|_\Gamma^2+|u_n^{j-1}-\bar u|_\Gamma^2-|u_n^j-\bar u|_\Gamma^2).$$

A similar calculus provides the estimation of $(u_n^o-u_n^{-1},y_n^o-\bar y)_\Gamma$ for $n=1,2,\dots$:

$$\begin{aligned}\gamma\Big(u_n^{-1}-u_n^o,\bar y-y_n^o\Big)_\Gamma\ge\quad&\Big(\alpha+\frac{\rho_2}{2}\Big)|u_n^o-u_n^{-1}|_\Gamma^2+\frac{\rho_2-\gamma}{2}|y_{n-1}-u_n^{-1}|_\Gamma^2\\&+\frac{\gamma}{2}|u_n^o-\bar u|_\Gamma^2-\frac{\gamma}{2}|u_n^{-1}-\bar u|_\Gamma^2.\end{aligned}\tag{4.11}$$

We henceforth assume $n\ge1$. We obtain

$$\begin{aligned}\gamma\sum_{j=0}^{k_n}\Big(u_n^j-u_n^{j-1},y_n^j-\bar y\Big)_\Gamma\ge\Big(\alpha+\frac{\rho_2}{2}\Big)|u_n^o-u_{n-1}|_\Gamma^2+\frac{\rho_2-\gamma}{2}|y_{n-1}-u_{n-1}|_\Gamma^2\\+\frac{\gamma}{2}|u_n^o-\bar u|_\Gamma^2-\frac{\gamma}{2}|u_{n-1}-\bar u|_\Gamma^2+\alpha\sum_{j=1}^{k_n}|u_n^j-u_n^{j-1}|_\Gamma^2\\-\frac{\gamma}{2}\sum_{j=1}^{k_n}(|y_n^{j-1}-u_n^{j-1}|_\Gamma^2+|u_n^{j-1}-\bar u|_\Gamma^2-|u_n^j-\bar u|_\Gamma^2).\end{aligned}\tag{4.12}$$

We finally get for $k_n\ge1$

$$\sum_{j=0}^{k_n}(|y_n^j-\bar y|_\Omega^2+\alpha|u_n^j-\bar u|_\Gamma^2)+\frac{k_n+1}{2\rho_1}|q_{n+1}-\bar q|_\Omega^2+\frac{k_n+1}{2\rho_2}|r_{n+1}-\bar r|_\Gamma^2$$
$$+\frac{1}{2c}|\lambda_{n+1}-\bar\mu|_L^2+\Big(\gamma-\frac{\rho_1}{2}\Big)\sum_{j=0}^{k_n}|Ay_n^j-f|_\Omega^2+\frac{\gamma-\rho_2}{2}\sum_{j=0}^{k_n}|y_n^j-u_n^j|_\Gamma^2$$
$$+\frac{\gamma}{2}|u_n-\bar u|_\Gamma^2\le\frac{k_n+1}{2\rho_1}|q_n-\bar q|_\Omega^2+\frac{k_n+1}{2\rho_2}|r_n-\bar r|_\Gamma^2$$
$$+\frac{1}{2c}|\lambda_n-\bar\mu|_L^2+\frac{\gamma-\rho_2}{2}|y_{n-1}-u_{n-1}|_\Omega^2+\frac{\gamma}{2}|u_{n-1}-\bar u|_\Gamma^2.$$

Since $\rho_2\le\gamma$, we deduce that

$$|y_n-\bar y|^2_\Omega+\alpha|u_n-\bar u|^2_\Gamma+\frac{k_n+1}{2\rho_1}|q_{n+1}-\bar q|^2_\Omega+\frac{k_n+1}{2\rho_2}|r_{n+1}-\bar r|^2_\Gamma$$
$$+\left(\gamma-\frac{\rho_1}{2}\right)|Ay_n-f|^2_\Omega+\left(\gamma-\frac{\rho_2}{2}\right)|y_n-u_n|^2_\Gamma+\frac{1}{2c}|\lambda_{n+1}-\bar\mu|^2_\Omega+\frac{\gamma}{2}|u_n-\bar u|^2_\Gamma$$
$$\le\frac{k_n+1}{2\rho_1}|q_n-\bar q|^2_\Omega+\frac{k_n+1}{2\rho_2}|r_n-\bar r|^2_\Gamma+\frac{1}{2c}|\lambda_n-\bar\mu|^2_\Omega$$
$$+\frac{\gamma-\rho_2}{2}|y_{n-1}-u_{n-1}|^2_\Gamma+\frac{\gamma}{2}|u_{n-1}-\bar u|^2_\Gamma$$

if $k_n\ge 1$. Using (4.11) the same estimate follows for $k_n=0$. The final claim again follows with a telescoping argument. ∎

5. Numerical experiments.

5.1. Implementation. Numerical experiments were carried out for one- and two-dimensional problems. Since Algorithm $\mathcal{A}^{GS}$ is the simplest for implementation, we have used it for our tests. The discretization of the problem was done with finite-differences discretization schemes. The size of the grid was $\frac{1}{N}$ so that $L=\mathbb{R}^{N+1}$ for the one-dimensional case and $L=\mathbb{R}^{2(N+1)}$ for the two-dimensional case. Λ was chosen as the discretization operator with respect to the given equidistant grid.

The main difficulty that remains in applying Algorithm $\mathcal{A}^{GS}$ is given by the (unconstrained) minimization with respect to y. This was done via the adjoint state equation and results, for fixed u, q, and r, in the resolution of

$$\begin{aligned} A^*p &= y-z_d+c\left[y+\frac{\lambda}{c}-P_K\left(y+\frac{\lambda}{c}\right)\right] \quad \text{in } \Omega,\ p=0 \quad \text{on } \Gamma,\\ Ay &= f-\frac{q+p}{\gamma} \quad \text{in } \Omega,\ y=u-\frac{r}{\gamma}+\frac{1}{\gamma}\frac{\partial p}{\partial\nu_{A^*}} \quad \text{on } \Gamma, \end{aligned} \tag{5.1}$$

for p and y. Here

$$\frac{\partial p}{\partial\nu_{A^*}}$$

denotes the conormal derivative of p with respect of A^* (which is the adjoint operator of A). The coupled system (5.1) was solved via a descent algorithm combined with a relaxation method. The control function was computed using the L^∞-projection of $\frac{r+\alpha u_d+\gamma y}{\alpha+\gamma}$ on U_{ad}.

All numerical tests were carried out on a Hewlett-Packard workstation using the MATLAB package. For all examples that we report here, the required accuracy and stopping criteria were set to 10^{-6}.

5.2. Examples.

One-dimensional example. In this example we chose

- $\Omega=]0,1[$ and $N=30$; $A=-\Delta$ and $f(x)=-(x+2)\exp(x)$.
- $z_d\equiv -1,\ \alpha=0.1,\ u_d(0)=-2,\ u_d(1)=1$; $U_{ad}=[-3,3]$ and $K=\{\ Y\in L\ :\ -1.1\le Y\le 1\ \}$.

Note that z_d is quite close to the boundary of K.

In fact, as can be seen from Figure 1, the lower bound on the state is active. The active set is a singleton. In view of the fact that the influence of the boundary

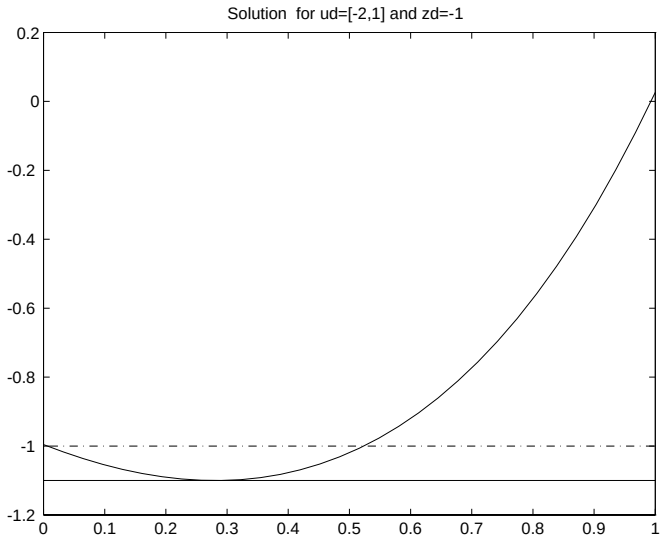


FIG. 1.

TABLE 1

c	γ	k_n (constant)	$\|\Delta y + f\|_\infty$	$\|y - v\|_\infty$	n	CPU units	min$[y-(-1.1)]$
10	1	10	4.8 e-07	4. e-07	58	1	4. e-10
10	1	1	9.3 e-07	6. e-07	154	2.17	2.5 e-06
10	1	100	2.2 e-07	5. e-07	13	1.35	-2 e-09
100	10	10	6.2 e-07	9. e-07	95	1.01	-1.3 e-11

control at $x = 0$ and $x = 1$ is restricted to the superposition of straight lines to the uncontrolled state, this is not surprising.

The numerical values for J and the control at the minimum are

$$J = 1.5862 \;\; 10^{-1} \text{ and } \bar{u}(0) = -9.9573 \;\; 10^{-1}, \quad \bar{u}(1) = 2.6314 \;\; 10^{-2}.$$

One of the main questions concerning the class of algorithms that we analyzed is the choice of the parameters ρ_i, c, and γ. From Table 1 we conclude that while the choice of the parameters certainly has an influence on the convergence properties of the algorithm, there is a wide range of parameters values for which convergence is achieved numerically, for this and other examples that we tested. In all calculations we chose $\rho_i = 1$. Some tests shows that the ratio $\frac{\gamma}{c} = \frac{1}{10}$ is a good one. For $(c, \gamma) = (1, 1)$, $(c, \gamma) = (100, 50)$, $(c, \gamma) = (1, 0.5)$ (all with $k_n = 10$ for all n), convergence is achieved, but it is slower than for those pairs that are presented in Table 1. From that table, as well as from other tests, it can also be seen that the auxiliary problem should be solved sufficiently accurately, before the Lagrange multipliers (q, r) for the state equation and the boundary condition are updated (see $k_n \equiv 1$). The values $(c, \gamma) = (10, 0.1)$ still with $\rho_i = 1$ led to divergence. This is not unexpected in view of the result of Theorem 4.3, which requires $\rho_2 \leq \gamma$.

Two-dimensional example. Now we consider

- $\Omega =]0, 1[\times]0, 1[$ and $N = 30$; $A = -\Delta$ and $f \equiv 20$.

TABLE 2

Iteration	$\|\Delta y + f\|_\infty$	$\|y - v\|_\infty$	J	min $(3.5 - y)$
0	4.688280e+00	1.223633e-02	2.414087e-01	−9.707107e-02
10	8.449125e-04	2.439992e-04	2.062097e-01	−1.030313e-02
50	2.819024e-05	3.966610e-06	2.083813e-01	−1.987233e-05
53	9.776863e-07	7.128897e-07	2.083877e-01	4.117863e-06

TABLE 3

N	10	20	30	40	50	60
$N^2 \cdot 10^{-2}$	1	4	9	16	25	36
CPU units	1	3.14	7	13	21.8	35.4

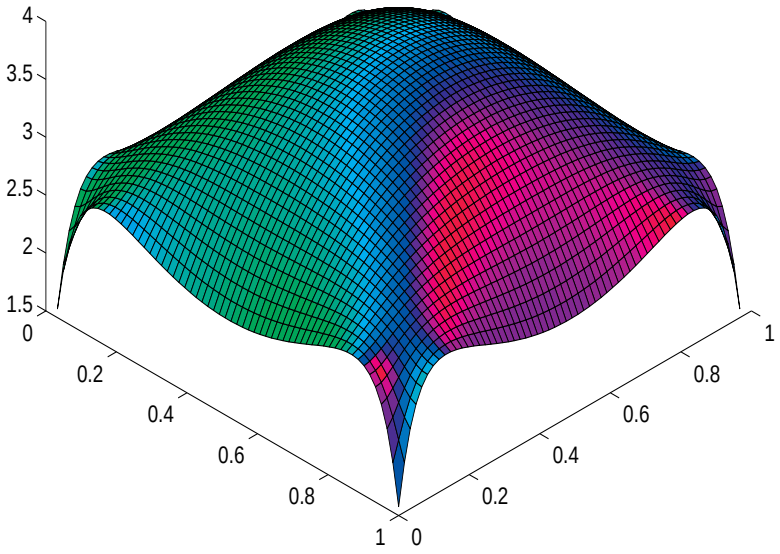


FIG. 2.

- $z_d(x_1, x_2) = 5[x_1(x_1 - 1) * x_2(x_2 - 1)] + 3,\ u_d \equiv 0$, and $\alpha = 0.01$.
- $U_{ad} = [-10, 10]$ and $K = \{\ y \in L \mid 0 \le y \le 3.5\}$.

The results for selected values during the iteration procedure are shown in Table 2. The effect of the discretization is given in Table 3: the CPU time is approximately a linear function of N^2. The optimal state and control (on one side of the domain) are given in Figures 2 and 3, respectively.

In this case the upper bound $y \le 3.5$ is active, while the lower bound $y \ge 0$ is not, except in some corners of the domain. We must admit, however, that the numerical values of y may not be accurate in the corners since no special treatment of the discontinuities of the conormals at the corners was incorporated in the code. The results were obtained with $c = 10,\ \gamma = 1$, and $k_n = 10$ for all n.

6. Conclusion. The augmented Lagrangian algorithms with splitting into state and control variables can effectively be used to solve state and control constrained optimization problems. For the first-order methods that are presented in this paper, the auxiliary problems in the inner loop must be solved sufficiently accurately before

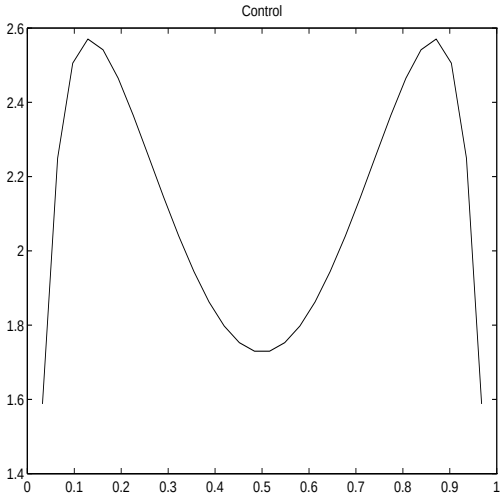


FIG. 3.

the Lagrange multipliers of the state equation and boundary condition are updated. Appropriate choices for the penalty parameters (here c and γ) and the step lengths ρ_i for the dual variables are easily determined since the algorithms are not particularly sensitive to them. It is our intention to also analyze second-order methods for the same class of problems.

REFERENCES

[1] M. BERGOUNIOUX, *On boundary state constrained control problems*, Numer. Funct. Anal. Optim., 14 (1993), pp. 515–543.

[2] M. BERGOUNIOUX AND K. KUNISCH, *Augmented Lagrangian Techniques for Elliptic State Constrained Optimal Control Problems*, Preprint 95-12, Université d'Orléans, 1995.

[3] J. F. BONNANS AND E. CASAS, *On the choice of the function space for some state constrained control problems*, Numer. Funct. Anal. Optim., 7 (1984-1985), pp. 333–348.

[4] E. CASAS, *Control of an elliptic problem with pointwise state constraints*, SIAM J. Control Optim., 24 (1986), pp. 1309–1322.

[5] M. FORTIN AND R. GLOWINSKI, *Méthodes de Lagrangien Augmenté—Applications à la Résolution de Problèmes aux Limites*, Méthodes Mathématiques pour l'Informatique, Dunod, Paris, 1982.

[6] K. ITO AND K. KUNISCH, *Augmented Lagrangian Methods for Nonsmooth Convex Optimization in Hilbert Spaces*, preprint, Berlin, 1994.

[7] J.L. LIONS AND E. MAGENES, *Problèmes aux Limites non Homogènes et Applications*, Gauthier-Villars, Paris, 1968.

SIAM J. CONTROL OPTIM.
Vol. 35, No. 5, pp. 1544–1556, September 1997

006

DETERMINING THE ACOUSTIC IMPEDANCE IN THE 1-D WAVE EQUATION VIA AN OPTIMAL CONTROL PROBLEM*

V. BARBU† AND N. H. PAVEL‡

Abstract. This paper deals with a coefficient optimal control problem for the one-dimensional (1-D) wave equation with nonhomogeneous boundary periodic inputs. A main concept is the notion of "weak solution" for the 1-D wave equation with T-periodic conditions (Definition 2.1). For $T = (2k+1)/p$ the weak solution (hyperbolic) operator $A(u)$ (see 2.20) has important properties such as closed range $R(A(u))$, nontrival null space $N(A(u))$ (resonant case)—see Propositions 2.1 and 2.2. However, the u-dependence of $R(A(u))$ and $N(A(u))$ gives rise to major difficulties. The maximum principle (Theorem 4.1) can be viewed as information (necessary conditions) on the optimal acoustic impedance function u^* (the impedance for which the corresponding seismic waves have a minimal effect given as a cost functional).

Key words. acoustic impedance, one-dimensional wave equation, periodic solutions, optimal control, singular control systems, resonant case, trace theorems, eigenvalues and eigenfunctions, Fourier series

AMS subject classifications. 49J20, 49K20, 35L05

PII. S0363012995283698

1. Introduction.

The propagation of seismic waves in a medium in the presence of a periodic seismic sourse g is described by the following wave equation:

$$u(x)y_{tt}(x,t) - (u(x)y_x(x,t))_x = 0, \quad x \in (0,1), t \in \mathbb{R}, \tag{1.1}$$

$$-u(0)y_x(0,t) = g, \quad y(1,t) = 0, \quad t \in \mathbb{R}, \tag{1.2}$$

with the T-periodic condition ($T > 0$)

$$y(x,t+T) = y(x,t), \quad y_t(x,t+T) = y_t(x,t) \quad \forall x \in [0,1], \forall t \in \mathbb{R}. \tag{1.3}$$

Here $u = (\rho\mu)^{1/2}$ is the acoustic impedance function (ρ is the density, μ is the elasticity modulus), $y(x,t)$ is the displacement at level x underground at time t, and g is a T-periodic seismic sourse at the surface $x = 0$. Throughout in this paper we assume that $u \in U$ with

$$U = \{u \in H^2(0,1); 0 < a \le u(x) < \infty; u(0) = u_0\} \tag{1.4}$$

as the set of all admissible impedances. The inverse problem in seismic consists in recovering the impedance distribution u from the observations of the displacements $y_0 = y_0(t)$ at the surface $x = 0$. The least square approach to this inverse problem leads to the following optimal control problem:

$$\begin{aligned} &\text{minimize } \int_0^T (y(0,t) - y_0(t))^2 dt + \gamma \int_0^1 ((u''(x))^2 + u^2(x))\, dx \\ &\text{subject to (1.1)–(1.3) and } u \in U, \text{ where } \gamma > 0. \end{aligned} \tag{1.5}$$

*Received by the editors May 22, 1995; accepted for publication (in revised form) June 19, 1996. http://www.siam.org/journals/sicon/35-5/28369.html

†Department of Mathematics, University of Iasi, 6600 Iasi, Romania (barbu@uaic.ro). This research was carried out while the author was visiting Ohio University.

‡Department of Mathematics, Ohio University, Athens, OH 45701 (npavel1@ohio.edu).

Due to the resonance phenomenon in (1.1)–(1.3), the solution to this problem is not unique and does not exist for all g, so (1.5) may be viewed as a singular control problem. Moreover, the dependence of the range and null space of the hyperbolic operator (1.1) on the control variable u represents a major difficulty in the study of (1.5). For this reason section 2 is entirely devoted to existence and continuous dependence of the weak solution of (1.1)–(1.3) on u. The main result is that if

$$T \in \{(2k+1)/p; k = 0, 1, \dots ; p = 1, 2, \dots\} = Q_0, \tag{1.6}$$

which is dense in Q^+ (the set of all positive rational numbers), then for (f, g) in a closed space $S(u)$ (see (2.18)) of $L^2((0,1)\times(0,T))\times L^2(0,T)$, the problem (1.1)–(1.3) (with f in the right-hand side of (1.1)) has modulo $N(u)$ (the null space of (1.1)) a unique weak solution which is continuous as a function of u. In this context, the analysis of the weak solution is related to the literature devoted to periodic solutions for 1-D hyperbolic problems (see, e.g., [6], [7]). See also [1], [3]. In sections 3 and 4, one gives existence and a maximum principle for the optimal control u^*. For equations (1.1)–(1.2) with Cauchy initial values, a related problem was previously studied by G. Chavent and coworkers (see [2], [8], and [9]). For a parabolic equation, a similar problem was recently studied by the authors in [4].

However, there is no any overlap between this paper and the previous papers mentioned above. In addition, the methods used here are completely different.

The "cost functional" in (1.5) can be viewed as "the effect" of seismic waves $y(x,t)$ corresponding to the impedance function $u(x)$ (which is determined by the density and the elasticity modulus at level x underground). The maximal principles (conditions (4.2) and (4.3)) give information on the optimal acoustic impedance u^*—the impedance for which the corresponding seismic waves have a minimal effect. Numerical determinations of u^* would be therefore important, although difficult. We will try it elsewhere.

2. The controlled periodic system. For the study of the optimal control problem we need results on the existence, regularity, and continuous dependence of the periodic solution y (of the problem below) on f, g, and u.

Precisely, in this section we are concerned with the problem

$$uy_{tt} - (uy_x)_x = f(x,t) \quad \text{in } Q = (0,1)\times(0,T), \tag{2.1}$$

$$-u(0)y_x(0,t) = g(t), \quad y(1,t) = 0, \quad t \in [0,T], \tag{2.2}$$

$$y(x,0) = y(x,T), \quad y_t(x,0) = y_t(x,T), \quad x \in [0,1], \tag{2.3}$$

where $f \in L^2(Q)$, $g \in L^2(0,T)$, and $u \in U$ (given as in (1.4)).

DEFINITION 2.1. *A function $y \in L^2(Q)$ is said to be a weak solution of* (2.1)–(2.3) *if*

$$\int_Q y(u\varphi_{tt} - (u\varphi_x)_x)\,dxdt - \int_0^T g(t)\varphi(0,t)\,dt = \int_Q f(x,t)\varphi(x,t)\,dxdt \tag{2.4}$$

for all T-periodic $\varphi \in H^2(Q)$ with $\varphi_x(0,t) = 0$, $\varphi(1,t) = 0$ $\forall t \in [0,T]$.

The following operation will be useful:

$$\begin{aligned} &A_0(u)\varphi = u\varphi_{tt} - (u\varphi_x)_x, \\ &D(A_0(u)) = \{\varphi \in H^2(Q); \varphi_x(0,t) = 0, \varphi(1,t) = 0, \\ &\qquad\qquad \varphi(x,0) = \varphi(x,T), t \in [0,T], x \in [0,1]\} \end{aligned} \tag{2.5}$$

for $u \in U$.

In order to characterize the set of all (f,g) for which there is a weak solution y, we shall use the following complete orthonormal system of eigenfunctions $\{\psi_m\varphi_n\}_{\substack{m\in\mathbb{Z}\\ n\in\mathbb{N}}}$ in $L^2(Q)$ with

$$\psi_m(t)=\frac{1}{\sqrt{T}}e^{i\mu_m t},\ m\in\mathbb{Z}\ \text{(the set of all integers)},\ \mu_m=\frac{2m\pi}{T} \tag{2.6}$$

[12, p. 88] and λ_n,φ_n given by the Sturm–Liouville problem

$$-\frac{1}{u}(u\varphi_n')_x=\lambda_n^2\varphi_n,\quad \varphi_n'(0)=0,\quad \varphi_n(1)=0,\quad n\in\mathbb{N},u\in U, \tag{2.6$'$}$$

where $\varphi_n'(x)=\frac{d}{dx}\varphi_n(x)$ and λ_n is increasingly convergent to $+\infty$. Clearly $\lambda_n=\lambda_n(u)$ and $\varphi_n=\varphi_n(u)$ depend on $u\in U$. The inner product in $L^2(0,1)$ is defined by

$$\langle F,G\rangle=\int_0^1 u(x)F(x)\bar{G}(x)\,dx,\quad u\in U. \tag{2.7}$$

Accordingly, the $L^2(Q)$-norm of the solution $y=y^u$ is

$$|y|_{L^2(Q)}=\int_Q u(x)|y(x,t)|^2dxdt=\int_0^T|y(\cdot,t)|^2_{L^2(0,1)}dt, \tag{2.8}$$

so $|\varphi_n|^2=\int_0^1 u(x)\varphi_n^2(x)\,dx=1$. For $u=1$, the eigenvalues λ_n are $(2n+1)\frac{\pi}{2}$, and the corresponding eigenfunctions are

$$\varphi_n(x)=\sqrt{2}\cos(2n+1)\frac{\pi}{2}x,\quad n\in\mathbb{N}.$$

For a general $u\in U$, it follows that $\lambda_n=\lambda_n^u$ have the form

$$\lambda_n=(2n+1)\frac{\pi}{2}+\frac{1}{n}\theta_n>0\quad\text{with } |\theta_n|=|\theta_n^u|\le M(u) \tag{2.9}$$

with M bounded on bounded subsets of U.

In order to prove (2.9), set $z_n(x)=\left(\frac{u(x)}{u(0)}\right)^{1/2}\varphi_n(x)$. Then z_n satisfies the Sturm–Liouville problem

$$\begin{aligned} &z_n''(x)+(\lambda_n^2+\eta_u(x))z_n(x)=0,\\ &-u'(0)z_n(0)+2u(0)z_n'(0)=0,\quad z_n(1)=0, \end{aligned} \tag{2.10}$$

with $\eta_u(x)=\frac{1}{2}\frac{u''}{u}-\frac{1}{4}(\frac{u'}{u})^2$. This implies [10, p. 262] that λ_n has the form indicated in (2.9) and

$$z_n(x)=\cos(2n+1)\frac{\pi}{2}x+\frac{1}{n}H_n(x) \tag{2.11}$$

with $|H_n(x)|+|H_n'(x)|\le M(u)$, $x\in[0,1]$, $n\in\mathbb{N}$. Therefore,

$$|\varphi_n(x)|\le M(u),\quad n\in\mathbb{N},x\in[0,1].$$

We now can easily characterize the null space $N(A_0(u))$ of A_0u. Precisely,

$$N(A_0(u))=\text{Span}\{\psi_m\varphi_n,\forall m\in\mathbb{Z},n\in\mathbb{N}\text{ with }\lambda_n=|\mu_m|\}. \tag{2.12}$$

Indeed, let $A_0(u)\varphi=0$ and let φ_{mn} be the Fourier coefficients of φ in $L^2(Q)$, i.e.,

$$\varphi=\sum_{m,n}\varphi_{mn}\psi_m\varphi_n,\quad \varphi_{mn}=\int_Q u\varphi\bar{\psi}_m\varphi_n dxdt. \tag{2.13}$$

Clearly, $A_0(u)\varphi = 0$ if and only if $(\lambda_n^2 - \mu_m^2)\varphi_{mn} = 0$.

Remark 2.1. For $T = \frac{2k+1}{p}$, $N(A_0(u))$ is finite dimensional. Indeed, $\lambda_n = |\mu_m|$ means $(2n+1)(2k+1) + \frac{2(2K+1)}{\pi n}\theta_n = 4|m|p$, which has at most a finite number of solutions (m,n) (if any). For example, if $\theta_n = 0$ for all n, then $N(A_0(u)) = 0$. If $u(x) = (Cx+C_1)^2$ with $C_1^2 = u_0$ and $C = \frac{u_1}{2\sqrt{u_0}}$, then one can prove that $N(A_0(u)) = 0$ for $T = 1/2$. Moreover, by (2.9) we see that dim $N(u)$ is bounded on bounded subsets of U. Note also that the maps $u \to \lambda_n(u)$ (and $u \to \varphi_n(u)$) are continuous from $H^2(0,1) \cap U$ into R (and $H^1(0,1)$). This follows by the fact that $\lambda_n(u)$ and $\varphi_n(u)$ are bounded on bounded sets of U and the eigenvalues λ_n of Sturm–Liouville problem (2.6′) are simple. As $A_0(u)$ is self-adjoint in $L^2(Q)$, we have

$$L^2(Q) = N(A_0(u)) \oplus \overline{R(A_0(u))}. \tag{2.14}$$

We will see below that actually $R(A_0(u))$ is closed. We are now in a position to study the existence and regularity of the weak solutions y as in Definition 2.1 via Fourier series. Denote by y_{mn} the Fourier coefficients of the component of y which is in $N(A_0(u))^\perp$ (the orthogonal of $N(A_0(u))$), i.e.,

$$y = \sum_{\lambda_n \neq |\mu_m|} y_{mn}\psi_m\varphi_n, \quad y_{mn} = \int_Q uy\bar{\psi}_m\varphi_n dxdt. \tag{2.15}$$

Similarly, the Fourier coefficients of $u^{-1}f$ in $L^2(Q)$ and of $g \in L^2(0,T)$ are

$$\tilde{f}_{mn} = (u^{-1}f)_{mn} = \int_Q f\bar{\psi}_m\varphi_n dxdt, \quad g_m = \int_0^T g(t)\bar{\psi}_m(t)\,dt \tag{2.16}$$

with $\sum_{\substack{m\in\mathbb{Z}\\ n\in\mathbb{N}}} |\tilde{f}_{mn}|^2 = \int_Q u(fu^{-1})^2 dx \le a^{-2}\int_Q uf^2 dxdt = a^{-2}|f|^2_{L^2(Q)}$; $|g|^2_{L^2(0,T)} = \sum_{m\in\mathbb{Z}} |g_m|^2$. For $\varphi = \bar{\psi}_m(t)\varphi_n(x)$, (2.4) implies

$$(\lambda_n^2 - \mu_m^2)y_{mn} = \tilde{f}_{mn} + \varphi_n(0)g_m. \tag{2.17}$$

Set (for $u \in U$)

$$\begin{aligned} S(u) = \{&(f,g); f \in L^2(Q), g \in L^2(0,T);\\ &\tilde{f}_{mn} + \varphi_n(0)g_m = 0 \text{ for all } (m,n) \text{ with } |\mu_m| = \lambda_n\},\\ S_0(u) = \{&f \in L^2(Q); \tilde{f}_{mn} = 0 \text{ for all } (m,n) \text{ with } \lambda_n = |\mu_m|\}. \end{aligned} \tag{2.18}$$

These depend on u as λ_n depend on u. In view of (2.17), a necessary condition for the existence of a weak solution y is $(f,g) \in S(u)$. We will see that this is also sufficient (Proposition 2.1 below). The following spaces are also needed:

(2.19)

$$\begin{aligned} &H^j_\pi(0,T) = \{g \in H^j(0,T); g^{(k)}(0) = g^{(k)}(T), k = 0,1,\dots,j-1\}, \quad j = 1,2,\\ &H^1_\pi(Q) = \{f \in H^1(Q); f(x,0) = f(x,T) \text{ a.e. } x \in (0,1)\},\\ &H^2_\pi(Q) = \{f \in H^2(Q); f(x,0) = f(x,T); f_t(x,0) = f_t(x,T), x \in (0,1)\}. \end{aligned}$$

Finally, for each $u \in U$ we define the operator $A(u) : D(A(u)) \to L^2(Q) \times L^2(0,T)$ by

$$A(u)y = (f,g), \tag{2.20}$$

where y is a weak solution corresponding to f and g in the sense of Definition 2.1. The main result of this section is given by the following proposition.

PROPOSITION 2.1. *Assume that $T = (2k+1)/p$ as in (1.6). Then for each $u \in U$, $A(u)$ is a closed operator with a closed range $R(A(u)) = S(u)$ (as in (2.18)). Moreover, $A^{-1}(u)$ is continuous and one to one from $S(u)$ onto $S_0(u) = N(A(u))^{\perp}$. The following estimates hold for $y = A^{-1}(u)(f,g)$, with $(f,g) \in R(A(u))$:*

$$|A^{-1}(u)(f,g)|_{L^2(Q)} \le C(|f|_{L^2(Q)} + |g|_{L^2(0,T)}), \tag{2.21}$$

$$|A^{-1}(u)(f,g)|_{H^1(Q)} \le C(|f|_{L^2(Q)} + |g|_{H^1(0,T)}). \tag{2.22}$$

If $f \in H^1_\pi(Q)$ and $g \in H^2_\pi(0,T)$, then $y \in H^2(Q)$, $t \to y(0,t) \in H^1(0,T)$, and

$$|A^{-1}(u)(f,g)|_{H^2(Q)} \le C(|f|_{H^1(Q)} + |g|_{H^2(0,T)}). \tag{2.22$'$}$$

Remark 2.2. Here in (2.21)–(2.22$'$), as well as throughout this paper, $C = C(u)$ (with $u \in U$) denotes several positive constants which are bounded on bounded subsets of U.

Proof of Proposition 2.1. Let $(f,g) \in R(A(u))$. In view of (2.17), the Fourier coefficients of $y = A^{-1}(u)(f,g) = \sum y_{mn}\psi_m\varphi_n$ are given by

$$y_{mn} = \frac{\tilde{f}_{mn} + \varphi_n(0)g_m}{\lambda_n^2 - \mu_m^2} \quad \text{for } \lambda_n \ne |\mu_m|. \tag{2.23}$$

A key part of the proof is the estimate

$$\inf_{\lambda_n \ne |\mu_m|} |\lambda_n - |\mu_m|| \ge C > 0. \tag{2.24}$$

Indeed, according to (2.6) and (2.9)

$$|\lambda_n - |\mu_m|| = \frac{\pi}{2T}(|4|m| - (2n+1)T - \theta_n^1|)$$

with $\theta_n^1 \to 0$ as $n \to \infty$. As $T = \frac{2k+1}{p}$, for some $p \in \mathbb{N}$ and $k = 0, 1, \dots$, we have

$$|\lambda_n - |\mu_m|| = b|4|m|p - (2n+1)(2k+1) - \theta_n^2|, \tag{2.25}$$

with $\theta_n^2 \to 0$ as $n \to \infty$, $b = \frac{\pi}{2(2k+1)}$, which yields (2.24). We also have

$$\lambda_n + |\mu_m| \ge C(n + |m|) \quad \text{for some } C > 0, \tag{2.26}$$

which follows from

$$\lambda + |\mu_m| = b(8p|m| + (2n+1)(2k+1) + \theta_n^2) \tag{2.27}$$

with $2k+1+\theta_n^2 > 0$ for sufficiently large n and $(2n+1)(2k+1)+\theta_n^2 = (2k+1)\frac{\pi}{2}\lambda_n > 0$ for all $n \in \mathbb{N}$. According to Parseval's formula, $|y|^2_{L^2(Q)} = \sum_{m,n} y^2_{mn}$, so (2.23) yields

$$|y|^2_{L^2(Q)} \le C \sum_{\lambda_n \ne |\mu_m|} \frac{|\tilde{f}_{mn}|^2 + |g_m|^2}{m^2 + n^2} \le C(|f|^2_{L^2(Q)} + |g|^2_{L^2(0,T)}), \tag{2.28}$$

which is just (2.21). If $g \in H^1_\pi(0,T)$, then the Fourier coefficient $\tilde{g}_m$ of g' is

$$\tilde{g}_m = \int_0^T g'(t)\bar{\psi}_m(t)\,dt = i\mu_m g_m, \tag{2.29}$$

so $|g'|^2_{L^2(0,T)} = \sum_m \mu_m^2 g_m^2 = \sum_m \tilde{g}_m^2$. Therefore

$$
\begin{aligned}
|y_t|^2_{L^2(Q)} &= \sum_{\lambda_n \neq |\mu_m|} \mu_m^2 y_{mn}^2 \\
&\leq C\left(\sum_{m,n} \frac{\mu_m^2 |\tilde{f}_{mn}|^2}{m^2+n^2} + \sum_{m,n} \frac{|\tilde{g}_m|^2}{m^2+n^2}\right) \leq C(|f|^2_{L^2(Q)} + |g'|^2_{L^2(0,T)}).
\end{aligned} \tag{2.30}
$$

In order to estimate y_x we need to recall that $|\varphi_n|^2_{L^2(0,1)} = \int_0^1 u(x)\varphi_n^2(x)\,dx = 1$ and that the distributional derivative y_x is given by

$$
y_x = \sum_{\lambda_n \neq |\mu_m|} y_{mn} \psi_m \varphi_n'. \tag{2.31}
$$

The system $\{\varphi_n'\}$ is orthogonal in $L^2(0,1)$ and

$$
|\varphi_n'|_{L^2(0,1)} = \int_0^1 u(x)(\varphi_n')^2 dx = -\int_0^1 \varphi_n (u(x)\varphi_n')_x dx = \lambda_n^2. \tag{2.31$'$}
$$

Therefore, we have

$$
\begin{aligned}
|y_x|^2_{L^2(Q)} &= \sum_{\lambda_n \neq \mu_m} \lambda_n^2 |y_{mn}|^2 \\
&\leq C \sum_{\lambda_n \neq \mu_m} \frac{\lambda_n^2 |\tilde{f}_{mn}|^2}{\lambda_n^2 + \mu_m^2} + C \sum_{\lambda_n \neq \mu_m} \frac{\lambda_n^2 |g_m|^2}{(\lambda_n - |\mu_m|)^2(\lambda_n + |\mu_m|)^2} \\
&= C(I_1 + I_2).
\end{aligned} \tag{2.32}
$$

Clearly, $I_1 \leq |f|^2_{L^2(Q)}$. Let us estimate I_2. First, for $m \neq 0$, we have $|g_m|^2 = \frac{|\tilde{g}_m|^2}{\mu_m^2}$ (by (2.29)), so

$$
I_2 = C \sum_{\substack{\lambda_n \neq \mu_m \\ m \neq 0}} \frac{\lambda_n^2}{\mu_m^2} \frac{|\tilde{g}_m|^2}{(\lambda_n + |\mu_m|)^2(\lambda_n - |\mu_m|)^2} + C \sum_n \frac{|g_0|^2}{\lambda_n^2} = C(I_3 + I_4). \tag{2.33}
$$

Since $|g_0|^2 \leq |g|^2_{L^2(0,T)}$, we have $I_4 \leq C|g|^2_{L^2(0,T)}$. Finally, for I_3 we must proceed as follows:

$$
\begin{aligned}
I_3 = &\sum_{\substack{|\lambda_n - |\mu_m|| < \varepsilon\lambda_n \\ \lambda_n \neq |\mu_m|}} \frac{\lambda_n^2 |\tilde{g}_m|^2}{\mu_m^2(\lambda_n + |\mu_m|)^2(\lambda_n - |\mu_m|)^2} \\
&+ \sum_{|\lambda_n - |\mu_m|| \geq \varepsilon\lambda_n} \frac{\lambda_n^2 |\tilde{g}_m|^2}{\mu_m^2(\lambda_n + |\mu_m|)(\lambda_n - |\mu_m|)^2} = I_5 + I_6
\end{aligned}
$$

with $0 < \varepsilon < 1$. Clearly $|\lambda_n - |\mu_m|| < \varepsilon\lambda_n$ yields $|1 - \frac{|\mu_m|}{\lambda_n}| < \varepsilon$, so $\frac{|\mu_m|}{\lambda_n} > 1 - \varepsilon$, i.e., $\frac{\lambda_n}{|\mu_m|} < \frac{1}{1-\varepsilon}$. Therefore

$$
I_5 \leq (1-\varepsilon)^{-2} C \sum_{m,n} \frac{|\tilde{g}_m|^2}{m^2+n^2} \leq C|g'|^2_{L^2(0,T)}.
$$

For I_6 a similar estimate holds. Indeed, $|\lambda_n - |\mu_m|| \geq \varepsilon\lambda_n$ yields

$$I_6 \leq \varepsilon^{-2} \sum_{\substack{m\neq 0\\ n\in\mathbb{N}}} \frac{|\tilde{g}_m|^2}{\mu_m^2(m^2+n^2)} \leq C|g'|_{L^2(0,T)},$$

so we conclude that

$$|y_x|^2 \leq C(|f|^2_{L^2(Q)} + |g|^2_{H^1(Q)}), \tag{2.34}$$

and therefore $y \in H^1(Q)$. It is easy to see that $y(x,0) = y(x,T)$ a.e. $x \in (0,1)$, i.e., $y \in H^1_\pi(Q)$. If $f \in H^1_\pi(Q)$ and $g \in H^2_\pi(0,T)$, then

$$\begin{aligned} |y_{tt}|^2_{L^2(Q)} &= \sum_{\lambda\neq|\mu_m|} \mu_m^4 |y_{mn}|^2 \\ &\leq C \sum_{\lambda\neq|\mu_m|} \left(\frac{\mu_m^2(\mu_m^2|\tilde{f}_{mn}|^2)6}{\lambda_n^2+\mu_m^2} + \frac{\mu_m^4|g_m|^2}{\lambda_n^2+\mu_m^2} \right) \\ &\leq C(|f_t|_{L^2(Q)} + |g''|^2_{L^2(0,T)}), \end{aligned} \tag{2.35}$$

where $|g''|^2_{L^2(0,T)} = \sum_m \mu_m^4 |g_m|^2$. To estimate the distributional derivative y_{tx} one repeats the above procedure, namely,

$$\begin{aligned} |y_{tx}|^2_{L^2(Q)} &= \sum_{\lambda_n\neq\mu_m} \lambda_n^2\mu_m^2 |y_{mn}|^2 \\ &\leq C \sum_{\lambda_n\neq\mu_m} \frac{\lambda_n^2(\mu_m^2|\tilde{f}_{mn}|^2)}{\lambda_n^2+\mu_m^2} + C \sum_{\lambda_n\neq\mu_m} \frac{\lambda_n^2(\mu_m^4|g_m|^2)}{\mu_m^2(\lambda_n^2+\mu_m^2)(\lambda_n-|\mu_m|)^2} \\ &\leq C(|f_t|^2_{L^2(Q)} + |g''|^2_{L^2(0,T)}). \end{aligned} \tag{2.36}$$

Therefore y_{tt} and y_{tx} belong to $H^2(Q)$. Finally, as $(uy_x)_x = uy_{tt} + f$ in $D'(Q)$ and $y_x, y_{tt} \in L^2(Q)$, it follows that $y \in H^2(Q)$. Clearly, the estimate (2.21) implies that the range $R(A(u))$ of $A(u)$ is closed in $L^2(Q)\times L^2(0,T)$. Definition 2.1 implies directly that $A(u)$ is closed for every $u \in U$. Moreover, $D(A(u))$ is dense in $L^2(Q)$. This is because $D(A(u))$ contains $\{\varphi \in H^2(Q); \varphi_x(0,t) = \varphi(1,t) = 0\}$. Since $R(A)$ is closed we have $L^2(Q) \times L^2(0,T) = R(A(u)) \oplus N(A^*(u))$; $R(A(u)) = (N(A^*(u)))^\perp$ (i.e., $R(A(u))$ is orthogonal on $N(A^*(u))$, where $A^* : L^2(Q) \times L^2(0,T) \to L^2(Q)$ is the adjoint of A. We also have $L^2(Q) = R(A^*(u)) \oplus N(A(u))$, with $N(A(u)) = N(A_0(u))$ and $(N(A_0(u)))^\perp = S_0(u)$, $u \in U$. Since $R(A(u)) = S(u)$ we conclude that

$$A^{-1}(u) \in L(S(u), S_0(u)) \quad \forall u \in U.$$

This completes the proof.

In particular it follows from (2.22), (2.22′), and trace theorems that

$$|y(0,\cdot)|_{L^2(0,T)} \leq C(u)(|f|_{L^2(Q)} + |g|_{H^1(0,T)}), \tag{2.37}$$

$$|y(0,\cdot)|_{H^1(0,T)} \leq C(u)(|f|_{H^1(Q)} + |g|_{H^2(0,T)}) \tag{2.38}$$

for $y \in A^{-1}(f,g)$ with $(f,g) \in R(A(u))$. Note also that since $\dim N(A(u)) < \infty$ it follows from Proposition 2.1 that if $g \in H^1(0,T)$, each weak solution $y \in L^2(Q)$ to (2.4) is in $H^1(Q)$, i.e.,

$$A(u)y = (f,g), (f,g) \in R(A(u)) \text{ with } g \in H^1(Q) \implies y \in H^1(Q), \tag{2.39}$$

$$|y|_{H^3(Q)} \leq C(u)|g|_{H^3(0,1)}. \tag{2.39′}$$

Proposition 2.2 below gives additional information on the dependence of $A^{-1}(u)(f,g)$ on bounded subsets B_r of U, where

$$B_r = \{u \in H^2(0,1) \cap U; |u|_{H^2(0,1)} \le r\}. \tag{2.40}$$

Moreover, if $g \in H^3_\pi(0,T)$ and $f = 0$, a simple inspection of (2.35) and (2.36) shows that y_{ttt} and y_{ttx} are in $L^2(Q)$, which implies (by (2.1)) that y_{txx} and y_{xxx} are also in $L^2(Q)$ and therefore $y \in H^3(Q)$. We also get the following proposition.

PROPOSITION 2.2. *Let $u \in B_r$. Then there are two orthogonal subspaces $S_1(u)$ and $S_2(u)$ of $S_0(u)$ such that*

(1) *$S_0(u) = S_1(u) \oplus S_2(u)$; $\dim S_1(u) < +\infty$.*

(2) *For any $u \in B_r$ and $(f,g) \in R(A(u))$, $y(u) = A^{-1}(u)(f,g) = y_1^u + y_2^u$, $y_1 \in S_1(u)$; $y_2 \in S_2(u)$ and y_2 satisfies the estimates (2.21), (2.22), and (2.22′) with a constant C independent of $u \in B_r$.*

Proof. Denote $k_{mn} = 4|m|p-(2n+1)(2k+1)$. By (2.25) we have (with $\theta_n(u) = \theta_n^2$ and $\lambda_n \ne |\mu_m|$)

$$|\lambda_n - |\mu_m|| = b|k_{mn} - \theta_n(u)|, \quad b = \frac{\pi}{2(2k+1)}, \quad \theta_n \to 0 \text{ as } n \to \infty. \tag{2.41}$$

There is $d > 0$ such that

$$|\theta_n(u)| \le d \quad \forall u \in B_r.$$

Set

$$\begin{aligned} &Z_1 = \{(m,n); m \in \mathbb{Z}, n \in \mathbb{N}, |k_{mn}| \ge d+1\}, \\ &Z_2 = \{(m,n); |k_{mn}| < d+1\} \quad (\text{so } Z_1 \cap Z_2 = \emptyset). \end{aligned} \tag{2.42}$$

Let N_0 be a positive integer such that

$$|\theta_n(u)| \le \frac{1}{2} \quad \forall u \in B_r \text{ and } n \ge N_0. \tag{2.42′}$$

Set

$$Z_1' = \{(m,n) \in Z_2, n < N_0\}; \quad Z_2'' = \{(m,n) \in Z_2; n \ge N_0\}. \tag{2.43}$$

Obviously

$$\begin{aligned} &Z_2 = Z_2' \cup Z_2'', \quad Z_2' \cap Z_2'' = \emptyset, \\ &|\lambda_n - |\mu_m|| \ge b \quad \forall (m,n) \in Z_1, \\ &|\lambda_n - |\mu_m|| \ge \frac{b}{2} \quad \forall m \in \mathbb{Z}, \forall n \ge N_0 \text{ (as } |k_{mn}| \ge 1). \end{aligned} \tag{2.44}$$

Therefore, with $\lambda_n = \lambda_n(u)$ we have

$$|\lambda_n - |\mu_m|| \ge \frac{b}{2} \quad \forall (m,n) \in Z_1 \cup Z_2'', \forall u \in B_r, \tag{2.45}$$

and the set Z_2' is finite. We now have

$$\begin{aligned} &S_1(u) = \text{Span}\{\varphi_n e^{i\mu_m t}; (m,n) \in Z_2'\}, \quad \dim S_1(u) < +\infty, \\ &S_2(u) = \text{Span}\{\varphi_n e^{i\mu_m t}; (m,n) \in Z_1 \cup Z_2''\}. \end{aligned} \tag{2.46}$$

In view of (2.23), (2.45), and (2.46) it follows that

$$y(u) = \sum_{m,n\in Z_2'} y_{mn}\varphi_n\psi_m + \sum_{(m,n)\in Z_1\cup Z_2''} y_{mn}\varphi_n\psi_m \qquad (\psi_m = e^{i\mu_m t})$$
$$= y^1(u) + y^2(u), \quad y^1(u)\in S_1(u), \quad y^2(u)\in S_2(u),$$

and (as $|\varphi_n(0)\le C\ \forall u\in B_r$)

$$|y^2(u)|^2_{L^2(Q)} \le C_r \sum_{(m,n)\in Z_1\cup Z_2''} \frac{|\tilde f_{mn}|^2 + |g_m|^2}{m^2+n^2} \le C_r(|f|^2_{L^2(Q)} + |g|^2_{L^2(0,T)}) \tag{2.47}$$

with C_r independent of $u\in B_r$. The proof is complete.

Remark 2.3. If $y = A^{-1}(f,g)$ is in $H^1_\pi(Q)$, then it is readily seen that $y(1,t)=0$ a.e. $t\in(0,T)$ and (2.4) reduces to

$$\int_Q u(y_t\varphi_t - y_x\varphi_x)\,dx\,dt + \int_0^T g(t)\varphi(0,t)\,dt + \int_Q f\varphi\,dx\,dt = 0\ \forall\varphi\in V, \tag{2.48}$$

where $V = \{\varphi\in H^1_\pi(Q);\varphi(1,t)=0 \text{ a.e. } t\in(0,T)\}$.

3. The existence of optimal controllers. We are now in a position to define precisely the problem (1.5), namely,

$$\text{(P)}\quad \text{minimize}\left\{\int_0^T |y(0,t)-y_0(t)|^2dt + \gamma\int_0^1((u''(x))^2+u^2(x))dx;\right.$$
$$\left. A(u)y = (0,g), \quad u\in U\right\}, \tag{3.1}$$

where $\gamma>0, g\in H^1_\pi(0,T), y_0\in L^2(0,T), T=\frac{2k+1}{p}\in Q_0$ (as in (1.6)), and $A(u)$ as in Proposition 2.1. The control function $u\in U$ is said to be admissible if $(0,g)\in R(A(u)) = S(u)$, defined by (2.18) (Proposition 2.1), i.e., if $g_m=0$ for all (m,n) for which $|\mu_m|=\lambda_n$. Equivalently, u is admissible if the problem (1.1)–(1.3) has a weak solution (in the sense of Definition 2.1). An example of admissible control is $u(x)\equiv a$. Indeed, in this case the eigenvalues λ_n given by (2.6′) are $\lambda_n = (2n+1)\frac{\pi}{2}$, i.e., $\theta_n=0$ for all $n=1,2,\dots$, and according to Remark 2.1 and (2.18), the null space $N(A_0(u))=0$ so $R(A(u)) = L^2(Q)\times L^2(0,T)$ for $u=a$ and therefore $(0,g)\in R(A(a))$.

THEOREM 3.1. *The problem* (3.1) *has at least one solution* $(y^*,u^*)\in H^1_\pi(Q)\times U$. *If* $g\in H^2_\pi(0,T)$, *then* $y^*\in H^2_\pi(Q)$.

Proof. Let $(y_k,u_k)\in H^1_\pi(Q)\times U$ be a minimizing sequence, i.e.,

$$d \le \int_0^T |y_k(0,t)-y_0(t)|^2dt + \gamma|u_k''|^2_{L^2(0,1)} + \gamma|u|^2_{L^2(0,1)} \le d + \frac{1}{k}, \quad k\in\mathbb{N}, \tag{3.2}$$

where $d=$ infimum of problem (3.1) and $A(u_k)y_k = (0,g)$. Relabeling if necessary, we may assume that $u_k\to u^*$ and $u_k'\to u^{*\prime}$ in $C[0,1]$ as $k\to\infty$. We need to show that one can pass to limit for $k\to+\infty$ in (3.2) and

$$\int_Q u_k(y_{kt}\varphi_t - y_{kx}\varphi_x)dxdt = -\int_0^T g(t)\varphi(0,t)dt \quad \forall v\in V \tag{3.3}$$

with V defined as in (2.48), where $y_{kt} = (y_k)_t$, $y_{kx} = (y_k)_x$ ((3.3) is the meaning of $A(u_k)y_k = (0, g)$ as mentioned in (2.48)). On the basis of Propositions 2.1 and 2.2, y_k can be uniquely written as

$$y_k = y_k^1 + y_k^2 + y_k^3, \quad y_k^1 \in S_1(u_k), y_k^2 \in S_2(u_k), y_k^3 \in N(A(u_k)). \tag{3.4}$$

This is because, by (3.2) $u_k \in B_r$ for some suitable r. We know only that the trace $t \to y_k(0,t)$ is bounded in $L^2(0,T)$ (by (3.2)) and that $y_k = A^{-1}(u_k)(0,g)$, which yields

$$|y_k^2|_{H^1(Q)} \le C \quad \text{for all } k = 1, 2, \dots . \tag{3.5}$$

It follows that $t \to y_k^j(0,t), j = 1,2,3$, are bounded in $L^2(0,T)$, which will imply the boundedness of y_k^1 and y_k^3 in $H^2(Q)$ (as dim $S_1(u) \le N_1$ and dim $N(A(u_k)) \le N_2$ for all k, with N_1 and N_2 independent of k). Indeed,

$$N(A(u)) = \text{Span}\{\varphi_n \psi_m; (m,n) \text{ with } \lambda_n = |\mu_m|\} \tag{3.6}$$

(by (2.6) and (2.17)), and consequently

$$y_k^3(0,t) = \sum_{\substack{m,n \\ \lambda_n = |\mu_m|}} a_{mn}^k \varphi_n^k(0) e^{i\mu_m t}, \tag{3.7}$$

where $\lambda_n = \lambda_n^k$ and φ_n^k are the eigenvalues and eigenfunctions corresponding to $u = u_k$. A key remark is that for each n, $\varphi_n^k(0) \to \varphi_n^*(0)$ for $k \to +\infty$, where λ_n^* and φ_n^* are the eigenvalues and eigenfunctions corresponding to $u = u^*$. Therefore $\varphi_n^*(0) \ne 0$ (as $\varphi_n^*(0) = 0$ jointly $\varphi_n^{*\prime}(0) = 0$ would imply $\varphi_n^*(x) \equiv 0$ on $[0,1]$) since φ_n^* is an eigenfunction:

$$-(u^* \varphi_{nx}^*)_x = -\lambda_n^* u^* \varphi_n^*, \quad \varphi_n^{*\prime}(0) = \varphi_n^*(1) = 0.$$

By (3.7) with $y_k^3(0,t)$ in $L^2(0,T)$ and it follows that $|a_{mn}^k \varphi_n^k(0)| \le C$ for all k and (m,n) with $\lambda_n = |\mu_m|$. As there are only a finite number (independent of k) of such pairs (m,n), and $|\varphi_n^k(0)| \ge C_1 > 0$ for all k and a finite number of n, it follows that $|a_{mn}^k| \le C_2$ for all $k \in \mathbb{N}$ and (m,n) with $\lambda_n = |\mu_m|$. The conclusion is that y_k^3 is bounded in $H^2(Q)$. Similarly, by (2.23) and (2.46)

$$y_k^1(0,t) = \sum_{(m,n) \in Z_2'} \theta_{mn}^k |\varphi_n^k(0)|^2 g_m e^{i\mu_m t}, \quad \theta_{mn}^k = ((\lambda_n^k)^2 - \mu_m^2)^{-1} \tag{3.8}$$

with $|\varphi_n^k(0)| \ge C_1 > 0 \ \forall k \in \mathbb{N}$ and $(m,n) \in Z_2'$ (which is finite). We can also assume $g_m \ne 0$ in (3.8). Fix m with $(m,n) \in Z_2'$. By (3.8) it follows that

$$\left| \sum_{(m,n) \in Z_2'} \theta_{mn}^k (|\varphi_n^k(0)|^2 \right| \le C_3 \quad \text{for all } k \text{ as } t \to y_k^1(0,t) \tag{3.9}$$

is bounded in $L^2(0,T)$ independently of k. Finally, (3.9) implies $|\theta_{mn}^k| \le C_4$ for all k and $(m,n) \in Z_2'$. Indeed, the sequence $n \to \lambda_n^k$ is strictly increasing, so $|\theta_{mn}^k| \to +\infty$ as $k \to +\infty$ can occur only for at most one n which would be in conflict with (3.9). (Say that $n_1 < n_2$ are such that

$$\lambda_{n_1}^k < |\mu_m| < \lambda_{n_2}^k \quad (m, n_j) \in Z_2', \quad j = 1, 2.$$

Letting $k \to +\infty$ we can have (at most) either $\lambda^*_{n_1} < |\mu_m| \le \lambda^*_{n_2}$ or $\lambda^*_1 \le |\mu_m| < \lambda^*_{n_2}$ as $\lambda^*_{n_1} < \lambda^*_{n_2}$.)

Therefore, (3.9) proves the boundedness of the Fourier coefficients of $y^1_k(0,t)$; i.e., y^1_k is also bounded in $H^2(Q)$. We now can pass to limit in (3.2) and (3.3) for $k \to +\infty$. This is because (relabeling if necessary) we have $y_k \to y^*$ weakly in $H^1(Q)$ and strongly in $L^2(Q)$; $y_k(0,t) \to y^*(0,t)$ strongly in $L^2(0,T)$. $u_k \to u^*$ in $C^1([0,1])$. Therefore $(y^*, u^*) \in H^1(Q) \times U$ is an optimal pair.

4. The maximum principle. Throughout this (and the next) section one assumes that

$$T = \frac{2k+1}{p} \quad \text{as in (1.6)}; \quad y_0 \in H^2_\pi(0,T), g \in H^3_\pi(0,T). \tag{4.1}$$

The main result of this section is the following theorem.

THEOREM 4.1 (the maximum principle). *Let (y^*, u^*) be an optimal pair of problems* (3.1) *such that $N(A(u^*)) = \{0\}$. Then there is p in $H^1_\pi(Q)$ such that*

$$\begin{aligned} &u^* p_{tt} - (u^* p_x)_x = 0 \quad \text{in } Q, \\ &-u_0 p_x(0,t) = y^*(0,t) - y_0(t), \quad p(1,t) = 0, t \in (0,T), \end{aligned} \tag{4.2}$$

$$\int_0^1 \int_0^T (y^*_t p_t - y^*_x p_x)(u^* - u)\, dxdt \le \gamma \int_0^1 (u^*)''(u^* - u)'' dx + \gamma \int_0^1 u^*(u^* - u) dx \quad \forall u \in U. \tag{4.3}$$

Remark 4.1. Problem (4.2) should be viewed of course in the sense of Definition 2.1. However, since $p \in H^1_\pi(Q)$, this problem can be equivalently written as indicated by (2.48), i.e.,

$$\int_Q u^*(x)(p_t\varphi_t - p_x\varphi_x)\, dx\, dt = -\int_0^T (y^*(0,t) - y_0(t))\varphi(0,t)\, dt \quad \forall \varphi \in V. \tag{4.4}$$

Note that (in view of Proposition 2.1), $y^* \in H^1_\pi(Q)$. Equation (4.3) is a variational inequality which can be equivalently written as

$$\int_0^T (y^*_t(x,t)p_t(x,t) - y^*_x(x,t)p_x(x,t))\, dt \in N_U(u^*) + B(u^*), \tag{4.5}$$

where N_U is the normal cone to U in $(H^2(0,1))'$ and $B : H^2(0,1) \to (H^2(0,1))'$ is defined by

$$\langle Bu, v\rangle = \gamma \int_0^1 (u''v'' + uv) dx, \quad u, v \in H^2(0,1). \tag{4.6}$$

Formally, y^* is the solution to the following free boundary problem:

$$\begin{aligned} &\gamma u^* + \gamma u^{*(4)} = \int_0^T (y^*_t p_t - y^*_x p_x)\, dt \quad \text{in } \{x \in [0,1); u^*(x) > a\}, \\ &\gamma u^* + \gamma u^{*(4)} \ge \int_0^T (y^*_t p_t - y_x p_x)\, dt \quad \text{in } (0,1), \\ &u^*(0) = u_0, \quad u^{*\prime\prime}(0) = u''(1) = 0, \quad (u^*)'''(1) = 0. \end{aligned} \tag{4.7}$$

Anyway, the solution $u = u^*$ to (4.5) is given by $u^* = \lim_{\lambda \downarrow 0} u_\lambda$ strongly in $H^1(0,1)$ and weakly in $H^2(0,1)$, where u_λ is the solution of the problem

$$\begin{aligned} &u_\lambda^{\mathrm{IV}} + \beta_\lambda(u_\lambda - a) = h \quad \text{a.e. in } (0,1), \\ &u_\lambda(0) = u_0, \quad u_\lambda''(0) = u_\lambda''(1) = 0, \quad u_\lambda'''(1) = 0, \end{aligned} \tag{4.8}$$

where $\beta_r = -\lambda^{-1} r^-$ for $r \in R$ and $h(x) = \int_0^T (y_t^* p_t - y_x^* p_x) dt$ (see [5, p. 132]).

Proof of Theorem 4.1. The fact that $N(A(u^*)) = \{0\}$ implies by (2.18) that $S(u^*) = R(A(u^*)) = L^2(Q) \times L^2(0,T)$, so in view of Proposition 2.1, there is $p \in H_\pi^2(Q)$ such that $A(u^*)p = (0, y^*(0,\cdot) - y_0)$, which is just (4.2). In order to derive the variational inequality (4.3) we need the tangent cone $TK(u^*, y^*)$ to K at (u^*, y^*), where

$$K = \{(u,y) \in U \times H_\pi^2(Q); A(u)y = (0,g)\}. \tag{4.9}$$

First we prove that the hypothesis $N(A(u^*)) = 0$ implies (for each $w = u - u^*$, $u \in U$)

$$N(A(u^* + \varepsilon w)) = \{0\} \quad \text{for some } \varepsilon \downarrow 0. \tag{4.9$'$}$$

Indeed, by (2.12) and Proposition 2.1,

$$N(A(u^* + \varepsilon w)) = \mathrm{Span}\{\varphi_n(u^* + \varepsilon w) e^{i\mu_m t}; \lambda_n(u^* + \varepsilon w) = |\mu_m|\}. \tag{4.10}$$

If (4.9′) were not true, then there would exist (m,n) and a sequence $\varepsilon_p \to 0$ as $p \to +\infty$ such that $\lambda_n(u^* + \varepsilon_p w) = |\mu_m|$ (as $N(A(u))$ is finite dimensional for any $u \in U$). Passing to limit for $p \to \infty$ and taking into account the continuity of $u \to \lambda_n(u)$, we get $\lambda_n(u^*) = |\mu_m|$, which is in conflict with $N(A(u^*)) = 0$ (which means that there are no (m,n) such that $\lambda_n(u^*) = |\mu_m|$). As we mentioned above, (4.10) implies $R(u^* + \varepsilon w) = L^2(q) \times L^2(0,T)$, so the equation

$$A(u^* + \varepsilon w) z_\varepsilon = (-A_0(w) y^*, 0) \tag{4.11}$$

has a unique solution $z_\varepsilon \in H^2(Q)$. This is because $y^* \in H^3(Q)$ (by (2.39′)), so $A_0(w) y^* \in H^1(Q)$, and by (2.22′) it follows that $z_\varepsilon \in H^2(Q)$ and z_ε is bounded in $H^2(Q)$. We may assume that $z_\varepsilon \rightharpoonup z$ weakly in $H^1(Q)$, and letting $\varepsilon \downarrow 0$, (4.11) yields

$$A(u^*) z = (-A_0(w) y^*, 0). \tag{4.12}$$

Combining (4.11) and (4.12) one obtains

$$A(u^*)(z_\varepsilon - z) = -\varepsilon A(w) z_\varepsilon = (-\varepsilon A_0(w) z_\varepsilon, 0). \tag{4.13}$$

This implies $z_\varepsilon - z = \theta_\varepsilon \to 0$ as $\varepsilon \downarrow 0$ in $H^2(Q)$ and $(u^* + \varepsilon w, y^* + \varepsilon(z + \theta_\varepsilon)) \in TK(u^*, y^*)$, i.e., $A(u^* + \varepsilon w)(y^* + \varepsilon(z + \theta_\varepsilon)) = (0, g)$. Therefore,

$$\begin{aligned} &\int_0^T (y^*(0,t) - y_0(t))^2 dt + \gamma |u^*|_{H^2(0,1)}^2 \\ &\quad \le \int_0^T ((y^* + \varepsilon(z + \theta_\varepsilon))(0,t) - y_0(t))^2 + \gamma |u^* + \varepsilon w|_{H^2(0,1)}^2 \quad \forall \varepsilon > 0 \end{aligned}$$

with $|u^*|_{H^2(0,1)}^2 = |u^{*\prime\prime}|_{L^2(0,1)}^2 + |u^*|_{L^2(0,1)}^2$. This implies

$$\int_0^T (y^*(0,t) - y_0(t)) z(0,t) dt + \gamma \int_0^1 ((u^*)'' w'' + u^* w) dx \ge 0; \tag{4.14}$$

on the other hand p satisfies

$$\begin{aligned}\int_Q pA_0(u^*)z\,dx\,dt &= \int_0^T (y^*(0,t) - y_0(t))z(0,t)\,dt \\ &= -\int_Q pA_0(w)y^*\,dx\,dt = \int_Q (y_t^* p_t - y_x^* p_x)w\,dx\,dt.\end{aligned} \tag{4.15}$$

Clearly, (4.15) and (4.14) imply (4.3), which completes the proof.

Acknowledgments. We would like to thank the referee and the associate editor in charge of this paper (Professor John A. Burns) for some useful suggestions.

REFERENCES

[1] A. R. Aftabizadeh, N. H. Pavel, and Y. K. Huang, *Anti-periodic oscillations of some second-order differential equations and optimal control problems*, J. Comput. Appl. Math., 52 (1994), pp. 3–21.
[2] A. Bamberger, G. Chavent, and P. Lailly, *About the stability of the inverse problem in the one dimensional wave equation; applications to the interpretation of seismic profiles,* Appl. Math. Optim., 5 (1979), pp. 1–47.
[3] V. Barbu and N. H. Pavel, *Optimal control problems with two point boundary conditions,* J. Optim. Theory Appl., 77 (1993), pp. 51–57.
[4] V. Barbu and N. H. Pavel, *Optimal control of thermal conductivity of a rod under periodic conditions,* Ricerche Mat., (1997), to appear.
[5] V. Barbu and T. Precupanu, *Convexity and Optimization in Banach Spaces*, D. Reidel, Dordrecht, 1986.
[6] H. Brezis, *Periodic solutions of nonlinear vibrating strings and duality principles,* Bull. AMS, 8 (1983), pp. 409–426.
[7] H. Brezis and L. Nirenberg, *Forced vibrations for a nonlinear wave equation,* Comm. Pure. Appl. Math., 31 (1978), pp. 1–30.
[8] G. Chavent, *About the stability of the optimal control solution of inverse problem,* in Inverse and Improperly Posed Problems in Differential Equations, G. Arger, ed., Akademie Verlag, Berlin, 1979.
[9] G. Chavent and A. Bamberger, *Sur le problème inverse en seismique,* Ann. Sci. Math. Quebec, 1 (1977), pp. 153–174.
[10] E. Kamke, *Differentialgleichungen Lösungsmethoden und Lösungen*, Chelsea, New York, 1948.
[11] J. L. Lions, *Controle Optimal des Systèmes Distribuées Singulières*, Dunod, Paris, 1983.
[12] K. Yosida, *Functional Analysis*, 6th ed., Springer-Verlag, New York, 1980.

SIAM J. CONTROL OPTIM.
Vol. 35, No. 5, pp. 1557–1573, September 1997

007

HOMOGENIZATION OF AN OPTIMAL CONTROL PROBLEM*

S. KESAVAN† AND J. SAINT JEAN PAULIN‡

Abstract. We consider an optimal control problem in which both the state equation and the cost functional have rapidly oscillating coefficients (characterized respectively by matrices A_ε and B_ε, where ε is a small parameter). We make no periodicity assumption. We study the limit of the problem when $\varepsilon \to 0$ and work in the framework of H-convergence. We prove that the limit satisfies a problem similar to the original one but with matrices A_0 (the H-limit of A_ε) and $B^\sharp$ (which is a perturbation of the H-limit B_0 of B_ε). We also study some particular cases. This paper extends former results obtained by Kesavan and Vanninathan in the periodic case.

Key words. homogenization, H-convergence, optimal control

AMS subject classifications. 35B, 49J

PII. S0363012994271843

1. Introduction.

The aim of this paper is to discuss the homogenization of an optimal control problem in which both the state equation (given by a second-order elliptic boundary value problem) and the cost function (involving a Dirichlet-type integral of the state function) have rapidly oscillating coefficients.

Let Ω be a bounded open set in $\mathbb{R}^n$. For any given constants $\alpha_M > \alpha_m > 0$, we denote by $\mathcal{M}(\alpha_m, \alpha_M, \Omega)$ the set of all $n \times n$ matrices $A = (a_{ij})$ whose entries are functions on Ω such that

$$\alpha_m \xi_i \xi_i \le a_{ij}(x)\xi_i\xi_j \le \alpha_M \xi_i \xi_i \quad \text{almost every (a.e.) } (x) \tag{1.1}$$

for all $\xi = (\xi_i) \in \mathbb{R}^n$. Here and throughout the sequel, we use the convention of summation over repeated indices.

Let $A \in \mathcal{M}(\alpha_m, \alpha_M, \Omega)$ and $B \in \mathcal{M}(\beta_m, \beta_M, \Omega)$, and assume that B is symmetric.

We define the optimal control problem as follows. Let $U_{ad} \subset L^2(\Omega)$ be a closed convex subset. Let $f \in L^2(\Omega)$ be a given function and $N > 0$ be a given constant. For $\theta \in U_{ad}$, the equation of state is given by

$$\begin{cases} -\mathrm{div}(A\nabla u) = f + \theta & \text{in } \Omega, \\ u = 0 & \text{on } \partial\Omega. \end{cases} \tag{1.2}$$

The state $u = u(\theta)$ is thus defined as the weak solution in $H_0^1(\Omega)$ of problem (1.2).

Then the cost function is given by

$$J(\theta) = \frac{1}{2}\int_\Omega (B\nabla u, \nabla u)dx + \frac{N}{2}\int_\Omega \theta^2 dx. \tag{1.3}$$

The optimal control $\theta^\star$ is the function in U_{ad} which minimizes $J(\theta)$ for $\theta \in U_{ad}$.

*Received by the editors July 19, 1994; accepted for publication (in revised form) June 20, 1996. http://www.siam.org/journals/sicon/35-5/27184.html

†The Institute of Mathematical Sciences, C.I.T. Campus, Taramani, Madras 600113, India (kesh@imsc.ernet.in). Permanent address: Tata Institute of Fundamental Research, Bangalore Centre, India. This research was done when S. Kesavan was at the Mathematics Department of the University of Metz.

‡Département de Mathématiques, Université de Metz, Ile du Saulcy, 57045 Metz cedex 01, France (sjpaulin@poncelet.univ-metz.fr).

The problem posed above is a standard one, and a discussion of this can be found in the book by Lions [3]. There exists a unique optimal control. The problem can be reduced to a system of equations by introducing the adjoint state p. Thus we get

$$\begin{cases} -\mathrm{div}(A\nabla u) = f + \theta & \text{in } \Omega, \\ \mathrm{div}({}^tA\nabla p - B\nabla u) = 0 & \text{in } \Omega, \end{cases} \tag{1.4}$$

where $u, p \in H_0^1(\Omega)$. The optimal control $\theta^\star$ is characterized by the variational inequality

$$\int_\Omega (p + N\theta^\star)(\theta - \theta^\star)dx \geq 0 \qquad \forall \theta \in U_{ad}\ . \tag{1.5}$$

The situation that we will be interested in is that in which, given a parameter $\varepsilon > 0$ which tends to zero, the matrices A and B above depend on ε. Thus we suppose that

$$A_\varepsilon \in \mathcal{M}(\alpha_m, \alpha_M, \Omega) \quad \text{and } B_\varepsilon \in \mathcal{M}(\beta_m, \beta_M, \Omega).$$

Then as usual, the optimal control $\theta_\varepsilon^\star$ exists and can be shown to be bounded in $L^2(\Omega)$. Thus (for a subsequence)

$$\theta_\varepsilon^\star \rightharpoonup \theta^\star \qquad \text{weakly in } L^2(\Omega),$$

and we would like to know whether $\theta^\star$ is an optimal control defined by a problem of the same type with matrices $A^\star$ and $B^\star$ and, if so, identify these matrices.

A special case of this situation was studied by Kesavan and Vanninathan [2]. They assumed that A_ε and B_ε are periodic, and they computed the matrices of the limiting problem. It turned out, as expected, that the matrix $A^\star$ was indeed the limit of the matrices A_ε in the topology of H-convergence (see Murat [4] or Tartar [6]) but the matrix $B^\star$ was a perturbation of the H-limit of the B_ε.

We will study the problem in the general case and identify the limiting problem. As in the case of Kesavan and Vanninathan [2], we will first study the homogenization of system (1.4) for a fixed θ in U_{ad}.

Given $A_\varepsilon \in \mathcal{M}(\alpha_m, \alpha_M, \Omega)$ and $B_\varepsilon \in \mathcal{M}(\beta_m, \beta_M, \Omega)$, it is easy to see that the corresponding solution $(u_\varepsilon, p_\varepsilon)$ of the system

$$\begin{cases} -\mathrm{div}(A_\varepsilon \nabla u_\varepsilon) = f + \theta & \text{in } \Omega, \\ \mathrm{div}({}^tA_\varepsilon \nabla p_\varepsilon - B_\varepsilon \nabla u_\varepsilon) = 0 & \text{in } \Omega, \\ u_\varepsilon = p_\varepsilon = 0 & \text{on } \partial\Omega \end{cases} \tag{1.6}$$

is bounded in $(H_0^1(\Omega))^2$, uniformly with respect to (w.r.t.) ε. Thus (for a subsequence, still denoted by the suffix ε)

$$\begin{aligned} u_\varepsilon &\rightharpoonup u_0 \qquad \text{weakly in } H_0^1(\Omega), \\ p_\varepsilon &\rightharpoonup p_0 \qquad \text{weakly in } H_0^1(\Omega). \end{aligned}$$

Since $\mathrm{div}(A_\varepsilon \nabla u_\varepsilon)$ is fixed, by the usual definition of the H-limit (see Murat [4]) we have

$$\begin{cases} \xi_\varepsilon \equiv A_\varepsilon \nabla u_\varepsilon \rightharpoonup A_0 \nabla u_0 & \text{weakly in } L^2(\Omega), \\ -\mathrm{div}(A_0 \nabla u_0) = f + \theta, \end{cases} \tag{1.7}$$

where $A_\varepsilon \xrightarrow{H} A_0$. By the H-limit of problem (1.6) we mean that in addition

$$\begin{cases} z_\varepsilon \equiv {}^t A_\varepsilon \nabla p_\varepsilon - B_\varepsilon \nabla u_\varepsilon \rightharpoonup z_0 & \text{weakly in } L^2(\Omega), \\ \text{div } (z_0) = 0. \end{cases} \tag{1.8}$$

We would like to express z_0 in a form similar to that of z_ε to identify the corresponding matrices and, if possible, to express them in terms of A_0 and B_0, the H-limits of A_ε and B_ε, respectively.

The plan of the paper is as follows. In section 2, we will briefly discuss the one-dimensional case which gives a nice formula for the limiting coefficients in terms of the H-limits. In section 3, we will prove that the limit satisfies a problem similar to the original problem. In particular, we will show that

$$z_0 = {}^t A_0 \nabla p_0 - B^\sharp \nabla u_0,$$

where A_0 is the H-limit of A_ε and $B^\sharp$ is a perturbation of B_0, the H-limit of B_ε. Section 4 is reserved for the study of some properties of $B^\sharp$.

2. The one-dimensional case. Let $0 < \alpha_m \le a_\varepsilon \le \alpha_M$ and $0 < \beta_m \le b_\varepsilon \le \beta_M$ on an interval $(c, d) \subset \mathbb{R}$. System (1.6) now reads as follows:

$$\begin{cases} -\dfrac{d}{dx}\left(a_\varepsilon(x)\dfrac{du_\varepsilon}{dx}\right) = f + \theta & \text{in } (c,d), \\ \dfrac{d}{dx}\left(a_\varepsilon(x)\dfrac{dp_\varepsilon}{dx} - b_\varepsilon(x)\dfrac{du_\varepsilon}{dx}\right) = 0 & \text{in } (c,d), \end{cases} \tag{2.1}$$

with u_ε and p_ε vanishing at c and d.

THEOREM 2.1. *Suppose that the functions a_0 and $b^\sharp$ are such that*

$$\begin{cases} \dfrac{1}{a_\varepsilon} \rightharpoonup \dfrac{1}{a_0} & \text{weakly } \star \text{ in } L^\infty(c,d), \\ b^\sharp = \dfrac{a_0^2}{g_0}, \end{cases} \tag{2.2}$$

where

$$\frac{1}{g_\varepsilon} \equiv \frac{b_\varepsilon}{a_\varepsilon^2} \rightharpoonup \frac{1}{g_0} \qquad \text{weakly } \star \text{ in } L^\infty(c,d). \tag{2.3}$$

Then the following convergences hold:

$$\begin{aligned} u_\varepsilon &\rightharpoonup u_0 \qquad \text{weakly in } H_0^1(c,d), \\ p_\varepsilon &\rightharpoonup p_0 \qquad \text{weakly in } H_0^1(c,d), \end{aligned}$$

where u_0 and p_0 satisfy the equations

$$\begin{cases} -\dfrac{d}{dx}\left(a_0\dfrac{du_0}{dx}\right) = f + \theta & \text{in } (c,d), \\ \dfrac{d}{dx}\left(a_0\dfrac{dp_0}{dx} - b^\sharp\dfrac{du_0}{dx}\right) = 0 & \text{in } (c,d). \end{cases} \tag{2.4}$$

Proof. Step 1. We define g_ε by (2.3) and set

$$\begin{cases} \xi_\varepsilon = a_\varepsilon \dfrac{du_\varepsilon}{dx}, \\ \xi_0 = a_0 \dfrac{du_0}{dx}. \end{cases}$$

It is easy to see that $\{u_e\}$ and $\{p_\varepsilon\}$ are bounded in $H_0^1(c,d)$ and hence possess weakly convergent subsequences. Working with such a subsequence (again denoted by ε) we define u_0 and p_0 to be the respective limits. By the usual arguments of H-convergence, it is obvious that $\xi_\varepsilon \rightharpoonup \xi_0$ weakly in $L^2(c,d)$ and that u_0 satisfies the first equation of (2.4).

Step 2. We now observe that the second equation in (2.1) implies that

$$a_\varepsilon \frac{dp_\varepsilon}{dx} - b_\varepsilon \frac{du_\varepsilon}{dx} = c_\varepsilon, \qquad \text{a constant},$$

and that the sequence $\{c_\varepsilon\}$ is bounded. We can thus assume that $c_\varepsilon \to c_0$ (after extracting a further subsequence if necessary). We write

$$\begin{cases} \dfrac{dp_\varepsilon}{dx} = \dfrac{c_\varepsilon}{a_\varepsilon} + \dfrac{b_\varepsilon}{a_\varepsilon}\ \dfrac{du_\varepsilon}{dx} \\ \qquad\quad = \dfrac{c_\varepsilon}{a_\varepsilon} + \dfrac{b_\varepsilon}{a_\varepsilon^2}\ \xi_\varepsilon. \end{cases} \tag{2.5}$$

Step 3. We wish to pass to the limit in this relation. Obviously

$$\frac{c_\varepsilon}{a_\varepsilon} \rightharpoonup \frac{c_0}{a_0} \text{ in } L^\infty \text{ weak } \star\,. \tag{2.6}$$

Moreover, since

$$-\frac{d\xi_\varepsilon}{dx} = f + \theta,$$

then $\xi_\varepsilon \in H^1(c,d)$ and

$$\|\xi_\varepsilon\|_{H^1(c,d)} \le C,$$

where C is a constant independent of ε. Note that this also holds when we have $\theta_\varepsilon^\star$ instead of θ since $\theta_\varepsilon^\star$ is bounded in $L^2(c,d)$ independently of ε. Thus there exists $\xi_0 \in H^1(c,d)$ such that, up to a subsequence,

$$\xi_\varepsilon \rightharpoonup \xi_0 \quad \text{weakly in } H^1(c,d) \quad \text{and} \quad \xi_\varepsilon \to \xi_0 \quad \text{strongly in } L^2(c,d).$$

Hence, using (2.3), we have

$$\frac{b_\varepsilon}{a_\varepsilon^2}\ \xi_\varepsilon \rightharpoonup \frac{\xi_0}{g_0}.$$

Thus

$$\frac{dp_0}{dx} = \frac{c_0}{a_0} + \frac{\xi_0}{g_0} = \frac{c_0}{a_0} + \frac{a_0}{g_0}\ \frac{du_0}{dx}.$$

Hence

$$a_0 \frac{dp_0}{dx} - \frac{a_0^2}{g_0}\ \frac{du_0}{dx} = c_0, \qquad \text{a constant},$$

which implies the second relation in (2.4). ∎

Remark 2.2. The limits u_0 and p_0 are not unique in general because the weak $\star$ limits $\frac{1}{a_0}$ and $\frac{1}{g_0}$ may not be unique.

Example 2.3. If $a_\varepsilon(x) = a(x/\varepsilon)$ and $b_\varepsilon(x) = b(x/\varepsilon)$, where a and b are periodic functions on, say, $[0,1]$, we have

$$\begin{cases} a_0 = \left[m\left(\dfrac{1}{a}\right)\right]^{-1}, \\ b^\sharp = \dfrac{a_0^2}{g_0}, \quad g_0 = \left[m\left(\dfrac{b}{a^2}\right)\right]^{-1}, \end{cases}$$

where $m(h) = \int_0^1 h(y)dy$ for a periodic function h on $[0,1]$. In this case, the limits u_0 and p_0 are unique.

Remark 2.4. One could try to imitate this proof in higher dimensions. If we set

$$L_\varepsilon = A_\varepsilon B_\varepsilon^{-1}\ {}^t A_\varepsilon,$$

then

$$\begin{cases} A_\varepsilon \xrightarrow{H} A_0, \\ L_\varepsilon \xrightarrow{H} L_0, \end{cases}$$

and one might expect $B^\sharp$ to be given by ${}^t A_0 (L_0)^{-1}\ A_0$. Unfortunately the arguments analogous to those of steps 2 and 3 are no longer valid. Indeed we will show in the next section, by means of an example, that $B^\sharp$ is not, in general, equal to ${}^t A_0 (L_0)^{-1} A_0$. □

3. Identification of the H-limit. We assume henceforth that $\Omega \subset \mathbb{R}^n$ is a bounded open set.

THEOREM 3.1. *Assume that $A_\varepsilon \in \mathcal{M}(\alpha_m, \alpha_M, \Omega)$ and $B_\varepsilon \in \mathcal{M}(\beta_m, \beta_M, \Omega)$. Assume also that A_0 and B_0 are H-limits of A_ε and B_ε. Let $(u_\varepsilon, p_\varepsilon)$ be the solution of system* (1.6). *Then $u_\varepsilon \rightharpoonup u_0$ and $p_\varepsilon \rightharpoonup p_0$ weakly in $H_0^1(\Omega)$. Furthermore*

$$z_\varepsilon \equiv {}^t A_\varepsilon \nabla p_\varepsilon - B_\varepsilon \nabla u_\varepsilon \rightharpoonup z_0 \text{ weakly in } L^2(\Omega),$$

where

$$z_0 = {}^t A_0 \nabla p_0 - B^\sharp \nabla u_0. \tag{3.1}$$

The matrix A_0 is the H-limit of $\{A_\varepsilon\}$ and the matrix $B^\sharp$ depends only on $\{B_\varepsilon\}$ and $\{A_\varepsilon\}$. It is given by formula (3.10) *below.*

Remark 3.2. In the case of the control problem, B_ε is a symmetric matrix. However, to study system (1.6), B_ε need not be taken to be symmetric. Also in order to prove that $B^\sharp$ is symmetric when the B_ε are symmetric, we will need to study the limit first without assuming the symmetry of the B_ε.

Proof of Theorem 3.1. *Step* 1. It is obvious from (1.6) that $\{u_\varepsilon\}$ and $\{p_\varepsilon\}$ are bounded sequences in $H_0^1(\Omega)$ and that $\{z_\varepsilon\}$ is bounded in $L^2(\Omega)$. Thus we assume (working, as usual, with convergent subsequences) that $u_\varepsilon \rightharpoonup u_0, p_\varepsilon \rightharpoonup p_0$ weakly in $H_0^1(\Omega)$, and $z_\varepsilon \rightharpoonup z_0$ weakly in $L^2(\Omega)$. We also have $\xi_\varepsilon \equiv A_\varepsilon \nabla u_\varepsilon$ bounded uniformly w.r.t. ε in $L^2(\Omega)$ and, by usual homogenization results,

$$\xi_\varepsilon \rightharpoonup \xi_0 \qquad \text{weakly in } L^2(\Omega)$$

with

$$\xi_0 = A_0 \nabla u_0, \tag{3.2}$$

where A_0 is the H-limit of $\{A_\varepsilon\}$. Thus we also have

$$-\operatorname{div}(A_0 \nabla u_0) = f + \theta. \tag{3.3}$$

Step 2. We define various test functions to be used in the identification of z_0. Let $e_k \in \mathbb{R}^n$ be the kth standard basis vector. Then we define X_k^ε, Y_k^ε, and ψ_k^ε in $H^1(\Omega)$ as follows:

$$\begin{cases} X_k^\varepsilon \rightharpoonup 0 & \text{weakly in } H^1(\Omega), \\ \operatorname{div}(A_\varepsilon \nabla(-X_k^\varepsilon + x_k)) \to \operatorname{div}(A_0 e_k) & \text{strongly in } H^{-1}(\Omega), \end{cases} \tag{3.4}$$

$$\begin{cases} Y_k^\varepsilon \rightharpoonup 0 & \text{weakly in } H^1(\Omega), \\ \operatorname{div}({}^t B_\varepsilon \nabla(-Y_k^\varepsilon + x_k)) \to \operatorname{div}(B_0 e_k) & \text{strongly in } H^{-1}(\Omega), \end{cases} \tag{3.5}$$

where B_0 is the usual H-limit of $\{B_\varepsilon\}$,

$$\begin{cases} \psi_k^\varepsilon \text{ bounded uniformly w.r.t. } \varepsilon \text{ in } H_0^1(\Omega), \\ \operatorname{div}\,({}^t A_\varepsilon \nabla \psi_k^\varepsilon + {}^t B_\varepsilon \nabla(-X_k^\varepsilon + x_k)) = 0. \end{cases} \tag{3.6}$$

Here x_k denotes the function mapping $x \in \mathbb{R}^n$ to its kth coordinate. The sequence $\{\psi_k^\varepsilon\}$ is bounded, and so it converges weakly in $H^1(\Omega)$ (up to a subsequence). We denote its limit by ψ_k^0.

Let us detail how these test functions can be built. Define X_k^ε by

$$\begin{cases} \operatorname{div}(A_\varepsilon \nabla(-X_k^\varepsilon + x_k)) = \operatorname{div}(A_0 e_k) & \text{in } \Omega, \\ X_k^\varepsilon = 0 & \text{on } \partial\Omega. \end{cases}$$

Multiplication of this relation by X_k^ε, integration by parts, and use of the fact that $A_\varepsilon \in \mathcal{M}(\alpha_m, \alpha_M, \Omega)$ imply that

$$\|X_k^\varepsilon\|_{H^1(\Omega)} \le C,$$

where the constant C is independent of ε. Hence, up to a subsequence,

$$X_k^\varepsilon \rightharpoonup X_k \quad \text{weakly in } H^1(\Omega)$$

and therefore

$$(-X_k^\varepsilon + x_k) \rightharpoonup (-X_k + x_k) \quad \text{weakly in } H^1(\Omega).$$

Then a well-known H-convergence result (see Murat [4, Theorem 1]) implies that

$$A_\varepsilon \nabla(-X_k^\varepsilon + x_k) \rightharpoonup A_0 \nabla(-X_k + x_k) \quad \text{weakly in } (L^2(\Omega))^n.$$

Hence X_k is the solution of

$$\begin{cases} \operatorname{div}(A_0 \nabla(-X_k + x_k)) = \operatorname{div}(A_0 e_k) & \text{in } \Omega, \\ X_k = 0 & \text{on } \partial\Omega; \end{cases}$$

that is,

$$X_k = 0 \quad \text{on } \Omega.$$

This holds for any subsequence X_k^ε, so this is true for the whole sequence X_k^ε.

Similar arguments obviously hold for (3.5).

To establish relation (3.6), let us define for instance ψ_k^ε by

$$\begin{cases} \operatorname{div}({}^tA_\varepsilon \nabla \psi_k^\varepsilon) = \operatorname{div}({}^tB_\varepsilon \nabla(-X_k^\varepsilon + x_k)) & \text{in } \Omega, \\ \psi_k^\varepsilon = 0 \quad \text{on } \partial\Omega. \end{cases}$$

Using the assumptions $A_\varepsilon \in \mathcal{M}(\alpha_m, \alpha_M, \Omega)$ and $B_\varepsilon \in \mathcal{M}(\beta_m, \beta_M, \Omega)$ and the fact that X_k^ε is bounded in $H^1(\Omega)$ independently of ε, we derive

$$\|\psi_k^\varepsilon\|_{H^1(\Omega)} \le C,$$

where the constant C is independent of ε. Thus relation (3.6) is established.

Let us point out that there are several others procedures to build functions X_k^ε, Y_k^ε, and ψ_k^ε satisfying relations (3.4)–(3.6). We detailed one of these for the reader's convenience.

Step 3. Let $\varphi \in \mathcal{D}(\Omega)$ be an arbitrary function. We multiply the second equation in (1.6) by $\varphi(-X_k^\varepsilon + x_k)$ and integrate by parts. Thus we get

$$\begin{aligned} 0 = &- \int_\Omega ({}^tA_\varepsilon \nabla p_\varepsilon - B_\varepsilon \nabla u_\varepsilon) \,.\, (\nabla\varphi)(-X_k^\varepsilon + x_k)dx \\ &- \int_\Omega ({}^tA_\varepsilon \nabla p_\varepsilon) \,.\, \nabla(-X_k^\varepsilon + x_k)\varphi \, dx + \int_\Omega (B_\varepsilon \nabla u_\varepsilon) \,.\, \nabla(-X_k^\varepsilon + x_k)\varphi \, dx \\ = &- \int_\Omega z_\varepsilon \,.\, (\nabla\varphi)(-X_k^\varepsilon + x_k)dx - \int_\Omega \nabla p_\varepsilon \,.\, A_\varepsilon \nabla(-X_k^\varepsilon + x_k)\varphi \, dx \\ &+ \int_\Omega \nabla u_\varepsilon \,.\, {}^tB_\varepsilon \nabla(-X_k^\varepsilon + x_k)\varphi \, dx, \end{aligned}$$

which yields

$$\begin{aligned} 0 = &- \int_\Omega z_\varepsilon \,.\, (\nabla\varphi)(-X_k^\varepsilon + x_k)dx + \int_\Omega p_\varepsilon \operatorname{div}\,(A_\varepsilon \nabla(-X_k^\varepsilon + x_k))\varphi \, dx \\ &+ \int_\Omega p_\varepsilon \, A_\varepsilon \nabla(-X_k^\varepsilon + x_k) \,.\, \nabla\varphi \, dx + \int_\Omega \nabla u_\varepsilon \,.\, {}^tB_\varepsilon \nabla(-X_k^\varepsilon + x_k)\varphi \, dx. \end{aligned} \tag{3.7}$$

Now, the first equation of (1.6) when multiplied by $\varphi\psi_k^\varepsilon$ and integrated by parts gives

$$\begin{aligned} \int_\Omega (f+\theta)\psi_k^\varepsilon\varphi \, dx &= \int_\Omega A_\varepsilon \nabla u_\varepsilon \,.\, (\nabla\varphi)\psi_k^\varepsilon dx + \int_\Omega A_\varepsilon \nabla u_\varepsilon \,.\, (\nabla\psi_k^\varepsilon)\varphi \, dx \\ &= \int_\Omega \xi_\varepsilon \,.\, (\nabla\varphi)\psi_k^\varepsilon dx + \int_\Omega \nabla u_\varepsilon \,.\, {}^tA_\varepsilon(\nabla\psi_k^\varepsilon)\varphi \, dx \\ &= \int_\Omega \xi_\varepsilon \,.\, (\nabla\varphi)\psi_k^\varepsilon dx - \int_\Omega u_\varepsilon \operatorname{div}({}^tA_\varepsilon \nabla\psi_k^\varepsilon)\varphi \, dx - \\ &\quad - \int_\Omega u_\varepsilon {}^tA_\varepsilon \nabla\psi_k^\varepsilon \,.\, \nabla\varphi \, dx \\ &= \int_\Omega \xi_\varepsilon \,.\, (\nabla\varphi)\psi_k^\varepsilon dx + \int_\Omega u_\varepsilon \operatorname{div}({}^tB_\varepsilon \nabla(-X_k^\varepsilon + x_k))\varphi \, dx \\ &\quad - \int_\Omega u_\varepsilon {}^tA_\varepsilon \nabla\psi_\varepsilon^k \,.\, \nabla\varphi \, dx, \end{aligned}$$

which yields

$$\begin{aligned}\int_\Omega (f+\theta)\psi_k^\varepsilon \varphi\ dx = &\int_\Omega \xi_\varepsilon\ .\ (\nabla\varphi)\psi_k^\varepsilon dx - \int_\Omega \nabla u_\varepsilon\ .\ {}^tB_\varepsilon \nabla(-X_k^\varepsilon + x_k)\varphi\ dx \\ &- \int_\Omega u_\varepsilon\, {}^tB_\varepsilon \nabla(-X_k^\varepsilon + x_k)\ .\ \nabla\varphi dx \\ &- \int_\Omega u_\varepsilon\, {}^tA_\varepsilon \nabla\psi_k^\varepsilon\ .\ \nabla\varphi\ dx.\end{aligned} \tag{3.8}$$

Adding (3.7) and (3.8), we get

$$\begin{cases}\displaystyle\int_\Omega (f+\theta)\psi_k^\varepsilon\varphi\ dx = -\int_\Omega z_\varepsilon\ .\ (\nabla\varphi)(-X_k^\varepsilon + x_k)dx \\ \qquad\qquad + \displaystyle\int_\Omega p_\varepsilon \mathrm{div}\ (A_\varepsilon\nabla(-X_k^\varepsilon + x_k))\varphi\ dx \\ \qquad\qquad + \displaystyle\int_\Omega p_\varepsilon A_\varepsilon \nabla(-X_k^\varepsilon + x_k)\ .\ \nabla\varphi\ dx + \int_\Omega \xi_\varepsilon\ .\ (\nabla\varphi)\psi_k^\varepsilon\ dx \\ \qquad\qquad - \displaystyle\int_\Omega u_\varepsilon[\ {}^tB_\varepsilon\nabla(-X_k^\varepsilon + x_k) + {}^tA_\varepsilon\nabla\psi_\varepsilon^k]\ .\ \nabla\varphi\ dx.\end{cases} \tag{3.9}$$

Step 4. We can pass to the limit as $\varepsilon \to 0$ in (3.9) since each of the terms in the right-hand side is a product of two sequences, one converging weakly and the other strongly in $L^2(\Omega)$ (since the injection of $H_0^1(\Omega)$ in $L^2(\Omega)$ is compact). In the second term we have a weak convergence in $H_0^1(\Omega)$ and a strong convergence in $H^{-1}(\Omega)$. Thus,

$$\begin{aligned}\int_\Omega (f+\theta)\psi_k^0\varphi\ dx = &-\int_\Omega z_0\ .\ (\nabla\varphi)\ x_k\ dx + \int_\Omega p_0\ \mathrm{div}(A_0e_k)\varphi\ dx \\ &+ \int_\Omega p_0(A_0e_k)\ .\ \nabla\varphi\ dx + \int_\Omega \xi_0\ .\ (\nabla\varphi)\psi_k^0\ dx \\ &- \int_\Omega u_0 \lim({}^tB_\varepsilon\nabla(-X_k^\varepsilon + x_k) + {}^tA_\varepsilon\nabla\psi_k^\varepsilon)\ .\ \nabla\varphi\ dx \\ = &\int_\Omega \mathrm{div}(z_0)\ (\varphi x_k)dx + \int_\Omega z_0\ .\ (\nabla x_k)\ \varphi\ dx + \int_\Omega p_0\ \mathrm{div}\ (A_0e_k)\varphi\ dx \\ &- \int_\Omega p_0\ \mathrm{div}(A_0e_k)\varphi\ dx - \int_\Omega \nabla p_0\ .\ A_0e_k\ \varphi\ dx \\ &- \int_\Omega \mathrm{div}(\xi_0)\ \psi_k^0\varphi\ dx - \int_\Omega \xi_0.(\nabla\psi_k^0)\ \varphi\ dx \\ &+ \int_\Omega \nabla u_0\ .\ \lim(\ {}^tB_\varepsilon\nabla(-X_k^\varepsilon + x_k) + {}^tA_\varepsilon\nabla\psi_k^\varepsilon)\varphi\ dx \\ &+ \int u_0\ \mathrm{div}\big(\lim(\ {}^tB_\varepsilon\nabla(-X_k^\varepsilon + x_k) + {}^tA_\varepsilon\nabla\psi_k^\varepsilon)\big)\varphi\ dx.\end{aligned}$$

Recall that

$$-\ \mathrm{div}\ \xi_0 = -\ \mathrm{div}(A_0\nabla u_0) = f + \theta.$$

Moreover

$$\mathrm{div}\ z_\varepsilon = 0 = \mathrm{div}(\ {}^tB_\varepsilon\ \nabla(-X_k^\varepsilon + x_k) + {}^tA_\varepsilon\nabla\psi_k^\varepsilon),$$

and so the same is true for the divergence of their weak limits in $L^2(\Omega)$. Thus we get, on eliminating φ,

$$z_0 \,.\, e_k = ({}^tA_0\nabla p_0) \,.\, e_k - \lim\big({}^tB_\varepsilon\nabla(-X_k^\varepsilon + x_k) + {}^tA_\varepsilon(\nabla\psi_k^\varepsilon)\big).\nabla u_0 + ({}^tA_0\nabla\psi_k^0) \,.\, \nabla u_0,$$

which we can write as

$$z_0 = {}^tA_0\nabla p_0 - B^\sharp\nabla u_0,$$

which completes the proof of the theorem. □

We would like to express $B^\sharp$ as a perturbation of the usual H-limit B_0. Hence we write

$${}^tB_\varepsilon\nabla(-X_k^\varepsilon + x_k) = {}^tB_\varepsilon\nabla(-X_k^\varepsilon + Y_k^\varepsilon) + {}^tB_\varepsilon\nabla(-Y_k^\varepsilon + x_k).$$

By (3.5) the second term on the right converges to B_0e_k. Thus, up to eventual subsequences

$$B^\sharp e_k = B_0e_k + \lim[\,{}^tA_\varepsilon\nabla\psi_k^\varepsilon - {}^tA_0\nabla\psi_k^0] + \lim[\,{}^tB_\varepsilon(Y_k^\varepsilon - X_k^\varepsilon)]. \tag{3.10}$$

To return to the control problem that we started with, we have the following result.

THEOREM 3.3. *Set $A_\varepsilon \in \mathcal{M}(\alpha_m, \alpha_M, \Omega)$ and $B_\varepsilon \in \mathcal{M}(\beta_m, \beta_M, \Omega)$ with B_ε symmetric. Let $\theta_\varepsilon^\star$ be the optimal control for the problem whose state equation is given by*

$$\begin{cases} -\operatorname{div}(A_\varepsilon\nabla u_\varepsilon) = f + \theta & \text{in } \Omega, \\ u_\varepsilon = 0 \quad \text{on } \partial\Omega \end{cases}$$

for $\theta \in U_{ad}$, and let the cost function be given by

$$J_\varepsilon(\theta) = \frac{1}{2}\int_\Omega (B_\varepsilon\nabla u_\varepsilon, \nabla u_\varepsilon)dx + \frac{N}{2}\int_\Omega \theta^2 dx.$$

Then

$$\theta_\varepsilon^\star \to \theta_0^\star \qquad \text{strongly in } L^2(\Omega),$$

and $\theta_0^\star$ is the optimal control for the corresponding problem defined by the matrices A_0 and $B^\sharp$. We also have

$$J_\varepsilon(\theta_\varepsilon^\star) \to J_0(\theta_0^\star) \equiv \frac{1}{2}\int_\Omega (B^\sharp\nabla u_0^\star, \nabla u_0^\star)\; dx + \frac{N}{2}\int_\Omega (\theta_0^\star)^2\; dx.$$

Proof. It is again obvious that $\theta_\varepsilon^\star$ is bounded in $L^2(\Omega)$ and so converges to some $\theta_0^\star$ weakly in $L^2(\Omega)$. If $u_\varepsilon^\star$ is the corresponding state function and $p_\varepsilon^\star$ the adjoint state function, we can repeat the proof of Theorem 3.1. (The fact that θ is replaced by $\theta_\varepsilon^\star$ in the first equation poses no problem.) Thus we get that $u_\varepsilon^\star \rightharpoonup u_0^\star$ and $p_\varepsilon^\star \rightharpoonup p_0^\star$ weakly in $H_0^1(\Omega)$ and that

$$\begin{cases} -\operatorname{div}(A_0\nabla u_0^\star) = f + \theta_0^\star & \text{in } \Omega, \\ \operatorname{div}({}^tA_0\nabla p_0^\star - B^\sharp\nabla u_0^\star) = 0 & \text{in } \Omega. \end{cases}$$

Further, we have, for every $\theta \in U_{ad}$,

$$\int_\Omega (p_\varepsilon^\star + N\theta_\varepsilon^\star)(\theta - \theta_\varepsilon^\star)dx \geq 0,$$

and we easily pass to the limit to get

$$\int_\Omega (p_0^\star + N\theta_0^\star)(\theta - \theta_0^\star)dx \geq 0.$$

Here we used the fact that $\underline{\lim} \int_\Omega \theta_\varepsilon^2 \, dx \geq \int_\Omega \theta_0^{\star 2} \, dx$. Further (as will be proven in the next section) $B^\sharp$ is both elliptic and symmetric. This proves that $\theta_0^\star$ is the optimal control for the problem involving A_0 and $B^\sharp$.

Moreover

$$\theta_\varepsilon^\star = \mathrm{proj}_{U_{ad}}(-p_\varepsilon/N) \quad \text{and} \quad \theta_0^\star = \mathrm{proj}_{U_{ad}}(-p_0/N),$$

where $\mathrm{proj}_{U_{ad}}$ denotes the projection on the set U_{ad}. Since

$$p_\varepsilon \to p_0^\star \quad \text{strongly in } L^2(\Omega)$$

and the projection is an L^2-contraction, we deduce that

$$\theta_\varepsilon^\star \to \theta_0^\star \quad \text{strongly in } L^2(\Omega).$$

Now

$$\begin{aligned}
\int_\Omega (B_\varepsilon \nabla u_\varepsilon^\star, \nabla u_\varepsilon^\star) \, dx &= -\int_\Omega \mathrm{div}\ (B_\varepsilon \nabla u_\varepsilon^\star)\ u_\varepsilon^\star \, dx \\
&= -\int_\Omega \mathrm{div}(\ {}^t A_\varepsilon \nabla p_\varepsilon^\star)\ u_\varepsilon^\star \, dx \\
&= \int_\Omega\ {}^t A_\varepsilon \nabla p_\varepsilon^\star \ .\ \nabla u_\varepsilon^\star \, dx \\
&= \int_\Omega \nabla p_\varepsilon^\star \ .\ A_\varepsilon \nabla u_\varepsilon^\star \, dx \\
&= \int_\Omega (f + \theta_\varepsilon^\star)\ p_\varepsilon^\star \, dx \\
&\to \int_\Omega (f + \theta_0^\star)\ p_0^\star \, dx = \int_\Omega (B^\sharp \nabla u_0^\star, \nabla u_0^\star) \, dx.
\end{aligned}$$

The last equality follows by retracing the above steps. This, together with the strong convergence of $\theta_\varepsilon^\star$, gives the convergence of $J_\varepsilon(\theta_\varepsilon^\star)$ to $J_0(\theta_0^\star)$. □

To conclude this section we will compute the matrix $B^\sharp$ in some special cases.

Example 3.4 (the periodic case). Let Y denote the unit cube $[0,1]^n$ in $\mathbb{R}^n$. Choose $A \in \mathcal{M}(\alpha_m, \alpha_M, Y)$ and $B \in \mathcal{M}(\beta_m, \beta_M, Y)$ such that their coefficients are all periodic. We also assume that B is symmetric. We define A_ε and B_ε on $\mathbb{R}^n$ by extending A and B by periodicity on a grid of size ε. More precisely, we define

$$a_{ij}^\varepsilon(x) = a_{ij}(x/\varepsilon)\ ,\ b_{ij}^\varepsilon(x) = b_{ij}(x/\varepsilon)$$

on $[0,\varepsilon]^n$ and then extend them to the whole of $\mathbb{R}^n$ by periodicity. Restricting the functions to Ω, we thus get $A_\varepsilon \in \mathcal{M}(\alpha_m, \alpha_M, \Omega)$ and $B_\varepsilon \in \mathcal{M}(\beta_m, \beta_M, \Omega)$.

Defining X_k^ε, Y_k^ε, and ψ_k^ε as in (3.4)–(3.6), we observe the following:

(i) Since A_0 is now a constant matrix, $\operatorname{div}(A_0 e_k) = 0$. In fact, we can define X_k^ε by

$$X_k^\varepsilon = X_k(x/\varepsilon),$$

where

$$\begin{cases} \operatorname{div}(A\nabla(-X_k + y_k)) = 0 & \text{in } Y, \\ X_k \text{ is } Y\text{-periodic}, \ \int_Y X_k dy = 0. \end{cases} \tag{3.11}$$

(ii) Similarly, $Y_k^\varepsilon(x) = Y_k(x/\varepsilon)$, where

$$\begin{cases} \operatorname{div}\ (B\nabla(-Y_k + y_k)) = 0 & \text{in } Y, \\ Y_k \text{ is } Y\text{-periodic}, \ \int_Y Y_k dy = 0. \end{cases} \tag{3.12}$$

(iii) Finally $\psi_k^\varepsilon(x) = \psi_k(x/\varepsilon)$, where $\psi_k(y)$ satisfies

$$\begin{cases} \operatorname{div}({}^t A\nabla\psi_k + B\nabla(-X_k + y_k)) = 0 & \text{in } Y, \\ \psi_k \text{ is } Y\text{-periodic}, \ \int_Y \psi_k dy = 0. \end{cases} \tag{3.13}$$

In the above definitions, y_k denotes the projection onto the kth coordinate of $y \in Y$.

Since the means of the test function are zero, the limits of X_k^ε, $Y_k^\varepsilon, \psi_k^\varepsilon$ are all zero. In particular, $\psi_k^0 = 0$. Thus

$$\begin{aligned} (B^\sharp)_{jk} &= \lim\Big[b_{ij}^\varepsilon \frac{\partial}{\partial x_i}(-X_k^\varepsilon + x_k) + a_{ij}^\varepsilon \frac{\partial \psi_k^\varepsilon}{\partial x_i}\Big] \\ &= \int_Y \Big[b_{ij}(y)\ \frac{\partial}{\partial y_i}\ (-X_k + y_k) + a_{ij}(y)\frac{\partial \psi_k}{\partial y_i}\Big] dy \end{aligned}$$

since $f(x/\varepsilon) \rightharpoonup \int_Y f(y)dy$ in $L^\infty(\Omega)$-weak $*$ for Y-periodic f. Hence

$$(B^\sharp)_{jk} = \int_Y b_{ij}(y)\frac{\partial}{\partial y_i}(-Y_k + y_k)dy + \int_Y \Big[a_{ij}\frac{\partial \psi_k}{\partial y_i} - b_{ij}\frac{\partial(X_k - Y_k)}{\partial y_i}\Big] dy.$$

Now

$$\int_Y b_{ij}(y)\frac{\partial}{\partial y_i}\ (-Y_k + y_k)dy = \int_Y b_{il}\frac{\partial(-Y_k + y_k)}{\partial y_i}\ \frac{\partial(-Y_j + y_j)}{\partial y_l} dy = (B_0)_{jk}$$

as per classical computations in homogenization (see, for example, Bensoussan, Lions,

and Papanicolaou [1]). The second term in $(B^\sharp)_{jk}$ is evaluated as follows:

$$\begin{aligned}
\int_Y a_{ij}\frac{\partial \psi_k}{\partial y_i}dy &= \int_Y a_{il}\frac{\partial \psi_k}{\partial y_i}\ \frac{\partial y_j}{\partial y_l}\ dy \\
&= \int_Y a_{il}\frac{\partial \psi_k}{\partial y_i}\ \frac{\partial X_j}{\partial y_l}\ dy \quad \text{(using (3.11))} \\
&= \int_Y b_{il}\frac{\partial (X_k - y_k)}{\partial y_i}\ \frac{\partial X_j}{\partial y_l}\ dy \quad \text{(using (3.13))} \\
&= \int_Y b_{il}\frac{\partial (X_k - Y_k)}{\partial y_i}\ \frac{\partial X_j}{\partial y_l}\ dy \quad \text{(using (3.12))} \\
&= \int_Y b_{il}\ \frac{\partial}{\partial y_i}(X_k - Y_k)\frac{\partial}{\partial y_l}(X_j - Y_j)dy + \int_Y b_{il}\frac{\partial}{\partial y_i}(X_k - Y_k)\frac{\partial Y_j}{\partial y_l}\ dy \\
&= \int_Y b_{il}\frac{\partial}{\partial y_i}\ (X_k - Y_k)\frac{\partial}{\partial y_l}\ (X_j - Y_j)dy \\
&\quad + \int_Y b_{ij}\frac{\partial}{\partial y_i}\ (X_k - Y_k)dy \quad \text{(using (3.12))}.
\end{aligned}$$

Thus,

$$\int_Y a_{ij}\frac{\partial \psi_k}{\partial y_i}\ dy - \int_Y b_{ij}\frac{\partial}{\partial y_i}\ (X_k - Y_k)dy = \int_Y b_{il}\frac{\partial}{\partial y_i}\ (X_k - Y_k)\ \frac{\partial}{\partial y_l}\ (X_j - Y_j)\ dy.$$

Finally we get

$$(B^\sharp)_{jk} = (B_0)_{jk} + \int_Y b_{il}\ \frac{\partial}{\partial y_i}\ (X^k - Y^k)\ \frac{\partial}{\partial y_l}\ (X^j - Y^j)dy, \tag{3.14}$$

which was announced by Kesavan and Vanninathan [2]. It is immediate from the above form that $B^\sharp$ is both symmetric and elliptic. □

Example 3.5. (the layered material). This time we assume that $n = 2$, that A and B are diagonal, and that a_{ii} and b_{ii}, $(i = 1, 2)$ depend periodically on x_1 and do not depend on x_2. Thus $a^\varepsilon_{ii}(x) = a_{ii}(x_1/\varepsilon)$ and $b^\varepsilon_{ii}(x) = b_{ii}(x_1/\varepsilon)$, where $a_{ii}(y)$ and $b_{ii}(y)$ are periodic on $[0, 1]$. In this case (see Murat [4])

$$A_0 = \begin{pmatrix} \dfrac{1}{m(1/a_{11})} & 0 \\ 0 & m(a_{22}) \end{pmatrix}, \tag{3.15}$$

where $m(\ .\ )$ denotes the mean, i.e., the integral over $[0, 1]$.

It is now easy to see that $X^\varepsilon_1,\ Y^\varepsilon_1$, and ψ^ε_1 are functions in x_1 alone defined by

$$X^\varepsilon_1(x) = X_1\left(\frac{x_1}{\varepsilon}\right), \quad Y^\varepsilon_1(x) = Y_1\left(\frac{x_1}{\varepsilon}\right), \quad \psi^\varepsilon_1(x) = \psi_1\left(\frac{x_1}{\varepsilon}\right),$$

where

$$\begin{cases}
-\dfrac{d}{dy_1}(a_{11}(y_1)\dfrac{d}{dy_1}(-X_1 + y_1)) = 0 \quad \text{in }]0, 1[, \\
-\dfrac{d}{dy_1}(b_{11}(y_1)\dfrac{d}{dy_1}(-Y_1 + y_1)) = 0 \quad \text{in }]0, 1[, \\
-\dfrac{d}{dy_1}(a_{11}(y_1)\dfrac{d\psi_1}{dy_1} + b_{11}(y_1)\dfrac{d}{dy_1}(-X_1 + y_1)) = 0 \quad \text{in }]0, 1[.
\end{cases} \tag{3.16}$$

Also, $X_2^\varepsilon = Y_2^\varepsilon = \psi_2^\varepsilon = 0$. Thus we get

$$B^\sharp = \begin{pmatrix} \dfrac{m\Big(b_{11}/a_{11}^2\Big)}{m\Big(1/a_{11}\Big)^2} & 0 \\ 0 & m(b_{22}) \end{pmatrix} \tag{3.17}$$

as per usual calculations. See the one-dimensional case in section 2 for this form of $(B^\sharp)_{11}$. One can also directly obtain this formula from the previous example. ∎

Example 3.5 also illustrates the fact that the generalization of the one-dimensional formula proposed in section 2 does not work. Indeed

$$L_\varepsilon = A_\varepsilon B_\varepsilon^{-1} A_\varepsilon = \begin{pmatrix} (a_{11}^\varepsilon)^2/b_{11}^\varepsilon & 0 \\ 0 & (a_{22}^\varepsilon)^2/b_{22}^\varepsilon \end{pmatrix}.$$

Thus

$$L_0 = \begin{pmatrix} \dfrac{1}{m(b_{11}/a_{11}^2)} & 0 \\ 0 & m\Big(a_{22}^2/b_{22}\Big) \end{pmatrix}$$

and so

$$A_0 L_0^{-1} A_0 = \begin{pmatrix} \dfrac{m(b_{11}/a_{11}^2)}{(m(1/a_{11}))^2} & 0 \\ 0 & \dfrac{m(a_{22})^2}{m\Big(a_{22}^2/b_{22}\Big)} \end{pmatrix},$$

which is not equal to $B^\sharp$.

4. Properties of the matrix $B^\sharp$. In this section we will study some of the properties of the matrix $B^\sharp$. In particular we will see that the ellipticity and symmetry of the B_ε are preserved.

THEOREM 4.1. *Let $A_\varepsilon \in \mathcal{M}(\alpha_m, \alpha_M, \Omega)$ and $B_\varepsilon \in \mathcal{M}(\beta_m, \beta_M, \Omega)$. Let $B^\sharp$ be the matrix obtained as indicated in Theorem* 3.1. *Consider the problem*

$$\begin{cases} -\operatorname{div}(A_\varepsilon \nabla u_\varepsilon) = f + \theta & \text{in } \Omega, \\ \operatorname{div}({}^t A_\varepsilon \nabla \widetilde{p}_\varepsilon - {}^t B_\varepsilon \nabla u_\varepsilon) = 0 & \text{in } \Omega, \\ u_\varepsilon = \widetilde{p}_\varepsilon = 0 & \text{on } \partial\Omega. \end{cases} \tag{4.1}$$

Then

$$\widetilde{z}_\varepsilon \equiv {}^t A_\varepsilon \nabla \widetilde{p}_\varepsilon - {}^t B_\varepsilon \nabla u_\varepsilon \ \rightharpoonup\ {}^t A_0 \nabla \widetilde{p}_0 - {}^t(B^\sharp)\nabla u_0 \qquad \text{weakly in } L^2(\Omega).$$

Proof. Step 1. Clearly, by Theorem 3.1, we obtain a matrix $B_1^\sharp$ such that

$$\widetilde{z}_\varepsilon = {}^t A_\varepsilon \nabla \widetilde{p}_\varepsilon - {}^t B_\varepsilon \nabla u_\varepsilon \ \rightharpoonup {}^t A_0 \nabla \widetilde{p}_0 - B_1^\sharp \nabla u_0 = \widetilde{z}_0.$$

We have to show that $B_1^\sharp = {}^t(B^\sharp)$.

Let $F \in H^{-1}(\Omega)$ be an arbitrary distribution. We define $(w_\varepsilon, q_\varepsilon) \in (H_0^1(\Omega))^2$ by

$$\begin{cases} -\operatorname{div}(A_\varepsilon \nabla w_\varepsilon) = F & \text{in } \Omega, \\ \operatorname{div}({}^t A_\varepsilon \nabla q_\varepsilon - B_\varepsilon \nabla w_\varepsilon) = 0 & \text{in } \Omega. \end{cases} \tag{4.2}$$

If we set

$$\widehat{z}_\varepsilon = {}^t A_\varepsilon \nabla q_\varepsilon - B_\varepsilon \nabla w_\varepsilon,$$

then $\widehat{z}_\varepsilon \rightharpoonup \widehat{z}_0$ weakly in $L^2(\Omega)$ and

$$\widehat{z}_0 = {}^t A_0 \nabla q_0 - B^\sharp \nabla w_0,$$

where $q_\varepsilon \rightharpoonup q_0$ and $w_\varepsilon \rightharpoonup w_0$ weakly in $H_0^1(\Omega)$. We also know that

$$\begin{cases} \widehat{\xi}_\varepsilon = A_\varepsilon \nabla w_\varepsilon \rightharpoonup A_0 \nabla w_0 = \widehat{\xi}_0, \\ \xi_\varepsilon = A_\varepsilon \nabla u_\varepsilon \rightharpoonup A_0 \nabla u_0 = \xi_0 \end{cases} \tag{4.3}$$

weakly in $L^2(\Omega)$.

Step 2. Since $\operatorname{div}\widetilde{z}_\varepsilon = \operatorname{div}\widehat{z}_\varepsilon = 0$, by the div-curl lemma of compensated compactness (see Murat [5]) we have

$$\begin{cases} \widetilde{z}_\varepsilon \,.\, \nabla w_\varepsilon \to \widetilde{z}_0 \,.\, \nabla w_0 & \text{in } \mathcal{D}'(\Omega), \\ \widehat{z}_\varepsilon \,.\, \nabla u_\varepsilon \to \widehat{z}_0 \,.\, \nabla u_0 & \text{in } \mathcal{D}'(\Omega). \end{cases}$$

Thus

$$\widetilde{z}_\varepsilon \,.\, \nabla w_\varepsilon - \widehat{z}_\varepsilon \,.\, \nabla u_\varepsilon \longrightarrow \widetilde{z}_0 \,.\, \nabla w_0 - \widehat{z}_0 \,.\, \nabla u_0 \quad \text{in } \mathcal{D}'(\Omega). \tag{4.4}$$

We now evaluate the left-hand side of (4.4) in a different manner.

$$\begin{aligned} \widetilde{z}_\varepsilon \,.\, \nabla w_\varepsilon - \widehat{z}_\varepsilon \,.\, \nabla u_\varepsilon &= {}^t A_\varepsilon \nabla \widetilde{p}_\varepsilon \,.\, \nabla w_\varepsilon - {}^t A_\varepsilon \nabla q_\varepsilon \,.\, \nabla u_\varepsilon \\ &\quad - {}^t B_\varepsilon \nabla u_\varepsilon \,.\, \nabla w_\varepsilon + B_\varepsilon \nabla w_\varepsilon \,.\, \nabla u_\varepsilon \\ &= \nabla \widetilde{p}_\varepsilon \,.\, A_\varepsilon \nabla w_\varepsilon - \nabla q_\varepsilon \,.\, A_\varepsilon \nabla u_\varepsilon \\ &= \nabla \widetilde{p}_\varepsilon \,.\, \widehat{\xi}_\varepsilon - \nabla q_\varepsilon \,.\, \xi_\varepsilon \\ &\to \nabla \widetilde{p}_0 \,.\, \widehat{\xi}_0 - \nabla q_0 \,.\, \xi_0 \end{aligned}$$

once again by compensated compactness. Thus by (4.4) we get

$$\begin{aligned} \widetilde{z}_0 \,.\, \nabla w_0 - \widehat{z}_0 \,.\, \nabla u_0 &= \nabla \widetilde{p}_0 \,.\, \widehat{\xi}_0 - \nabla q_0 \,.\, \xi_0 \\ &= \nabla \widetilde{p}_0 \,.\, A_0 \nabla w_0 - \nabla q_0 \,.\, A_0 \nabla u_0 \\ &= {}^t A_0 \nabla \widetilde{p}_0 \,.\, \nabla w_0 - {}^t A_0 \nabla q_0 \,.\, \nabla u_0 \\ &= \widetilde{z}_0 \,.\, \nabla w_0 - \widehat{z}_0 \,.\, \nabla u_0 \\ &\quad + \; B_1^\sharp \nabla u_0 \,.\, \nabla w_0 - B^\sharp \nabla w_0 \,.\, \nabla u_0. \end{aligned}$$

Hence we have

$$B_1^\sharp \nabla u_0 \,.\, \nabla w_0 = B^\sharp \nabla w_0 \,.\, \nabla u_0 = {}^t(B^\sharp) \nabla u_0 \,.\, \nabla w_0 \quad \text{in } \mathcal{D}'(\Omega).$$

As can be observed, the matrices $B^\sharp$ and $B_1^\sharp$ depend only on the A_ε and B_ε. Thus we can choose u_0 and w_0 arbitrarily. In particular, if $\omega \subset\subset \Omega$ is any relatively compact open set, we can find, for any $\lambda, \mu \in \mathbb{R}^n$, elements u_0 and w_0 in $H_0^1(\Omega)$ such that

$$\nabla u_0 = \lambda \quad \text{and} \quad \nabla w_0 = \mu \quad \text{in } \omega.$$

Thus

$$(B_1^\sharp - {}^t(B^\sharp))\lambda \,.\, \mu = 0 \quad \text{in} \quad \mathcal{D}'(\omega),$$

and so for any $\omega \subset\subset \Omega$, we have $B_1^\sharp = {}^t(B^\sharp)$ in ω. Hence

$$B_1^\sharp = {}^t(B^\sharp) \quad \text{in} \quad \Omega,$$

which completes the proof. ∎

The above theorem immediately gives the following result.

COROLLARY 4.2. *If B_ε is symmetric for each $\varepsilon > 0$, then $B^\sharp$ is symmetric.* ∎

We now turn to the ellipticity of $B^\sharp$.

In the periodic case we see immediately that $B^\sharp$ has the same ellipticity constant as B_0, which is β_m itself. We will now see that this is the case even without the periodicity assumption.

THEOREM 4.3. *Let $B_\varepsilon \in \mathcal{M}(\beta_m, \beta_M, \Omega)$. Then the ellipticity constant of $B^\sharp$ is also β_m, so for some $\widetilde{\beta}_M > 0$, we have*

$$B^\sharp \in \mathcal{M}(\beta_m, \widetilde{\beta}_M, \Omega).$$

Proof. Let $F \in H^{-1}(\Omega)$ be an arbitrary distribution. We define

$$(w_\varepsilon, q_\varepsilon) \in (H_0^1(\Omega))^2$$

as in (4.2). Now

$$\begin{aligned} B_\varepsilon \nabla w_\varepsilon \,.\, \nabla w_\varepsilon &= {}^t A_\varepsilon \nabla q_\varepsilon \,.\, \nabla w_\varepsilon - \widehat{z}_\varepsilon . \nabla w_\varepsilon \\ &= \nabla q_\varepsilon \,.\, A_\varepsilon \nabla w_\varepsilon - \widehat{z}_\varepsilon \,.\, \nabla w_\varepsilon \\ &= \nabla q_\varepsilon \,.\, \widehat{\xi}_\varepsilon - \widehat{z}_\varepsilon \,.\, \nabla w_\varepsilon, \end{aligned}$$

where $\widehat{z}_\varepsilon$ and $\widehat{\xi}_\varepsilon$ are as in Step 1 of the proof of Theorem 4.1. Then, once again by compensated compactness,

$$\begin{aligned} B_\varepsilon \nabla w_\varepsilon \,.\, \nabla w_\varepsilon \xrightarrow{\mathcal{D}'(\Omega)} & \nabla q_0 \,.\, \widehat{\xi}_0 - \widehat{z}_0 \,.\, \nabla w_0 \\ &= \nabla q_0 \,.\, A_0 \nabla w_0 - \widehat{z}_0 \,.\, \nabla w_0 \\ &= {}^t A_0 \nabla q_0 \,.\, \nabla w_0 - \widehat{z}_0 \,.\, \nabla w_0 \\ &= B^\sharp \nabla w_0 \,.\, \nabla w_0. \end{aligned}$$

Thus if $\varphi \geq 0$ is any function in $\mathcal{D}(\Omega)$,

$$\begin{aligned} \beta_m \int_\Omega \varphi |\nabla w_0|^2 \, dx &\leq \beta_m \, \underline{\lim} \int_\Omega \varphi |\nabla w_\varepsilon|^2 \, dx \\ &\leq \lim \int_\Omega (B_\varepsilon \nabla w_\varepsilon \,.\, \nabla w_\varepsilon) \varphi \, dx \\ &= \int_\Omega (B^\sharp \nabla w_0 \,.\, \nabla w_0) \varphi \, dx. \end{aligned}$$

We can now choose, for any $\lambda \in \mathbb{R}^n$, an element w_0 such that $\nabla w_0 = \lambda$ on the support of φ so that

$$(B^\sharp \lambda, \lambda) \geq \beta_m |\lambda|^2 \quad \text{in } \mathcal{D}'(\Omega),$$

which proves the theorem. □

Remark 4.4. We said nothing about the upper bound $\widetilde{\beta}_M$ for the norm of $B^\sharp$. In the case of H-convergence the upper bound for the norm of B_0 is dominated by β_M^2/β_m, and in the case of symmetric matrices, by β_M itself. However, in the case of $B^\sharp$, the upper bound cannot be expressed in terms of the ellipticity constant and upper bound of B_ε alone, as can be seen even in the one-dimensional case.

Indeed, in the one-dimensional case, let us take $b_\varepsilon = b$ constant and a_ε periodic. Thus

$$B^\sharp = b\ \frac{m(1/a^2)}{(m(1/a))^2}.$$

Let a be a step function, say,

$$a = a_F\ \chi_F + a_G\ \chi_G,$$

where the sets F and G are such that $F \cup G = [0,1]$ and $\overset{\circ}{F} \cap \overset{\circ}{G} = \emptyset$. For any set E, we denote by χ_E its characteristic function and $|E|$ its measure. Then

$$m(1/a^2) = |F|\ \frac{1}{a_F^2} + |G|\ \frac{1}{a_G^2},$$
$$(m(1/a))^2 = \left(|F|\frac{1}{a_F} + |G|\ \frac{1}{a_G}\right)^2.$$

Thus

$$\frac{m(1/a^2)}{(m(1/a))^2} = \frac{|F|\ a_G^2 + |G|\ a_F^2}{(|F|\ a_G + |G|\ a_F)^2}.$$

If $a_G \to +\infty$, the above quantity tends to $\frac{1}{|F|}$. Thus by choosing a_G large and $|F|$ small we can always arrive at

$$B^\sharp >> b = \frac{\beta_M^2}{\beta_m}$$

(since $\beta_M = \beta_m = b$). □

The above considerations show that the upper bounds should depend on the constant α_m and α_M as well as on β_m and β_M. But we have no conjecture to offer on the form of this dependence.

Acknowledgments. The authors gratefully acknowledge the referee, who made useful suggestions including a simplification of the proof of Theorem 2.1 and remarks which helped us improve the paper.

REFERENCES

[1] A. Bensoussan, J. L. Lions, and G. Papanicolaou, *Asymptotic Analysis for Periodic Structures*, North Holland, Amsterdam, 1978.

[2] S. Kesavan and M. Vanninathan, *L'homogénéisation d'un problème de contrôle optimal*, C.R. Acad. Sci. Paris Ser. A-B, 285 (1977), pp. 441–444.

[3] J. L. Lions, *Sur le contrôle optimal de systèmes gouvernés par des équations aux dérivées partielles*, Dunod, Paris, 1968.
[4] F. Murat, *H-convergence, séminaire d'analyse fonctionnelle et numérique*, 1977/78, Alger, mimeographed notes.
[5] F. Murat, *Compacité par compensation*, Ann. Scuola Norm. Sup. Pisa Cl. Sci. (4), 5 (1978), pp. 489–567.
[6] L. Tartar, *Quelques remarques sur l'homogénéisation*, in Functional Analysis and Numerical Analysis, Proc. Japan-France Seminar 1976, H. Fujita, ed., Japanese Society for the Promotion of Science, 1978, pp. 468–482.

SIAM J. CONTROL OPTIM.
Vol. 35, No. 5, pp. 1574–1590, September 1997

008

LOCALLY DISTRIBUTED CONTROL AND DAMPING FOR THE CONSERVATIVE SYSTEMS*

KANGSHENG LIU†

Abstract. In this paper we note the equivalence between exact controllability and exponential stabilizability for an abstract conservative system with bounded control. This enables us to establish a frequency domain characterization for the exact controllability/uniform exponential decay property of second-order elastic systems, such as the wave equation and the Petrovsky equation, with (locally) distributed control/damping. A piecewise multiplier method for frequency domain is introduced. For several classes of PDEs on regions which are not necessarily smooth, we obtain a sufficient condition for the subregion on which the application of control/damping will yield the exact controllability/uniform exponential decay property. This result provides useful information for designing the location of controllers/dampers for distributed systems with a law of conservation.

Key words. conservative partial differential equation, exact controllability, damping, uniform exponential decay property

AMS subject classifications. 93B05, 93D15, 35B37, 35B40

PII. S0363012995284928

1. Introduction.

Many examples of conservative partial differential equations (CPDEs) modelling wave propagations, quantum phenomena, and mechanical vibrations can be found in the engineering and physics literature. Boundary control/damping for CPDEs has been studied extensively (cf. Russell [Ru1] and Lions [Li1], [Li2]). The consideration of locally distributed control of CPDEs seems to have been initiated by Lagnese [La] in 1983. The general question is how to choose the location of the control/damping subregion so that the exact controllability and/or uniform exponential decay property (UEDP) for a CPDE can be achieved. Much work has been done on this problem for special regions, such as one-dimensional, rectangular, and spherical cases (see Lagnese [La], Chen et al. [CFNS], Haraux [Ha1], [Ha2], Ho [Ho], Jaffard [J], Kim [Ki1], [Ki2], and Komornik [Ko]). For a general n-dimensional region with appropriately smooth boundary, Zuazua [Li2, Chap. 7], [Zu] showed that applying distributed control on an ϵ-neighborhood of a part of the boundary with certain geometric properties is sufficient for the exact controllability of the wave equation and the Petrovsky equation.

As indicated by Bardos, Lebeau, and Rauch [BLR1], Ralston's result [Ra] suggests that for exact controllability/UEDP, control/damping should be applied to a subregion which satisfies the "geometric optics condition," that is, that each ray, reflected at the boundary in the usual way, meets the subregion. This principle was also noted by Lagnese [La] in contrapositive form. Indeed, Bardos, Lebeau, and Rauch [BLR2] showed by means of microlocal analysis that the "geometric optics condition" is necessary and sufficient for exact controllability/UEDP of a second-order hyperbolic PDE with locally distributed control/damping when the coefficients and the boundary are of class C^∞. The C^∞ condition recently has been relaxed by Burq [Bu1], by proving that the C^3 condition is sufficient.

*Received by the editors April 21, 1995; accepted for publication (in revised form) June 20, 1996. This research was supported in part by the National Natural Science Foundation of China.
http://www.siam.org/journals/sicon/35-5/28492.html
†Center for Mathematical Sciences, Zhejiang University, Hangzhou, 310027, China (ksl@math.zju.edu.cn).

In practice, the problem of optimal location of the controllers/dampers for the system modelled by the CPDE is more significant. When choosing as the cost function for this optimal design problem the integral of the energy over infinite time, we must select the admissible locations of the controllers/dampers so that exact controllability/UEDP can be achieved. This also leads to our problem.

In this paper, by introducing the frequency domain inequality and the piecewise multiplier method for frequency domain, we get a sufficient condition for the subregion on which the application of control/damping will yield exact controllability/UEDP of any of the wave equation, the Petrovsky equation, and the Schrödinger equation on a not necessarily smooth region. We illustrate the applicability of our result with a simple example. Consider the wave equation on a triangular region. It follows readily from our result that applying control/damping on an ϵ-neighborhood (in the region) of a straight line passing through any vertex and across the triangle is sufficient for exact controllability/UEDP (see Remark 4.3 for other examples). This provides a sufficient number of admissible locations for the optimal location problem. We point out that this example is not covered under the cases studied in Bardos, Lebeau, and Rauch [BLR2], Burq [Bu1], and Zuazua [Li2, Chap. 7]. Our sufficient condition is formulated in terms of general sets and vectors in $\mathbb{R}^N$, like the result in Zuazua [Li2, Chap. 7]. The "geometric optics condition" is formulated in terms of complicated vector bundles. Thus, for a nondisk region, it is not easy to check whether the "geometric optics condition" is satisfied, although it is easy to make a conjecture directly, based on its meaning in optics.

In section 2 we exhibit in a semigroup framework the equivalence between controllability and stabilizability for a conservative system, as well as necessary and sufficient conditions for both. The regularity of the control steering the system from a smoother state to zero state is also discussed. In section 3, the results in section 2 are applied to a second-order elastic system with control/damping. A frequency domain characterization for exact controllability is given. We consider the wave equation in section 4, and the Schrödinger and Petrovsky equations in section 5. A sufficient geometric control condition and piecewise multiplier techniques are developed. It is discovered that exact controllability of the wave equation with Dirichlet boundary condition implies exact controllability of the Petrovsky equation with the simply supported boundary condition, and also that the latter is equivalent to exact controllability of the Schrödinger equation with Dirichlet boundary condition.

Some consequences of our results (Remarks 4.3(b), 5.4, and 5.5) give answers to several conjectures posed by Chen et al. in [CFNS]. We also point out that the frequency domain inequality introduced in this paper has been applied to the Maxwell equation and the Kirchhoff plate-like equation to establish exact internal controllability [Zh], [LY].

2. Equivalence between controllability and stabilizability; necessary and sufficient conditions. Let $\mathcal{H}$ and $\mathcal{U}$ be Hilbert spaces. Consider the control system $(\mathcal{A}, \mathcal{B})$,

$$y(u,t) = e^{t\mathcal{A}}y_0 + \int_0^t e^{(t-s)\mathcal{A}}\mathcal{B}u(s)ds, \tag{2.1}$$

where $\mathcal{A}$ generates a C_0-semigroup $e^{t\mathcal{A}}$ on $\mathcal{H}$, $\mathcal{B} \in \mathcal{L}(\mathcal{U};\mathcal{H})$, $y_0 \in \mathcal{H}$.

DEFINITION 2.1. *The system $(\mathcal{A}, \mathcal{B})$ is said to be exactly controllable on $[0,T]$ if for every $y_0, y_1 \in \mathcal{H}$ there exists $u(\cdot) \in L^2(0,T;\mathcal{U})$ such that $y(u,0) = y_0$, $y(u,T) = y_1$; it is said to be exponentially stabilizable if there exists $\mathcal{K} \in \mathcal{L}(\mathcal{H};\mathcal{U})$ such that $\mathcal{A}+\mathcal{B}\mathcal{K}$ generates an exponentially stable C_0-semigroup on $\mathcal{H}$.*

Remark 2.2. When $\mathcal{A}$ is the infinitesimal generator of a C_0-group on $\mathcal{H}$, (uniform) exact controllability on some $[0,T]$, defined in Definition 2.1, is equivalent to exact null-controllability on $[0,T]$ ($y_1 \equiv 0$ in Definition 2.1) and to exact controllability on $[0,\infty)$. (For every $y_0, y_1 \in \mathcal{H}$ there exist $T > 0$ and $u(\cdot) \in L^2(0,T;\mathcal{U})$ such that $y(u,0) = y_0$, $y(u,T) = y_1$; see Zabczyk's Proposition 1 in [Za].) Thus, in this case we can generally say that the system $(\mathcal{A},\mathcal{B})$ is exactly controllable.

THEOREM 2.3. *Let $\mathcal{A}^* = -\mathcal{A}$, $\mathcal{B} \in \mathcal{L}(\mathcal{U};\mathcal{H})$. Then the following propositions are equivalent:*

(a) *The system $(\mathcal{A},\mathcal{B})$ is exponentially stabilizable with an arbitrary prefixed exponential decay rate.*
(b) *The system $(\mathcal{A},\mathcal{B})$ is exponentially stabilizable.*
(c) *For every positive-definite self-adjoint $\mathcal{S} \in \mathcal{L}(\mathcal{U})$ the operator $\mathcal{A} - \mathcal{B}\mathcal{S}\mathcal{B}^*$ generates an exponentially stable C_0-semigroup on $\mathcal{H}$.*
(d) *The system $(\mathcal{A},\mathcal{B})$ is exactly controllable.*
(e) *(observability inequality) There exist $T, \delta > 0$ such that*

$$\int_0^T \|\mathcal{B}^* e^{t\mathcal{A}} y\|^2 dt \geq \delta \|y\|^2 \quad \forall\ y \in \mathcal{H}. \tag{2.2}$$

(f) *The following frequency domain condition holds:*

$$i\mathbb{R} \subset \rho(\mathcal{A} - \mathcal{B}\mathcal{B}^*), \quad \textit{the resolvent set of } \mathcal{A} - \mathcal{B}\mathcal{B}^*, \tag{2.3}$$

$$\sup\{\|(\lambda - \mathcal{A} + \mathcal{B}\mathcal{B}^*)^{-1}\| \mid \lambda \in i\mathbb{R}\} < +\infty. \tag{2.4}$$

Proof. (a)⇒(b) This is obvious.

(b)⇒(c) For every positive-definite self-adjoint $\mathcal{S} \in \mathcal{L}(\mathcal{U})$ we consider the following algebraic inner product Riccati equation in $\mathcal{H}$:

$$\langle Py, -\mathcal{A}z\rangle + \langle -\mathcal{A}y, Pz\rangle + \langle \mathcal{C}y, \mathcal{C}z\rangle - \langle P\mathcal{B}\mathcal{R}^{-1}\mathcal{B}^* Py, z\rangle = 0, \quad P \geq 0, \tag{2.5}$$

for any $y, z \in D(\mathcal{A})$, where $\mathcal{C} = \mathcal{S}^{\frac{1}{2}}\mathcal{B}^*, \mathcal{R}^{-1} = \mathcal{S}$. It is obvious that $P = I$ is a solution. By (b) there is a $\mathcal{K} \in \mathcal{L}(\mathcal{H};\mathcal{U})$ such that $\mathcal{A} + \mathcal{B}\mathcal{K}$, and therefore its adjoint operator $-\mathcal{A} + \mathcal{K}^*\mathcal{S}^{-\frac{1}{2}}\mathcal{C}$, generates an exponentially stable C_0-semigroup on $\mathcal{H}$. This says that $(\mathcal{C}, -\mathcal{A})$ is detectable. It follows from Zabczyk's Theorem 1^o in [Za] that $-\mathcal{A} - \mathcal{B}\mathcal{R}^{-1}\mathcal{B}^* I$, and, hence, its adjoint operator $\mathcal{A} - \mathcal{B}\mathcal{S}\mathcal{B}^*$, generates an exponentially stable C_0-semigroup on $\mathcal{H}$.

(c)⇒(d) This follows from Russell's "controllability via stabilizability" [Ru1], [Ru2] and also from Theorem 2.4 in this section.

(d)⇒(e) This follows from $\|e^{t\mathcal{A}}\| = 1$ and the well-known necessary and sufficient condition for exact controllability.

(e)⇒(a) See Slemrod [S].

(c)⇒(f)⇒(b) Proposition (f) is equivalent to exponential stability of the semigroup $e^{t(\mathcal{A} - \mathcal{B}\mathcal{B}^*)}$, from the result in Gearhart [Ge] or Huang [Hu] or Prüss [Pr]. □

THEOREM 2.4. *Let $\mathcal{A}$ be the infinitesimal generator of a C_0-group on $\mathcal{H}$, and $\omega_0(-\mathcal{A}) = \lim_{t\to+\infty} t^{-1} \ln \|e^{-t\mathcal{A}}\|$, the type of $e^{-t\mathcal{A}}$. Let $\mathcal{B} \in \mathcal{L}(\mathcal{U};\mathcal{H})$. Then the system $(\mathcal{A},\mathcal{B})$ is exactly controllable if and only if there exist $\mathcal{K} \in \mathcal{L}(\mathcal{H};\mathcal{U}), \mu > \omega_0(-\mathcal{A})$, and $M_\mu > 0$ such that*

$$\left\| e^{t(\mathcal{A} + \mathcal{B}\mathcal{K})} \right\| \leq M_\mu e^{-\mu t} \quad \forall\ t \geq 0. \tag{2.6}$$

Proof. Also see Slemrod [S] for the case of necessity. Let

$$J = J(T) = e^{-T\mathcal{A}}e^{T(\mathcal{A}+\mathcal{B}\mathcal{K})}. \tag{2.7}$$

From (2.6) there exists $T > 0$ large enough such that $\|J\| = \|J(T)\| < 1$. Thus $(I-J)^{-1} \in \mathcal{L}(\mathcal{H})$. For every $y_0 \in \mathcal{H}, t \in [0,T]$, set

$$z_1(t) = e^{t(\mathcal{A}+\mathcal{B}\mathcal{K})}(I-J)^{-1}y_0, \quad z_2(t) = e^{t\mathcal{A}}[I-(I-J)^{-1}]y_0. \tag{2.8}$$

From the formula for perturbations of C_0-semigroups [Pa] we have

$$y(t) \equiv z_1(t) + z_2(t) = e^{t\mathcal{A}}y_0 + \int_0^t e^{(t-s)\mathcal{A}}\mathcal{B}\mathcal{K}z_1(t)ds. \tag{2.9}$$

Observing that $y(0) = y_0$ and

$$y(T) = e^{T\mathcal{A}}[J(I-J)^{-1} + I - (I-J)^{-1}]y_0 = 0, \tag{2.10}$$

we conclude that the control

$$u(\cdot) = \mathcal{K}z_1(\cdot) = \mathcal{K}e^{\cdot(\mathcal{A}+\mathcal{B}\mathcal{K})}(I-J)^{-1}y_0 \in C([0,T];\mathcal{U}) \tag{2.11}$$

steers the system $(\mathcal{A},\mathcal{B})$ from the initial state y_0 to rest at time T. □

It should be noted that μ in (2.6) can be negative when $\omega_0(-\mathcal{A}) < 0$.

Theorem 2.4 is an improvement of a known result on exact controllability of the system with time reversibility (see Theorem 3.14 in the survey article [PZ]). Our proof is suggested by Russell's "controllability via stabilizability" method. Taking advantage of the structure of the control in (2.11), we get the following regularity result.

THEOREM 2.5. *Let $\mathcal{A}$ and $\mathcal{B}$ be the same as in Theorem* 2.4. *If the system $(\mathcal{A},\mathcal{B})$ is exactly controllable then, for every $y_0 \in D(\mathcal{A})$, there exists $u(\cdot) \in C^1([0,T_1];\mathcal{U})$ with $T_1 > 0$ independent of y_0 such that $y(u,T_1) = 0$ and $y(u,\cdot) \in C^1([0,T_1];\mathcal{H}) \cap C([0,T_1];D(\mathcal{A}))$ strictly satisfies the differential equation in $\mathcal{H}$,*

$$\begin{cases} \dot{y}(t) = \mathcal{A}y(t) + \mathcal{B}u(t), & t \in [0,T_1], \\ y(0) = y_0. \end{cases} \tag{2.12}$$

Proof. From Theorem 2.4 there exist $\mathcal{K} \in \mathcal{L}(\mathcal{H};\mathcal{U}), \mu > \omega_0(-\mathcal{A})$, and $M_\mu > 0$ such that (2.6) holds. Let $J(T)$ and $u(\cdot)$ be defined by (2.7) and (2.11), respectively. Since

$$\begin{aligned} &(\mu+\mathcal{A})J(T)(\mu+\mathcal{A})^{-1} \\ &= e^{-T\mathcal{A}}(\mu+\mathcal{A})(\lambda-\mathcal{A}-\mathcal{B}\mathcal{K})^{-1}e^{T(\mathcal{A}+\mathcal{B}\mathcal{K})}(\lambda-\mathcal{A}-\mathcal{B}\mathcal{K})(\mu+\mathcal{A})^{-1} \end{aligned}$$

for some $\lambda > -\mu$ and all $T > 0$, and since

$$(\mu+\mathcal{A})(\lambda-\mathcal{A}-\mathcal{B}\mathcal{K})^{-1}, \ (\lambda-\mathcal{A}-\mathcal{B}\mathcal{K})(\mu+\mathcal{A})^{-1} \in \mathcal{L}(\mathcal{H}), \tag{2.13}$$

there exists a fixed $T_1 > 0$, sufficiently large, such that

$$\|J_1\| < 1, \quad \|(\mu+\mathcal{A})J_1(\mu+\mathcal{A})^{-1}\| < 1 \tag{2.14}$$

where $J_1 = J(T_1)$. Thus, for every $y_0 \in D(\mathcal{A})$,

$$(I-J_1)^{-1}y_0 = (\mu+\mathcal{A})^{-1}[I-(\mu+\mathcal{A})J_1(\mu+\mathcal{A})^{-1}]^{-1}(\mu+\mathcal{A})y_0 \in D(\mathcal{A}), \tag{2.15}$$

and, therefore, the control

$$u(\cdot) = \mathcal{K}e^{\cdot(\mathcal{A}+\mathcal{B}\mathcal{K})}(I-J_1)^{-1}y_0 \in C^1([0,T_1];\mathcal{H}) \tag{2.16}$$

satisfies all the requirements from the proof of Theorem 2.4 and from the well-known result on the regularity of the mild solution of (2.12). □

3. Frequency domain characterization for the exact controllability of second-order systems. Let H be a Hilbert space with the norm $\|\cdot\|$ and the inner product $\langle\cdot,\cdot\rangle$. We consider the second-order system with control

$$\ddot{w}(t) + Aw(t) = Bu(t), \quad t > 0, \tag{3.1}$$

$$w(0) = w_0, \quad \dot{w}(0) = w_1, \tag{3.2}$$

where A is a positive-definite, self-adjoint operator with defined domain $D(A) \subset H$, and $B \in \mathcal{L}(\mathcal{U}; H)$. It is clear that $V = D(A^{\frac{1}{2}})$ is also a Hilbert space with norm $\|\cdot\|_V = \|A^{\frac{1}{2}}\cdot\|$. Introduce the state space

$$\mathcal{H} = D(A^{\frac{1}{2}}) \times H \tag{3.3}$$

with the inner product induced by the energy norm

$$\langle (f_1, g_1), (f_2, g_2)\rangle_{\mathcal{H}} = \langle A^{\frac{1}{2}} f_1, A^{\frac{1}{2}} f_2\rangle + \langle g_1, g_2\rangle \tag{3.4}$$

for $(f_j, g_j) := (f_j, g_j)^T \in \mathcal{H},\ j = 1, 2$. (Throughout this paper we omit the symbol of transposition and use a row vector to denote the actual column vector.) It is obvious that $\mathcal{H}$ is a Hilbert space with inner product (3.4). In $\mathcal{H}$, define

$$\mathcal{A} = \begin{bmatrix} 0 & I \\ -A & 0 \end{bmatrix} \quad \text{with} \quad D(\mathcal{A}) = D(A) \times D(A^{\frac{1}{2}}), \tag{3.5}$$

and define $\mathcal{B} :\ \mathcal{U} \to \mathcal{H}$,

$$\mathcal{B}u = (0,\ Bu) \quad \forall\ u \in \mathcal{U}, \tag{3.6}$$

$$y_0 = (w_0, w_1). \tag{3.7}$$

Then we have that $\mathcal{A}^* = -\mathcal{A},\ \mathcal{B} \in \mathcal{L}(\mathcal{U}, \mathcal{H})$. We also have the following lemma showing that (2.1) is the variation-of-parameters formula for the first-order system reduced from (3.1).

LEMMA 3.1. *Let $\mathcal{A}, \mathcal{B}, y_0$ be defined as above and $y(u,t) = (w(t), w_1(t))$ be given by* (2.1). *If $y_0 \in \mathcal{H}$ and $u(\cdot) \in L^2_{loc}(0,\infty;\mathcal{U})$, then*

$$w(\cdot) \in C^1([0,\infty); H) \cap C([0,\infty); V),\ \ \dot{w}(\cdot) = w_1(\cdot) \tag{3.8}$$

satisfies the variational evolution equation

$$\frac{d}{dt}\langle \dot{w}(t), v\rangle + \langle A^{\frac{1}{2}} w(t), A^{\frac{1}{2}} v\rangle = \langle Bu(t), v\rangle\ \ \forall\, v \in V,\ \ a.e.\ \ t \in (0,\infty), \tag{3.9}$$

with initial state (3.2). *The associated energy is conservative when $u(\cdot) = 0$, i.e.,*

$$E(t) \equiv \frac{1}{2}\left(\|A^{\frac{1}{2}} w(t)\|^2 + \|\dot{w}(t)\|^2\right) = E(0) \quad \forall\ t \geq 0. \tag{3.10}$$

Moreover, if $y_0 \in D(\mathcal{A})$ and $u(\cdot) \in C^1([0,\infty);\mathcal{U})$, then a classical solution of the Cauchy problem (3.1)–(3.2) *satisfies*

$$w(\cdot) \in C^2([0,\infty); H) \cap C^1([0,\infty); V) \cap C([0,\infty); D(A)). \tag{3.11}$$

Proof. The lemma follows straightforwardly from C_0-semigroup theory and a standard limit argument. We omit the details. □

For every positive-definite, self-adjoint $\mathcal{K} \in \mathcal{L}(\mathcal{U})$, it can also be verified that the closed-loop system in $\mathcal{H}$,

$$\dot{y}(t) = (\mathcal{A} - \mathcal{B}\mathcal{K}\mathcal{B}^*)y(t) = \begin{bmatrix} 0 & I \\ -A & -B\mathcal{K}B^* \end{bmatrix} y(t), \tag{3.12}$$

is the first-order reduction of the damped second-order system in H,

$$\ddot{w}(t) + B\mathcal{K}B^*\dot{w}(t) + Aw(t) = 0, \tag{3.13}$$

which results from setting $y(t) = (w(t), \dot{w}(t))$.

The second-order system (3.1) is said to be *exactly controllable in* $\mathcal{H}$ if its first-order reduction, $(\mathcal{A}, \mathcal{B})$, is exactly controllable. The damped system (3.13) is said to possess the UEDP of energy if there exist $\mu > 0, M_\mu > 0$ such that

$$E(t) \equiv \frac{1}{2}\left(\|A^{\frac{1}{2}}w(t)\|^2 + \|\dot{w}(t)\|^2\right) \leq M_\mu e^{-\mu t}E(0) \quad \forall\ t \geq 0 \tag{3.14}$$

for every solution $w(\cdot)$ of (3.13) with initial state $(w(0), \dot{w}(0)) \in \mathcal{H}$. Obviously, the system (3.13) possesses UEDP of energy if and only if $\mathcal{A} - \mathcal{B}\mathcal{K}\mathcal{B}^*$ generates an exponentially stable C_0-semigroup on $\mathcal{H}$. As an immediate consequence of Theorem 2.3, we have Theorem 3.2.

THEOREM 3.2. *The following statements are equivalent:*

(a) *The second-order system* (3.1) *is exactly controllable in* $\mathcal{H}$.

(b) *The damped system* (3.13) *possesses UEDP of energy.*

(c) *(observability inequality) There exist* $T, \delta > 0$ *such that*

$$\int_0^T \|B^*\dot{w}(t)\|^2 dt \geq 2\delta E(0) \tag{3.15}$$

for every solution $w(\cdot)$ *of* (3.1) *with* $u(\cdot) = 0$ *and the initial state* $(w(0), \dot{w}(0)) \in \mathcal{H}$.

To characterize the frequency domain condition we need the following lemma.

LEMMA 3.3. *Let* $\mathcal{A}_B \equiv \mathcal{A} - \mathcal{B}\mathcal{B}^* = \left[\begin{smallmatrix} 0 & I \\ -A & -BB^* \end{smallmatrix}\right]$, *and define, in* H, $\Delta(\lambda) = \lambda^2 + \lambda BB^* + A$ *with* $D(\Delta(\lambda)) = D(A)$ *for* $\lambda \in \mathbb{C}$. *Then*

(a) $\lambda \in \rho(\mathcal{A}_B)$ *if and only if* $\Delta(\lambda)$ *has inverse* $\Delta^{-1}(\lambda) \in \mathcal{L}(H)$. *Moreover, for* $\lambda \in \rho(\mathcal{A}_B)$,

$$(\lambda - \mathcal{A}_B)^{-1} = \begin{bmatrix} \Delta^{-1}(\lambda)(\lambda + BB^*) & \Delta^{-1}(\lambda) \\ \lambda\Delta^{-1}(\lambda)(\lambda + BB^*) - I & \lambda\Delta^{-1}(\lambda) \end{bmatrix}; \tag{3.16}$$

(b) $0 \in \rho(\mathcal{A}_B)$, *and* $\mathcal{A}_B$ *has compact resolvent whenever* A *has;*

(c) *Let* $\Lambda = i(-\alpha, \alpha) \subset \rho(\mathcal{A}_B)$ *for some* $\alpha > 0$. *If there exists a constant* $M > 0$ *such that*

$$\sup\left\{\|A^{\frac{1}{2}}\Delta^{-1}(\lambda)\|_{\mathcal{L}(H)} \mid \lambda \in \Lambda\right\} \leq M, \tag{3.17}$$

then there exists a corresponding constant $C = C(A, B, M) > 0$ *such that*

$$\sup\left\{\|(\lambda-\mathcal{A}_B)^{-1}\|_{\mathcal{L}(\mathcal{H})} \mid \lambda\in\Lambda\right\}\le C. \tag{3.18}$$

Proof. (a) For $\lambda\in\mathbb{C}$, $(f,g)\in\mathcal{H}$, $(u,v)\in D(\mathcal{A}_B)$, we consider the resolvent equation

$$(\lambda-\mathcal{A}_B)(u,v)=(f,g). \tag{3.19}$$

This equation is equivalent to

$$\begin{cases} v=\lambda u-f,\\ \Delta(\lambda)u=g+(\lambda+BB^*)f. \end{cases} \tag{3.20}$$

If $\lambda\in\rho(\mathcal{A}_B)$, then for every $(f,g)\in\mathcal{H}$ there exists a unique $(u,v)\in D(\mathcal{A}_B)=D(A)\times D(A^{\frac12})$ such that (3.19) holds and, hence, that (3.20) holds. Taking $f=0$, we find that $\mathrm{Ker}\Delta(\lambda)=\{0\}$ and $R(\Delta(\lambda))=H$. From the closed graph theorem we have $\Delta^{-1}(\lambda)\in\mathcal{L}(H)$, and therefore (3.16) follows from (3.20). Conversely, if $\Delta^{-1}(\lambda)\in\mathcal{L}(H)$, then (3.20) yields (3.16). Thus $(\lambda-\mathcal{A}_B)^{-1}\in\mathcal{L}(H), \lambda\in\rho(\mathcal{A}_B)$.

(b) This follows from $\Delta^{-1}(0)=A^{-1}$ and (3.16).

(c) If (3.17) holds, then we have

$$\|A^{\frac12}\Delta^{-1}(\lambda)f\|\le M\|f\| \quad\forall\ \lambda\in\Lambda,\ \ f\in H. \tag{3.21}$$

Let $u=\Delta^{-1}(\lambda)f$. Since

$$\lambda^2\|u\|^2+\langle\lambda BB^*u,u\rangle+\|A^{\frac12}u\|^2=\langle f,u\rangle,$$

we have

$$\begin{aligned}\|\lambda u\|^2&\le\|A^{\frac12}u\|^2+\|\lambda u\|\,\|BB^*A^{-\frac12}\|\,\|A^{\frac12}u\|+\|A^{-\frac12}\|\,\|A^{\frac12}u\|\,\|f\|\\ &\le\frac12\|\lambda u\|^2+\left(2+\frac12\|BB^*A^{-\frac12}\|^2\right)\|A^{\frac12}u\|^2+\frac14\|A^{-\frac12}\|^2\|f\|^2,\end{aligned}$$

where $\|\cdot\|$ also denotes the norm in $\mathcal{L}(H)$. Thus there is a positive constant $M_2=M_2(A,B,M)$ such that

$$\|\lambda\Delta^{-1}(\lambda)f\|\le M_2\|f\| \quad\forall\ \lambda\in\Lambda,\ \ f\in H. \tag{3.22}$$

Observing that

$$[I-\lambda\Delta^{-1}(\lambda)(\lambda+BB^*)]^*=[\Delta^{-1}(\lambda)A]^*=A\Delta^{-1}(\bar\lambda)\in\mathcal{L}(H),$$

we obtain

$$\|[\lambda\Delta^{-1}(\lambda)(\lambda+BB^*)-I]A^{-\frac12}f\|\le M\|f\| \quad\forall\ \lambda\in\Lambda,\ \ f\in H. \tag{3.23}$$

Let $u_1=\lambda\Delta^{-1}(\lambda)A^{-\frac12}f$. We then have

$$\begin{aligned}\|\lambda u_1\|&\le M\|f\|+\|\lambda\Delta^{-1}(\lambda)\|\,\|BB^*A^{-\frac12}f\|+\|A^{-\frac12}f\|\\ &\le(M+M_2\|BB^*A^{-\frac12}\|+\|A^{-\frac12}\|)\|f\|\end{aligned}$$

and

$$\lambda^2\|u_1\|^2 + \langle \lambda BB^* u_1, u_1\rangle + \|A^{\frac{1}{2}}u_1\|^2 = \langle \lambda A^{-\frac{1}{2}} f, u_1\rangle.$$

Therefore, there is a constant $M_1 = M_1(A, B, M) > 0$ such that

$$\|\lambda A^{\frac{1}{2}}\Delta^{-1}(\lambda)A^{-\frac{1}{2}} f\| = \|A^{\frac{1}{2}}u_1\| \le M_1\|f\| \quad \forall\ \lambda \in \Lambda,\ \ f \in H. \tag{3.24}$$

Combining (3.16), (3.21)–(3.24), we know that there exists $C = C(A, B, M) > 0$ such that (3.18) holds. The proof is complete. □

THEOREM 3.4. *The system* (3.1) *is exactly controllable in* $V \times H$ *if and only if there exists a constant* $\delta > 0$ *such that the following frequency domain inequality holds:*

$$\|(\omega^2 - A)u\| + \|\omega BB^* u\| \ge \delta\|A^{\frac{1}{2}}u\| \quad \forall\ \omega \in \mathbb{R},\ \ u \in D(A), \tag{3.25}$$

or equivalently, there exists $\delta' > 0$ *such that*

$$\|(\omega^2 - A)u\| + \|\omega BB^* u\| \ge \delta'\|\omega u\| \quad \forall\ \omega \in \mathbb{R},\ \ u \in D(A). \tag{3.26}$$

Proof. It is easy to verify the equivalence of (3.25) and (3.26) by a contradiction argument. The inequality (3.25) is also equivalent to

$$\|\Delta(i\omega)u\| \ge \delta_0\|A^{\frac{1}{2}}u\| \quad \forall\ \omega \in \mathbb{R},\ \ u \in D(A) \tag{3.27}$$

for some $\delta_0 > 0$, where Δ is the same as in Lemma 3.3. In fact, (3.27) implies (3.25) immediately. On the other hand, if (3.27) is not valid, then there exist $\omega_n \in \mathbb{R}, u_n \in D(A)$ with $\|A^{\frac{1}{2}}u_n\| = 1$ such that $\Delta(i\omega_n)u_n = (-\omega_n^2 + i\omega_n BB^* + A)u_n \to 0$ in H. Observing that

$$-\|\omega_n u_n\|^2 + 1 = \mathrm{Re}\langle \Delta(i\omega_n)u_n, u_n\rangle \to 0, \tag{3.28}$$

we have

$$\langle BB^*\omega_n u_n, \omega_n u_n\rangle = \mathrm{Im}\langle \Delta(i\omega_n)u_n, \omega_n u_n\rangle \to 0. \tag{3.29}$$

This implies that $\omega_n BB^* u_n \to 0$ in H. Hence, $(-\omega_n^2 + A)u_n \to 0$ in H and (3.25) is not valid, either. Thus, from Theorem 2.3, we need only prove that the inequality (3.27) is necessary and sufficient for (2.3)–(2.4) with $\mathcal{A}$ and $\mathcal{B}$ defined by (3.5)–(3.6).

From Lemma 3.3(a), (2.3)–(2.4) obviously implies (3.27). On the other hand, from Lemma 3.3(c) it suffices to show that if (3.27) holds, then $i\mathbb{R} \subset \rho(\mathcal{A}_B)$, where $\mathcal{A}_B = \mathcal{A} - \mathcal{B}\mathcal{B}^*$. Since $0 \in \rho(\mathcal{A}_B)$, there exists $\alpha_0 > 0$ such that $i(-\alpha_0, \alpha_0) \subset \rho(\mathcal{A}_B)$. Let

$$\Lambda_n = i\left(-\alpha_0 - \frac{n}{2C}\ ,\ \alpha_0 + \frac{n}{2C}\right),\ \ n = 0, 1, 2, \ldots, \tag{3.30}$$

where $C = C(A, B, \frac{1}{\delta_0})$ is determined by Lemma 3.3(c) with $\Lambda := \Lambda_0$. Suppose $\Lambda_n \subset \rho(\mathcal{A}_B)$. Then by (3.27) we have (3.18) with $\Lambda := \Lambda_n$ and C independent of n. For any $\lambda \in \Lambda_{n+1}$, there is $\mu \in \Lambda_n$ such that $|\lambda - \mu| < \frac{2}{3C}$. This, combined with (3.18), yields

$$\left[I + (\lambda - \mu)(\mu - \mathcal{A}_B)^{-1}\right]^{-1} \in \mathcal{L}(\mathcal{H}), \tag{3.31}$$

and hence,

$$(\lambda - \mathcal{A}_B)^{-1} = (\mu - \mathcal{A}_B)^{-1}\left[I + (\lambda - \mu)(\mu - \mathcal{A}_B)^{-1}\right]^{-1} \in \mathcal{L}(\mathcal{H}). \tag{3.32}$$

We have proved that $\Lambda_n \subset \rho(\mathcal{A}_B)$ implies that $\Lambda_{n+1} \subset \rho(\mathcal{A}_B)$. By the induction principle we know that

$$\Lambda_n \subset \rho(\mathcal{A}_B) \quad \forall\, n = 0, 1, 2, \ldots. \tag{3.33}$$

Therefore, $i\mathbb{R} \subset \rho(\mathcal{A}_B)$. The proof is complete. □

In the next two sections we will use the following comparison theorem for exponential stability of C_0-semigroups of contractions on *Hilbert* spaces.

THEOREM 3.5. *Suppose that L generates an exponentially stable C_0-semigroup of contractions on a Hilbert space H. If $F \in \mathcal{L}(H)$ satisfies*

(F1) $\mathrm{Re}\langle Fy, y\rangle \le 0 \ \forall\ y \in H$,

(F2) $\mathrm{Re}\langle Fy_n, y_n\rangle \to 0 \Rightarrow \|Fy_n\| \to 0$ *for any sequence* $\{y_n\}_{n\in\mathbf{N}}$ *in* H,

then the semigroup $e^{t(L+F)}$ is also exponentially stable.

We omit the proof because the frequency domain condition [Hu] can be verified by following the contradiction argument in [CFNS]. If F is symmetric and nonpositive, the conditions (F1) and (F2) are naturally satisfied.

4. Control and damping for the wave equation. Consider the wave equation with locally distributed control/damping

$$\ddot{w}(t) - \Delta w(t) = \chi_G(x)u(x,t) \quad \text{in} \quad \Omega \times \mathbb{R}^+, \tag{4.1}_c$$

$$\ddot{w}(t) + d(x)\dot{w}(t) - \Delta w(t) = 0 \quad \text{in} \quad \Omega \times \mathbb{R}^+ \tag{4.1}_d$$

with the boundary and initial conditions

$$w = 0 \quad \text{on} \quad \partial\Omega \times \mathbb{R}^+, \tag{4.2}$$

$$w(x,0) = w_0(x), \quad \dot{w}(x,0) = w_1(x), \quad x \in \Omega, \tag{4.3}$$

where Ω is a bounded open subset in $\mathbb{R}^N$ with the Lipschitz boundary $\partial\Omega$, $G \subset \Omega$ is an L-measurable set, $\chi_G(\cdot)$ is the characteristic function of the set G, $0 \le d(\cdot) \in L^\infty(\Omega)$. A measurable subset D of supp$d(\cdot)$ is said to be an effective damping region of $(4.1)_d$ if there is a constant $d_0 > 0$ such that $d(x) \ge d_0$ on D. The system $(4.1)_c$–(4.2) can be rewritten in the form (3.1) by setting

$$H = \mathcal{U} = L^2(\Omega), \quad Bu = \chi_G(\cdot)u(\cdot), \ \ u \in \mathcal{U},$$
$$A = -\Delta \quad \text{with} \quad D(A) = \{w \in H_0^1(\Omega) \mid \Delta w \in H\}.$$

We refer the reader to [A] and [Gr] for a discussion of Sobolev spaces. It is easy to see that $A^* = A \ge 0$, $V \equiv D(A^{\frac{1}{2}}) = H_0^1(\Omega)$, and

$$\|A^{\frac{1}{2}}v\| = \| \, |\nabla v| \, \| \ge C(\Omega)\|v\| \quad \forall\ v \in V, \tag{4.4}$$

where $C(\Omega)$ is the Poincaré constant. Thus, from the Rellich–Kondrachov embedding theorem we know that $A^{-\frac{1}{2}}, A^{-1}$ are compact operators on H. Moreover, setting $B_d \in \mathcal{L}(H)$, $B_d v = d(\cdot)v(\cdot)$ for $v \in H$, we can rewrite $(4.1)_d$–(4.2) as

$$\ddot{w}(t) + B_d\dot{w}(t) + Aw(t) = 0, \quad t > 0. \tag{4.5}$$

G is said to be a control region giving exact controllability of $(4.1)_c$–(4.2) if the corresponding system (3.1) is exactly controllable in $\mathcal{H} = V \times H$. An effective damping region D is said to be the one giving the UEDP of $(4.1)_d$–(4.2) if the corresponding (4.5) has the UEDP of energy, defined analogously by (3.14). Applying Theorems 3.2 and 3.5 we immediately obtain the following theorem.

THEOREM 4.1. *Every $D \supset G$ is an effective damping region giving UEDP of $(4.1)_d$–(4.2) if G is a control region giving exact controllability of $(4.1)_c$–(4.2).*

We now assume that the following geometric conditions hold on Ω and G:

(g, Ω): Ω is either convex or of class $C^{1,1}$;

(g, G): There exist open sets $\Omega_j \subset \Omega$ with Lipschitz boundary $\partial\Omega_j$ and points $x_0^j \in \mathbb{R}^N$, $j = 1, 2, \ldots, J$, such that $\Omega_i \cap \Omega_j = \emptyset$ for any $1 \le i < j \le J$ and

$$G \supset \Omega \bigcap \mathcal{N}_\epsilon \left[\left(\bigcup_{j=1}^{J} \Gamma_j \right) \bigcup \left(\Omega \setminus \bigcup_{j=1}^{J} \Omega_j \right) \right] \tag{4.6}$$

for some $\epsilon > 0$ where

$$\mathcal{N}_\epsilon[S] := \bigcup_{x \in S} \{ y \in \mathbb{R}^N \mid |y - x| < \epsilon \} \quad \text{for } S \subset \mathbb{R}^N, \tag{4.7}$$

$$\Gamma_j = \{ x \in \partial\Omega_j \mid (x - x_0^j) \cdot \nu^j(x) > 0 \} \tag{4.8}$$

with $\nu^j(x)$, the unit normal vector of $\partial\Omega_j$ at x pointing towards the exterior of Ω_j, defined a.e. on $\partial\Omega_j$ and being in $L^\infty(\partial\Omega_j;\ \mathbb{R}^N)$ (cf. [A], [Gr]).

THEOREM 4.2. *If the geometric conditions (g, Ω) and (g, G) are satisfied, then G is a control region giving exact controllability of $(4.1)_c$–(4.2).*

Remark 4.3. We list several pairs of regions, (Ω, G), which satisfy the geometric conditions (g, Ω) and (g, G).

(a) Let $x = (y, z)$ with $y \in \mathbb{R}^{n-1}$ and $z \in \mathbb{R}$, Ω^{-1} be an open subset in $\mathbb{R}^{n-1}$, $f_1(y)$ and $f_2(y)$ be real-valued functions defined on $\overline{\Omega^{-1}}$, $f_1(y) < f_2(y)$ for all $y \in \Omega^{-1}$, and $f_1(y) = f_2(y) = 0$ for all $y \in \partial\Omega^{-1}$. Let $\Omega = \{(y, z) \mid y \in \Omega^{-1},\ f_1(y) < z < f_2(y)\}$ satisfy (g, Ω), which implies $f_1, f_2 \in C^{0,1}_{loc}(\Omega^{-1}) \cap C(\overline{\Omega^{-1}})$. Then $G = \Omega \cap \mathcal{N}_\epsilon[\{(y, z) | z = 0\}]$ for any $\epsilon > 0$ satisfies (g, G) with $J = 2$, $\Omega_1 = \{(y, z) \mid y \in \Omega^{-1},\ f_1(y) < z < -\epsilon/4\}$, $\Omega_2 = \{(y, z) \mid y \in \Omega^{-1}, \epsilon/4 < z < f_2(y)\}$, and $x_0^j = (y_0, z_j)$, where $y_0 \in \mathbb{R}^{n-1}$, $z_1 < 0$ small enough, and $z_2 > 0$ large enough.

Let $G = \Omega \cap \mathcal{N}_\epsilon[S]$ with $S = \{(y, z) | z = 0\}$ and any $\epsilon > 0$. We note the invariability of (g, Ω) and (g, G) under translation and rotation of coordinates, and we show some specific cases of (a) as follows:

(a1) Ω is a convex subset in $\mathbb{R}^2$ and S is one of the longest diameters of Ω;

(a2) Ω is an ellipsoid and S is one of the principal planes of Ω;

(a3) Ω is a convex quadrilateral in plane and S is either one of the straight lines passing through opposite vertices of Ω.

(b) Let Ω be a rectangle and S be two straight lines which are parallel to adjacent sides of the rectangle, respectively, and which intersect at a point in Ω. Then $G = \Omega \cap \mathcal{N}_\epsilon[S]$, for any $\epsilon > 0$, satisfies (g, G) with $J = 4$, Ω_j being subrectangles consisting of $\partial\Omega$ and S, x_0^j being vertices of the rectangle. Thus from Theorems 4.1 and 4.2 we know that the conjecture posed in [CFNS, pp. 288–289] is true.

(c) When $J = 1$ and $\Omega_1 = \Omega$ is of class C^2, Theorem 4.2 was given by Zuazua [Li2, Chap. 7].

Let

$$x = (x_1, x_2, \dots, x_n), \quad x_0^j = (x_{01}^j, x_{02}^j, \dots, x_{0n}^j),$$
$$m^j = (m_1^j, m_2^j, \dots, m_n^j), \quad m_k^j = x_k - x_{0k}^j, \ \ k = 1, 2, \dots, n,$$
$$D_k u = \frac{\partial u}{\partial x_k}, \quad \nu^j = (\nu_1^j, \nu_2^j, \dots, \nu_n^j), \ \ j = 1, 2, \dots, J.$$

We use the summation convention for repeated lower indices k and l. Then we recall that Green's formula is valid on any bounded open subset Ω' of $\mathbb{R}^N$ with Lipschitz boundary (see [Gr]), i.e.,

$$\int_{\Omega'} (vD_k u + uD_k v)dx = \int_{\partial\Omega'} uv\nu_k' d\sigma \quad \forall \ u, v \in H^1(\Omega'). \tag{4.9}$$

For every real-valued $\phi^j \in C_0^\infty(\mathbb{R}^N)$ and complex-valued $u \in H^2(\Omega_j)$, we apply (4.9) to get

$$\begin{aligned} \mathrm{Re}\int_{\Omega_j} \phi^j u m_k^j D_k \bar{u} dx &= \frac{1}{2}\int_{\partial\Omega_j} \phi^j(\nu^j \cdot m^j)|u|^2 d\sigma \\ &\quad -\frac{1}{2}\int_{\Omega_j} (n\phi^j + m_k^j D_k\phi^j)|u|^2 dx, \\ \mathrm{Re}\int_{\Omega_j} \phi^j m_k^j D_l u D_l D_k \bar{u} dx &= \frac{1}{2}\int_{\partial\Omega_j} \phi^j(\nu^j \cdot m^j)|\nabla u|^2 d\sigma \\ &\quad -\frac{1}{2}\int_{\Omega_j} (n\phi^j + m_k^j D_k\phi^j)|\nabla u|^2 dx. \end{aligned} \tag{4.10}$$

Hence,

$$\begin{aligned} \mathrm{Re}\int_{\Omega_j} \phi^j(\Delta u) m_k^j D_k \bar{u} dx &= \int_{\Omega_j} \phi^j \left(\frac{n}{2} - 1\right)|\nabla u|^2 dx + \mathrm{Re} X_j \\ &+ \ \mathrm{Re}\int_{\partial\Omega_j} \phi^j \left[(\nu^j \cdot \nabla u) m_k^j D_k \bar{u} - \frac{1}{2}(\nu^j \cdot m^j)|\nabla u|^2\right] d\sigma, \end{aligned} \tag{4.11}$$

where

$$X_j = \int_{\Omega_j} m_k^j \left(\frac{1}{2} D_k\phi^j |\nabla u|^2 - D_l\phi^j D_l u D_k \bar{u}\right) dx, \quad j = 1, \dots, J. \tag{4.12}$$

Proof of Theorem 4.2. We shall verify the frequency domain inequality (3.25) for A and B defined in this section. From the regularity results on elliptic problems [Gr] we know that the geometric condition (g, Ω) implies

$$D(A) = H^2(\Omega) \cap H_0^1(\Omega). \tag{4.13}$$

If (3.25) fails, there exist $\omega_p \in \mathbb{R}$, $u_p \in D(A)$, $p \in \mathbf{N}$, with

$$\|A^{\frac{1}{2}} u_p\|^2 = \int_\Omega |\nabla u_p|^2 dx = 1 \tag{4.14}$$

such that $\|(\omega_p^2 - A)u_p\| + \|\omega_p BB^* u_p\| \to 0$ as $p \to \infty$; i.e.,

$$\int_G |\omega_p u_p|^2 \, dx \to 0, \tag{4.15}$$

$$\omega_p^2 u_p + \Delta u_p \equiv f_p \to 0 \quad \text{in } L^2(\Omega). \tag{4.16}$$

Then, (4.4), (4.14), and (4.16) imply

$$\left| \frac{\omega_p}{C(\Omega)} \right|^2 \geq \int_\Omega |\omega_p u_p|^2 \, dx \to 1. \tag{4.17}$$

Let

$$S = \left(\bigcup_{j=1}^J \Gamma_j \right) \bigcup \left(\Omega \setminus \bigcup_{j=1}^J \Omega_j \right), \quad Q = \mathcal{N}_{\epsilon_o}[S] \tag{4.18}$$

for some $0 < \epsilon_0 < \epsilon$. Choose a fixed $\xi \in C_0^\infty(\mathbb{R}^N)$ satisfying $0 \leq \xi \leq 1$, $\xi = 1$ on $\overline{Q}$, and $\text{supp}\xi \subset \mathcal{N}_\epsilon[S]$. Then we have $\Omega \cap \text{supp}\xi \subset \Omega \cap \mathcal{N}_\epsilon[S] \subset G$ and $\xi = 1$ on $G_0 \equiv \Omega \cap Q$. It follows from (4.14)–(4.17) that

$$\begin{aligned} \int_{G_0} |\nabla u_p|^2 dx &\leq \int_\Omega \xi |\nabla u_p|^2 dx \\ &\leq |\langle f_p, \, \xi u_p \rangle| + \int_G (\xi |\omega_p u_p|^2 + |(\nabla \xi \cdot \nabla u_p)\bar{u}_p|) \, dx \to 0, \end{aligned} \tag{4.19}$$

$$\lim_{p \to \infty} \int_{\Omega \setminus G_0} |\nabla u_p|^2 dx = \lim_{p \to \infty} \int_{\Omega \setminus G_0} |\omega_p u_p|^2 dx = 1. \tag{4.20}$$

For Ω_j we can choose $0 \leq \phi^j \in C_0^\infty(\mathbb{R}^N)$ such that $\phi^j = 1$ on $\overline{\Omega}_j \cap Q^C$ and $\text{supp}\phi^j \subset \mathcal{N}_{\epsilon_1}[S]^C$, $j = 1, \ldots, J$, for some $0 < \epsilon_1 < \epsilon_0$. Here the notation Y^C means the relative complement $\mathbb{R}^N \setminus Y$ of any subset Y of $\mathbb{R}^N$. It is then easy to see that

$$\begin{cases} \displaystyle\int_{\Omega_j} m_k^j D_k \phi^j |\omega_p u_p|^2 dx = \int_{\Omega_j \cap Q} m_k^j D_k \phi^j |\omega_p u_p|^2 dx \to 0, \\ X_{j,p} \to 0, \end{cases} \tag{4.21}$$

where $X_{j,p} := X_j$ is defined by (4.12) with $u := u_p$. Multiplying (4.16) by $\phi^j m_k^j D_k \bar{u}_p$, then integrating on Ω_j, by (4.10)–(4.15), (4.19), (4.21) and the characteristics of ϕ^j, we obtain

$$\begin{aligned} \int_{\Omega_j \cap Q^C} |\nabla u_p|^2 dx &= \frac{n}{2} \int_{\Omega_j \cap Q^C} (|\nabla u_p|^2 - |\omega_p u_p|^2) dx + o(1)_j \\ &+ \int_{\partial \Omega_j} \phi^j (\nu^j \cdot \nabla u_p) m_k^j D_k \bar{u}_p \, d\sigma - \frac{1}{2} \int_{\partial \Omega_j} \phi^j (\nu^j \cdot m^j)(|\nabla u_p|^2 - |\omega_p u_p|^2) d\sigma, \end{aligned} \tag{4.22}$$

where $o(1)_j \to 0$ as $p \to \infty$, $j = 1, 2, \ldots, J$. Let $S_j = \Gamma_j \cup (\partial \Omega_j \cap \Omega)$. Then $u_p = 0$ on $\partial \Omega_j \setminus S_j \subset \partial \Omega \cap \Gamma_j^C$. This, when combined with (4.8), leads to

$$\nabla u_p = (\nu^j \cdot \nabla u_p) \nu^j \quad \text{and} \quad (m^j \cdot \nu^j) \leq 0 \quad \text{a.e. on } \partial \Omega_j \setminus S_j. \tag{4.23}$$

Applying (4.23) and $\phi^j = 0$ on $S_j \subset S$, we know that (4.22) implies

$$\int_{\Omega_j \cap Q^C} |\nabla u_p|^2 dx \le \frac{n}{2} \int_{\Omega_j \cap Q^C} (|\nabla u_p|^2 - |\omega_p u_p|^2) dx + o(1)_j. \tag{4.24}$$

It is easy to verify that $\Omega = \bigcup_{j=1}^J (\Omega_j \cup G_0)$, $\Omega \setminus G_0 = \bigcup_{j=1}^J (\Omega_j \cap Q^C)$. Hence, it follows from (4.24) that

$$\begin{aligned} \int_{\Omega \setminus G_0} |\nabla u_p|^2 dx &= \sum_{j=1}^J \int_{\Omega_j \cap Q^C} |\nabla u_p|^2 dx \\ &\le \frac{n}{2} \int_{\Omega \setminus G_0} (|\nabla u_p|^2 - |\omega_p u_p|^2) dx + o(1). \end{aligned} \tag{4.25}$$

This contradicts (4.20). The proof is complete. □

From the interior regularity of the solution of an elliptic equation [GT], the argument from (4.14) to (4.22) is also valid for $S = \partial\Omega$, $\Omega_j = \Omega$ even if both (g, Ω) and (g, G) are false. Therefore, we have the following theorem.

THEOREM 4.4. *$\Omega \cap \mathcal{N}_\epsilon(\partial\Omega)$ is always a control region giving exact controllability of $(4.1)_c$–(4.2).*

On the other hand, we next give a negative result which shows that the location of the control region is more essential than its size for exact controllability.

THEOREM 4.5. *Let Ω_1, Ω_2 be bounded open sets in $\mathbb{R}^m, \mathbb{R}^n$, respectively, and $\Omega = \Omega_1 \times \Omega_2 \subset \mathbb{R}^{m+n}$. Let G_1 be any nonempty open subset of Ω_1 and let Ω_2 have Lipschitz boundary. Then a control region G satisfying $G \cap (G_1 \times \Omega_2) = \emptyset$ is never the one giving exact controllability of $(4.1)_c$–(4.2).*

Proof. It is well known that the eigenproblem

$$-\Delta_y \psi = \omega^2 \psi, \quad \omega \in \mathbb{R}, \psi \in H_0^1(\Omega_2), \tag{4.26}$$

has a sequence of solutions (ω_p^2, ψ_p) such that

$$\omega_p^2 \to +\infty \quad \text{and} \quad \int_{\Omega_2} |\psi_p|^2 dy = 1. \tag{4.27}$$

Choose a fixed $\phi(x) \in C_0^\infty(\Omega_1)$ such that

$$\text{supp}\phi \subset G_1 \quad \text{and} \quad \int_{\Omega_1} |\phi|^2 dx = 1. \tag{4.28}$$

Set $u_p(x, y) = \phi(x)\psi_p(y)$. We then have

$$\begin{aligned} u_p \in D(A), \ \ \|u_p\|_\Omega = 1, \ \ BB^* u_p = \chi_G u_p = 0 \ \ \text{in } \Omega, \\ \|(\omega_p^2 - A)u_p\|_\Omega^2 = \|(\omega_p^2 + \Delta_x + \Delta_y)u_p\|_\Omega^2 = \int_{\Omega_1} |\Delta_x \phi|^2 dx. \end{aligned}$$

Hence, the frequency domain inequality (3.26) fails. □

Remark 4.6. When $\dim \Omega_2 = n = 1$, Theorem 4.5 can be deduced from Haraux's Proposition 1.4.1 in [Ha3] by means of Lions's HUM [Li1], [Li2]. We see that the size of the control region G can be arbitrarily close to that of the whole region Ω in Theorem 4.5.

Remark 4.7. From Theorem 4.2 in this section and the result in [BLR2] or [Bu1] we can conclude that the condition (g, G) implies the "geometric optics condition" when $\partial\Omega$ is sufficiently smooth. However, we have not been able to prove this by using geometric argument, except when Ω is the disk. To our knowledge, the following problem remains open: is a control region G whose closure is contained in the open convex set Ω with smooth boundary never the one giving exact controllability of $(4.1)_c$–(4.2)?

Remark 4.8. Thanks to the fact that in applications the frequency domain inequality (3.25) involves only elliptic problems, the regularity condition (g, Ω) posed on Ω can be replaced by the condition that Ω is either a curvilinear polygon of class $C^{1,\alpha}$ or an open set in $\mathbb{R}^N$ of class $C^{1,\alpha}$ for some $0 < \alpha < 1$ (Bian, Chen, and Liu [BCL]).

5. Control and damping for the Schrödinger and Petrovsky equations. Consider the controlled/damped Schrödinger equation with Dirichlet boundary condition

$$(Sch)_c \qquad \begin{cases} \dot{w} + i\Delta w = \chi_G \cdot u & \text{in } \Omega \times \mathbb{R}^+, \\ w = 0 & \text{on } \partial\Omega \times \mathbb{R}^+, \end{cases}$$

$$(Sch)_d \qquad \begin{cases} \dot{w} + i\Delta w - d(x)w = 0 & \text{in } \Omega \times \mathbb{R}^+, \\ w = 0 & \text{on } \partial\Omega \times \mathbb{R}^+ \end{cases}$$

and the controlled/damped Petrovsky equation with the simply supported boundary condition

$$(Pet)_c \qquad \begin{cases} \ddot{w} + \Delta^2 w = \chi_G \cdot u & \text{in } \Omega \times \mathbb{R}^+, \\ w = \Delta w = 0 & \text{on } \partial\Omega \times \mathbb{R}^+, \end{cases}$$

$$(Pet)_d \qquad \begin{cases} \ddot{w} + d(x)\dot{w} + \Delta^2 w = 0 & \text{in } \Omega \times \mathbb{R}^+, \\ w = \Delta w = 0 & \text{on } \partial\Omega \times \mathbb{R}^+. \end{cases}$$

Using the same notation as in section 4 we can rewrite the four above equations as the following abstract equations in H:

$$(Sch)'_c \qquad \dot{w}(t) = iAw(t) + Bu(t), \quad t > 0,$$

$$(Sch)'_d \qquad \dot{w}(t) = (iA - B_d)w(t), \quad t > 0,$$

$$(Pet)'_c \qquad \ddot{w}(t) + A^2 w(t) = Bu(t), \quad t > 0,$$

$$(Pet)'_d \qquad \ddot{w}(t) + B_d\dot{w}(t) + A^2 w(t) = 0, \quad t > 0.$$

By an argument analogous to those in section 4, we first have the following theorem.

THEOREM 5.1. *Every $D \supset G$ is an effective damping region giving UEDP of $(Sch)_d$, respectively, $(Pet)_d$, if G is a control region giving exact controllability of $(Sch)_c$, respectively, $(Pet)_c$.*

We show the relationship between exact controllability of the two systems $(Pet)_c$, $(Sch)_c$ and $(4.1)_c$–(4.2) by the following more general theorem.

THEOREM 5.2. *Consider the abstract systems $(Sch)'_c$, $(Pet)'_c$, and* (3.1) *with the same operators A and B. Then exact controllability of* (3.1) *in $D(A^{\frac{1}{2}}) \times H$ implies*

exact controllability of $(Pet)'_c$ *in* $D(A) \times H$*, and the latter is equivalent to exact controllability of* $(Sch)'_c$*.*

Proof. If the system (3.1) is exactly controllable in $D(A^{\frac{1}{2}}) \times H$, then from Theorem 3.4 there exists a constant $\delta > 0$ such that

$$\|\omega^{-1}(\omega^2 - A)u\| + \|BB^*u\| \geq \delta\|u\| \quad \forall\ u \in D(A),\ \ 0 \neq \omega \in \mathbb{R}. \tag{5.1}$$

Observing that $\omega^4 - A^2 = (\omega^2 + A)(\omega^2 - A)$ and

$$|\omega|^{-1}(\omega^2 + A) \geq 2\|A^{-\frac{1}{2}}\|^{-1}I \quad \forall\ 0 \neq \omega \in \mathbb{R},$$

we have that (5.1) implies

$$\|\omega^{-2}(\omega^4 - A^2)u\| + \|BB^*u\| \geq \delta_1\|u\|, \quad \forall\ u \in D(A^2),\ \ 0 \neq \omega \in \mathbb{R}, \tag{5.2}$$

where $\delta_1 = \delta \min\{1, 2\|A^{-\frac{1}{2}}\|^{-1}\}$. Again, from Theorem 3.4, we see that the system $(Pet)'_c$ is exactly controllable in $D(A) \times H$.

According to Theorem 2.3(e), the system $(Sch)'_c$ is exactly controllable if and only if there exist $T, \delta > 0$ such that

$$\int_0^T \|B^* e^{t(iA)} w_0\|^2 dt \geq \delta\|w_0\|^2 \quad \forall\ w_0 \in H. \tag{5.3}$$

Inequality (5.3) is equivalent to

$$\int_0^T \|B^* A e^{t(iA)} w_0\|^2 dt \geq \delta\|Aw_0\|^2 \quad \forall\ w_0 \in D(A^2) \tag{5.4}$$

because $D(A^2)$ is dense in $D(A)$ with the norm $\|A \cdot\|$. Set $w(t) = e^{t(iA)}w_0$ for $w_0 \in D(A^2)$. Then $\dot{w}(t) = iAe^{t(iA)}w_0$ and $\ddot{w}(t) = -A^2 w(t)$ for $t \geq 0$. Therefore, from Theorem 3.2(c) we find that (5.4) holds for some $T, \delta > 0$ if $(Pet)'_c$ is exactly controllable in $D(A) \times H$. The inverse proposition follows readily from the inequality

$$\|\omega^{-2}(\omega^4 - A^2)u\| \geq \|i(\omega^2 - A)u\| \quad \forall\ u \in D(A^2),\ \ 0 \neq \omega \in \mathbb{R},$$

and the frequency domain characterization for exact controllability. ∎

Remark 5.3. Liu and Yu [LY] have shown by means of the frequency domain inequality (3.25) that exact controllability of the Kirchhoff plate equation with the simply supported-like boundary condition is equivalent to exact controllability of the wave equation with Dirichlet boundary condition.

Remark 5.4. From Theorems 4.4, 5.1, and 5.2 we know that the conjecture about $(Sch)_d$ posed in [CFNS, p. 284] is valid and that the assumption they made concerning the convexity of Ω is not necessary.

Remark 5.5. It follows from Theorems 5.1 and 5.2 and Jaffard's result [J] that the damped system $(Sch)_d$ has UEDP when Ω is a rectangle and when the effective damping region D has nonempty interior. (Compare with Theorem 4.5; also see Burq [Bu2] for exact boundary controllability in the absence of the "geometric optics condition.") This gives a negative answer to the conjecture posed in [CFNS, p. 279].

Remark 5.6. For the case in which B is a nonnegative self-adjoint operator on H, Theorem 3.2 was proved by Haraux [Ha1]. This special case is sufficient for Theorems 4.1 and 5.1.

Acknowledgment. The author would like to thank Professor Xunjing Li for his helpful discussions.

REFERENCES

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

[BCL] B. Bian, S. Chen, and K. Liu, *Exact Controllability of the Wave Equation on a Nonsmooth Region by Locally Distributed Controls*, preprint.

[BLR1] C. Bardos, G. Lebeau, and J. Rauch, *Sharp sufficient conditions for the observation, control, and stabilization of waves from the boundary*, SIAM J. Control Optim., 30 (1992), pp. 1024–1065.

[BLR2] C. Bardos, G. Lebeau, and J. Rauch, *Un example d'utilisation des notions de propagation pour le contrôle et la stabilisation des les problèmes hyperboliques*, Rend. Sem. Mat. Univ. Pol. Torino, 1988, Spec. Issue, (1989), pp. 11–31.

[Bu1] N. Burq, *Contrôle de l'équation des ondes dans des ouverts peu réguliers*, Preprint 1094, Centre de Mathématiques, Ecole Polytechnique, 91128 Palaiseau, France.

[Bu2] N. Burq, *Contrôle de l'équation des plaques en présence d'obstacles strictement convexes*, Mém. Soc. Math. France, 55 (1993), 126 pp.

[CFNS] G. Chen, S. A. Fulling, F. J. Narcowich, and S. Sun, *Exponential decay of energy of evolution equations with locally distributed damping*, SIAM J. Appl. Math., 51 (1991), pp. 266–301.

[Ge] L. M. Gearhart, *Spectral theory for contraction semigroups on Hilbert space*, Trans. Amer. Math. Soc., 236 (1978), pp. 385–394.

[Gr] P. Grisvard, *Elliptic Problems in Nonsmooth Domain*, Monographs Studies Math. 24, Pitman, Boston, MA, 1985.

[GT] D. Gilbarg and N. S. Trudinger, *Elliptic Partial Differential Equations of Second Order*, Springer-Verlag, New York, 1977.

[Ha1] A. Haraux, *Une remarque sur la stabilisation de certains systèms du deuxième ordre en temps*, Portugal Math., 46 (1989), pp. 245–258.

[Ha2] A. Haraux, *Séries lacunaires et contrôle semi-interne des vibrations d'une plaque rectangulaire*, J. Math. Pures Appl., 68 (1989), pp. 457–465.

[Ha3] A. Haraux, *On a completion problem in the theory of distributed control of wave equations*, in Nonlinear Partial Differential Equations and Their Applications, Collège de France Seminar Volume X, H. Brezis and J. L. Lions, eds., Longman, Harlow, UK, 1991.

[Ho] L. F. Ho, *Exact controllability of the one-dimensional wave equation with locally distributed control*, SIAM J. Control Optim., 28 (1990), pp. 733–748.

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

[J] S. Jaffard, *Contrôle interne exact des vibrations d'une plaque rectangulaire*, Portugal Math., 47 (1990), pp. 423–429.

[Ki1] J. Kim, *Exponential decay of the energy of a one-dimensional nonhomogeneous medium*, SIAM J. Control Optim., 29 (1991), pp. 368–380.

[Ki2] J. Kim, *Exact internal controllability of a one-dimensional aeroelastic plate*, Appl. Math. Optim., 24 (1991), pp. 99–111.

[Ko] V. Komornik, *On the exact internal controllability of a Petrowsky system*, J. Math. Pures Appl., 71 (1992), pp. 331–342.

[La] J. Lagnese, *Control of wave processes with distributed controls supported on a subregion*, SIAM J. Control Optim., 21 (1983), pp. 68–85.

[Li1] J. L. Lions, *Exact controllability, stabilization and perturbations for distributed systems*, SIAM Rev., 30 (1988), pp. 1–68.

[Li2] J. L. Lions, *Contrôlabilité Exacte, Perturbations et Stabilisation de Systèmes Distribués*, Tome 1, Rech. Math. Appl. 8, Masson, Paris, 1988.

[LY] K. Liu and X. Yu, *Equivalence between Exact Internal Controllability of the Kirchhoff Plate-like Equation and the Wave Equation*, preprint.

[Pa] A. Pazy, *Semigroups of Linear Operators and Applications to Partial Differential Equations*, Springer-Verlag, New York, 1983.

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

[PZ] A. J. Pritchard and J. Zabczyk, *Stability and stabilizability of infinite dimensional systems*, SIAM Rev., 23 (1981), pp. 25–52.

[Ra] J. V. Ralston, *Solution of the wave equation with localized energy*, Comm. Pure Appl. Math., 22 (1969), pp. 807–823.

[Ru1] D. L. Russell, *Controllability and stabilizability theory for linear PDE's: Recent progress and open questions*, SIAM Rev., 20 (1978), pp. 639–739.

[Ru2] D. L. Russell, *Exact boundary value controllability theorems for wave and heat processes in star-complemented regions*, in Differential Game and Control Theory, E. O. Roxin, P. T. Liu, and R. Sternberg, eds., Marcel Dekker, New York, 1974, pp. 291–319.

[S] M. Slemrod, *A note on complete controllability and stabilizability for linear control systems in Hilbert space*, SIAM J. Control, 12 (1974), pp. 500–508.

[Za] J. Zabczyk, *Remarks on the algebraic Riccati equation in Hilbert space*, Appl. Math. Optim., 2 (1976), pp. 251–258.

[Zh] Q. Zhou, *Exact Internal Controllability of Maxwell's Equations*, preprint UTMS95-60, University of Tokyo, Japan, Dec. 1995.

[Zu] E. Zuazua, *Contrôlabilité exacte en un temps arbitrairement petit de quelques modèles de plaques*, Appendix I in Contrôlabilité Exacte, Perturbations et Stabilisation de Systèmes Distribués, Tome 1, Rech. Math. Appl. 8, J. L. Lions, Masson, Paris, 1988.

SIAM J. CONTROL OPTIM.
Vol. 35, No. 5, pp. 1591–1613, September 1997

009

RAPID BOUNDARY STABILIZATION OF LINEAR DISTRIBUTED SYSTEMS*

VILMOS KOMORNIK†

Abstract. We prove that under rather general assumptions an exactly controllable problem is uniformly stabilizable with arbitrarily prescribed decay rates. Our approach is direct and constructive and avoids many of the technical difficulties associated with the usual methods based on Riccati equations. We give several applications for the wave equation and for Petrovsky systems.

Key words. observability, controllability, stabilizability by feedback, partial differential equation, wave equation, Petrovsky system

AMS subject classifications. 35L05, 35Q72, 93B05, 93B07, 93C20, 93D15

PII. S0363012996301609

1. Introduction.

Let Ω be a nonempty bounded open set in $\mathbb{R}^n$ having a boundary Γ of class C^2, and consider the following problem:

$$y'' - \Delta y = 0 \quad \text{in} \quad \Omega \times (0,\infty), \tag{1.1}$$

$$y(0) = y_0 \quad \text{and} \quad y'(0) = y_1 \quad \text{in} \quad \Omega, \tag{1.2}$$

$$y = u \quad \text{on} \quad \Gamma \times (0,\infty). \tag{1.3}$$

Considering u as a control function, a natural problem is to seek stabilizing feedback laws $u = F(y, y')$. In order to motivate our work, let us recall the following theorem of Lions [31, Theorem 10.1] (below, we shall denote by ν the outward unit normal vector to Γ).

THEOREM. *There exist two linear maps*

$$P : H^{-1}(\Omega) \to H_0^1(\Omega), \qquad Q : L^2(\Omega) \to H_0^1(\Omega) \tag{1.4}$$

and two positive constants M and ω such that, putting

$$u = \partial_\nu (Py' + Qy), \tag{1.5}$$

the problem (1.1), (1.2), (1.3), (1.5) *is well posed in $\mathcal{H} := L^2(\Omega) \times H^{-1}(\Omega)$, and its solutions satisfy the estimates*

$$\|(y(t), y'(t))\|_{\mathcal{H}} \le M \|(y_0, y_1)\|_{\mathcal{H}} e^{-\omega t} \tag{1.6}$$

for all $t \ge 0$ and for all $(y_0, y_1) \in \mathcal{H}$.

Several remarks are in order.

(a) This theorem was proved by applying a general and systematic approach based on the optimal control theory. Several other results of this type were also given in [31], and many more were provided by Lasiecka and Triggiani in [27]. See also [4], [5], and [42] for related results.

*Received by the editors April 8, 1996; accepted for publication June 25, 1996. Some results of this paper were announced earlier without proof in [*C. R. Acad. Sci. Paris Sér.* I *Math.*, 321 (1995), pp. 433–437, 581–586].

http://www.siam.org/journals/sicon/35-5/30160.html

†Institut de Recherche Mathématique Avancée, Université Louis Pasteur et C.N.R.S., 7, rue René Descartes, 67084 Strasbourg Cédex, France (komornik@math.u-strasbg.fr).

(b) This approach does not lead to any estimate of the constants ω and M: we have uniform exponential decay but with unknown speed.

(c) The existence of stabilizing operators P and Q was established without constructing them: they are obtained (in principle) by solving infinite-dimensional Riccati equations.

(d) Compared with the related problem of exact controllability, for which a simple and elegant approach (the Hilbert uniqueness method (HUM)) was introduced in [31], the above method of stabilization is technically much more involved. From a purely esthetic point of view, there is no reason for the stabilization problem to be much more difficult than the controllability problem.

In this paper we propose another approach for the stabilization, which is similar in spirit to the HUM. This method is as general as the former one; however, it provides stronger results with simpler proofs. For example, in section 4 this method will give the following improvement of Lions's theorem.

THEOREM. *Fix $\omega > 0$ arbitrarily (large). Then there exist two linear maps P and Q satisfying* (1.4) *and a positive constant M such that the problem* (1.1), (1.2), (1.3), (1.5) *is well posed in $\mathcal{H} := L^2(\Omega) \times H^{-1}(\Omega)$ and its solutions satisfy the estimates* (1.6) *for all $t \geq 0$ and for all $(y_0, y_1) \in \mathcal{H}$.*

The main novelty is that ω may be prescribed as large as we like. Furthermore, we shall construct the operators P and Q explicitly, and we shall also obtain an estimate of the constant M.

Let us remark at this point that many works were devoted to the construction of explicit boundary feedback laws and to the proof of exponential decay by the multiplier method or by microlocal analysis; see, e.g., [2], [3], [6], [14], [19], [24], [25], [31], [35], [39], [40]. Sometimes this approach led to good decay rate estimates with relatively elementary proofs; see, e.g., [16], [20], [21]. For example, for the wave equation above, the usual feedback law has the form

$$\begin{gathered} y = 0 \quad \text{on} \quad \Gamma_0 \times (0, \infty), \\ \partial_\nu y + ay + by' = 0 \quad \text{on} \quad \Gamma_1 \times (0, \infty), \end{gathered}$$

where Γ_0, Γ_1 is a partition of Γ and $a, b \in L^\infty(\Gamma_1)$ are given nonnegative functions. With this type of feedback (to our knowledge) no decay estimate was obtained with $\omega > 1/(2R)$, where R is the radius of the smallest open ball containing Ω (except the trivial one-dimensional case, however; see, e.g., the introduction of [19]). In any case, it follows from a recent result of Koch and Tataru [14] that in dimension $n \geq 2$ we cannot have arbitrarily large decay rate ω by using this type of feedback.

In order to explain the novelty of our approach, let us briefly recall Lions's proof of the above theorem. It consists of three steps.

First step: Observability of the dual problem. Consider the problem (1.1)–(1.3) with $u = 0$. Then the "observation" of the normal derivative of the solution during some time T allows one to determine the initial data y_0, y_1 provided T is sufficiently large. In technical terms, we have the equivalence of two norms:

$$\|(y_0, y_1)\|_{H_0^1(\Omega) \times L^2(\Omega)} \sim \|\partial_\nu y\|_{L^2(\Gamma \times (0,T))}.$$

Second step: Controllability of the primal problem. By duality, it follows from the preceding result that for any given initial data $(y_0, y_1) \in L^2(\Omega) \times H^{-1}(\Omega)$ there exists a function $u \in L^2(\Gamma \times (0, T))$ such that the solution of (1.1)–(1.3) satisfies

$$y(T) = y'(T) = 0 \quad \text{in} \quad \Omega.$$

Third step: Stabilization of the primal problem. Consider the "cost function"

$$J(y,u) := \int_0^\infty \int_\Omega y^2 \; dx \; dt + \int_0^\infty \int_\Gamma u^2 \; d\Gamma \; dt$$

defined on the set of pairs (y,u) satisfying (1.1)–(1.3). Thanks to the preceding step, this cost function has finite values, and hence its infimum is achieved at a unique pair (y,u). Applying the theory of optimal control, one can show that y and u are connected by (1.4) and (1.5). Finally, the estimate (1.6) follows by applying a theorem of Datko from the theory of semigroups.

Our main idea is to shortcut the implication chain

$$\text{observability} \Rightarrow \text{controllability} \Rightarrow \text{stabilization}$$

and to prove directly the implication

$$\text{observability} \Rightarrow \text{stabilization}.$$

As we shall see, our construction is analogous to the HUM, applied in [31] and [32] for the proof of the implication

$$\text{observability} \Rightarrow \text{controllability}.$$

The plan of the paper is as follows. For the reader's convenience and in order to compare our approach with the HUM, in the next section we recall briefly some basic notions and results (in the form needed later) on observability and controllability.

Section 3 is devoted to the proof of an abstract stabilization theorem.

In the rest of the paper (sections 4–7) we shall apply this theorem to the boundary stabilization of the wave equation and of a Petrovsky system.

2. Observability and controllability. A review. Consider a linear evolutionary problem

$$x' = Ax + Bu, \qquad x(0) = x_0, \tag{2.1}$$

where A is a densely defined, closed linear operator in some Hilbert space H, and B is a bounded linear operator of another Hilbert space G into $D(A^*)'$. Let us also consider the dual problem

$$\varphi' = -A^*\varphi, \qquad \varphi(0) = \varphi_0, \qquad \psi = B^*\varphi, \tag{2.2}$$

where A^*, B^* denote the adjoints of A and B. In control-theoretical terminology, B is a control operator, u is a control, and B^* is an observability operator.

Assume that the following three hypotheses are satisfied (we denote by G', H' the (anti)dual spaces of G and H).

(H1) The operator A^* generates a *group* e^{sA^*} in H';

(H2) There exist a *bounded* linear operator $E \in L(G,H)$ and a number $\lambda \in \mathbf{C}$ such that $B^* = E^*(A+\lambda I)^*$;

(H3) There exist two positive numbers T' and c' such that

$$\|\psi\|_{L^2(0,T';G')} \le c' \|\varphi_0\|_{H'}$$

for all $\varphi_0 \in D(A^*)$.

Let us explain these assumptions. In applications, hypothesis (H1) is usually satisfied for time-reversible problems.

Hypothesis (H2) can be considered as a weakening of the boundedness of B. Indeed, if B is bounded, then we may choose $E = (A + \lambda I)^{-1}B$, where $-\lambda$ is an arbitrary number in the resolvent set of A. In the bounded control case many results are available; see, e.g., [1], [11], [29], and [38]. On the other hand, in boundary control problems the operator B is usually unbounded, and this leads to serious difficulties. As was demonstrated in [27], the hypothesis (H2) is still satisfied for a large class of boundary control problems. (See also the applications in the second half of this paper.)

Finally, hypothesis (H3) is a regularity property: this corresponds to the so-called *direct inequality* in the terminology of HUM.

Let us begin by discussing the well-posedness of the problems (2.1) and (2.2).

It follows from (H1) that for every $\varphi_0 \in H'$ the equation in (2.2) has a unique solution $\varphi \in C(\mathbb{R}; H')$, and that $\varphi \in C^1(\mathbb{R}; H') \cap C(\mathbb{R}; D(A^*))$ if $\varphi_0 \in D(A^*)$. In particular, using (H2), we see that hypothesis (H3) is meaningful.

Furthermore, (H1)–(H3) imply the following strengthened version of (H3): for every $T > 0$, there exists a constant c_T such that the solutions of (2.2) satisfy the estimates

$$\|\psi\|_{L^2(-T,T;L^2(G'))} \le c_T \|\varphi_0\|_{H'} \tag{2.3}$$

for all $\varphi_0 \in D(A^*)$. Extending this inequality by continuity, we obtain that $\psi \in L^2_{loc}(\mathbb{R}; G')$ and ψ satisfies (2.3) for all $\varphi_0 \in H'$. This result also shows that the value of T' in (H3) has no importance: if it is satisfied for *some* $T' > 0$, then in fact it is satisfied for *every* $T' > 0$.

Next we *define* the solutions of (2.1) by transposition. Fix $x_0 \in H$ and $u \in L^2_{loc}(\mathbb{R}; G)$ arbitrarily. Multiply the equation in (2.1) by the solution φ of the equation in (2.2). Integrating by parts formally between 0 and $T \in \mathbb{R}$, we easily obtain the identity

$$\langle x(T), \varphi(T)\rangle_{H,H'} = \langle x_0, \varphi_0\rangle_{H,H'} + \int_0^T \langle u(s), \psi(s)\rangle_{G,G'}\, ds. \tag{2.4}$$

Hence we define a solution of (2.1) as a *continuous* function $x : \mathbf{R} \to H$ satisfying the identity (2.4) for all $\varphi_0 \in H'$ and for all $T \in \mathbb{R}$. This definition is justified by the following lemma.

LEMMA 2.1. *Given $x_0 \in H$ and $u \in L^2_{loc}(\mathbb{R}; G)$ arbitrarily, the problem* (2.1) *has a unique solution. Moreover, we have the estimates*

$$\|x\|_{L^\infty(-T,T;H)} \le c_T(\|x_0\|_H + \|u\|_{L^2(-T,T;G)}) \tag{2.5}$$

for all $T > 0$.

Proof. Thanks to the inequality (2.3), the right-hand side of (2.4) defines a bounded linear form of $\varphi_0 \in H'$. It follows from hypothesis (H1) that the map $\varphi_0 \mapsto \varphi(T)$ is an automorphism of H'; therefore the right-hand side of (2.4) is also a bounded linear form of $\varphi(T) \in H'$. Since $H'' = H$, we conclude the existence of a unique $x(T) \in H$ satisfying (2.4).

Since the bounded linear form of φ_0 defined by the right-hand side of (2.4) and the automorphism $\varphi_0 \mapsto \varphi(T)$ both depend continuously on $T \in \mathbb{R}$, the function $x : \mathbb{R} \to H$ is also continuous.

Finally, the estimates (2.5) follow easily from (2.4). □

Next we introduce some notions of controllability and observability.

DEFINITIONS. Let $T > 0$.

(1) The problem (2.1) is called *exactly null-controllable* in time T if for every given initial state $x_0 \in H$ there exists a function $u \in L^2_{loc}(\mathbb{R}; G)$ such that

$$\|u\|_{L^2(0,T;G)} \leq c\|x_0\|_H,$$

and that the solution of (2.1) satisfies the final condition $x(T) = 0$. (We say that the *control* u drives the system to rest in time T.)

(2) The problem (2.1) is called *exactly controllable* in time T if for every *pair* of data $x_0, x_1 \in H$ there exists a function $u \in L^2_{loc}(\mathbb{R}; G)$ such that

$$\|u\|_{L^2(0,T;G)} \leq c'(\|x_0\|_H + \|x_1\|_H),$$

and that the solution of (2.1) satisfies the final condition $x(T) = x_1$. (The control u drives the system from the initial state x_0 to the state x_1 in time T.)

(3) The problem (2.2) is called *exactly observable* in time T if

$$\|\varphi_0\|_{H'} \leq c''\|\psi\|_{L^2(0,T;G')} \tag{2.6}$$

for all $\varphi_0 \in D(A^*)$.

Observe that (2.6) is the inverse inequality to that in hypothesis (H3). Contrary to hypothesis (H3), in (2.6) the value of T is important. One can readily verify the existence of a number $0 \leq T_0 \leq \infty$ such that (2.6) is satisfied for all $T > T_0$ and is not satisfied for any $T < T_0$.

The following important result, essentially proved in [7], extends some classical results from the finite-dimensional case (see [29]) and from the case of bounded control operators (see [1]).

THEOREM 2.2. *Assume* (H1)–(H3). *Then, for any given $T > 0$, the following three properties are equivalent.*

(a) *The problem* (2.2) *is exactly observable in time T;*

(b) *The problem* (2.1) *is exactly null-controllable in time T;*

(c) *The problem* (2.1) *is exactly controllable in time T.*

Proof. (a) $\Rightarrow$ (b). Thanks to hypotheses (H1)–(H3) the formula

$$\langle \Lambda\varphi_0, \psi_0\rangle_{H,H'} := \int_0^T \langle JB^*e^{-sA^*}\varphi_0, B^*e^{-sA^*}\psi_0\rangle_{G,G'}\, ds,\ \varphi_0, \psi_0 \in D(A^*),$$

where $J : G' \to G$ denotes the canonical Riesz isomorphism, defines a positive definite self-adjoint operator Λ in $L(H', H)$. Applying the Riesz–Fréchet theorem it follows that Λ is an isomorphism of H' onto H. Now given $x_0 \in H$ arbitrarily, the control

$$u(s) := -JB^*e^{-sA^*}\Lambda^{-1}x_0$$

drives x_0 to rest in time T. Indeed, using the formula (2.4), we have

$$\langle x(T), \varphi(T)\rangle_{H,H'}$$

$$= \langle x_0, \varphi_0\rangle_{H,H'} + \int_0^T \langle u(s), \psi(s)\rangle_{G,G'}\, ds$$

$$= \langle x_0, \varphi_0\rangle_{H,H'} - \int_0^T \langle JB^*e^{-sA^*}\Lambda^{-1}x_0, B^*e^{-sA^*}\varphi_0\rangle_{G,G'}\, ds$$
$$= \langle x_0, \varphi_0\rangle_{H,H'} - \langle \Lambda\Lambda^{-1}x_0, \varphi_0\rangle_{H,H'}$$
$$= 0$$

for all $\varphi_0 \in H'$, and hence $x(T) = 0$. (Note that $\varphi(T)$ runs over H' if φ_0 does.)

(b) $\Rightarrow$ (c). Given $x_0, x_1 \in H$ arbitrarily, let u be a control driving $x_0 - e^{-TA}x_1$ to rest in time T. Since the zero control drives $e^{-TA}x_1$ to x_1 in time T, it follows that the control u drives x_0 to x_1 in time T.

(c) $\Rightarrow$ (a). Fix x_0 arbitrarily and let $u \in L^2(0,T;G)$ be a control satisfying

$$\|u\|_{L^2(0,T;G)} \le c\|x_0\|_H$$

and driving the system (2.1) to rest in time T. Given $\varphi_0 \in H'$ arbitrarily, by (2.4) the solution of (2.2) satisfies the equality

$$\langle x_0, \varphi_0\rangle_{H,H'} = -\int_0^T \langle u(s), \psi(s)\rangle_{G,G'}\, ds.$$

Hence

$$|\langle x_0, \varphi_0\rangle_{H,H'}| \le c\|x_0\|_H\|\psi\|_{L^2(0,T;G')}$$

for all $x_0 \in H$, and therefore

$$\|\varphi_0\|_{H'} \le c\|\psi\|_{L^2(0,T;G')},$$

proving the exact observability of (2.2). □

In [31] and [32] Lions developed a general and systematic approach for the study of exact controllability of linear distributed systems, the so-called HUM. It is based on the implication (a) $\Rightarrow$ (b) of the preceding theorem.

3. Observability and stabilizability. We continue to study the problems (2.1) and (2.2) under the assumptions (H1)–(H3). Moreover, we shall also assume that the problem (2.2) is exactly observable in some time $T > 0$; in other words, we assume that the following hypothesis (H4) is also satisfied.

(H4) There exist two positive numbers T and c such that

$$\|\varphi_0\|_{H'} \le c\|\psi\|_{L^2(0,T;G')}$$

for all $\varphi_0 \in D(A^*)$.

(This is the so-called *inverse inequality* in the terminology of HUM.)

Given an arbitrary real number $\omega > 0$, set $T_\omega = T + (2\omega)^{-1}$, $e_\omega(s) = e^{-2\omega s}$ if $0 \le s \le T$, and $e_\omega(s) = 2\omega e^{-2\omega T}(T_\omega - s)$ if $T \le s \le T_\omega$. It follows from assumptions (H1)–(H4) that the formula (see a note at the end of this paper)

$$\langle \Lambda_\omega\varphi_0, \psi_0\rangle_{H,H'} := \int_0^{T_\omega} e^\omega(s)\langle JB^*e^{-sA^*}\varphi_0, B^*e^{-sA^*}\psi_0\rangle_{G,G'}\, ds, \tag{3.1}$$

where $J : G' \to G$ denotes the canonical Riesz isomorphism, defines a positive definite self-adjoint operator $\Lambda_\omega \in L(H', H)$. Hence Λ_ω is an isomorphism of H' onto H, and the formula

$$\|x\|_\omega := \langle \Lambda_\omega^{-1}x, x\rangle_{H',H}^{1/2}$$

defines an *equivalent* norm in H.

THEOREM 3.1. *Assume* (H1)–(H4) *for some* $T > 0$. *Fix* $\omega > 0$ *arbitrarily and set*

$$F = -JB^*\Lambda_\omega^{-1}. \tag{3.2}$$

Then the operator $A + BF$ *generates a strongly continuous* group *in* H, *and the solutions of the closed-loop problem*

$$x' = Ax + BFx, \qquad x(0) = x_0 \tag{3.3}$$

satisfy the estimate

$$\|x(t)\|_\omega \le \|x_0\|_\omega e^{-\omega t}, \qquad \forall t > 0, \tag{3.4}$$

for all $x_0 \in H$ *and for all* $t \ge 0$.

Remark. The operator $A + BF$ is not an infinitesimal generator of a strongly continuous group in H: it is a densely defined restriction of such a generator, naturally related to the problem (3.3) in a sense explained in [8, pp. 99–100].

In order to explain the main ideas, let us first give a formal proof. (In fact, this proof is entirely correct in the finite-dimensional case.) We write $\Lambda_\omega \in L(H', H)$ in the following form:

$$\Lambda_\omega := \int_0^{T_\omega} e_\omega(s) e^{-sA} BJB^* e^{-sA^*}\, ds.$$

Fix $x_0 \in H$ arbitrarily and consider the solution of (3.2), (3.3). By a simple computation we obtain the following identity:

$$\frac{d}{dt}\langle \Lambda_\omega^{-1} x, x\rangle_{H',H} = \langle \Lambda_\omega^{-1} x, (A\Lambda_\omega + \Lambda_\omega A^* - 2BJB^*)\Lambda_\omega^{-1} x\rangle_{H',H}. \tag{3.5}$$

Since $2\omega e_\omega \le -e_\omega^1$, we have

$$\begin{aligned} A\Lambda_\omega + \Lambda_\omega A^* + 2\omega\Lambda_\omega &\le -\int_0^{T_\omega} \frac{d}{ds}\left(e_\omega(s) e^{-sA} BJB^* e^{-sA^*}\right) ds \\ &= BJB^*. \end{aligned} \tag{3.6}$$

Hence we obtain that

$$A\Lambda_\omega + \Lambda_\omega A^* - 2BJB^* \le -2\omega\Lambda_\omega.$$

(It means that the right-hand side minus the left-hand side is positive semidefinite.) Therefore, we deduce from the identity (3.5) the following inequality:

$$\frac{d}{dt}\langle \Lambda_\omega^{-1} x, x\rangle_{H',H} \le -2\omega \langle \Lambda_\omega^{-1} x, x\rangle_{H',H}.$$

Hence

$$\|x(t)\|_\omega^2 \le \|x_0\|_\omega^2 e^{-2\omega t}$$

for all $t \ge 0$, and the estimate (3.4) follows.

In the general case, the above proof is incorrect because even the strong solutions of (3.2), (3.3) are not sufficiently smooth so as to justify the computations. We shall overcome this difficulty by working with an equivalent integral equation.

We modify the equality (3.6) as follows. Fix $\varphi_0 \in D((A^*)^2)$ arbitrarily and consider the solution φ of (2.2). We have

$$-\|B^*\varphi_0\|_{G'}^2 = \int_0^{T_\omega} \frac{d}{ds}\left(e_\omega(s)\|B^*\varphi(s)\|_{G'}^2\right)\, ds$$

$$= \int_0^{T_\omega} e'_\omega(s)\|B^*\varphi(s)\|_{G'}^2\, ds - \langle \Lambda_\omega\varphi_0, A^*\varphi_0\rangle_{H,H'} - \langle A^*\varphi_0, \Lambda_\omega\varphi_0\rangle_{H',H},$$

and hence

$$\begin{aligned} &-\|B^*\varphi_0\|_{G'}^2 \\ &= \int_0^{T_\omega} e'_\omega(s)\|B^*\varphi(s)\|_{G'}^2\, ds - \langle \Lambda_\omega\varphi_0, A^*\varphi_0\rangle_{H,H'} - \langle A^*\varphi_0, \Lambda_\omega\varphi_0\rangle_{H',H} \end{aligned} \tag{3.7}$$

for all $\varphi_0 \in D(A^*)$ by an obvious density argument. Identifying H' with H, thanks to hypothesis (H3) there exists a nonnegative bounded self-adjoint operator $C \in L(H,H)$ (defined as a square root) such that

$$\|C\Lambda_\omega\varphi_0\|_H^2 = -\int_0^{T_\omega} e'_\omega(s)\|B^*\varphi(s)\|_{G'}^2\, ds \tag{3.8}$$

for all $\varphi_0 \in D(A^*)$. Then we conclude from (3.7) that Λ_ω satisfies the *algebraic Riccati equation*

$$A\Lambda_\omega + \Lambda_\omega A^* - BJB^* + \Lambda_\omega C^*C\Lambda_\omega = 0. \tag{3.9}$$

Thanks to hypothesis (H1)–(H3) we may apply a theorem of Flandoli [8, pp. 99–100] to conclude that Λ_ω^{-1} satisfies the *dual algebraic Riccati equation*

$$\Lambda_\omega^{-1}A + A^*\Lambda_\omega^{-1} - \Lambda_\omega^{-1}BJB^*\Lambda_\omega^{-1} + C^*C = 0 \tag{3.10}$$

in the following weak sense: the operator $A + BF = A - BJB^*\Lambda_\omega^{-1}$ generates (in a sense explained in [8]) a strongly continuous group $U(s)$ in H, and

$$\begin{aligned} \Lambda_\omega^{-1} &= U(t-s)^*\Lambda_\omega^{-1}U(t-s) \\ &+ \int_s^t U(\tau-s)^*(C^*C + \Lambda_\omega^{-1}BJB^*\Lambda_\omega^{-1})U(\tau-s)\, d\tau \end{aligned} \tag{3.11}$$

for all $t, s \in \mathbb{R}$. (In [8] the hypothesis (H2) is used with $\lambda = 0$, but his proof extends easily to the general case.)

Let us explain at least formally, for the convenience of the reader, the equivalence of (3.10) and (3.11). Fixing $s \in \mathbb{R}$ arbitrarily, and denoting by $f(t)$ the difference of the right-hand and left-hand sides of (3.11), we obviously have $f(s) = 0$. Hence (3.11) holds if and only if the derivative $f'(t)$ vanishes identically. Now, using the relation $U' = (A - BJB^*\Lambda_\omega^{-1})U$, a straightforward computation shows that

$$f'(t) = U(t-s)^*(\Lambda_\omega^{-1}A + A^*\Lambda_\omega^{-1} - \Lambda_\omega^{-1}BJB^*\Lambda_\omega^{-1} + C^*C)U(t-s).$$

Hence $f'(t)$ vanishes identically if and only if the equality (3.10) is satisfied.

Since we obviously have (see (3.8))

$$C^*C + \Lambda_\omega^{-1}BJB^*\Lambda_\omega^{-1} \geq C^*C \geq 2\omega\Lambda_\omega^{-1},$$

we deduce from (3.11) the inequality

$$\Lambda_\omega^{-1} \geq U(t-s)^* \Lambda_\omega^{-1} U(t-s) + 2\omega \int_s^t U(\tau - s)^* \Lambda_\omega^{-1} U(\tau - s)\ d\tau$$

for all $t \geq s$.

Now fix $x_0 \in H$ arbitrarily and consider the solution $x = x(t)$ of (3.2), (3.3). It follows from the preceding inequality that

$$\langle \Lambda_\omega^{-1} x(s), x(s) \rangle_{H',H} \geq \langle \Lambda_\omega^{-1} x(t), x(t) \rangle_{H',H} + 2\omega \int_s^t \langle \Lambda_\omega^{-1} x(\tau), x(\tau) \rangle_{H',H}\ d\tau$$

for all $t \geq s$. If we can infer from this estimate the inequality

$$\langle \Lambda_\omega^{-1} x(t), x(t) \rangle_{H',H} \leq \langle \Lambda_\omega^{-1} x_0, x_0 \rangle_{H',H} e^{-2\omega t} \tag{3.12}$$

for all $t \geq 0$, then the proof of Theorem 3.1 can be completed as above. Therefore, it only remains to verify the following simple lemma.

LEMMA 3.2. *Let $f : [0, +\infty) \to \mathbb{R}$ be a continuous function, satisfying for some constant $\omega > 0$ the inequalities*

$$f(s) \geq f(t) + 2\omega \int_s^t f(\tau)\ d\tau \tag{3.13}$$

for all $0 \leq s < t < +\infty$. Then

$$f(t) \leq f(0) e^{-2\omega t} \tag{3.14}$$

for all $t \geq 0$.

Proof of the lemma. Let us first assume that f is continuously differentiable. Dividing the inequality (3.13) by $t - s$ and letting $s \to t$ we obtain easily that

$$f'(t) + 2\omega f(t) \leq 0$$

for all $t \geq 0$. Hence the function $e^{2\omega t} f(t)$ is nonincreasing (its derivative is ≤ 0) and the estimate (3.14) follows.

In the general case we approximate the function f by the sequence of continuously differentiable functions f_k defined by

$$f_k(t) = k \int_t^{t+k^{-1}} f(s)\ ds, \qquad t \geq 0, \qquad k = 1, 2, \ldots.$$

Each function f_k satisfies the inequalities (3.13) with the same constant ω. Therefore $f_k(t) \leq f_k(0) e^{-2\omega t}$ for all $t \geq 0$ and for all k. Letting $k \to +\infty$, the lemma follows. □

COROLLARY 3.3. *Assume* (H1), (H2) *and assume that for some positive constants T, ω, c_1, and c_2 the following inequalities are satisfied:*

$$c_1 \|\varphi_0\|_{H'}^2 \leq \langle \Lambda_\omega \varphi_0, \varphi_0 \rangle_{H,H'} \leq c_2 \|\varphi_0\|_{H'}^2 \tag{3.15}$$

for all $\varphi_0 \in D(A^*)$. Then the operator $A - BJB^* \Lambda_\omega^{-1}$ generates (in the sense of [8]*) a strongly continuous* group *in H, and the solutions of the closed-loop problem*

$$x' = Ax - BJB^* \Lambda_\omega^{-1} x, \qquad x(0) = x_0$$

satisfy the estimate

$$\|x(t)\|_H \le (c_2/c_1)\|x_0\|_H e^{-\omega t} \qquad \forall t > 0, \tag{3.16}$$

for all $x_0 \in H$ *and for all* $t \ge 0$.

Proof. It follows from (3.15) that hypotheses (H3) and (H4) are also satisfied. Since Λ_ω is self-adjoint, (3.15) implies

$$c_1\|\varphi_0\|_{H'} \le \|\Lambda_\omega \varphi_0\|_{H'} \le c_2\|\varphi_0\|_{H'} \tag{3.17}$$

for all $\varphi_0 \in H'$. Therefore, we have

$$\|x_0\|_\omega^2 = \langle \Lambda_\omega^{-1} x_0, x_0\rangle_{H',H} \le \|\Lambda_\omega^{-1}\|\|x_0\|_H^2 \le c_1^{-1}\|x_0\|_H^2$$

and

$$\|x(t)\|_\omega^2 = \langle \Lambda_\omega^{-1} x(t), x(t)\rangle_{H',H} \ge c_1\|\Lambda_\omega^{-1}x(t)\|_{H'}^2 \ge c_1 c_2^{-2}\|x(t)\|_H^2.$$

Substituting these inequalities into (3.4), the inequality (3.16) follows.

Remarks. (1) Let us emphasize the following facts. First of all, given $\omega > 0$ *arbitrarily large*, the proof of Theorem 3.1 allows us to *construct* a feedback law leading to exponential decay with rate ω. Second, in the usual applications (as in the second half of this paper) the constants c_1, c_2 may be computed explicitly, and hence we can obtain explicit estimates of the type (3.16).

(2) A similar method was used earlier by Lukes [34] in the finite-dimensional case. His proof did not extend to the infinite-dimensional case.

(3) Subsequently, Slemrod [38] proved a variant of Theorem 3.1 by assuming instead of (H2) and (H3) that the operator B is bounded. (We recall that this assumption excludes the applications to boundary control.)

(4) There is another, earlier theorem of Lasiecka and Triggiani [27]: under the hypotheses (H1)–(H4) there exists a linear operator F from H into G such that $A+BF$ generates (in a weakened sense) a strongly continuous semigroup of bounded linear operators in H, and there are two positive constants M and ω such that the solutions of the closed-loop problem (3.3) satisfy the estimates

$$\|x(t)\|_H \le M\|x_0\|_H e^{-\omega t} \qquad \forall t > 0,$$

for all $x_0 \in H$ and for all $t \ge 0$. Unlike the results of this section, they did not obtain F explicitly, and they did not obtain estimates of ω and M. On the other hand, they proved the existence of feedback controls for a large class of cost functions. (See the next remark in this respect.)

(5) Let us note that the feedback F of Theorem 3.1 corresponds to the minimization of the cost function

$$\int_0^\infty \|Cx(t)\|_H^2 + \|u(t)\|_G^2 \, dt \tag{3.18}$$

on the set of pairs (x, u) satisfying (2.1). Indeed, applying the optimal control theory as in [27], [29], or [31], one can readily verify that this minimization problem leads to the feedback $F = -JB^*P$ where P is the (unique) solution of the algebraic Riccati equation

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

It follows from (3.10) that $P = \Lambda_\omega^{-1}$ satisfies this equation, and our assertion follows.

This remark also explains the simplicity of our proof as compared to the usual Riccati approach. Indeed, the Riccati approach works for a large class of cost functions. Being only interested in the construction of feedbacks with good decay properties, we have chosen the very particular cost function (3.18) (which is different from those applied in [31]). Since for this cost function the solution of the corresponding Riccati equation may be found explicitly, we can in fact avoid all difficulties related to the resolution of this equation in the infinite-dimensional case.

4. Application to the wave equation I. Let us consider the problem

$$\begin{cases} y'' - \Delta y = 0 \quad \text{in} \quad \Omega \times \mathbb{R}, \\ y = 0 \quad \text{on} \quad \Gamma_0 \times \mathbb{R}, \\ y = u \quad \text{on} \quad \Gamma_1 \times \mathbb{R}, \\ y(0) = y_0 \quad \text{and} \quad y'(0) = y_1 \quad \text{in} \quad \Omega, \end{cases} \tag{4.1}$$

where Ω is a bounded open set in $\mathbb{R}^n$, Γ_1 is an open subset of its boundary Γ, and $\Gamma_0 = \Gamma \backslash \Gamma_1$. We shall denote by ν the outward unit normal vector to Γ.

Let us assume that Γ is analytic and that Γ_1 satisfies the *geometrical control condition* of Bardos, Lebeau, and Rauch [2], [3]: there exists a positive number T such that every ray of geometrical optics in $\overline{\Omega}$ hits Γ_1 at a nondiffractive point in some time $\leq T$. We refer to [2], [3] for the detailed explication and analysis of this important condition, practically necessary and sufficient for the uniform stabilizability of hyperbolic problems. This condition is obviously satisfied if $\Gamma_1 = \Gamma$; it suffices to choose the diameter of Ω for T.

We shall prove the following theorem.

THEOREM 4.1. *Fix an arbitrarily large positive number ω. Then there exist two bounded linear maps*

$$P : H^{-1}(\Omega) \to H_0^1(\Omega), \qquad Q : L^2(\Omega) \to H_0^1(\Omega) \tag{4.2}$$

and a constant M such that, putting

$$u = \frac{\partial}{\partial_\nu}(Py' + Qy), \tag{4.3}$$

the problem (4.1), (4.3) *is well posed in $\mathcal{H} := L^2(\Omega) \times H^{-1}(\Omega)$ and its solutions satisfy the estimates*

$$\|(y, y')(t)\|_{\mathcal{H}} \leq M \|(y_0, y_1)\|_{\mathcal{H}} e^{-\omega t} \tag{4.4}$$

for all $t \geq 0$ and for all $(y_0, y_1) \in \mathcal{H}$.

Remarks. (1) Assume instead of the geometrical control condition that Γ is only of class C^2 but that there exists a point $x^0 \in \mathbb{R}^n$ such that Γ_1 contains all points x of Γ satisfying $(x - x^0) \cdot \nu(x) > 0$. Then the above theorem remains valid: we shall indicate the necessary modification during the proof.

(2) Theorem 4.1 improves (in the sense mentioned in the introduction) some earlier results of Lions [31] and Lasiecka and Triggiani [27].

Turning to the proof of the theorem, let us first consider the following (dual) problem:

$$\begin{cases} \xi'' - \Delta \xi = 0 \quad \text{in} \quad \Omega \times \mathbb{R}, \\ \xi = 0 \quad \text{on} \quad \Gamma \times \mathbb{R}, \\ \xi(0) = \xi_0 \quad \text{and} \quad \xi'(0) = \xi_1 \quad \text{in} \quad \Omega, \\ \psi = \partial_\nu \xi|_{\Gamma_1} \quad \text{in} \quad \mathbb{R}. \end{cases} \tag{4.5}$$

Putting $\varphi = (\xi, \xi')$, $\varphi_0 = (\xi_0, \xi_1)$, and introducing the linear operators A^* and B^* by the formulae

$$D(A^*) = D(B^*) = (H^2(\Omega) \cap H_0^1(\Omega)) \times H_0^1(\Omega),$$
$$A^*(\eta_0, \eta_1) = -(\eta_1, \Delta\eta_0),$$
$$B^*(\eta_0, \eta_1) = \partial_\nu \eta_0|_{\Gamma_1},$$

we may rewrite (4.5) in the form (2.2):

$$\varphi' = -A^*\varphi, \qquad \varphi(0) = \varphi_0, \qquad \psi = B^*\varphi.$$

Let us show that choosing $H' = H_0^1(\Omega) \times L^2(\Omega)$ and $G' = L^2(\Gamma_1)$, the assumptions (H1)–(H4) of Theorem 3.1 are satisfied. In what follows, we shall identify $L^2(\Omega)$ and $L^2(\Gamma_1)$ with their respective duals, so that $G := G'' = L^2(\Gamma_1)$ and $H := H'' = H^{-1}(\Omega) \times L^2(\Omega)$.

It is well known that A^* generates a group in H'; see, e.g., [33].

To prove (H2) let us introduce the Dirichlet map $D : L^2(\Gamma_1) \to L^2(\Omega)$ defined by

$$\begin{cases} -\Delta Du = 0 & \text{in} \quad \Omega, \\ Du = 0 & \text{on} \quad \Gamma_0, \\ Du = u & \text{on} \quad \Gamma_1. \end{cases}$$

It is well known (see, e.g., [33]) that D is a bounded linear map of $L^2(\Gamma_1)$ into $H^{1/2}(\Omega)$. Let us define a bounded linear map $E \in L(G, H)$ by the formula $Eu = (0, -Du)$, $u \in G$. Now, given any $(\eta_0, \eta_1) \in D(A^*)$, we have for every $u \in H_0^{3/2}(\Gamma_1)$ the following equality:

$$\begin{aligned} \langle E^*A^*(\eta_0, \eta_1), u\rangle_{G',G} &= \langle A^*(\eta_0, \eta_1), Eu\rangle_{H',H} \\ &= \langle (-\eta_1, -\Delta\eta_0), (0, -Du)\rangle_{H',H} \\ &= \int_\Omega (\Delta\eta_0) Du \, dx \\ &= \int_\Omega \eta_0(\Delta Du) \, dx + \int_\Gamma (\partial_\nu\eta_0) Du - \eta_0(\partial_\nu Du) \, d\Gamma. \end{aligned}$$

(The last step is correct because $u \in H_0^{3/2}(\Gamma_1)$ implies $Du \in H^2(\Omega)$.) Since $\Delta Du \equiv 0$ in Ω, $\eta_0 \equiv 0$ on Γ, $Du \equiv 0$ on Γ_0, and $Du \equiv u$ on Γ_1, we conclude that

$$\langle E^*A^*(\eta_0, \eta_1), u\rangle_{G',G} = \langle \partial_\nu\eta_0, u\rangle_{G',G} = \langle B^*(\eta_0, \eta_1), u\rangle_{G',G}.$$

Using the density of $H_0^{3/2}(\Gamma_1)$ in $G = L^2(\Gamma_1)$, the equality $B^* = E^*A^*$ follows.

Rewriting the assumptions (H3) and (H4) for the original problem (4.5), we have to prove the existence of a positive number T and of two positive constants c_1, c_2 such that for every $(\xi_0, \xi_1) \in (H^2(\Omega) \cap H_0^1(\Omega)) \times H_0^1(\Omega)$ the solution of (4.5) satisfies the inequalities

$$c_1\|(\xi_0, \xi_1)\|^2_{H_0^1(\Omega)\times L^2(\Omega)} \le \int_0^T \int_{\Gamma_1} |\partial_\nu \xi|^2 \, d\Gamma \, dt \le c_2\|(\xi_0, \xi_1)\|^2_{H_0^1(\Omega)\times L^2(\Omega)}. \tag{4.6}$$

The second inequality of (4.6) was proven in [26] and [30]. The first one was established in [12] under the stronger geometrical condition mentioned in Remark (2)

after the formulation of Theorem 4.1. The estimates of T were subsequently improved in [31] and [15]. Finally, the general case was settled by Bardos, Lebeau, and Rauch [2], [3].

Since hypotheses (H1)–(H4) are all satisfied, we may apply Theorem 3.1 with A and B defined by $A := A^{**}$ and $B := B^{**}$.

Let us explicitly write the resulting closed-loop problem (3.2), (3.3). Consider the solutions of (4.1) and (4.5). We have, at least formally,

$$\begin{aligned} 0 &= \int_0^T \int_\Omega (y'' - \Delta y)\xi \; dx \; dt \\ &= \Big[\int_\Omega y'\xi - y\xi' \; dx\Big]_0^T + \int_0^T \int_\Omega y(\xi'' - \Delta\xi) \; dx \; dt \\ &\quad + \int_0^T \int_\Gamma -(\partial_\nu y)\xi + y(\partial_\nu \xi) \; d\Gamma \; dt. \end{aligned}$$

Using (4.1) and (4.5) this equality reduces to

$$0 = \Big[\int_\Omega y'\xi - y\xi' \; dx\Big]_0^T + \int_0^T \int_{\Gamma_1} u(\partial_\nu \xi) \; d\Gamma \; dt.$$

Putting $x = (-y', y)$, $x_0 = (-y_1, y_0)$, $\varphi = (\xi, \xi')$, and $\varphi_0 = (\xi_0, \xi_1)$ this equality may be rewritten as (2.4). Since (4.5) is equivalent to (2.2), we conclude that (by definition) (4.1) is equivalent to (2.1).

Furthermore, writing the operator

$$\Lambda_\omega^{-1} : H^{-1}(\Omega) \times L^2(\Omega) \to H_0^1(\Omega) \times L^2(\Omega)$$

in the matrix form

$$\Lambda_\omega^{-1} = \begin{pmatrix} P & -Q \\ -R & S \end{pmatrix},$$

we have

$$u = -JB^*\Lambda_\omega^{-1} x = \frac{\partial}{\partial_\nu}(Py' + Qy)$$

on Γ_1, and (4.3) follows. □

Let us end this section by writing more explicitly the feedback constructed in our proof. Given $(\xi_0, \xi_1) \in H_0^1(\Omega) \times L^2(\Omega)$ arbitrarily, first we solve the problem (4.5), and then the problem

$$\begin{cases} z'' - \Delta z = 0 \quad \text{in} \quad \Omega \times (0, T_\omega), \\ z(T_\omega) = z'(T_\omega) = 0 \quad \text{in} \quad \Omega, \\ z = e_\omega(s)\psi \quad \text{on} \quad \Gamma \times (0, T_\omega). \end{cases}$$

Thanks to the inequalities (4.6) the formula

$$\Lambda_\omega(\xi_1, \xi_0) = (-z(0), z'(0))$$

defines an isomorphism of $H_0^1(\Omega) \times L^2(\Omega)$ onto $H^{-1}(\Omega) \times L^2(\Omega)$.

Now our feedback law is the following: for each $t > 0$ we set $u(t) = \partial_\nu \xi_0$ where $(\xi_1, \xi_0) = \Lambda_\omega^{-1}(-y(t), y'(t))$.

5. Application to the wave equation II. We consider in this section the problem

$$\begin{cases} y'' - \Delta y = 0 \quad \text{in} \quad \Omega \times \mathbb{R}, \\ \partial_\nu y = u \quad \text{on} \quad \Gamma \times \mathbb{R}, \\ y(0) = y_0 \quad \text{and} \quad y'(0) = y_1 \quad \text{in} \quad \Omega \end{cases} \tag{5.1}$$

in an open ball Ω of radius R in $\mathbb{R}^n$. As in the preceding section, we denote by Γ the boundary of Ω and by ν the outward unit normal vector to Γ.

It follows from a theorem of Graham and Russell [9] (see also [18] for another proof) that for any fixed $T > 2R$ the formula

$$(\xi_0, \xi_1) \mapsto \|\xi\|_{L^2(0,T;L^2(\Gamma))},$$

where ξ denotes the solution of the homogeneous problem

$$\begin{cases} \xi'' - \Delta \xi = 0 \quad \text{in} \quad \Omega \times \mathbb{R}, \\ \partial_\nu \xi = 0 \quad \text{on} \quad \Gamma \times \mathbb{R}, \\ \xi(0) = \xi_0 \quad \text{and} \quad \xi'(0) = \xi_1 \quad \text{in} \quad \Omega, \end{cases} \tag{5.2}$$

defines a Euclidean *norm* in $H^1(\Omega) \times L^2(\Omega)$. Moreover, all the norms $\|\cdot\|_{L^2(0,T;L^2(\Gamma))}$ ($T > 2R$) are equivalent. Completing $H^1(\Omega) \times L^2(\Omega)$ with respect to any of these norms, we obtain a Hilbert space which we shall denote by H'.

As usual, let us identify $L^2(\Omega)$ with its dual. By a theorem proved in [32, Théorème III.1.6, p. 167] we also have the inclusions

$$H^1(\Omega) \times L^2(\Omega) \subset H' \subset L^2(\Omega) \times (H^1(\Omega))',$$

and therefore (putting $H := H''$),

$$L^2(\Omega) \times H^1(\Omega) \subset H \subset (H^1(\Omega))' \times L^2(\Omega).$$

(In [32] H' is denoted by F.) Setting for commodity

$$\mathcal{H} = \{(y_0, y_1) \mid (y_1, y_0) \in H\},$$

we shall prove the following theorem.

THEOREM 5.1. *Given $\omega > 0$ arbitrarily large, there exists a bounded linear operator*

$$\mathcal{F} : \mathcal{H} \to L^2(\Omega) \tag{5.3}$$

and a constant M such that putting

$$u = \mathcal{F}(y, y') \tag{5.4}$$

the problem (5.1), (5.4) *is well posed in $\mathcal{H}$, and its solutions satisfy the estimates*

$$\|(y, y')(t)\|_{\mathcal{H}} \le M \|(y_0, y_1)\|_{\mathcal{H}} e^{-\omega t} \tag{5.5}$$

for all $t \ge 0$ and for all $(y_0, y_1) \in \mathcal{H}$.

Remarks. (1) Using the (algebraical and topological) inclusion

$$H^1(\Omega) \times L^2(\Omega) \subset \mathcal{H},$$

it follows that the restriction of $\mathcal{F}$ to $H^1(\Omega)\times L^2(\Omega)$ has the form $\mathcal{F}(y,y')=Py'+Qy$ with two suitable bounded linear operators $P: L^2(\Omega)\to L^2(\Omega)$ and $Q: H^1(\Omega)\to L^2(\Omega)$.

(2) Unlike the other applications of this paper, here the estimate (5.5) involves a nontraditional space $\mathcal{H}$, which is not characterized completely by using the usual Sobolev spaces. (But we have at least two-sided estimates of this space.) This is a well-known feature of control problems with Neumann action, due to the absence of direct and inverse inequalities involving the same Sobolev spaces; cf., e.g., [32].

(3) There are many former results where estimates of this type were obtained for general domains with particular choices of the operators P and Q; see, e.g., [6], [16], [24], [35], [36], [39], [40]. However, in these results we never had $\omega > 1/(2R)$ where R denotes the radius of the smallest ball containing Ω. (Estimates with $\omega = 1/(2R)$ were obtained in [16].) The fact that arbitrarily large decay rates can be achieved by boundary feedbacks of this type seems to be new.

Let us turn to the proof of Theorem 5.1. Consider the problem

$$\begin{cases} \xi'' - \Delta\xi = 0 \quad \text{in} \quad \Omega\times\mathbb{R}, \\ \partial_\nu \xi = 0 \quad \text{on} \quad \Gamma\times\mathbb{R}, \\ \xi(0)=\xi_0 \quad \text{and} \quad \xi'(0)=\xi_1 \quad \text{in} \quad \Omega, \\ \psi = \xi|_\Gamma \quad \text{in} \quad \mathbb{R}. \end{cases} \tag{5.6}$$

Putting $\varphi=(\xi,\xi')$, $\varphi_0=(\xi_0,\xi_1)$ and introducing the linear operators A^* and B^* by the formulas

$$\begin{gathered} D(A^*)=D(B^*)=\{\eta_0\in H^2(\Omega)\mid \partial_\nu\eta_0=0 \ \text{ on } \ \Gamma\}\times H^1(\Omega), \\ A^*(\eta_0,\eta_1)=-(\eta_1,\Delta\eta_0), \\ B^*(\eta_0,\eta_1)=\eta_0|_\Gamma, \end{gathered}$$

we may rewrite (5.6) in the form (2.2):

$$\varphi' = -A^*\varphi, \qquad \varphi(0)=\varphi_0, \qquad \psi = B^*\varphi.$$

It is well known (see, e.g., [33]) that A^* generates a strongly continuous group in $H^1(\Omega)\times L^2(\Omega)$. It follows from the above-mentioned result of Graham and Russell that A^* also generates a strongly continuous group in H'. Hence hypothesis (H1) is satisfied.

Next we show that choosing $G=L^2(\Gamma)$, the hypothesis (H2) is also satisfied. (We shall identify G with its dual.) For this we introduce the Neumann map $N: L^2(\Gamma)\to H^1(\Omega)$ defined by

$$\begin{cases} -\Delta Nu + Nu = 0 \qquad \text{in} \qquad \Omega, \\ \partial_\nu Nu = u \qquad \text{on} \qquad \Gamma, \end{cases}$$

and then the operator $E\in L(G,H)$ defined by $Eu=(Nu,Nu)$. (We note that N is in fact a bounded linear operator of $L^2(\Gamma)$ into $H^{3/2}(\Omega)$.)

Given $(\eta_0,\eta_1)\in D(A^*)$ and $u\in G\cap H^{1/2}(\Gamma)$ arbitrarily, we have

$$\begin{aligned}
&\langle E^*(A+I)^*(\eta_0,\eta_1),u\rangle_{G',G}\\
&=\langle (A+I)^*(\eta_0,\eta_1),Eu\rangle_{H',H}\\
&=\langle (\eta_0-\eta_1,\eta_1-\Delta\eta_0),(Nu,Nu)\rangle_{H',H}\\
&=\int_\Omega (\eta_0-\Delta\eta_0)Nu\ dx\\
&=\int_\Omega \eta_0(-\Delta Nu+Nu)\ dx+\int_\Gamma -(\partial_\nu\eta_0)Nu+\eta_0(\partial_\nu Nu)\ d\Gamma\\
&=\int_\Gamma \eta_0 u\ d\Gamma\\
&=\langle B^*(\eta_0,\eta_1),u\rangle_{G',G}.
\end{aligned}$$

Since $G\cap H^{1/2}(\Gamma)$ is dense in G, we conclude that $E^*(A+I)^*=B^*$.

Finally, hypotheses (H3) and (H4) are satisfied by the *definition* of H'.

We may now apply Theorem 3.1 with $A:=A^{**}$ and $B:=B^{**}$. Let us determine the resulting closed-loop problem. If y solves (5.1) and ξ solves (5.6), then (at least formally)

$$\begin{aligned}
0&=\int_0^T\int_\Omega (y''-\Delta y)\xi\ dx\ dt\\
&=\Big[\int_\Omega y'\xi-y\xi'\ dx\Big]_0^T+\int_0^T\int_\Omega y(\xi''-\Delta\xi)\ dx\ dt\\
&\quad+\int_0^T\int_\Gamma -(\partial_\nu y)\xi+y(\partial_\nu\xi)\ d\Gamma\ dt.
\end{aligned}$$

Using (5.1) and (5.6), it follows that

$$0=\Big[\int_\Omega y'\xi-y\xi'\ dx\Big]_0^T+\int_0^T\int_\Gamma -u\xi\ d\Gamma\ dt.$$

Putting $x=(y',-y)$, $x_0=(y_1,-y_0)$, $\varphi=(\xi,\xi')$, and $\varphi_0=(\xi_0,\xi_1)$, this is equivalent to (2.4). Since (5.6) is equivalent to (2.2), we conclude that (by definition) (5.1) is equivalent to (2.1).

It follows easily from the boundedness of the operator $\Lambda_\omega^{-1}:H\to H'$ and from the definition of B^* that $u=-JB^*\Lambda_\omega^{-1}x$ may also be written as $u=\mathcal{F}(y,y')$ for a suitable bounded linear operator $\mathcal{F}:\mathcal{H}\to L^2(\Omega)$. □

6. Application to a Petrovsky system I. Let us consider the problem

$$\begin{cases} y''+\Delta^2 y=0 \quad \text{in} \quad \Omega\times\mathbb{R},\\ y=\partial_\nu y=0 \quad \text{on} \quad \Gamma_0\times\mathbb{R},\\ y=0 \quad \text{and} \quad \partial_\nu y=u \quad \text{on} \quad \Gamma_1\times\mathbb{R},\\ y(0)=y_0 \quad \text{and} \quad y'(0)=y_1 \quad \text{in} \quad \Omega, \end{cases} \tag{6.1}$$

where Ω is a bounded open set in $\mathbb{R}^n$, Γ_1 is an open subset of its boundary Γ, $\Gamma_0=\Gamma\backslash\Gamma_1$, and ν denotes the outward unit normal vector to Γ.

Let us assume that Γ is of class C^4 and that there exists a point $x^0\in\mathbb{R}^n$ such that Γ_1 contains all points x of Γ which satisfy the inequality $(x-x^0)\cdot\nu(x)>0$.

We shall prove the the following theorem.

THEOREM 6.1. *Fix an arbitrarily large positive number ω. Then there exist two bounded linear maps*

$$P : H^{-2}(\Omega) \to H_0^2(\Omega), \qquad Q : L^2(\Omega) \to H_0^2(\Omega) \tag{6.2}$$

and a constant M such that, putting

$$u = \Delta(Py' + Qy), \tag{6.3}$$

the problem (6.1), (6.3) *is well posed in $\mathcal{H} := L^2(\Omega) \times H^{-2}(\Omega)$, and its solutions satisfy the estimates*

$$\|(y, y')(t)\|_{\mathcal{H}} \le M\|(y_0, y_1)\|_{\mathcal{H}} e^{-\omega t} \tag{6.4}$$

for all $t \ge 0$ and for all $(y_0, y_1) \in \mathcal{H}$.

Remark. Theorem 6.1 improves some earlier results of Lasiecka and Triggiani [27], obtained by the Riccati equation approach.

Turning to the proof of the theorem, let us first consider the following problem:

$$\begin{cases} \xi'' + \Delta^2 \xi = 0 \quad \text{in} \quad \Omega \times \mathbb{R}, \\ \xi = \partial_\nu \xi = 0 \quad \text{on} \quad \Gamma \times \mathbb{R}, \\ \xi(0) = \xi_0 \quad \text{and} \quad \xi'(0) = \xi_1 \quad \text{in} \quad \Omega, \\ \psi = \Delta\xi|_{\Gamma_1} \quad \text{in} \quad \mathbb{R}. \end{cases} \tag{6.5}$$

Putting $\varphi = (\xi, \xi')$, $\varphi_0 = (\xi_0, \xi_1)$ and introducing the linear operators A^* and B^* by the formulas

$$\begin{gathered} D(A^*) = D(B^*) = (H^4(\Omega) \cap H_0^2(\Omega)) \times H_0^2(\Omega), \\ A^*(\eta_0, \eta_1) = (-\eta_1, \Delta^2 \eta_0), \\ B^*(\eta_0, \eta_1) = \Delta\eta_0|_{\Gamma_1}, \end{gathered}$$

we may rewrite (6.5) in the abstract form (2.2):

$$\varphi' = -A^*\varphi, \qquad \varphi(0) = \varphi_0, \qquad \psi = B^*\varphi.$$

We are going to show that, choosing $H' = H_0^2(\Omega) \times L^2(\Omega)$ and $G' = L^2(\Gamma_1)$, the assumptions (H1)–(H4) of Theorem 3.1 are satisfied. As before, we identify $L^2(\Omega)$ and $L^2(\Gamma_1)$ with their respective duals, so that $G := G'' = L^2(\Gamma_1)$ and $H := H'' = H^{-2}(\Omega) \times L^2(\Omega)$.

It is well known that A^* generates a group in H'; see, e.g., [33].

To prove (H2) let us introduce the Dirichlet map $D : L^2(\Gamma_1) \to L^2(\Omega)$ now by

$$\begin{cases} \Delta^2 Du = 0 \quad \text{in} \quad \Omega, \\ Du = \partial_\nu Du = 0 \quad \text{on} \quad \Gamma_0, \\ Du = 0 \quad \text{and} \quad \partial_\nu Du = u \quad \text{on} \quad \Gamma_1. \end{cases}$$

We recall from [33] that D is a bounded linear map. Let us define $E \in L(G, H)$ by the formula $Eu = (0, -Du)$, $u \in G$. Now, for every $(\eta_0, \eta_1) \in D(A^*)$ and $u \in H_0^{5/2}(\Gamma_1)$,

we have the following equality:

$$\begin{aligned}
&\langle E^*A^*(\eta_0,\eta_1),u\rangle_{G',G}\\
&=\langle A^*(\eta_0,\eta_1),Eu\rangle_{H',H}\\
&=\langle(-\eta_1,\Delta^2\eta_0),(0,-Du)\rangle_{H',H}\\
&=-\int_\Omega(\Delta^2\eta_0)Du\ dx\\
&=-\int_\Omega\eta_0(\Delta^2Du)\ dx\\
&+\int_\Gamma-(\partial_\nu\Delta\eta_0)Du+(\Delta\eta_0)(\partial_\nu Du)-(\partial_\nu\eta_0)(\Delta Du)+\eta_0(\partial_\nu\Delta Du)\ d\Gamma.
\end{aligned}$$

(The last step is correct because $u\in H_0^{5/2}(\Gamma_1)$ implies $Du\in H^4(\Omega)$.) Using the definition of Du and the property $\eta_0=\partial_\nu\eta_0\equiv 0$ on Γ, we conclude that

$$\langle E^*A^*(\eta_0,\eta_1),u\rangle_{G',G}=\langle\Delta\eta_0,u\rangle_{G',G}.$$

Using the density of $H_0^{5/2}(\Gamma_1)$ in $G=L^2(\Gamma_1)$, the equality $B^*=E^*A^*$ follows.

Rewriting the assumptions (H3) and (H4) for the original problem (6.5), we have to prove the existence of a positive number T and of two positive constants c_1, c_2 such that for every $(\xi_0,\xi_1)\in(H^4(\Omega)\cap H_0^2(\Omega))\times H_0^2(\Omega)$ the solution of (6.5) satisfies the inequalities

$$c_1\|(\xi_0,\xi_1)\|^2_{H_0^2(\Omega)\times L^2(\Omega)}\le\int_0^T\int_{\Gamma_1}|\Delta\xi|^2\ d\Gamma\ dt\le c_2\|(\xi_0,\xi_1)\|^2_{H_0^2(\Omega)\times L^2(\Omega)}. \tag{6.6}$$

These inequalities were established in [31]. (See also [15] for a better estimate of T by elementary means, and [41] and [19, Theorem 6.7] for two different proofs of these estimates for *any* positive T.)

We may now apply Theorem 3.1 with A and B defined by $A:=A^{**}$ and $B:=B^{**}$.

Let us explicitly write the resulting closed-loop problem (3.2), (3.3). Consider the solutions of (6.1) and of (6.5). We have

$$\begin{aligned}
0&=\int_0^T\int_\Omega(y''+\Delta^2y)\xi\ dx\ dt\\
&=\Big[\int_\Omega y'\xi-y\xi'\ dx\Big]_0^T+\int_0^T\int_\Omega y(\xi''+\Delta^2\xi)\ dx\ dt\\
&+\int_0^T\int_\Gamma(\partial_\nu\Delta y)\xi-(\Delta y)\partial_\nu\xi+(\partial_\nu y)\Delta\xi-y(\partial_\nu\Delta\xi)\ d\Gamma\ dt.
\end{aligned}$$

Using (6.1) and (6.5) this equality reduces to

$$0=\Big[\int_\Omega y'\xi-y\xi'\ dx\Big]_0^T+\int_0^T\int_{\Gamma_1}u\Delta\xi\ d\Gamma\ dt.$$

Putting $x=(-y',y)$, $x_0=(-y_1,y_0)$, $\varphi=(\xi,\xi')$, and $\varphi=(\xi_0,\xi_1)$, this equality may be rewritten as (2.4). Since (6.5) is equivalent to (2.2), we conclude that (6.1) is equivalent to (2.1).

Writing the operator

$$\Lambda_\omega^{-1} : H^{-2}(\Omega) \times L^2(\Omega) \to H_0^2(\Omega) \times L^2(\Omega)$$

in the matrix form

$$\Lambda_\omega^{-1} = \begin{pmatrix} P & -Q \\ -R & S \end{pmatrix},$$

we have

$$u = -JB^*\Lambda_\omega^{-1}x = \Delta(Py' + Qy)$$

on Γ_1, and (6.3) follows. □

7. Application to a Petrovsky system II. Consider the problem

$$\begin{cases} y'' + \Delta^2 y = 0 \quad \text{in} \quad \Omega \times \mathbb{R}, \\ y = \Delta y = 0 \quad \text{on} \quad \Gamma_0 \times \mathbb{R}, \\ y = 0 \quad \text{and} \quad \Delta y = u \quad \text{on} \quad \Gamma_1 \times \mathbb{R}, \\ y(0) = y_0 \quad \text{and} \quad y'(0) = y_1 \quad \text{in} \quad \Omega, \end{cases} \tag{7.1}$$

where Ω is a bounded open set in $\mathbb{R}^n$, Γ_1 is an open subset of its boundary Γ, and $\Gamma_0 = \Gamma \backslash \Gamma_1$. We shall denote by ν the outward unit normal vector to Γ.

As in section 4, assume that Γ is analytic and that Γ_1 satisfies the *geometrical control condition* of Bardos, Lebeau, and Rauch [2], [3]: there exists a positive number T such that every ray of geometrical optics in $\overline{\Omega}$ hits Γ_1 at a nondiffractive point in some time $\leq T$. Then we have the following theorem.

THEOREM 7.1. *Fix $\omega > 0$ arbitrarily. There exist two bounded linear maps*

$$P : H^{-1}(\Omega) \to H_0^1(\Omega), \qquad Q : H_0^1(\Omega) \to H_0^1(\Omega) \tag{7.2}$$

and a constant M such that, putting

$$u = \frac{\partial}{\partial \nu}(Py' + Qy), \tag{7.3}$$

the problem (7.1), (7.3) *is well posed in $\mathcal{H} := H_0^1(\Omega) \times H^{-1}(\Omega)$ and its solutions satisfy the estimates*

$$\|(y, y')(t)\|_{\mathcal{H}} \leq M \|(y_0, y_1)\|_{\mathcal{H}} e^{-\omega t} \tag{7.4}$$

for all $t \geq 0$ and for all $(y_0, y_1) \in \mathcal{H}$.

Remarks. (1) Theorem 6.1 improves some earlier results of Lasiecka and Triggiani [27].

(2) The same conclusion holds if Γ is only of class C^4 but there exists a point $x^0 \in \mathbb{R}^n$ such that Γ_1 contains all points x of Γ which satisfy the inequality $(x - x^0) \cdot \nu(x) > 0$.

(3) Let us note that, contrary to the wave equation, the necessity of the geometrical control condition is far from clear here. (See [10], [13], and [17] for closely related exact *internal* controllability results in the "right spaces" in the absence of the geometrical control condition.)

For the proof, consider the following problem:

$$\begin{cases} \xi'' + \Delta^2 \xi = 0 \quad \text{in} \quad \Omega \times \mathbb{R}, \\ \xi = \Delta \xi = 0 \quad \text{on} \quad \Gamma \times \mathbb{R}, \\ \xi(0) = \xi_0 \quad \text{and} \quad \xi'(0) = \xi_1 \quad \text{in} \quad \Omega, \\ \psi = \partial_\nu \xi|_{\Gamma_1} \quad \text{in} \quad \mathbb{R}. \end{cases} \tag{7.5}$$

Putting $\varphi = (\xi, \xi')$, $\varphi_0 = (\xi_0, \xi_1)$, and introducing the linear operators A^* and B^* by the formulae

$$D(A^*) = D(B^*) = \{(\eta_0, \eta_1) \in H^3(\Omega) \times H^1(\Omega) \mid \eta_0 = \Delta\eta_0 = \eta_1 = 0 \text{ on } \Gamma\},$$
$$A^*(\eta_0, \eta_1) = (-\eta_1, \Delta^2\eta_0),$$
$$B^*(\eta_0, \eta_1) = \partial_\nu \eta_0|_{\Gamma_1},$$

we may rewrite (7.5) in the abstract form (2.2):

$$\varphi' = -A^*\varphi, \qquad \varphi(0) = \varphi_0, \qquad \psi = B^*\varphi.$$

Let us show that, choosing $H' = H_0^1(\Omega) \times H^{-1}(\Omega)$ and $G' = L^2(\Gamma_1)$, the assumptions (H1)–(H4) of Theorem 3.1 are satisfied. As usual, we identify $L^2(\Omega)$ and $L^2(\Gamma_1)$ with their respective duals, so that $G := G'' = L^2(\Gamma_1)$ and $H := H'' = H^{-1}(\Omega) \times H_0^1(\Omega)$.

It is well known that A^* generates a group in H'; see, e.g., [33].

To prove (H2) let us introduce the bounded linear map $D : L^2(\Gamma_1) \to H^2(\Omega) \cap H_0^1(\Omega)$ now by

$$\begin{cases} \Delta^2 Du = 0 & \text{in} \quad \Omega, \\ Du = \Delta Du = 0 & \text{on} \quad \Gamma_0, \\ Du = 0 \quad \text{and} \quad \Delta Du = u & \text{on} \quad \Gamma_1. \end{cases}$$

Let us define $E \in L(G, H)$ by the formula $Eu = (0, Du)$, $u \in G$. Now, given any $(\eta_0, \eta_1) \in D(A^*)$, we have for every $u \in H_0^{3/2}(\Gamma_1)$ the following equality:

$$\begin{aligned} &\langle E^*A^*(\eta_0, \eta_1), u\rangle_{G',G} \\ &= \langle A^*(\eta_0, \eta_1), Eu\rangle_{H',H} \\ &= \langle(-\eta_1, \Delta^2\eta_0), (0, Du)\rangle_{H',H} \\ &= \int_\Omega (\Delta^2\eta_0) Du \, dx \\ &= \int_\Omega \eta_0(\Delta^2 Du) \, dx \\ &+ \int_\Gamma (\partial_\nu \Delta\eta_0) Du - (\Delta\eta_0)(\partial_\nu Du) + (\partial_\nu\eta_0)(\Delta Du) - \eta_0(\partial_\nu \Delta Du) \, d\Gamma. \end{aligned}$$

(The last step is justified because $u \in H_0^{3/2}(\Gamma_1)$ implies $Du \in H^4(\Omega)$.) Using the definition of Du and the property $\eta_0 = \Delta\eta_0 \equiv 0$ on Γ, we conclude that

$$\langle E^*A^*(\eta_0, \eta_1), u\rangle_{G',G} = \langle \partial_\nu\eta_0, u\rangle_{G',G}.$$

Using the density of $H_0^{3/2}(\Gamma_1)$ in $G = L^2(\Gamma_1)$, the equality $B^* = E^*A^*$ follows.

Rewriting the assumptions (H3) and (H4) for the original problem (7.5), we have to prove the existence of a positive number T and of two positive constants c_1, c_2 such that for every $(\xi_0, \xi_1) \in D(A^*)$ the solution of (7.5) satisfies the inequalities

$$c_1\|(\xi_0, \xi_1)\|^2_{H_0^1(\Omega)\times H^{-1}(\Omega)} \le \int_0^T \int_{\Gamma_1} |\partial_\nu \xi|^2 \, d\Gamma \, dt \le c_2\|(\xi_0, \xi_1)\|^2_{H_0^1(\Omega)\times H^{-1}(\Omega)}. \tag{7.6}$$

The second inequality is essentially a special case of a theorem proved in [31]; see also [19, Theorem 2.10]. The first inequality was proved (improving some former

results of Lions [31]) in [19, Theorem 6.11] under the stronger geometrical assumption mentioned in Remark (2) above, and by Lebeau [28] in the general case. In the last two works these inequalities are established for every (arbitrarily small) positive T.

We may thus apply Theorem 3.1 with A and B defined by $A := A^{**}$ and $B := B^{**}$.

Let us show that the resulting closed-loop problem (3.2), (3.3) is equivalent to (7.1), (7.3). Consider the solutions of (7.1) and (7.5). We have

$$\begin{aligned} 0 &= \int_0^T \int_\Omega (y'' + \Delta^2 y)\xi \; dx \; dt \\ &= \Big[\int_\Omega y'\xi - y\xi' \; dx\Big]_0^T + \int_0^T \int_\Omega y(\xi'' + \Delta^2 \xi) \; dx \; dt \\ &+ \int_0^T \int_\Gamma (\partial_\nu \Delta y)\xi - (\Delta y)\partial_\nu \xi + (\partial_\nu y)\Delta\xi - y(\partial_\nu \Delta \xi) \; d\Gamma \; dt. \end{aligned}$$

Using (7.1) and (7.5), this equality reduces to

$$0 = \Big[\int_\Omega y'\xi - y\xi' \; dx\Big]_0^T - \int_0^T \int_{\Gamma_1} u\partial_\nu \xi \; d\Gamma \; dt.$$

Putting $x = (y', -y)$, $x_0 = (y_1, -y_0)$, $\varphi = (\xi, \xi')$, and $\varphi_0 = (\xi_0, \xi_1)$, this equality may be rewritten as (2.4). Since (7.5) is equivalent to (2.2), we conclude that (7.1) is equivalent to (2.1).

Writing the operator

$$\Lambda_\omega^{-1} : H^{-1}(\Omega) \times H_0^1(\Omega) \to H_0^1(\Omega) \times H^{-1}(\Omega)$$

in the matrix form

$$\Lambda_\omega^{-1} = -\begin{pmatrix} P & -Q \\ -R & S \end{pmatrix},$$

we have

$$u = -JB^*\Lambda_\omega^{-1} x = \partial_\nu(Py' + Qy)$$

on Γ_1, and (7.3) follows. □

Note added in proof. In an earlier version of this paper we used in Theorem 3.1 a different weight function (corresponding to the change of T_ω to T in the formula (3.1)). Frédéric Bourquin observed that in that case Flandoli's theorem cannot be applied unless the operator C is bounded. At the same time, he also suggested the use of the weight function appearing now in (3.1) in order to keep the formulation of the theorem and most of the proof unchanged. We are most grateful to him for allowing us to include his corrections in our paper.

The hypothesis (H2) of the paper is equivalent to hypothesis (H2′) below, which is easier to verify in the applications: instead of constructing a Dirichlet type map it suffices to invoke the elliptic regularity theory.

(H2′) We have $D(A^*) = D(B^*)$, and there exist two numbers $\lambda \in \mathbf{C}$ and $c \in \mathbf{R}$ such that

$$\|B^*\phi\| \le c\|(A + \lambda I)^*\phi\|$$

for all $\phi \in D(A^*)$.

Indeed, (H2) obviously implies (H2′) with $c = \|E^*\|$. Conversely, (H2′) implies that the formula

$$E^*(A + \lambda I^*)^*\phi := B^*\phi$$

defines a bounded linear map E^* from the range of $(A+\lambda I)^*$ in H' into G'. Extending it continuously to the closure of range $(A + \lambda I)^*$ and then defining E^* by zero on its orthogonal complement, we obtain an operator $E^* \in L(H', G')$ satisfying (H2). Finally we set $E := E^{**}$.

For example, in the example of section 4 we have

$$D(A^*) = D(B^*) = H^2(\Omega) \cap H_0^1(\Omega)$$

by definition and

$$\|B^*(\eta_0, \eta_1)\|_{G'} = \|\partial_\nu \eta_0\|_{L^2\Gamma_1)} \le c\|\eta_0\|_{H_2(\Omega)}$$
$$\le c\|\Delta\eta_0\|_{L^2(\Omega)} \le c\| - (\eta_1, \Delta\eta_0)\|_{H'} = \|A^*(\eta_0, \eta_1)\|_{H'},$$

proving (H2′).

Acknowledgments. Some results of this paper were announced earlier without proof in [22] and [23].

Let us remark that T. I. Seidman [37] constructed, using a different approach, stabilizing feedbacks leading to arbitrarily large decay rates in *augmented* state spaces.

The author is grateful to F. Bourquin, J. Lagnese, J.-L. Lions, O. J. Staffans, and G. Weiss for their useful advice concerning the presentation of our results.

REFERENCES

[1] A. V. Balakrishnan, *Applied Functional Analysis*, Springer-Verlag, New York, 1976.
[2] C. Bardos, G. Lebeau, and J. Rauch, *Contrôle et stabilisation dans les problèmes hyperboliques*, in Contrôlabilité exacte et stabilisation de systèmes distribuées, Vol. 1, J. L. Lions, Masson, Paris, 1988, Appendix II, pp. 492–537.
[3] C. Bardos, *Sharp sufficient conditions for the observation, control and stabilization of waves from the boundary*, SIAM J. Control Optim., 30 (1992), pp. 1024–1065.
[4] A. Benabdallah and M. Lenczner, *Stabilisation de l'équation des ondes par un contrôle optimal distribué: Estimation du taux de décroissance*, C. R. Acad. Sci. Paris Sér. I Math., 319 (1994), pp. 691–696.
[5] A. Bensoussan, G. Da Prato, M. C. Delfour, and S. K. Mitter, *Representation and Control of Infinite Dimensional Systems* I-II, Birkhäuser, Boston, 1992, 1993.
[6] G. Chen, *Energy decay estimates and exact boundary value controllability for the wave equation in a bounded domain*, J. Math. Pures Appl., 58 (1979), pp. 249–274.
[7] S. Dolecki and D. L. Russell, *A general theory of observation and control*, SIAM J. Control Optim., 15 (1977), pp. 185–220.
[8] F. Flandoli, *A New Approach to the L-Q-R Problem for Hyperbolic Dynamics with Boundary Control*, Lecture Notes in Control and Information Sciences 102, Springer-Verlag, Berlin, Heidelberg, New York, 1987, pp. 89–111.
[9] K. D. Graham and D. L. Russell, *Boundary value control of the wave equation in a spherical region*, SIAM J. Control Optim., 13 (1975), pp. 174–196.
[10] A. Haraux, *Séries lacunaires et contrôle semi-interne des vibrations d'une plaque rectangulaire*, J. Math. Pures Appl., 68 (1989), pp. 457–465.
[11] A. Haraux, *Une remarque sur la stabilisation de certains systèmes du deuxième ordre en temps*, Portugal Math., 46 (1989), pp. 246–257.
[12] L. F. Ho, *Observabilité frontière de l'équation des ondes*, C. R. Acad. Sci. Paris Sér. I Math., 302 (1986), pp. 443–446.
[13] S. Jaffard, *Contrôle interne exact des vibrations d'une plaque rectangulaire*, Portugal Math., 47 (1990), pp. 423–429.

[14] H. Koch and D. Tataru, *On the spectrum of hyperbolic semigroups*, Comm. Partial Differential Equations, 20 (1995), pp. 901–937.
[15] V. Komornik, *Contrôlabilité exacte en un temps minimal*, C. R. Acad. Sci. Paris Sér. I Math., 304 (1987), pp. 223–225.
[16] V. Komornik, *Rapid boundary stabilization of the wave equation*, SIAM J. Control Optim., 29 (1991), pp. 197–208.
[17] V. Komornik, *On the exact internal controllability of a Petrowsky system*, J. Math. Pures Appl., 71 (1992), pp. 331–342.
[18] V. Komornik, *On the zeros of Bessel type functions and applications to exact controllability problems*, Asymptotic Analy., 5 (1991), pp. 115–128.
[19] V. Komornik, *Exact Controllability and Stabilization: The Multiplier Method*, Masson–John Wiley, Paris, 1994.
[20] V. Komornik, *Boundary stabilization, observation and control of Maxwell's equations*, Panamer. Math. J., 4 (1994), pp. 47–61.
[21] V. Komornik, *Boundary Stabilization of Linear Elasticity Systems*, Lecture Notes in Pure and Appl. Math. 174, Marcel Dekker, New York, 1995, pp. 135–146.
[22] V. Komornik, *Stabilisation frontière rapide de systèmes distribués linéaires*, C. R. Acad. Sci. Paris Sér. I Math., 321 (1995), pp. 433–437.
[23] V. Komornik, *Stabilisation rapide de problèmes d'évolution linéaires*, C. R. Acad. Sci. Paris Sér. I Math., 321 (1995), pp. 581–586.
[24] V. Komornik and E. Zuazua, *A direct method for the boundary stabilization of the wave equation*, J. Math. Pures Appl., 69 (1990), pp. 33–54.
[25] J. Lagnese, *Boundary Stabilization of Thin Plates*, SIAM Studies in Appl. Math., SIAM, Philadelphia, 1989.
[26] I. Lasiecka and R. Triggiani, *Regularity of hyperbolic equations under $L_2(0,T;L_2(\Gamma))$ boundary terms*, Appl. Math. Optim., 10 (1983), pp. 275–286.
[27] I. Lasiecka and R. Triggiani, *Differential and Algebraic Riccati Equations with Applications to Boundary/Point Control Problems: Continuous Theory and Approximation Theory*, Lecture Notes in Control and Inform. Sci. 164, Springer-Verlag, Berlin, Heidelberg, New York, 1991.
[28] G. Lebeau, *Contrôle de l'équation de Schrödinger*, J. Math. Pures Appl., 71 (1992), pp. 267–291.
[29] E. B. Lee and L. Markus, *Foundations of Optimal Control Theory*, John Wiley, New York, 1967.
[30] J.-L. Lions, *Contrôle des systèmes distribués singuliers*, Gauthiers-Villars, Paris, 1983.
[31] J.-L. Lions, *Exact controllability, stabilizability, and perturbations for distributed systems*, SIAM Rev., 30 (1988), pp. 1–68.
[32] J.-L. Lions, *Contrôlabilité exacte et stabilisation de systèmes distribués*, Vol. 1, Masson, Paris, 1988.
[33] J.-L. Lions and E. Magenes, *Problèmes aux limites non homogènes et applications* I–III, Dunod, Paris, 1968–1970.
[34] D. L. Lukes, *Stabilizability and optimal control*, Funkcial. Ekvac., 11 (1968), pp. 39–50.
[35] P. Martinez, *Stabilisation frontière de l'équation des ondes dans des domaines polygonaux*, C. R. Acad. Sci. Paris Sér. I Math., 322 (1996), pp. 365–370.
[36] D. L. Russell, *Controllability and stabilizability theory for linear partial differential equations. Recent progress and open questions*, SIAM Rev., 20 (1978), pp. 639–739.
[37] T. I. Seidman, Private communication, 1995.
[38] M. Slemrod, *A note on complete controllability and stabilizability for linear control systems in Hilbert space*, SIAM J. Control, 12 (1974), pp. 500–508.
[39] T. L. R. Tcheugoué, *Deux remarques sur la stabilisation de l'équation des ondes en dimension* 2, C. R. Acad. Sci. Paris Sér. I Math., 319 (1994), pp. 585–588.
[40] R. Triggiani, *Wave equation on a bounded domain with boundary dissipation: An operator approach*, J. Math. Anal. Appl., 137 (1989), pp. 438–461.
[41] E. Zuazua, *Contrôlabilité exacte en un temps arbitrairement petit de quelques modèles de plaques*, in Contrôlabilité exacte et stabilisation de systèmes distribuées, Vol. 1, J. L. Lions, Masson, Paris, 1988, Appendix I, pp. 465–491.
[42] E. Zuazua, *Sur l'optimalité des feedbacks de stabilisation*, C. R. Acad. Sci. Paris Sér. I Math., 309 (1989), pp. 547–552.

SIAM J. CONTROL OPTIM.
Vol. 35, No. 5, pp. 1614–1637, September 1997

010

BOUNDARY CONTROLLABILITY OF A LINEAR HYBRID SYSTEM ARISING IN THE CONTROL OF NOISE*

SORIN MICU† AND ENRIQUE ZUAZUA‡

Abstract. We consider a simple model arising in the control of noise. We assume that the two-dimensional cavity $\Omega = (0,1) \times (0,1)$ is occupied by an elastic, inviscid, compressible fluid. The potential ϕ of the velocity field satisfies the linear wave equation. The boundary of Ω is divided into two parts, Γ_0 and Γ_1. The first one, Γ_0, is flexible and occupied by a vibrating string that obeys the one-dimensional wave equation. On Γ_0 the continuity of the normal velocities of the fluid and the string is imposed. The subset Γ_1 of the boundary is assumed to be rigid, and therefore, the normal velocity of the fluid vanishes. This constitutes a conservative system of two coupled wave equations in dimensions two and one, respectively.

The control (an elastic force or an exterior source of noise) is assumed to act on the flexible part of the boundary. We are interested on the controllability problem: given a large enough control time, what are the initial conditions we can drive to the equilibrium by means of, say, L^2-controls? By using Fourier series the problem is decomposed into an infinite number of one-dimensional control problems that we solve by classical methods that combine the Hilbert uniqueness method, multiplier techniques, and Ingham-type inequalities. Putting these one-dimensional results together, we give a precise characterization of the space of controllable data in terms of Fourier series.

Key words. boundary control, hyperbolic system, aeromechanic structure interaction

AMS subject classifications. 35B37, 93C20, 73K70

PII. S0363012996297972

1. Introduction.

Let Ω be the two-dimensional square $\Omega = (0,1) \times (0,1) \subset \mathbb{R}^2$.

We assume that Ω is filled with an elastic, inviscid, compressible fluid whose velocity field $\vec{v}$ is given by the potential $\phi = \phi(x,y,t) : \vec{v} = \nabla\phi$. By linearization we assume that the potential ϕ satisfies the linear wave equation in $\Omega \times (0,\infty)$.

The boundary $\Gamma = \partial\Omega$ of Ω is divided into two parts: $\Gamma_0 = \{(x,0) : x \in (0,1)\}$ and $\Gamma_1 = \Gamma \backslash \Gamma_0$. The subset Γ_1 is assumed to be rigid, and we impose zero normal velocity of the fluid on it. The subset Γ_0 is supposed to be flexible and occupied by a flexible string that vibrates under the pressure of the fluid on the plane where Ω lies. The displacement of Γ_0, described by the scalar function $W = W(x,t)$, obeys the one-dimensional wave equation. On the other hand, on Γ_0 we impose the continuity of the normal velocities of the fluid and the string. The string is assumed to satisfy Neumann boundary conditions on its extremes. All deformations are supposed to be small enough that linear theory applies. Under natural initial conditions for ϕ and W the linear motion of this system is described by means of the following coupled wave equations:

*Received by the editors January 31, 1996; accepted for publication (in revised form) July 1, 1996. http://www.siam.org/journals/sicon/35-5/29797.html

†Departamento de Matemática Aplicada, Facultad de Ciencias Matemáticas, Universidad Complutense, 28040 Madrid, Spain, and Facultatea de Matematica-Informatica, Universitatea din Craiova, Craiova 1100, Romania (sorin@sunma4.mat.ucm.es).

‡Departamento de Matemática Aplicada, Facultad de Ciencias Matemáticas, Universidad Complutense, 28040 Madrid, Spain (zuazua@sunma4.mat.ucm.es). The research of this author was supported by grant PB93-1203 of the DGICYT (Spain) and CHRX-CT94-0471 of the European Union.

$$
(1.1)\qquad \begin{cases} \phi_{tt} - \Delta\phi = 0 & \text{in} \quad \Omega \times (0,\infty), \\ \frac{\partial \phi}{\partial \nu} = 0 & \text{on} \quad \Gamma_1 \times (0,\infty), \\ \frac{\partial \phi}{\partial y} = -W_t & \text{on} \quad \Gamma_0 \times (0,\infty), \\ W_{tt} - W_{xx} + \phi_t = 0 & \text{on} \quad \Gamma_0 \times (0,\infty), \\ W_x(0,t) = W_x(1,t) = 0 & \text{for} \quad t > 0, \\ \phi(0) = \phi^0, \phi_t(0) = \phi^1 & \text{in} \quad \Omega, \\ W(0) = W^0, W_t(0) = W^1 & \text{on} \quad \Gamma_0. \end{cases}
$$

By ν we denote the unit outward normal to Ω.

In (1.1) we have chosen to take the various parameters of the system to be equal to one.

The system (1.1) is well posed in the energy space $\mathcal{X} = H^1(\Omega) \times L^2(\Omega) \times H^1(\Gamma_0) \times L^2(\Gamma_0)$ for the variables (ϕ, ϕ_t, W, W_t). The energy

$$
(1.2)\qquad E(t) = \frac{1}{2} \int_\Omega \left[|\, \nabla\phi \,|^2 + |\, \phi_t \,|^2 \right] dxdy + \frac{1}{2} \int_{\Gamma_0} \left[|\, W_x \,|^2 + |\, W_t \,|^2 \right] dx
$$

remains constant along trajectories.

We study the controllability of system (1.1) under the action of an exterior force or source of noise on the flexible part of the boundary Γ_0. The control is given by a scalar function $\beta = \beta(x,t)$, and the controlled system reads as follows:

$$
(1.3)\qquad \begin{cases} \phi_{tt} - \Delta\phi = 0 & \text{in} \quad \Omega \times (0,\infty), \\ \frac{\partial \phi}{\partial \nu} = 0 & \text{on} \quad \Gamma_1 \times (0,\infty), \\ \frac{\partial \phi}{\partial y} = -W_t & \text{on} \quad \Gamma_0 \times (0,\infty), \\ W_{tt} - W_{xx} + \phi_t = \beta & \text{on} \quad \Gamma_0 \times (0,\infty), \\ W_x(0,t) = W_x(1,t) = 0 & \text{for} \quad t > 0, \\ \phi(0) = \phi^0, \phi_t(0) = \phi^1 & \text{in} \quad \Omega, \\ W(0) = W^0, W_t(0) = W^1 & \text{on} \quad \Gamma_0. \end{cases}
$$

It is easy to see that the equilibria of these systems are of the form

$$
(1.4)\qquad (\phi, \phi_t, W, W_t) = (c_1, 0, c_2, 0),
$$

c_1 and c_2 being constant functions.

In view of the finite speed of propagation of the wave equation satisfied by ϕ, the geometry of Ω, and the support of the control β (the subset Γ_0 of the boundary of Ω), the minimal controllability time for system (1.3) is $T_0 = 2$.

We choose the control β to be in the space $H^{-2}(0,T;L^2(\Gamma_0))$. Of course, this is an arbitrary choice and many others make sense. However, this is the most natural one when solving the control problem by means of Lions's Hilbert uniqueness method (HUM) (see [10]), as we will do.

The problem of controllability can be formulated as follows: given $T > 2$, find the space of initial data $(\phi^0, \phi^1, W^0, W^1)$ that can be driven to an equilibrium of the form (1.4) in time T by means of a suitable control $\beta \in H^{-2}(0,T;L^2(\Gamma_0))$.

The control set Γ_0 does not satisfy the necessary geometric conditions for controllability given by Bardos, Lebeau, and Rauch in [6]. Indeed, any segment of the form $\{(x,\ell) : x \in (0,1)\}$ with $0 < \ell < 1$ constitutes a ray of geometric optics that never intersects the control region Γ_0. Therefore, we cannot expect the space of controllable initial data to be an energy space.

In this paper we give a complete characterization of the controllable space in terms of Fourier series. This space consists of initial data whose Fourier coefficients, roughly, decay exponentially as the frequency increases.

The Fourier analysis of the system is possible because of the boundary conditions we have chosen for W. Indeed, W is assumed to satisfy Neumann-type boundary conditions, which are compatible with those of ϕ, to develop solutions in Fourier series.

Indeed, let us decompose the control β, the solutions ϕ, W, and the initial data in the following way:

$$\begin{cases} \beta = \displaystyle\sum_{n=0}^{\infty} \beta_n(t)\cos(n\pi x), \\ \Phi = \displaystyle\sum_{n=0}^{\infty} \psi_n(y,t)\cos(n\pi x), \quad (\phi^0,\phi^1) = \displaystyle\sum_{n=0}^{\infty} (\psi_n^0(y),\psi_n^1(y))\cos(n\pi x), \\ W = \displaystyle\sum_{n=0}^{\infty} V_n(t)\cos(n\pi x), \quad (W^0,W^1) = \displaystyle\sum_{n=0}^{\infty} (V_n^0,V_n^1)\cos(n\pi x). \end{cases} \tag{1.5}$$

With this decomposition, system (1.3) can be split into the following sequence of one-dimensional controlled systems for $n = 0, 1, \ldots$:

$$\begin{cases} \psi_{n,tt} - \psi_{n,yy} + n^2\pi^2\psi_n = 0 & \text{for} \quad (y,t) \in (0,1)\times(0,\infty), \\ \psi_{n,y}(1,t) = 0 & \text{for} \quad t > 0, \\ \psi_{n,y}(0,t) = -V_t(t) & \text{for} \quad t > 0, \\ V_{n,tt}(t) + n^2\pi^2 V_n(t) + \psi_{n,t}(0,t) = \beta_n(t) & \text{for} \quad t > 0, \\ \psi_n(0) = \psi_n^0, \psi_{n,t}(0) = \psi_n^1 & \text{in} \quad (0,1), \\ V_n(0) = V_n^0, V_{n,t}(0) = V_n^1. \end{cases} \tag{1.6}$$

First we will study the controllability of system (1.6) by using classical methods that combine HUM, multiplier techniques, and Ingham-type inequalities (see [9] and [8]). Combining these one-dimensional results with the Fourier decomposition (1.5), the controllability result for system (1.3) will be proved. Although the techniques we use are well known, the obtainment of sharp estimates for the controls requires the use of them in a rather refined way.

The control β we obtain is of the form $\beta = \frac{\partial^2}{\partial t^2}\gamma$, with $\gamma \in L^2(\Gamma_0 \times (0,T))$ having compact support in time. Therefore $\int_0^T \beta = 0$. Taking this fact into account, it is easy to see that the constants c_1, c_2 of the equilibrium that we reach at time $t = T$ are determined a priori by the initial data. Indeed, integrating the first equation of (1.3) in Ω we obtain that $\int_\Omega \phi_t dxdy - \int_{\Gamma_0} W dx$ remains constant in time. Therefore, necessarily,

$$c_2 = \int_{\Gamma_0} W^0 dx - \int_\Omega \phi^1 dxdy. \tag{1.7}$$

On the other hand, integrating the equation satisfied by W on $\Gamma_0 \times (0,T)$ and taking into account that $\int_0^T \beta = 0$, we deduce that

$$\int_{\Gamma_0} W_t(T)dx + \int_{\Gamma_0} \phi(x,0,T)dx = \int_{\Gamma_0} W^1 dx + \int_{\Gamma_0} \phi^0(x,0)dx,$$

and therefore,

$$c_1 = \int_{\Gamma_0} \left(W^1 + \phi^0(x,0)\right) dx. \tag{1.8}$$

In terms of the Fourier coefficients (1.5) these constants can be written in the following way:

$$c_1 = V_0^1 + \psi_0^0(0), \quad c_2 = V_0^0 - \int_0^1 \psi_0^1(y)dy. \tag{1.9}$$

Therefore, the constants c_1 and c_2 of the equilibrium we may reach are uniquely determined by the Fourier coefficients of the initial data corresponding to the frequency $n = 0$ in the x-variable.

This fact is related to the different nature of systems (1.6) for $n = 0$ and $n \geq 1$. While for any $n \geq 1$, system (1.6) is exactly controllable to zero at any time $T > 2$, when $n = 0$ we can control the system to the equilibrium given by (1.9) only in terms of the initial data.

The system under consideration can be viewed as a hybrid system coupling a fluid with an elastic structure. From a mathematical point of view the system couples a two-dimensional wave equation with a one-dimensional one. This type of system is rather common when studying the vibrations of structures connecting several flexible bodies of different dimensions. Examples of this type can be found, for instance, in [11], [7], and [16]. However, in all these cases the coupling is of a different nature since the continuity of displacements is imposed, but not the continuity of normal velocities.

The model under consideration is inspired in and related to that of Banks et al. in [5]. However, there are some important differences between these two models. In [5] the flexible part of the boundary Γ_0 is occupied by a flexible damped beam instead of a flexible string. But the main difference is related to the nature of the controls. In [5] the control acts on the system through a finite number of piezoceramic patches located on Γ_0. This restricts very much the set of admissible controls, which are essentially second derivatives of Heaviside functions, and much weaker controllability results have to be expected. In [5] the controllability problem is not addressed. Instead, they consider a quadratic optimal control problem. More recently, in [3], a Riccati equation for the optimal control is derived. The problem of the controllability of one-dimensional beams with piezoelectric actuators has been successfully addressed by Tucsnak [17]. However, to our knowledge, there are no rigorous results on the controllability of fluid-structure systems under such controls. To our knowledge the present paper represents the first attempt to solve the controllability problem for the two-dimensional system although, as we said above, we do not address the problem in which the control is made through piezoelectric patches.

In [13], the authors have addressed the problem of the feedback stabilization of system (1.3) with a damping term concentrated on Γ_0. The results in [13] show that, in such a situation, every trajectory converges towards an equilibrium as time goes to infinity, but the decay rate is not uniform. A more detailed discussion on the lack of uniform decay can be found in [12] and [15]. More recently, in [2], the system introduced in [5] has been considered, with the condition $\frac{\partial \Phi}{\partial \nu} = -W_t$ on the continuity of the velocity fields replaced by a dissipative condition of the form $\frac{\partial \Phi}{\partial \nu} = -W_t + \Phi_t$. In [2] it is proven that when Ω is a general smooth bounded domain and the subset Γ_0 of the boundary is sufficiently large (in the spirit of the geometric conditions arising in the boundary stabilization of the wave equation), then the energy decays uniformly to zero. In [13] the existence of periodic solutions of this dissipative system on the presence of a periodic source of noise acting on the system through the flexible part of the boundary is considered too. Due to the very weak effect that the damping located

on Γ_0 has on the fluid inside Ω, in order to guarantee the existence of such periodic solutions of finite energy, the exterior source of noise has to be assumed to belong to a rather small class of functions with rapidly decreasing Fourier coefficients. In this sense, this result is very close to the controllability one we present in this paper. For a detailed discussion, see [12].

The rest of the paper is organized as follows. In section 2 we rigorously present the main results of this paper and make a discussion on their optimality. In section 3 we address the one-dimensional control problem (1.6). First, distinguishing the cases $n = 0$ and $n \geq 1$, we derive the necessary observability inequalities. Then, applying HUM, the one-dimensional controllability result is deduced. In section 4, combining the results of the previous one, we derive the controllability result for system (1.3).

In an appendix at the end of the paper we give a detailed proof of an Ingham-type inequality that provides explicit estimates of the constants appearing in it.

2. The main results: Statements and discussion. As we said in the introduction, the controllability problem of system (1.3) is reduced to study the one-parameter family of one-dimensional systems (1.6). When $n \geq 1$ we have the following controllability result for (1.6).

THEOREM 2.1. *Let $\mathcal{Y}$ be the space $H^1(0,1) \times L^2(0,1) \times \mathbb{R} \times \mathbb{R}$. Assume that $T > 2$ and $n \geq 1$. Then, for any $(\psi^0, \psi^1, V^0, V^1) \in \mathcal{Y}$, there exists a control $\beta \in H^{-2}(0,T)$ with compact support such that the solution (ψ, V) of (1.6) satisfies*

$$\psi(T) = \psi_t(T) \equiv 0 \text{ in } (0,1), \quad V(T) = V_t(T) = 0. \tag{2.1}$$

Remark 1. In the statement of Theorem 2.1 and below, we drop the index n from the unknowns (ψ, V) to simplify the notation.

The solution (ψ, V) is defined by transposition. Therefore, (2.1) has to be understood in a suitable weak sense. We will return to this question in the proof of the theorem.

The proof of Theorem 2.1 provides the continuous dependence of the control β on the initial data. More precisely,

$$\|\beta\|^2_{H^{-2}(0,T)} \leq C_n \left\{ \|(\psi^1, \psi^0, V^1, V^0)\|^2_{\mathcal{Y}'} + \mid \psi^0(0) \mid^2 \right\} \tag{2.2}$$

for any initial data $(\psi^0, \psi^1, V^0, V^1)$, as in the statement of Theorem 2.1. By $\mathcal{Y}'$ we denote the dual of the space $\mathcal{Y}$, $(\mathcal{Y}' = (H^1(0,1)) \times L^2(0,1) \times \mathbb{R}^2)$. The constant C_n in (2.2) will be evaluated in the next section (see also Remark 4).

As we said in the introduction, when $n = 0$ one can not expect the same controllability result due to the conservation of the quantities (1.9) along the trajectories. In this case the controllability result reads as follows.

THEOREM 2.2. *Assume that $T > 2$ and $n = 0$. Then, for any $\left(\psi^0, \psi^1, V^0, V^1\right) \in \mathcal{Y}$, there exists a control $\beta \in H^{-2}(0,T)$ with compact support such that the solution (ψ, V) of (1.6) satisfies*

$$\psi(T) = V^1 + \psi^0(0), \psi_t(T) = 0 \text{ in } (0,1),\ V(T) = V^0 - \int_0^1 \psi^1 dy, V_t(T) = 0. \tag{2.3}$$

Remark 2. This result asserts that, when $n = 0$, any solution of (1.6) can be driven to an equilibrium configuration which is determined a priori by the initial data.

Let us now state the controllability results for the two-dimensional system (1.3).

We use the Fourier decomposition method described in the introduction. Thus we develop the initial data $\left(\phi^0, \phi^1, W^0, W^1\right)$ to be controlled in Fourier series:

$$\left\{\begin{array}{ll} \phi^0 = \displaystyle\sum_{n=0}^{\infty} \psi_n^0(y)\cos(n\pi x), & \phi^1 = \displaystyle\sum_{n=0}^{\infty} \psi_n^1(y)\cos(n\pi x), \\ W^0 = \displaystyle\sum_{n=0}^{\infty} V_n^0 \cos(n\pi x), & W^1 = \displaystyle\sum_{n=0}^{\infty} V_n^1 \cos(n\pi x). \end{array}\right. \tag{2.4}$$

We assume that for every $n = 0, 1, \ldots$ the initial data satisfy the assumptions of Theorems 2.1 and 2.2. We set

$$\rho_n^0 = \psi_n^0, \quad \rho_n^1 = -\psi_n^1 + V_n^0\delta_0, \quad \mu^0 = -V_n^0, \mu_n^1 = V_n^1 + \psi_n^0(0). \tag{2.5}$$

We introduce the following space of initial data:

$$\begin{aligned} H = \Bigg\{ & \left(\phi^0, \phi^1, W^0, W^1\right) \in \mathcal{X} : \\ & \sum_{n=0}^{\infty} C_n \left\| \left(\rho_n^1, \rho_n^0, \mu_n^1, \mu_n^0\right)\right\|_{\mathcal{Y}'}^2 = \left\|\left(\phi^0, \phi^1, W^0, W^1\right)\right\|_H^2 < \infty \Bigg\}, \end{aligned} \tag{2.6}$$

where the constants C_n are those appearing in (2.2).

THEOREM 2.3. *Assume that $T > 2$. Then, for every initial data $\left(\phi^0, \phi^1, W^0, W^1\right)$ in H, there exists a control $\beta \in H^{-2}(0, T; L^2(0,1))$ such that the solution (ϕ, W) of (1.3) satisfies*

$$\left\{\begin{array}{l} \phi(T) \equiv \mu^1 = \displaystyle\int_0^1 W^1(x)dx + \int_0^1 \psi^0(x,0)dx,\ \phi_t(T) \equiv 0, \\ W(T) \equiv \langle \rho^1, 1\rangle = \displaystyle\int_0^1 W^0(x)dx - \int_0^1\int_0^1 \psi^1(x,y)dxdy,\ W_t(T) \equiv 0. \end{array}\right. \tag{2.7}$$

Moreover, there exists a constant $C > 0$ such that

$$\|\beta\|_{H^{-2}(0,T;L^2(0,1))} \leq C \| \left(\phi^0, \phi^1, W^0, W^1\right) \|_H. \tag{2.8}$$

Remark 3. The control time $T > 2$ is optimal. Indeed, when $T < 2$ it is easy to see that the set of controllable data is not dense in the space of finite energy data. Actually, when $T < 2$, none of the one-dimensional problems (1.6) is approximately controllable; i.e., the space of controllable data is not even dense in $\mathcal{Y}'$.

Remark 4. The developments of this article allow us to show that $C_n = \mathcal{O}\left(e^{n^\alpha}\right)$ as $n \to \infty$ for any $\alpha > 1$. Thus, roughly speaking, the Fourier coefficients in the x-variable have to decay exponentially to guarantee the controllability. Let us explain in more detail why this result is natural.

From the definition (2.6) of H and from the fact that C_n grows exponentially, it is clear that there is no Sobolev space that might be contained in H (observe that Sobolev spaces correspond roughly to polynomial weights C_n). But this is known a priori. Indeed, as we said in the introduction, our control problem does not verify the geometric control property given in [6], and as a consequence of this, no Sobolev space of initial data may be exactly controllable with β in $H^{-2}(0, T; L^2(0,1))$.

After the first version of this paper was written, Allibert in [1] obtained some complementary results. In [1], he proved that for any ε there exists $T(\varepsilon) > 0$ such

that system (1.3) is controllable in time $T(\varepsilon)$ for all initial data in the space $H(\varepsilon)$ which is defined as in (2.6) but with $C_n = \exp(\varepsilon n)$ as $n \to \infty$. Thus, the result in [1] shows, roughly, that as $t \to \infty$, the system is controllable in a larger and larger class of analytic functions. The results in [1] are an extension of previous results by the same author on the controllability of the classical wave equation in the square Ω and with control in Γ_0. Observe that all these problems have in common the fact that the geometric control condition of [6] is not satisfied. The structure of the set of controllable data in those situations is mainly unknown.

Since the constants C_n in our estimates are of order e^{n^α}, we can control all the initial data which belong to the Gevrey classes of exponent $\alpha > 1$ in the x-variable.

Remark 5. Finally, let us mention that if a second control $\alpha \in L^2(0,T)$ is allowed to act in the system through the condition of continuity of the velocity fields

$$\frac{\partial \Phi}{\partial y} = -W_t + \alpha \text{ in } \Gamma_0 \times (0,T), \tag{2.9}$$

the same result holds with $C_n = \mathcal{O}(n^4 e^{2n\pi})$. This is a consequence of Proposition 3.2 below. From the proof of Proposition 3.2 it follows that this constant is sharp. However, introducing controls of the form (2.9) does not seem to be realistic. This is the reason for using only the control β, which requires important additional developments.

3. Controllability of the one-dimensional systems. This section is devoted to proving the controllability results for the one-dimensional systems (1.6) that are necessary to derive the controllability of system (1.3). In section 3.1, by using classical multiplier techniques, we derive some hidden regularity results. In section 3.2, with the same techniques, we get the first observability inequalities. In section 3.3, by using Ingham's inequalities, we obtain a refined version of these observability inequalities. Finally, in sections 3.4 and 3.5, we apply HUM and prove the controllability result for (1.6).

3.1. Hidden regularity. Let us consider the system

$$\begin{cases} \eta_{tt} - \eta_{yy} + n^2\pi^2\eta = f & \text{in} \quad (0,1)\times(0,T), \\ \eta_y(1) = 0 & \text{for} \quad t \in (0,T), \\ \eta_y(0) = u_t & \text{for} \quad t \in (0,T), \\ u_{tt} + n^2\pi^2 u - \eta_t(0) = g & \text{for} \quad t \in (0,T), \\ \eta(0) = \eta^0, \eta_t(0) = \eta^1 & \text{in} \quad (0,1), \\ u(0) = u^0, u_t(0) = u^1. \end{cases} \tag{3.1}$$

System (3.1) is the adjoint of (1.6). The unknowns are $\eta = \eta(y,t)$ and $u = u(t)$. Of course, since the coefficients of the system depend on $n = 0, 1 \ldots$, solutions (η, u) depend on n too. However, in order to simplify the notations, we will not use the index n to distinguish the solutions of (3.1) for the different values of n.

The energy space for system (3.1) is the Hilbert space $\mathcal{Y} = H^1(0,1) \times L^2(0,1) \times \mathbb{R} \times \mathbb{R}$.

It is easy to see that for any $(\eta^0, \eta^1, u^0, u^1) \in \mathcal{Y}$ and $(f,g) \in L^1(0,T; L^2(0,1)\times\mathbb{R})$ system (3.1) has a unique solution in the class

$$\eta \in C\left([0,T]; H^1(0,1)\right) \cap C^1\left([0,T]; L^2(0,1)\right); \qquad u \in C^1([0,T];\mathbb{R}). \tag{3.2}$$

In other words, $(\eta, \eta_t, u, u_t) \in C\left([0,T]; \mathcal{Y}\right)$.

The energy of the system

$$F(t) = \frac{1}{2}\int_0^1 \left[\mid \eta_t \mid^2 + \mid \eta_y \mid^2 + n^2\pi^2\eta^2 \right] dy + \frac{1}{2}\left[\mid u_t \mid^2 + n^2\pi^2 \mid u \mid^2 \right] \tag{3.3}$$

satisfies

$$\frac{dF(t)}{dt} = \int_0^1 f(y,t)\eta_t(y,t)dy + g(t)u_t(t). \tag{3.4}$$

Therefore, when $f \equiv 0$ and $g \equiv 0$, the energy F remains constant along trajectories.

We observe that when $n \geq 1$, the square root of F defines a norm in $\mathcal{Y}$ equivalent to the canonical norm $\| \cdot \|_{\mathcal{Y}}$ of $\mathcal{Y}$:

$$\|(u,v,w,z,)\|_{\mathcal{Y}} = \left[\int_0^1 \left(\mid u_y \mid^2 + \mid u \mid^2 + \mid v \mid^2 \right) dy + w^2 + z^2 \right]^{1/2}. \tag{3.5}$$

However, when $n = 0$, this is not the case. Actually, for $n = 0, (\eta, u) = (c_1, c_2)$ with c_1, c_2 real constants are stationary solutions of (3.1) with $f \equiv 0, g \equiv 0$ for which the energy F vanishes.

We have the following "hidden regularity" result.

PROPOSITION 3.1. *For any $T > 0$ there exists a constant $C(T) > 0$ independent of $n = 0, 1, \ldots$ such that*

$$\begin{aligned} &\left(\int_0^T \mid u_{tt} \mid dt \right)^2 + \int_0^T \left[\mid u_t \mid^2 + (1 + n^4\pi^4)u^2 + (1 + n^2\pi^2)\eta^2(0,t) \right] dt \\ &\leq C\left(n^4 + 1\right) \left[\| \left(\eta^0, \eta^1, u^0, u^1\right) \|^2_{\mathcal{Y}} + \|f\|^2_{L^1(0,T;L^2(0,1))} + \|g\|^2_{L^1(0,T)} \right] \end{aligned} \tag{3.6}$$

for any $(\eta^0, \eta^1, u^0, u^1) \in \mathcal{Y}$, $f \in L^1(0,T;L^2(0,1))$, and $g \in L^1(0,T)$.

If $g \in L^2(0,T)$, then $u \in H^2(0,T)$, and we also have

$$\begin{aligned} &\int_0^T \mid u_{tt} \mid^2 dt \\ &\leq C(n^4 + 1) \left[\|(\eta^0, \eta^1, u^0, u^1)\|^2_{\mathcal{Y}} + \|f\|^2_{L^1(0,T;L^2(0,1))} + \|g\|^2_{L^2(0,T)} \right]. \end{aligned} \tag{3.7}$$

Remark 6. This proposition shows that u is more smooth than what (3.2) guarantees. This is due to the structure of the second-order (in time) equations that u satisfies. The fact that the constant in (3.6) and (3.7) does not depend on the index n is worth mentioning.

Proof of Proposition 3.1. It is enough to consider smooth solutions since a classical density argument allows us to extend inequalities (3.6) and (3.7) to any solution with finite right-hand side. We use a classical multiplier technique (see, for instance, [10]). We multiply the first equation in (3.1) by $(1-y)\eta_y$ and integrate over $(0,1) \times (0,T)$. Integrating by parts we obtain

$$\begin{aligned} &\frac{1}{2}\int_0^T \left[\mid \eta_t \mid^2 + \mid \eta_y \mid^2 - n^2\pi^2\eta^2 \right](0,t)dt = -\int_0^1 \eta_t(1-y)\eta_y dy \Big|_0^T \\ &+\frac{1}{2}\int_0^T \int_0^1 \left[\eta_t^2 + \eta_y^2 - n^2\pi^2\eta^2 \right] dydt + \int_0^T \int_0^1 f(1-y)\eta_y dydt = X. \end{aligned}$$

In this identity we use the notation $L\mid_0^T = L(T) - L(0)$. The right-hand side of this identity can be easily bounded as follows:

$$\mid X \mid \leq \frac{1}{2}\int_0^1 \left[\eta_t^2 + \eta_y^2\right](y,0)dy + \frac{1}{2}\int_0^1 \left[\eta_t^1 + \eta_y^2\right](y,T)dy + \int_0^T F(t)dt$$

$$+\frac{1}{2}\left[\|f\|^2_{L^1(0,T;L^2(0,1))} + \|\eta_y\|^2_{L^\infty(0,T;L^2(0,1))}\right] \leq F(0) + F(T) + \int_0^T F(t)dt$$

$$+\|F(t)\|_{L^\infty(0,T)} + \frac{1}{2}\|f\|^2_{L^1(0,T;L^2(0,1))} \leq C\left[\|F\|_{L^\infty(0,T)} + \|f\|^2_{L^1(0,T;L^2(0,1))}\right],$$

with $C > 0$ independent of n.

Below, if some constant in the inequalities depends on n, we will make it explicit by means of an index n on that constant.

On the other hand, from identity (3.4) and using Gronwall's inequality, it is easy to deduce that

$$\|F\|^2_{L^\infty(0,T)} \leq C\left[\|f\|^2_{L^1(0,T;L^2(0,1))} + \|g\|^2_{L^1(0,T)} + F(0)\right].$$

Since $H^1(0,1)$ is continuously embedded in $C([0,1];\mathbb{R})$ we also have

$$\int_0^T \eta^2(0,t)dt \leq C\int_0^T F(t)dt \leq C\left[\|f\|^2_{L^1(0,T;L^2(0,1))} + \|g\|^2_{L^1(0,T)} + F(0)\right].$$

Combining these inequalities we deduce that

$$\int_0^T \left[\mid \eta_t \mid^2 + \mid \eta_y \mid^2 + n^2\pi^2\eta^2\right](0,t)dt$$

$$\leq C(n^2+1)\left[\|(\eta^0,\eta^1,u^0,u^1)\|^2_{\mathcal{Y}} + \|f\|^2_{L^1(0,T;L^2(0,1))} + \|g\|^2_{L^1(0,T)}\right]. \tag{3.8}$$

On the other hand,

$$n^4\pi^4\int_0^T u^2(t)dt \leq 2n^2\pi^2\int_0^T F(t)dt$$

$$\leq Cn^4\left[\|(\eta^0,\eta^1,u^0,u^1)\|^2_{\mathcal{Y}} + \|f\|^2_{L^1(0,T;L^2(0,1))} + \|g\|^2_{L^1(0,T)}\right]. \tag{3.9}$$

Inequalities (3.6) and (3.7) are a direct consequence of (3.8) and (3.9) and the coupling conditions between η and u given in system (3.1), i.e.,

$$\eta_y(0,t) = u_t(t); u_{tt}(t) = g(t) + \eta_t(0,t) - n^2\pi^2 u(t) \text{ for } t \in (0,T). \tag{3.10}$$ □

3.2. Observability inequalities. In this paragraph we consider the adjoint system (3.1) in the particular case where $f \equiv 0$ and $g \equiv 0$. More precisely, assume that η and u solve

$$\begin{cases} \eta_{tt} - \eta_{yy} + n^2\pi^2\eta = 0 & \text{in} \quad (0,1)\times(0,T), \\ \eta_y(1,t) = 0 & \text{for} \quad t \in (0,T), \\ \eta_y(0,t) = u_t(t) & \text{for} \quad t \in (0,T), \\ u_{tt}(t) + n^2\pi^2 u(t) - \eta_t(0,t) = 0 & \text{for} \quad t \in (0,T), \\ \eta(0) = \eta^0, \eta_t(0) = \eta^1 & \text{in} \quad (0,1), \\ u(0) = u^0, u_t(0) = u^1. \end{cases} \tag{3.11}$$

We have the following observability result.

PROPOSITION 3.2. *For any $T > 2$ there exists a constant $C > 0$ which is independent of $n = 0, 1, \ldots$ such that*

$$2F(0) + \|\eta^0\|^2_{L^2(0,1)} + |u^0|^2 \leq Ce^{2n\pi} \int_0^T \left[|u_{tt}|^2 + |u_t|^2 + (1+n^4\pi^4)|u|^2 + (1+n^2\pi^2)|\eta(0,t)|^2 \right] dt \tag{3.12}$$

for any solution of (3.11).

Remark 7. Let $\rho : (0,T) \to [0,1]$ be a nonnegative smooth function with compact support and $\rho \equiv 1$ in $(\varepsilon, T-\varepsilon)$ with $\varepsilon > 0$ small enough such that $T - 2\varepsilon > 2$. In view of the time invariance of system (3.11), we deduce that

$$2F(\varepsilon) + \|\eta(\varepsilon)\|^2_{L^2(0,1)} + |u(\varepsilon)|^2 \leq Ce^{2n\pi} \int_0^T \rho(t) \left[|u_{tt}|^2 + (1+n^4\pi^4)|u|^2 + (1+n^2\pi^2)|\eta(0,t)|^2 \right] dt.$$

Using the conservation of energy, we deduce that

$$\begin{aligned} \|(\eta^0, \eta^1, u^0, u^1)\|^2_{\mathcal{Y}} &\leq 2F(0) + \|\eta^0\|^2_{L^2(0,1)} + |u^0|^2 \\ &\leq Ce^{2n\pi} \int_0^T \rho(t) \left[|u_{tt}|^2 + |u_t|^2 + (1+n^4\pi^4)|u|^2 \right. \\ &\qquad \left. + (1+n^2\pi^2)|\eta(0,t)|^2 \right] dt. \end{aligned} \tag{3.13}$$

This estimate will allow us to construct controls with compact support in time.

Proof of Proposition 3.2. The proof of this result is obtained by means of a genuinely one-dimensional method which consists roughly on viewing the wave equation in (3.11) as being an evolution equation with respect to y, while t plays the role of the space variable. This argument was used in [18] when studying the controllability of the one-dimensional semilinear wave equation.

For any $0 \leq y \leq 1$, we define

$$G(y) = \frac{1}{2} \int_y^{T-y} \left[|\eta_t|^2 + |\eta_y|^2 + n^2\pi^2 |\eta|^2 \right](y,t) dt.$$

We have

$$G(0) = \frac{1}{2} \int_0^T \left[|\eta_t|^2 + |\eta_y|^2 + n^2\pi^2 |\eta|^2 \right](0,t) dt. \tag{3.14}$$

On the other hand,

$$\begin{aligned} G'(y) = &\int_y^{T-y} \left[|\eta_{yy}\eta_y + \eta_{ty}\eta_t + n^2\pi^2\eta_y\eta| \right](y,t) dt \\ &- \frac{1}{2} \sum_{t=y,T-y} \left[|\eta_y(y,t)|^2 + |\eta_t(y,t)|^2 + n^2\pi^2 |\eta(y,t)|^2 \right] \end{aligned}$$

and

$$\int_y^{T-y} \eta_{ty}(y,t)\eta_t(y,t) dt = - \int_y^{T-y} \eta_y(y,t)\eta_{tt}(y,t) dt + \eta_y(y,t)\eta_t(y,t) \Big|_{t=y}^{t=T-y}.$$

Therefore

$$G'(y) = \int_y^{T-y} \left[\eta_{yy} - \eta_{tt} + n^2\pi^2\eta\right] \eta_y(y,t)dt + \eta_y(y,t)\eta_t(y,t)\Big|_{t=y}^{t=T-y}$$

$$- \frac{1}{2} \sum_{t=y,T-y} \left[\mid \eta_y(y,t) \mid^2 + \mid \eta_t(y,t) \mid^2 + n^2\pi^2 \mid \eta(y,t) \mid^2\right]. \tag{3.15}$$

Using the first equation in (3.11) we have that

$$\int_y^{T-y} \left[\eta_{yy} - \eta_{tt} + n^2\pi^2\eta\right] \eta_y(y,t)dt = 2n^2\pi^2 \int_y^{T-y} \eta\eta_y(y,t)dt,$$

and on the other hand,

$$\eta_y(y,t)\eta_t(y,t)|_{t=y}^{t=T-y} - \frac{1}{2} \sum_{t=y,T-y} \left[\mid \eta_y(y,t) \mid^2 + \mid \eta_t(y,t) \mid^2 + n^2\pi^2 \mid \eta(y,t) \mid^2\right] \leq 0.$$

Combining these identities with (3.15), we deduce that

$$G'(y) \leq 2n^2\pi^2 \int_y^{T-y} \eta\eta_y(y,t)dt$$

$$\leq n\pi \int_y^{T-y} \left[\mid \eta_y \mid^2 + n^2\pi^2 \mid \eta \mid^2\right](y,t)dt \leq 2n\pi G(y).$$

Thus $G(y) \leq e^{2n\pi}G(0)$, for all $y \in (0,1)$, and therefore $\int_0^1 G(y) \leq e^{2n\pi}G(0)$.

In particular,

$$(T-2)F(T) = \int_1^{T-1} F(t)dt$$

$$= \frac{1}{2}\int_1^{T-1} \left\{\left[\int_0^1 \mid \eta_y \mid^2 + \mid \eta_t \mid^2 + n^2\pi^2\eta^2\right] dy + \mid u_t \mid^2 + n^2\pi^2u^2\right\} dt$$

$$\leq \int_0^1 G(y)dy + \frac{1}{2}\int_1^{T-1} \left[\mid u_t \mid^2 + n^2\pi^2u^2\right] dt$$

$$\leq \frac{e^{2n\pi}}{2} \int_0^T \left[\mid \eta_y \mid^2 + \mid \eta_t \mid^2 + n^2\pi^2\eta^2\right](0,t)dt + \frac{1}{2}\int_0^T \left[\mid u_t \mid^2 + n^2\pi^2u^2\right] dt. \tag{3.16}$$

Using the relations (3.11) at $y = 0$, we deduce that (3.12) holds when $n \geq 1$.

When $n = 0$, it is sufficient to add in (3.16) the extra quantity $\int_0^T [\mid \eta \mid^2 (0,t) + \mid u \mid^2 (t)]dt$ to deduce that (3.12) holds in that case too. □

Remark 8. When $n = 0$, inequality (3.12) shows that

$$\|\eta_y^0\|^2_{L^2(0,1)} + \|\eta^1\|^2_{L^2(0,1)} + \mid u^1 \mid^2 \leq C \int_0^T \left[\mid u_{tt} \mid^2 + \mid u_t \mid^2\right] dt. \tag{3.17}$$

This inequality does not provide any estimate on u^0. This is related to the fact that, when $n = 0$, system (1.6) cannot be driven exactly to zero but rather to the equilibrium given by the constants c_1, c_2 in (1.9).

3.3. Improved observability inequalities. The goal of this section is to obtain observability inequalities of the form (3.12) but in which the only term appearing on the right-hand side is $\int_0^T \mid u_{tt} \mid^2 dt$. As we will see, this is related to the controllability of system (1.6) using the sole control β. We have the following theorem.

THEOREM 3.3. *Assume that $T > 2$. Then*

(i) *for any $n \geq 1$ there exists a constant $C = C(T, n) > 0$ such that*

$$\|(\eta^0, \eta^1, u^0, u^1)\|^2_{\mathcal{Y}} \leq C(T, n) \int_0^T \mid u_{tt} \mid^2 dt \tag{3.18}$$

for any solution of (3.11). *Moreover, $C(T, n) = \mathcal{O}\left(e^{n^\alpha}\right)$, for any $\alpha > 1$.*

(ii) *if $n = 0$ there exists a constant $C = C(T) > 0$ such that*

$$\|\eta^0_y\|^2_{L^2(0,1)} + \mid u^1 \mid^2 \leq C(T) \int_0^T \mid u_{tt} \mid^2 dt \tag{3.19}$$

for any solution of (3.11).

Remark 9. As observed in Remark 7, in estimates (3.18) and (3.19) one can replace the right-hand side by the quantity $\int_0^T \rho(t) \mid u_{tt}(t) \mid^2 dt$, where ρ is a smooth nonnegative function with compact support in $(0, T)$ and such that $\rho \equiv 1$ in $(\varepsilon, T - \varepsilon)$ with $\varepsilon > 0$ small enough such that $T - 2\varepsilon > 2$.

To prove Theorem 3.3 we need the following refined version of a result by Haraux [8] on nonharmonic Fourier series.

THEOREM 3.4. *Let $f = f(t)$ be of the form $f(t) = \sum_{n \in \mathbb{Z}} a_n e^{i\lambda_n t}$, where λ_n is a sequence of real numbers. We assume that there exist $N \in \mathbb{N}, \gamma > 0$, and $\gamma_\infty > 0$ such that*

$$\lambda_{n+1} - \lambda_n \geq \gamma_\infty > 0 \text{ if } \mid n \mid > N, \tag{3.20}$$

$$\lambda_{n+1} - \lambda_n \geq \gamma > 0 \text{ for any } n \in \mathbb{Z}. \tag{3.21}$$

Let $J = [0, T] \subset \mathbb{R}$ be a finite interval with $T > \frac{2\pi}{\gamma_\infty}$. Then, there exist two positive constants $C_1, C_2 > 0$ such that

$$C_1 \sum_{n \in \mathbb{Z}} \mid a_n \mid^2 \leq \int_J \mid f(t) \mid^2 dt \leq C_2 \sum_{n \in \mathbb{Z}} \mid a_n \mid^2 \tag{3.22}$$

for all $(a_n)_n \in l^2$.

More precisely $C_1 = C_1(2N + 1)$ and $C_2 = C_2(2N + 1)$, where $C_i(j), i = 1, 2$, are given by the following recurrent formulas:

$$\begin{cases} C_1(j+1) = \left[\left(\dfrac{2C_2(j)}{\mid J \mid} + 1\right) \dfrac{4}{C_1(j)(\mid J \mid \gamma_\infty - 2\pi)^2 \gamma^2} + \dfrac{2}{\mid J \mid}\right]^{-1}, \\ C_2(j+1) = 2\left(\mid J \mid (j+1) + C_2(0)\right), \; j = 0, 1, \ldots, \end{cases} \tag{3.23}$$

and $C_1(0), C_2(0)$ are such that (3.22) *holds in the particular case in which $\gamma_\infty = \gamma > 0$.*

Remark 10. (a) When $\gamma_\infty = \gamma$, a sequence on the conditions of Theorem 3.4 satisfies $\lambda_{n+1} - \lambda_n \geq \gamma > 0\,, \forall n \in \mathbb{Z}$. In this particular case the classical result by Ingham [9] shows the existence of $c_1, c_2 > 0$ such that (3.22) holds when $\mid J \mid > \frac{2\pi}{\gamma}$.

Theorem 3.4 allows us to deduce that (3.22) holds when the length of the interval J is smaller. Indeed, it suffices that $\mid J \mid > 2\pi/\gamma_\infty, \gamma_\infty$ being the "asymptotic gap" of the sequence $\{\lambda_n\}$, which is in general larger than γ. This relaxed gap condition was shown to be sufficient for (3.22) in [4]. Later, Haraux in [8] gave a constructive proof which allows us to give explicit estimates on the constants C_1 and C_2. Following the construction in [8], one can easily see that (3.23) suffices. In the appendix at the end of this paper, we give all the details of this construction.

(b) Clearly, the constants C_1 and C_2 degenerate as $N \to \infty$. More precisely, $C_2(N) = \mathcal{O}(N)$ while $\gamma^{2N}(C_1(N))^{-1} = \mathcal{O}\left(e^{N^\alpha}\right)$ for any $\alpha > 1$. Indeed, we have

$$(C_1(N))^{-1} \leq \frac{2C_2(N)}{|J|}\frac{4}{(\mid J \mid \gamma_\infty - 2\pi)^2\gamma^2}(C_1(N-1))^{-1} = \frac{16N(C_1(N-1))^{-1}}{(\mid J \mid \gamma_\infty - 2\pi)^2\gamma^2}.$$

Hence

$$\gamma^{2N}(C_1(N))^{-1} \leq \left(\frac{16}{(\mid J \mid \gamma_\infty - 2\pi)^2\gamma^2}\right)^N N!\,(C_1(0))^{-1} \leq e^{N^\alpha}\,(C_1(0))^{-1}.$$

In order to apply Theorem 3.4 and deduce that Theorem 3.3 holds, we need precise estimates on the spectrum of (3.11). We look for solutions of (3.11) in separated variables of the form $(\eta, u) = e^{\nu t}(\varphi(y), \omega)$ with $\varphi = \varphi(y)$ and $\omega \in \mathbb{R}$. Due to the conservative character of the system, we know that all eigenvalues ν are purely imaginary. On the other hand, the spectrum is symmetric with respect to the real axis. Thus, for any $n = 0, 1, \ldots$ there exists a sequence of eigenvalues $\nu_{n,m}$ with $\overline{\nu}_{n,m} = -\nu_{n,m} = \nu_{-n,m}$.

We have the following estimates.

THEOREM 3.5 (see [12] and [14]). *For any $n = 0, 1, \ldots$ and $m \in \mathbb{Z}$ such that $\mid m \mid > n$ we have*

$$\begin{cases} \left|\nu_{n,m} - \sqrt{m^2+n^2}\pi i\right| & \leq \dfrac{24}{\sqrt{m^2+n^2}\pi} \quad \text{if } m > n, \\ \left|\nu_{n,m} + \sqrt{m^2+n^2}\pi i\right| & \leq \dfrac{24}{\sqrt{m^2+n^2}\pi} \quad \text{if } m < -n. \end{cases} \tag{3.24}$$

Remark 11. This theorem shows that, for sufficiently high frequencies, the eigenvalues of (3.11) are uniformly close to the eigenvalues $\lambda = \pm\sqrt{m^2+n^2}\pi i$ of the wave equation with Neumann boundary conditions

$$\begin{cases} \eta_{tt} - \eta_{yy} + n^2\pi^2\eta = 0 & \text{in} \quad (0,1)\times(0,\infty), \\ \eta_y(0,t) = \eta_y(1,t) = 0 & \text{for} \quad t > 0. \end{cases} \tag{3.25}$$

Clearly, system (3.25) corresponds to the decomposition of the wave equation with Neumann boundary conditions in the square Ω following the development (1.5) in Fourier series. In other words, Theorem 3.5 asserts that the spectrum of the adjoint system of (1.1), i.e.,

$$\begin{cases} \phi_{tt} - \Delta\phi = 0 & \text{in} \quad \Omega\times(0,\infty), \\ \frac{\partial\phi}{\partial\nu} = 0 & \text{on} \quad \Gamma_1\times(0,\infty), \\ \frac{\partial\phi}{\partial y} = W_t & \text{on} \quad \Gamma_0\times(0,\infty), \\ W_{tt} - W_{xx} - \phi_t = 0 & \text{on} \quad \Gamma_0\times(0,\infty), \\ W_x(0,t) = W_x(1,t) = 0 & \text{for} \quad t > 0, \end{cases}$$

at high frequencies is uniformly close to the eigenvalues of the wave equation with Neumann boundary conditions on the whole boundary of the cavity Ω:

$$\begin{cases} \phi_{tt} - \Delta\phi = 0 & \text{in} \quad \Omega \times (0,\infty), \\ \frac{\partial\phi}{\partial\nu} = 0 & \text{on} \quad \partial\Omega \times (0,\infty). \end{cases}$$

This means roughly that the effect of the flexible boundary in the interior of the cavity is negligible for high frequencies. However, it is worth mentioning that the high frequency asymptotics are of a different nature in the region $\mid m \mid \leq n$.

From Theorem 3.5 it is easy to get explicit bounds on the gaps γ and γ_∞ associated with the sequence $\{\nu_{n,m}\}_{m\in\mathbb{Z}}$ for each $n = 0, 1, \ldots$.

PROPOSITION 3.6. *Given any $n = 0, 1, \ldots$ and $0 < \delta < \pi$ we have*

$$\mid \nu_{n,m+1} - \nu_{n,m} \mid \geq \pi - \delta \tag{3.26}$$

for any m with $\mid m \mid \geq N(n,\delta)$, where

$$N(n,\delta) = \max\left[\sqrt{\frac{96}{\pi\delta} - n^2}, \frac{2n\pi}{\delta} - n - \frac{1}{2}\right]. \tag{3.27}$$

On the other hand,

$$\begin{cases} \mid \nu_{n,m+1} - \nu_{n,m} \mid \geq \frac{\pi}{4} \quad \forall m \in \mathbb{Z} & \text{if} \quad n = 0, 1, \\ \mid \nu_{n,m+1} - \nu_{n,m} \mid \geq \frac{\pi}{1+2n} \quad \forall m \in \mathbb{Z} & \text{if} \quad n \geq 2. \end{cases} \tag{3.28}$$

Furthermore, (3.22) *holds for functions f of the form*

$$f(t) = \sum_{m\in\mathbb{Z}^\star} a_{n,m} e^{-\nu_{n,m}t} + a_n^\star e^{-\nu_n^\star t} + a_n^{\star\star} e^{-\nu_n^{\star\star} t} \tag{3.29}$$

with $C_2 = C_2(2N(n,\delta)+1) = \mathcal{O}(n)$ and $(C_1)^{-1} = (C_1(2N(n,\delta)+1))^{-1}$. Moreover, $C_2 = \mathcal{O}(n)$ and $(C_1)^{-1} = \mathcal{O}\left(e^{n^\alpha}\right)$ for any $\alpha > 1$.

Proof. In view of (3.24) we have

$$\begin{aligned} &\mid \nu_{n,m+1} - \nu_{n,m} \mid \\ &\geq \pi \left|\sqrt{(m+1)^2+n^2} - \sqrt{m^2+n^2}\right| - \tfrac{24}{\pi}\left[\tfrac{1}{\sqrt{(m+1)^2+n^2}} + \tfrac{1}{\sqrt{m^2+n^2}}\right] \\ &\geq \tfrac{(2|m|+1)\pi}{(2|m|+1)+2n} - \tfrac{48}{\pi\sqrt{m^2+n^2}} \geq \pi - \left[\tfrac{48}{\pi\sqrt{m^2+n^2}} + \tfrac{2n\pi}{2|m|+1+2n}\right]. \end{aligned}$$

It is easy to see that when $\mid m \mid \geq N(n,\delta)$, where $N(n,\delta)$ is given by (3.27), then

$$\frac{48}{\pi\sqrt{m^2+n^2}} + \frac{2n\pi}{2 \mid m \mid +1+n} \leq \delta.$$

This concludes the proof of (3.26).

To prove (3.28) we observe that, for any $n = 0, 1, \ldots$, the eigenvalues $\nu_{n,m}$ with $m > 0$ are of the form

$$\nu_{n,m} = \sqrt{z_{n,m}^2 + n^2\pi^2}, \tag{3.30}$$

where $z_{n,m}$ are the zeros (ordered so that $z_{n,m}$ increases as m does) of the equation

$$tgz = \frac{z^2 + n^2\pi^2}{z^3}. \tag{3.31}$$

There are also two eigenvalues, which we denote by $\nu_n^\star$ and $\nu_n^{\star\star}$, that do not satisfy (3.30). Indeed, they are given by

$$\nu_n^\star = \sqrt{n^2\pi^2 - (z_n^\star)^2}, \tag{3.32}$$

where $z_n^\star$ is the unique real positive solution of

$$e^{2z} = \frac{z^3 - z^2 + n^2\pi^2}{z^3 + z^2 - n^2\pi^2}, \tag{3.33}$$

when $n \geq 1$ and $\nu_0^\star = 0$, and $\nu_n^{\star\star} = \overline{\nu_n^\star}$.

By analyzing the graphs of the functions in (3.31) and (3.33) it is easy to see that (3.28) holds. We refer to [14] for a detailed proof.

To finish the proof we have to apply Theorem 3.4 for $\gamma = \min\left\{\frac{\pi}{4}, \frac{\pi}{1+2n}\right\}$ and $\gamma_\infty = \pi - \delta$. We obtain that (3.22) holds for functions f of the form (3.29).

In order to evaluate the constants, we use the recurrent formulas (3.23). We have

$$C_2 = C_2(2N(n,\delta)+1) = 2(T(N(n,\delta)+1) + C_2(0)) = \mathcal{O}(n).$$

On the other hand,

$$(C_1)^{-1} = (C_1(2N(n,\delta)+1))^{-1} \leq \frac{8C_2(N(n,\delta)+1)(C_1(2N(n,\delta)))^{-1}}{T(T\gamma_\infty - 2\pi)^2\gamma^2}$$

$$\leq Mn^3(C_1(2N(n,\delta)))^{-1} \leq M^{2N(n,\delta)+1}((N(n,\delta)+1)!)^3(C_1(0))^{-1} \leq C(\alpha)e^{n^\alpha},$$

where M is a positive constant and $\alpha > 1$. □

Now we have all the ingredients we need to prove Theorem 3.3.

Proof of Theorem 3.3. Let us consider first the case $n \geq 1$. In view of Proposition 3.2 it is sufficient to show the existence of a constant C (depending on n and T) such that

$$\int_0^T \left[\mid u_t \mid^2 + n^4\pi^4 \mid u \mid^2 + n^2\pi^2 \mid \eta(0,t) \mid^2 \right] dt \leq C \int_0^T \mid u_{tt} \mid^2 dt \tag{3.34}$$

holds for any solution of (3.11).

Let $U(t) = (\eta(t), \eta_t(t), u(t), u_t(t))$ be the vector-valued unknown associated with (3.11) viewed as a first-order (in time) system. Let us denote by $\xi_\nu = (\xi_\nu^1, \xi_\nu^2, \xi_\nu^3, \xi_\nu^4)$ the vector-valued eigenfunction of system (3.11) associated with the eigenvalue ν.

The solutions η and u of (3.11) can be written as follows:

$$\eta(t) = \sum_{m\in\mathbb{Z}^\star} a_{n,m}e^{-\nu_{n,m}t}\xi_{n,m}^1 + a_n^\star e^{-\nu_n^\star t}\xi_{\nu_n^\star}^1 + a_n^{\star\star}e^{-\nu_n^{\star\star}t}\xi_{\nu_n^{\star\star}}^1,$$

$$u(t) = \sum_{m\in\mathbb{Z}^\star} a_{n,m}e^{-\nu_{n,m}t}\xi_{n,m}^3 + a_n^\star e^{-\nu_n^\star t}\xi_{\nu_n^\star}^3 + a_n^{\star\star}e^{-\nu_n^{\star\star}t}\xi_{\nu_n^\star}^3,$$

where the coefficients $\{a_{n,m}, a_n^\star, a_n^{\star\star}\}$ are those associated with the development of the initial data on the orthogonal basis generated by the eigenfunctions.

To get the bounds in (3.34) we first observe that

$$\eta(0,t)=\sum_{m\in\mathbb{Z}^\star} a_{n,m}e^{-\nu_{n,m}t}\xi^1_{n,m}(0)+a^\star_n e^{-\nu^\star_n t}\xi^1_{\nu^\star_n}(0)+a^{\star\star}_n e^{-\nu^{\star\star}_n t}\xi^1_{\nu^\star_n}(0)$$

and

$$\eta_t(0,t)=-\sum_{m\in\mathbb{Z}^\star} a_{n,m}\nu_{n,m}e^{-\nu_{n,m}t}\xi^1_{n,m}(0)-a^\star_n\nu^\star_n e^{-\nu^\star_n t}\xi^1_{\nu^\star_n}(0)-a^{\star\star}_n\nu^{\star\star}_n e^{-\nu^{\star\star}_n t}\xi^1_{\nu^\star_n}(0).$$

In view of Proposition 3.6 we can apply Theorem 3.4 to these series in any time interval $J=(0,T)$ with $T>2$. Therefore, taking into account that $\mid \nu^\star_n \mid = \min\{\mid \nu_{n,m}\mid, \mid \nu^\star_n\mid, \mid \nu^{\star\star}_n\mid\}$, we have

$$\int_0^T \mid \eta(0,t)\mid^2 dt \le C_2\left(\sum_{m\in\mathbb{Z}^\star}\mid a_{n,m}\xi^1_{n,m}(0)\mid^2+\mid a^\star_n\xi^1_{\nu^\star_n}(0)\mid^2+\mid a^{\star\star}_n\xi^1_{\nu^\star_n}(0)\mid^2\right)$$

$$\le \frac{C_2}{\mid\nu^\star_n\mid^2}\left(\sum_{m\in\mathbb{Z}^\star}\mid a_{n,m}\xi^1_{n,m}(0)\,\nu_{n,m}\mid^2+\mid a^\star_n\xi^1_{\nu^\star_n}(0)\,\nu^\star_n\mid^2+\mid a^{\star\star}_n\xi^1_{\nu^\star_n}(0)\,\nu^{\star\star}_n\mid^2\right)$$

$$\le \frac{C_2}{C_1\mid\nu^\star_n\mid^2}\int_0^T\mid\eta_t(0,t)\mid^2 dt.$$

On the other hand, from the equation that u satisfies in (3.11), we have

$$\int_0^T(\eta_t(0,t))^2\,dt\le 2\int_0^T\left[\mid u_{tt}\mid^2+n^4\pi^4\mid u\mid^2\right]dt.$$

Thus, in order to conclude (3.34), it is sufficient to show that

$$\int_0^T\left[\mid u_t\mid^2+n^4\pi^4u^2\right]dt\le C\int_0^T\mid u_{tt}\mid^2\,dt$$

holds. The argument we have used to bound $\int_0^T\mid\psi(0,t)\mid^2 dt$ allows us to show that

$$\int_0^T\mid u\mid^2 dt\le\frac{C_2}{C_1\mid\nu^\star_n\mid^4}\int_0^T\mid u_{tt}\mid^2 dt \text{ and } \int_0^T\mid u_t\mid^2 dt\le\frac{C_2}{C_1\mid\nu^\star_n\mid^2}\int_0^T\mid u_{tt}\mid^2 dt.$$

Combining these results we deduce that (3.34) holds with a constant C of the order of

$$C=\frac{C_2}{C_1}\left\{\frac{1}{\mid\nu^\star_n\mid^2}+\frac{n^4\pi^4}{\mid\nu^\star_n\mid^4}\left(1+\frac{2C_1}{C_1\mid\nu^\star_n\mid^2}\right)+\frac{2C_2}{C_1\mid\nu^\star_n\mid^2}\right\}, \tag{3.35}$$

where $C_1=C_1(2N+1), C_2=C_2(2N+1)$ are given by (3.23) with $N=N(n,\delta)$ as in (3.27), and $\delta>0$ such that $T=\frac{2\pi}{\pi-\delta}$.

We now proceed to estimate the constant C of (3.35). In [14] we prove that $\nu^\star_n\sim n\pi$. On the other hand, from Proposition 3.3, we have that $C_2(C_1)^{-1}=\mathcal{O}\left(e^{n^\alpha}\right)$. Finally, we obtain that $C=\mathcal{O}\left(e^{n^\alpha}\right)$ for all $\alpha>1$.

Let us now consider the case $n=0$. In view of (3.17) we have

$$\|\eta^0_y\|^2_{L^2(0,1)}+\|\eta^1\|^2_{L^2(0,1)}+\mid u^1\mid^2\le\frac{1}{T-2}\int_0^T\left[\mid u_{tt}\mid^2+2\mid u_t\mid^2\right]dt. \tag{3.36}$$

Therefore, it is sufficient to show that

$$\int_0^T \mid u_t \mid^2 dt \leq C \int_0^T \mid u_{tt} \mid^2 . \tag{3.37}$$

Proceeding as above, we see that (3.37) holds with $C = C_2/C_1 \mid \nu_{n,1} \mid^2$, where $C_1 = C_1(2N+1), C_2 = C_2(2N+1)$, and $N = N(0,\delta)$ with $\delta > 0$ such that $T = \frac{2\pi}{\pi - \delta}$. □

3.4. Controllability in one space dimension for $n \geq 1$: Proof of Theorem 2.1. In this section, applying HUM, we prove Theorem 2.1 as a consequence of the observability inequality (3.18).

Given any $(\eta^0, \eta^1, u^0, u^1) \in \mathcal{Y}$ we solve the adjoint system (3.11).

We fix some nonnegative smooth function $\rho : (0,T) \to \mathbb{R}$ with compact support such that $\rho \equiv 1$ in $(\varepsilon, T-\varepsilon)$ with $T - 2\varepsilon > 2$.

We then solve the backward system

$$\begin{cases} \psi_{tt} - \psi_{yy} + n^2\pi^2\psi = 0 & \text{in} \quad (0,1)\times(0,T), \\ \psi_y(1,t) = 0 & \text{for} \quad t \in (0,T), \\ \psi_y(0,t) = -V_t(t) & \text{for} \quad t \in (0,T), \\ V_{tt} + n^2\pi^2 V + \psi_t(0,t) = -\frac{d^2}{dt^2}(\rho(t)u_{tt}(t)) & \text{for} \quad t \in (0,T), \\ \psi(T) = \psi_t(T) = 0 & \text{in} \quad (0,1), \\ V(T) = V_t(T) = 0. \end{cases} \tag{3.38}$$

The solution of (3.38) is defined by transposition (see [10]). If we multiply in (3.38) by any solution $(\widetilde{\eta}, \widetilde{u})$ of (3.1) and integrate (formally) by parts, we obtain the following identity:

$$\int_0^T \rho(t)u_{tt}(t)\widetilde{u}_{tt}(t)dt + \int_0^T \int_0^1 \widetilde{f}\psi dydt - \int_0^T \widetilde{g}V dt = \int_0^1 [-\psi_t(0)\widetilde{\eta}(0)$$

$$+\psi(0)\widetilde{\eta}_t(0)]\, dy + V(0)\widetilde{\eta}(0,0) + \psi(0,0)\widetilde{u}(0) - V(0)\widetilde{u}_t(0) + V_t(0)\widetilde{u}(0). \tag{3.39}$$

Notice that in the obtainment of (3.39) we have used the fact that ρ and its first derivative vanish for $t = 0$ and T.

We adopt (3.39) as the definition of weak solution in the sense of transposition of (3.38). More precisely, we say that (ψ, V) solve (3.38) if (3.39) holds for any $(\widetilde{\eta}^0, \widetilde{\eta}^1, \widetilde{u}^0, \widetilde{u}^1) \in \mathcal{Y}$ and $(\widetilde{f}, \widetilde{g}) \in L^1(0,T; L^2(0,1)\times\mathbb{R})$.

We observe that (3.39) can be rewritten in the following way:

$$\begin{aligned} \int_0^T \rho(t)u_{tt}(t)\widetilde{u}_{tt}dt - \int_0^T \int_0^1 \widetilde{f}\psi dydt + \int_0^T \widetilde{g}V dt = &- \langle \psi_t(0) + V(0)\delta_0, \widetilde{\eta}(0)\rangle \\ &+ \langle \psi(0), \widetilde{\eta}_t(0)\rangle + (V_t(0) + \psi(0,0))\,\widetilde{u}(0) - V(0)\widetilde{u}_t(0), \end{aligned} \tag{3.40}$$

where $\langle\cdot,\cdot\rangle$ denotes both the duality pairing between $(H^1(0,1))'$ and $H^1(0,1)$ and the scalar product in $L^2(0,1)$, and $\delta_0 \in (H^1(0,1))'$ denotes the Dirac delta at $y = 0$.

We have the following existence and uniqueness result of solutions in the sense of transposition.

PROPOSITION 3.7. *System* (3.38) *has a unique solution in the sense of transposition. More precisely, for any solution* (η, u) *of* (3.11) *with initial data in* $\mathcal{Y}$*, there*

exists a unique $(\psi, V) \in C([0,T]; L^2(0,1)) \times L^2(0,T)$, $\rho^0 \in L^2(0,1)$, $\rho^1 \in (H^1(0,1))'$, $\mu^0 \in \mathbb{R}, \mu^1 \in \mathbb{R}$ *satisfying*

$$\begin{aligned}\int_0^T \rho(t) u_{tt}(t) \widetilde{u}_{tt} dt &= \int_0^T \int_0^1 \widetilde{f} \psi dy dt - \int_0^T \widetilde{g} V dt \\ &+ \left\langle \rho^1, \widetilde{\eta}(0) \right\rangle + \left\langle \rho^0, \widetilde{\eta}_t(0) \right\rangle + \mu^1 \widetilde{u}(0) + \mu^0 \widetilde{u}_t(0)\end{aligned} \tag{3.41}$$

for any solution $(\widetilde{\eta}, \widetilde{u})$ *of* (3.1) *with*

$$(\widetilde{\eta}^0, \widetilde{\eta}^1, \widetilde{u}^0, \widetilde{u}^1) \in \mathcal{Y}, \quad \widetilde{f} \in L^1\left(0,T; L^2(0,1)\right), \quad \widetilde{g} \in L^2(0,1). \tag{3.42}$$

Remark 12. In the identity (3.41), ρ^0, ρ^1, μ^0, and μ^1 play, respectively, the roles of $\psi(0), -\psi_t(0) + V(0)\delta_0, -V(0)$, and $V_t(0) + \psi(0,0)$. It is easy to see that, in the frame of smooth functions, there is a one to one correspondence between $\left(\rho^0, \rho^1, \mu^0, \mu^1\right)$ and $(\psi(0), \psi_t(0), V(0), V_t(0))$.

Proof of Proposition 3.7. In view of Proposition 3.1 we know that the map

$$\left(\widetilde{\eta}^0, \widetilde{\eta}^1, \widetilde{u}^0, \widetilde{u}^1, \widetilde{f}, \widetilde{g}\right) \longrightarrow \int_0^T \rho(t) u_{tt}(t) \widetilde{u}_{tt}(t) dt$$

is linear and continuous from $\mathcal{Y} \times L^1\left(0,T; L^2(0,1)\right) \times L^2(0,T)$ into $\mathbb{R}$. This implies the existence and uniqueness of $\left(\rho^1, \rho^0, \mu^1, \mu^0\right) \times (\psi, V) \in \mathcal{Y}' \times L^\infty\left(0,T; L^2(0,1)\right) \times L^2(0,T)$ such that (3.41) holds. Moreover, there exists a constant $C > 0$ such that

$$\begin{aligned}\|(\psi, V)\|_{L^\infty(0,T;L^2(0,1)) \times L^2(0,T)} + \| \left(\rho^1, \rho^0, \mu^1, \mu^0\right) \|_{\mathcal{Y}'} &\leq C \|u_{tt}\|_{L^2(0,T)} \\ &\leq C \| \left(\eta^1, \eta^0, u^1, u^0\right) \|_{\mathcal{Y}'}.\end{aligned} \tag{3.43}$$

The fact that $\psi \in C\left([0,T]; L^2(0,1)\right)$ can be deduced from (3.43) by a classical density argument. □

Remark 13. When the data of (3.11) are smooth, the solution (η, u) is smooth too. It is easy to see that (3.38) has a finite energy solution. In this case one can check that the element $\left(\rho^1, \rho^0, \mu^1, \mu^0\right) \in \mathcal{Y}'$ obtained in Proposition 3.7 is such that

$$\rho^0 = \psi(0), \quad \rho^1 = -\psi_t(0) + V(0)\delta_0, \quad \mu^0 = -V(0), \quad \mu^1 = V_t(0) + \psi(0,0).$$

By a density argument, one can then deduce that the solution (ψ, V) obtained in Proposition 3.7 is such that the traces

$$\psi\,|_{t=0}, -\psi_t + V\delta_0|_{t=0}, \quad -V\,|_{t=0}, V_t + \psi(0,t)|_{t=0}$$

are well defined and coincide with $\left(\rho^0, \rho^1, \mu^0, \mu^1\right)$.

The same argument allows us to show that the traces are also well defined at $t = T$. This suffices to assert that the weak solution of (3.38) we have constructed by transposition is at rest at $t = T$.

We can now complete the proof of Theorem 2.1. □

End of the proof of Theorem 2.1. In view of Proposition 3.7 and Remark 13, we can define a linear and continuous map Λ from $\mathcal{Y}$ into $\mathcal{Y}'$ such that

$$\Lambda\left(\eta^0, \eta^1, u^0, u^1\right) = \left(-\psi_t + V\delta_0\,|_{t=0}, \psi(0), V_t + \psi(0,t)\,|_{t=0}, -V\,|_{t=0}\right).$$

Taking, in (3.41), $\widetilde{f} \equiv 0, \widetilde{g} \equiv 0$ and $(\widetilde{\eta}, \widetilde{u}) = (\eta, u)$, we deduce that

$$\left\langle \Lambda\left(\eta^0, \eta^1, u^0, u^1\right), \left(\eta^0, \eta^1, u^0, u^1\right) \right\rangle = \int_0^T \rho(t) \left|u_{tt}(t)\right|^2 dt,$$

and in view of Theorem 3.3 and Remark 9, we deduce that there exists $C > 0$ such that

$$\left\langle \Lambda\left(\eta^0, \eta^1, u^0, u^1\right), \left(\eta^0, \eta^1, u^0, u^1\right)\right\rangle \geq C \left\| \left(\eta^0, \eta^1, u^0 u^1\right)\right\|_{\mathcal{Y}}^2.$$

Actually, $C = [C(T,n)]^{-1}$, where $C(T,n)$ is as in (3.18).

This implies that Λ is an isomorphism.

This shows that given any $\left(\rho^1, \rho^0, \mu^1, \mu^0\right) \in \mathcal{Y}'$ there exists $\left(\eta^0, \eta^1, u^0, u^1\right) = \Lambda^{-1}\left(\rho^1, \rho^0, \mu^1, \mu^0\right)$ such that the corresponding solution of (3.38) in the sense of transposition satisfies

$$\psi(0) = \rho^0, \quad -\psi_t + V\delta_0 \left|_{t=0} = \rho^1, \quad -V \right|_{t=0} = \mu^0, \quad V_t + \psi(0,t) \left|_{t=0} = \mu^1. \tag{3.44}$$

If we want this to be equivalent to the initial data of (1.6), we have to take

$$\rho^0 = \psi^0, \quad \rho^1 = -\psi^1 + V^0\delta_0, \quad \mu^0 = -V^0, \quad \mu^1 = V^1 + \psi^0(0). \tag{3.45}$$

This makes sense when the data $\left(\psi^0, \psi^1, V^0, V^1\right)$ are in $\mathcal{Y}$.

The control we have obtained is of the form $\beta = -\frac{d^2}{dt^2}(\rho u_{tt})$, where u corresponds to the solution (η, u) of (3.11) with data $\left(\eta^0, \eta^1, u^0, u^1\right) = \Lambda^{-1}\left(\rho^1, \rho^0, \mu^1, \mu^0\right)$, where $\left(\rho^0, \rho^1, \mu^0, \mu^1\right)$ is given by (3.44). From the identities above, we see that

$$\begin{aligned} \|\beta\|^2_{H^{-2}(0,T)} \leq \|\rho u_{tt}\|^2_{L^2(0,T)} &\leq C\| \left(\rho^1, \rho^0, \mu^1, \mu^0\right) \|^2_{\mathcal{Y}'} \\ &\leq C\left\{\| \left(\psi^1, \psi^0, V^1, V^0\right) \|^2_{\mathcal{Y}'} + \mid \psi^0(0) \mid^2\right\}, \end{aligned}$$

where $C = C(T,n)$ is the constant obtained in (3.18). □

Remark 14. In fact, in some sense, we obtain a stronger result, since we prove that we can control the problem (3.41) for any initial data $\left(\rho^1, \rho^0, \mu^1, \mu^0\right) \in \mathcal{Y}'$. In order to give an interpretation of the control problem in terms of the initial data $\left(\psi^0, \psi^1, V^0, V^1\right)$, we have to ensure that $\psi^0(0)$ makes sense. For this reason we consider that $\left(\psi^0, \psi^1, V^0, V^1\right) \in \mathcal{Y}$.

3.5. Controllability in one space dimension for $n = 0$: Proof of Theorem 2.2. First, we observe that proving Theorem 2.2 is equivalent to showing that for any initial data as in the statement of Theorem 2.2 and satisfying the further assumptions

$$V^1 + \psi^0(0) = 0, \quad V^0 - \int_0^1 \psi^1(y)dy = 0, \tag{3.46}$$

there exists a control β such that

$$\psi(T) = \psi_t(T) \equiv 0 \text{ in } (0,1), \quad V(T) = V_t(T) = 0. \tag{3.47}$$

Indeed, this is an immediate consequence of the remark made in the introduction, which shows that when β is of zero average the following identities hold:

$$V_t(T) + \psi(0,T) = V^1 + \psi^0(0), \quad V(T) - \int_0^1 \psi_t(y,T)dy = V^0 - \int_0^1 \psi^1(y)dy. \tag{3.48}$$

Thus, below we focus on initial data $\left(\psi^0, \psi^1, V^0, V^1\right)$ satisfying (3.46). For the adjoint system

$$\left\{\begin{array}{lll} \eta_{tt} - \eta_{yy} = 0 & \text{in} & (0,1)\times(0,T), \\ \eta_y(1) = 0 & \text{for} & t \in (0,T), \\ \eta_y(0) = u_t & \text{for} & t \in (0,T), \\ u_{tt} - \eta_t(0) = 0 & \text{for} & t \in (0,T), \\ \eta(0) = \eta^0, \eta_t(0) = \eta^1 & \text{in} & (0,1), \\ u(0) = u^0, u_t(0) = u^1 & & \end{array}\right. \tag{3.49}$$

we consider initial data in the following subspace $\mathcal{Y}_0$ of $\mathcal{Y}$:

$$\mathcal{Y}_0 = \left\{ \left(\eta^0, \eta^1, u^0, u^1\right) \in \mathcal{Y} : u^1 - \eta^0(0) = 0, \int_0^1 \eta^1 dy + u^0 = 0 \right\}. \tag{3.50}$$

It is easy to see that the subspace $\mathcal{Y}_0$ is invariant under the flow generated by (3.49).

Given $\left(\eta^0, \eta^1, u^0, u^1\right) \in \mathcal{Y}_0$, we solve first (3.49) and then the backward system

$$\left\{\begin{array}{lll} \psi_{tt} - \psi_{yy} = 0 & \text{in} & (0,1)\times(0,T), \\ \psi_y(1,t) = 0 & \text{for} & t \in (0,T), \\ \psi_y(0,t) = -V_t(t) & \text{for} & t \in (0,T), \\ V_{tt}(t) + \psi_t(0,t) = -\frac{d^2}{dt^2}\left(\rho(t) u_{tt}(t)\right) & \text{for} & t \in (0,T), \\ \psi(T) = \psi_t(T) = 0 & \text{in} & (0,1), \\ V(T) = V_t(T) = 0, & & \end{array}\right. \tag{3.51}$$

where ρ is as in the proof of Theorem 2.1.

Proceeding as in the proof of Proposition 3.7 one can show that (3.51) has a unique solution defined by transposition such that the traces (3.47) are well defined. On the other hand, integrating the equations in (3.51), we deduce that

$$\int_0^1 \rho^1(y) dy = 0, \quad \mu^1 = 0. \tag{3.52}$$

Let us denote by Z the subspace of $\mathcal{Y}'$ satisfying (3.52). More precisely,

$$Z = \left\{ \left(\rho^1, \rho^0, \mu^1, \mu^0\right) \in \mathcal{Y}' : \text{ (3.52) holds } \right\}. \tag{3.53}$$

It is easy to check that Z is actually the dual of $\mathcal{Y}_0$. Indeed, the dual of $\mathcal{Y}_0$ is a quotient space of $\mathcal{Y}'$, and there is a one-to-one correspondence between Z and this quotient space in the sense that, in Z, we have chosen the unique element of each class of the quotient space satisfying (3.52).

As in the proof of Theorem 2.1, we can define a linear and continuous operator $\Lambda : \mathcal{Y}_0 \longrightarrow Z$ that associates the trace $\left(\rho^1, \rho^0, \mu^1, \mu^0\right) \in Z$ in (3.41) with each $\left(\eta^0, \eta^1, u^0, u^1\right) \in \mathcal{Y}_0$.

We also have

$$\left\langle \Lambda\left(\eta^0, \eta^1, u^0, u^1\right), \left(\eta^0, \eta^1, u^0, u^1\right) \right\rangle = \int_0^T \rho(t) \left|u_{tt}(t)\right|^2 dt.$$

In view of Theorem 3.3 and Remark 9 we deduce the existence of a constant $C > 0$ such that

$$\left\langle \Lambda\left(\eta^0, \eta^1, u^0, u^1\right), \left(\eta^0, \eta^1, u^0, u^1\right) \right\rangle \geq C \left\| \left(\eta^0, \eta^1, u^0, u^1\right) \right\|_{\mathcal{Y}'}^2, \ \forall \left(\eta^0, \eta^1, u^0, u^1\right) \in \mathcal{Y}_0,$$

since the quantity $\left[\|\eta_y^0\|^2_{L^2(0,1)} + \|\eta^1\|^2_{L^2(0,1)} + |\ u^1\ |^2\right]^{1/2}$ defines a norm in $\mathcal{Y}_0$ which is equivalent to the norm induced by $\mathcal{Y}$.

We deduce that $\Lambda : \mathcal{Y}_0 \longrightarrow Z$ is an isomorphism.

Then, given initial data as in the statement of Theorem 2.2 and such that (3.46) holds, we define $\left(\rho^1, \rho^0, \mu^1, \mu^0\right) \in Z$ by (3.45). The control we are looking for is $\beta = -\frac{d^2}{dt^2}\left(\rho(t) u_{tt}(t)\right)$, where u is the second component of the solution (η, u) of (3.49) with initial data $\left(\eta^0, \eta^1, u^0, u^1\right) = \Lambda^{-1}\left(\rho^1, \rho^0, \mu^1, \mu^0\right)$.

This concludes the proof of Theorem 3.5. □

4. Controllability of the two-dimensional system: Proof of Theorem 2.3. In view of Theorems 2.1 and 2.2 for any $n = 0, 1, \ldots$, there exists a control $\beta_n \in H^{-2}(0,T)$ such that the solution (ψ_n, V_n) of (1.6) satisfies

$$\psi_n(T) \equiv \psi_{n,t}(T) = 0 \text{ in } (0,1), \quad V_n(T) = V_{n,t}(T) = 0 \tag{4.1}$$

for $n \geq 1$ and

$$\psi_0(T) = \mu^1, \quad \psi_{0,t}(T) = 0 \text{ in } (0,1), \quad V_0(T) = \left\langle \rho^1, 1 \right\rangle, \quad V_{0,t}(T) = 0 \tag{4.2}$$

when $n = 0$.

On the other hand,

$$\|\beta_n\|^2_{H^{-2}(0,T)} \leq C_n \left\| \left(\rho_n^1, \rho_n^0, \mu_n^1, \mu_n^0\right) \right\|^2_{\mathcal{Y}'}. \tag{4.3}$$

We construct the following control for the two-dimensional system:

$$\beta(x,t) = \sum_{n=0}^{\infty} \beta_n \cos(n\pi x). \tag{4.4}$$

We have, in view of (4.3),

$$\|\beta\|^2_{H^{-2}(0,T;L^2(0,1))} = \sum_{n=0}^{\infty} \|\beta_n(t)\|^2_{H^{-2}(0,T)}$$

$$\leq \sum_{n=0}^{\infty} C_n \left\| \left(\rho_n^1, \rho_n^0, \mu_n^1, \mu_n^0\right) \right\|^2_{\mathcal{Y}'} = \left\| \left(\psi^0, \psi^1, W^0, W^1\right) \right\|^2_H < \infty.$$

Therefore, $\beta \in H^{-2}\left(0,T; L^2(0,1)\right)$. On the other hand,

$$\psi(x,y,t) = \sum_{n=0}^{\infty} \psi_n(y,t) \cos(n\pi x), \quad W(x,t) = \sum_{n=0}^{\infty} V_n(t) \cos(n\pi x)$$

solves (1.3) with the control β given in (4.4), and satisfies (2.7) at time $t = T$.

This concludes the proof of this theorem. □

5. Appendix: Proof of Theorem 3.4. First, we recall a classical result due to Ingham.

THEOREM A (see Ingham [9, Thms. 1 and 2]). *Let $f = f(t)$ be of the form $f(t) = \sum_{n\in\mathbb{Z}} a_n e^{i\lambda_n t}$, where λ_n is a sequence of real numbers.*

We assume that there exists $\gamma > 0$ *such that*

$$\lambda_{n+1} - \lambda_n \geq \gamma, \ \forall n \in \mathbb{Z}. \tag{5.1}$$

Let $J = [0, T]$ *with* $T > \frac{2\pi}{\gamma}$. *Then there exist two positive constants* $C_1^0, C_2^0 > 0$ *such that*

$$C_1^0 \sum_{n\in\mathbb{Z}} |\, a_n \,|^2 \leq \int_J |\, f(t) \,|^2 \, dt \leq C_2^0 \sum_{n\in\mathbb{Z}} |\, a_n \,|^2 \tag{5.2}$$

for all $a_n \in \ell^2$

Remark 15. The constants C_1^0 and C_2^0 depend only on $T - \frac{2\pi}{\gamma}$.

To prove (3.22) we follow the ideas of Haraux [8], paying special attention to the evaluation of the constants appearing there.

The second inequality of (3.22) results, with $C_2 = 2C_2^0 + 2|J|(2N+1)$, immediately using Theorem A. Indeed we have

$$\int_J |f(t)|^2 \, dt = \int_J \left| \sum_{|n|>N} a_n e^{i\lambda_n t} + \sum_{|n|\leq N} a_n e^{i\lambda_n t} \right|^2$$

$$\leq 2 \int_J \left(\left| \sum_{|n|>N} a_n e^{i\lambda_n t} \right|^2 + \left| \sum_{|n|\leq N} a_n e^{i\lambda_n t} \right|^2 \right).$$

Now, applying Theorem A to the function $g(t) = \sum_{|n|\geq N} a_n e^{i\lambda_n t}$, we obtain

$$\int_J |f(t)|^2 \, dt \leq 2C_2^0 \sum_{|n|>N} |a_n|^2 + 2|J| \left(\sum_{|n|\leq N} |a_n| \right)^2$$

$$\leq 2C_2^0 \sum_{|n|>N} |a_n|^2 + 2|J|(2N+1) \sum_{|n|\leq N} |a_n|^2 \leq \left(2C_2^0 + 2|J|(2N+1)\right) \sum_{n\in\mathbb{Z}} |a_n|^2.$$

We now proceed to prove the first inequality of (3.22). We do this by induction in p, the number of indexes $n \in \mathbb{Z}$ for which $\lambda_{n+1} - \lambda_n < \gamma_\infty$.

If $p = 0$, the result follows from Theorem A with $C_1 = C_1(0) = C_1^0$. Suppose now that $p > 0$.

We write the function f in the form $f(t) = \sum_{n\neq 0} a_n e^{i\lambda_n t} + a_0 e^{i\lambda_0 t}$, where λ_0 is one of those values for which $\lambda_{n+1} - \lambda_n < \gamma_\infty$. Moreover, without loss of generality, we may suppose that $\lambda_0 = 0$ (since we can consider the function $f(t)e^{-i\lambda_0 t}$ instead of $f(t)$). We now apply the induction hypothesis for the function $g(t) = \sum_{n\neq 0} a_n e^{i\lambda_n t}$, and we obtain that

$$C_1(p-1) \sum_{n\neq 0} |a_n|^2 \leq \int_J |g(t)|^2 \leq C_2(p-1) \sum_{n\neq 0} |a_n|^2. \tag{5.3}$$

We know that $C_2(p-1) = 2C_2^0 + 2|J|(p-1)$. Let $\varepsilon > 0$ be such that $T' = T - \varepsilon > \frac{2\pi}{\gamma_\infty}$.

We have

$$\int_0^\varepsilon (f(t+\eta) - f(t)) \, d\eta = \sum_{n\neq 0} a_n \left(\frac{e^{i\lambda_n \varepsilon} - 1}{i\lambda_n} - \varepsilon \right) e^{i\lambda_n t} \quad \forall t \in [0, T'].$$

Applying the induction hypothesis to the function $h(t)=\int_0^\varepsilon (f(t+\eta)-f(t))\,d\eta$, we obtain that

$$C_1(p-1)\sum_{n\neq 0}\left|\frac{e^{i\lambda_n\varepsilon}-1}{i\lambda_n}-\varepsilon\right||a_n|^2\leq\int_0^{T'}\left|\int_0^\varepsilon (f(t+\eta)-f(t))\,d\eta\right|^2. \tag{5.4}$$

We now evaluate the coefficients $\dfrac{e^{i\lambda_n\varepsilon}-1}{i\lambda_n}-\varepsilon$. We have

$$\left|e^{i\lambda_n\varepsilon}-1-\varepsilon i\lambda_n\right|^2=|\cos(\lambda_n\varepsilon)-1|^2+|\sin(\lambda_n\varepsilon)-\lambda_n\varepsilon|^2$$

$$=4\sin^4\left(\frac{\lambda_n\varepsilon}{2}\right)+(\sin(\lambda_n\varepsilon)-\lambda_n\varepsilon)^2\geq\begin{cases}4\left(\frac{\lambda_n\varepsilon}{2}\operatorname{arctg}\left(\frac{\pi}{2}\right)\right)^4 & \text{if } |\lambda_n|\varepsilon\leq\frac{\pi}{2},\\ \left(|\lambda_n|\varepsilon-1\right)^2 & \text{if } |\lambda_n|\varepsilon>\frac{\pi}{2}.\end{cases}$$

Finally, taking into account that $|\lambda_n|\geq\gamma$, we obtain that, for ε small enough,

$$\left|\frac{e^{i\lambda_n\varepsilon}-1}{i\lambda_n}-\varepsilon\right|^2\geq\gamma^2\varepsilon^4.$$

We now return to (5.4), and we get that

$$\gamma^2\varepsilon^4C_1(p-1)\sum_{n\neq 0}|a_n|^2\leq\int_0^{T'}\left|\int_0^\varepsilon (f(t+\eta)-f(t))\,d\eta\right|^2. \tag{5.5}$$

On the other hand,

$$\int_0^{T'}\left|\int_0^\varepsilon (f(t+\eta)-f(t))\,d\eta\right|^2\leq\int_0^{T'}\varepsilon\int_0^\varepsilon|f(t+\eta)-f(t)|^2\,d\eta$$

$$\leq 2\varepsilon\int_0^{T'}\int_0^\varepsilon\left(|f(t+\eta)|^2+|f(t)|^2\right)d\eta\leq 2\varepsilon^2\int_0^T|f(t)|^2$$

$$+2\varepsilon\int_0^\varepsilon\int_0^{T'}|f(t+\eta)|^2\,dt\,d\eta\leq 4\varepsilon^2\int_0^T|f(t)|^2.$$

From (5.5) it follows that

$$\sum_{n\neq 0}|a_n|^2\leq\frac{4}{\varepsilon^2\gamma^2C_1(p-1)}\int_0^T|f(t)|^2. \tag{5.6}$$

Observe that

$$|a_0|^2=\left|f(t)-\sum_{n\neq 0}a_ne^{i\lambda_nt}\right|^2=\frac{1}{T}\int_0^T\left|f(t)-\sum_{n\neq 0}a_ne^{i\lambda_nt}\right|^2dt$$

$$\leq\frac{2}{T}\left(\int_0^T|f(t)|^2+\int_0^T\left|\sum_{n\neq 0}a_ne^{i\lambda_nt}\right|^2\right)\leq\frac{2}{T}\left(\int_0^T|f(t)|^2+C_2(p-1)\sum_{n\neq 0}|a_n|^2\right)$$

$$\leq\left(\frac{2}{T}+\frac{8C_2(p-1)}{T\varepsilon^2\gamma^2C_1(p-1)}\right)\int_0^T|f(t)|^2.$$

From (5.6) we get that

$$\sum_{n\in\mathbb{Z}} |a_n|^2 \le \left[\frac{4}{\varepsilon^2\gamma^2 C_1(p-1)}\left(\frac{2C_2(p-1)}{T}+1\right)+\frac{2}{T}\right]\int_0^T |f(t)|^2 .$$

We obtain the desired result and a recurrent formula to compute the constant $C_1(p)$:

$$C_1(p) = \left[\frac{4}{\varepsilon^2\gamma^2 C_1(p-1)}\left(\frac{2C_2(p-1)}{T}+1\right)+\frac{2}{T}\right]^{-1}. \qquad \square$$

Acknowledgments. This paper was part of the author's doctoral dissertation at Universidad Complutense de Madrid in February 1996. Most of this research was done while this author was visiting Universidad Complutense with the financial support of the European Tempus Program "Matarou." The authors acknowledge the coordinators of this program and, in particular, Doina Cioranescu for their continuous support.

REFERENCES

[1] B. Allibert, *Contrôle analytique d'un modèle hybride fluide-structure*, preprint, 1996.
[2] G. Avalos, *The exponential stability of a coupled hyperbolic/parabolic system arising in structural acoustics*, Abstract Appl. Anal., to appear.
[3] G. Avalos and I. Lasiecka, *A differential Riccati equation for the active control of a problem in structural acoustics*, J. Optim. Theory Appl., to appear.
[4] J. Ball and M. Slemrod, *Nonharmonic Fourier series and the stabilization of distributed semi-linear control systems*, Comm. Pure Appl. Math., XXXII (1979), pp. 555–587.
[5] H. T. Banks, W. Fang, R. J. Silcox, and R. C. Smith, *Approximation methods for control of acoustic/structure models with piezoceramic actuators*, J. Intelligent Material Systems Structures, 4 (1993), pp. 98–116.
[6] C. Bardos, G. Lebeau, and J. Rauch, *Sharp sufficient conditions for the observation, control and stabilization of waves from the boundary*, SIAM J. Control Optim., 30 (1992), pp. 1024–1065.
[7] S. Hansen and E. Zuazua, *Exact controllability and stabilization of a vibrating string with an interior point mass*, SIAM J. Control Optim., 33 (1995), pp. 1357–1391.
[8] A. Haraux, *Séries lacunaires et contrôle semi-interne des vibrations d'une plaque rectangulaire*, J. Math. Pures Appl., 68 (1989), pp. 457–465.
[9] A. E. Ingham, *Some trigonometrical inequalities with applications to the theory of series*, Math. Z., 41 (1936), pp. 367–369.
[10] J. L. Lions, *Contrôlabilité exacte, perturbations et stabilization de systèmes distribués. Tome 1. Contrôlabilité exacte*, Rech. Math. Appl. 8, Masson, Paris, 1988.
[11] W. Littman and L. Marcus, *Exact boundary controllability of a hybrid system of elasticity*, Arch. Rational Mech. Anal., 103 (1988), pp. 193–236.
[12] S. Micu, *Análisis de un modelo híbrido bidimensional fluido-estructura*, Ph.D. thesis, Universidad Complutense de Madrid, Spain, 1996.
[13] S. Micu and E. Zuazua, *Propriétés qualitatives d'un modèle hybride bi-dimensionnel intervenant dans le contrôle du bruit*, C. R. Acad. Sci. Paris Sér. I Math., 319 (1994), pp. 1263–1268.
[14] S. Micu and E. Zuazua, *Asymptotics for the Spectrum of a Fluid/Structure Hybrid System Arising in the Control of Noise*, preprint.
[15] S. Micu and E. Zuazua, *Stabilization and periodic solutions of a hybrid system arising in the control of noise*, in Proceedings of the IFIP TC7/WG-7.2 International Conference, Laredo, España, Lecture Notes in Pure and Applied Mathematics 174, Marcel Dekker, New York, 1996, pp. 219–230.
[16] J. L. Puel and E. Zuazua, *Exact controllability for a model of a multidimensional flexible structure*, Proc. Roy. Soc. Edinburgh Sect. A, 123 (1993), pp. 323–344.
[17] M. Tucsnak, *Regularity and exact controllability for a beam with piezoelectric actuators*, SIAM J. Control Optim., 34 (1996), pp. 922–930.
[18] E. Zuazua, *Exact controllability for the semilinear wave equation in one space dimension*, Ann. Inst. H. Poincaré Analyse Non Linéaire, 10 (1993), pp. 109–129.

SIAM J. CONTROL OPTIM.
Vol. 35, No. 5, pp. 1638–1652, September 1997

011

ON THE REGULARITY OF SEMIPERMEABLE SURFACES IN CONTROL THEORY WITH APPLICATION TO THE OPTIMAL EXIT-TIME PROBLEM (PART I)*

PIERRE CARDALIAGUET†

Abstract. In control theory, a semipermeable surface is an (in general nonsmooth) oriented surface that, on one hand, contains solutions (the so-called barrier solutions) of the controlled system and, on the other hand, may be crossed by the solutions of this system in only one direction. Without making any assumption on the regularity of the boundary of the semipermeable surface, we show that the barrier solutions contained in this semipermeable surface satisfy the Pontryagin principle, that this surface is a Lipschitz manifold, and that it is, locally, the graph of a semiconcave function. Applying these results to the optimal exit-time function from a given open set yields, without any controllability assumption at the boundary of the open set, that this function is semiconcave on an open dense subset of its domain.

Key words. semipermeable surfaces, differential inclusion, viability theory, minimal time function

AMS subject classifications. 49J24, 49J52, 49N60

PII. S0363012995287295

Introduction. Let

$$\begin{cases} x'(t) = f(x(t), u(t)),\ u(t) \in U, \\ x(0) = x_0 \end{cases} \tag{1}$$

be a controlled system with a hamiltonian defined by

$$H(x,p) := \inf_{u \in U(x)} \langle f(x,u), p \rangle.$$

A *smooth*, *semipermeable surface* is an oriented hypersurface S such that the outward normal p at each point $x \in S$ satisfies $H(x,p) = 0$. Such a surface S is called semipermeable because

- (α) S can be crossed in only one direction by the trajectories of the controlled system.
- (β) From any initial position $x_0 \in S$ at least one solution $x(\cdot)$ of (1) starts, and remains locally on S (namely, $\exists \tau > 0$ such that, $\forall t \in [0,\tau]$, $x(t) \in S$).

A solution satisfying condition (β) is called *a barrier solution*.

In many problems, one encounters closed sets (which are not necessarily smooth) with a boundary enjoying properties (α) and (β). We still say that their boundary is "a semipermeable surface." The aim of this work is to show that a closed set with semipermeable boundary enjoys some regularity properties. Namely, under suitable assumptions on f, the boundary of such a set is a Lipschitz (and even semiconcave) manifold, and the barrier solutions satisfy the Pontryagin principle.

To our knowledge, this problem has never been treated, although it is of great interest for qualitative and quantitative control problems (see the examples below). However, our work is related to several studies on the regularity of the value function

*Received by the editors June 5, 1995; accepted for publication (in revised form) July 1, 1996.
http://www.siam.org/journals/sicon/35-5/28729.html
†CEREMADE, URA CNRS 749, Université Paris-Dauphine, Place du Maréchal de Lattre de Tassigny, 75775 Paris cedex 16, France (cardaliaguet@ceremade.dauphine.fr).

of optimal control problems (see, for instance, [1], [8], [9], [11], [14], [19], [20] [21]). It is not easy to compare our results (which are of geometric nature) with those given in the previous references (which are concerned with the regularity of functions). For this reason, we illustrate our results through the study of the regularity of the optimal exit-time function.

(1) *The optimal exit-time function* θ_Ω from the open subset $\Omega \subset \mathbb{R}^N$ is defined by, $\forall x_0 \in \Omega$,

$$\theta_\Omega(x_0) := \inf \{t \geq 0 \mid \exists x(\cdot) \text{ solution to (1) such that } x(t) \notin \Omega\}.$$

Roughly speaking, $\theta_\Omega(x_0)$ is the minimal time any solution of the controlled system (1) starting from x_0 needs to leave Ω. The regularity of the optimal exit-time function is the aim of several works [23], [24], [6], [7], [9], [10]. In [9] Cannarsa and Sinestrari prove that θ_Ω is semiconcave on its (open) domain under the following assumptions: (a) f is smooth; (b) $\partial\Omega$ enjoys some regularity (roughly speaking, its curvature is bounded); (c) a "controllability condition" on the boundary of Ω is required, which ensures that θ_Ω is Lipschitz continuous. Thanks to conditions (b) and (c), θ_Ω is "smooth" in a neighborhood of $\partial\Omega$. Then condition (a) ensures, by using the Pontryagin principle, that this "smoothness" propagates along the (smooth) optimal trajectories.[1] So, in this method, the crucial points are, on one hand, the smoothness of θ_Ω at the boundary of Ω and, on the other hand, the propagation of this regularity.

Our method is, on the contrary, based on the local study of the epigraph of θ_Ω. Combining the results of [10] and of [25] yields that this epigraph has a semipermeable boundary for some dynamics Φ_f constructed from f. Thanks to the regularity results of semipermeable surfaces given in this paper, we prove, without conditions (b) and (c), that θ_Ω is (locally) Lipschitz and semiconcave on an open dense subset of its domain.

(2) *Boundary of the viability kernel:* The first definition of (nonsmooth) semipermeability is due to Quincampoix and appeared in the framework of viability theory (Aubin [3]). If $K \subset \mathbb{R}^N$ is a closed set, the *viability kernel* $\mathrm{Viab}_f(K)$ of K for f is

$$\mathrm{Viab}_f(K) := \left\{ x_0 \in K \mid \begin{array}{l} \exists x(\cdot) \text{ solution to (1)} \\ \text{such that } x(t) \in K \ \forall t \geq 0 \end{array} \right\}.$$

Under suitable assumptions, the viability kernel of K for f is a closed subset of K (see also [5], [15]). In [25], Quincampoix proves that the boundary of $\mathrm{Viab}_f(K)$ enjoys the semipermeability property in the interior of K.

(3) *Boundary of the reachable set:* The reachable set for f starting from a point $x_0 \in \mathbb{R}^N$ is the set of points y for which there exists a solution $x(\cdot)$ of (1) and a time $t \geq 0$ such that $x(t) = y$.

If, for instance, 0 belongs to the interior of $\bigcup_u f(x_0, u)$ and f is Lipschitz continuous, then the reachable set is open. Moreover, its boundary is semipermeable for $-f$ (see Quincampoix [26]).

This research is presented as follows. In the present paper, Part I, we give two equivalent definitions of the semipermeability, and we also prove that semipermeable boundaries are Lipschitz manifolds. Then we show that semipermeable surfaces are "smooth" along barrier solutions. We also explain how to recover the Pontryagin principle.

[1] A similar method based on the propagation of the regularity of the final data along optimal trajectories is also applied to Mayer's problem in [14] and to the Bolza problem in [11].

Part II, also in this issue, is devoted to the regularity results for the closed sets with semipermeable boundaries. We first show that the contingent cone to such closed sets is a union of half-spaces and enjoys some upper semicontinuity property. We then show, under a stronger assumption on the dynamics, that the boundaries of such closed sets are locally graphs of semiconcave functions. We complete this paper by applying these results to the case of the optimal exit-time functions.

1. Semipermeable boundaries.

1.1. Definition of semipermeability. Let us from now on replace the controlled system (1) by the differential inclusion

$$\begin{cases} x'(t) \in F(x(t)), \\ x(0) = x_0, \end{cases} \tag{2}$$

where $F(x) := \bigcup_{u\in U} f(x,u)$. The advantage of using differential inclusions instead of controlled systems lies in the fact that the regularity properties explained below depend on the geometrical properties of the sets $F(x)$ and not on its representation as a controlled system. Moreover, the formulation as differential inclusions simplifies the statements and the proofs of the results.

It is well known that, under conditions that we impose here, controlled system (1) has the same solutions as differential inclusion (2). We denote by $\mathcal{S}_F(x_0)$ the set of (Carathéodory) solutions of differential inclusion (2). With a set-valued map $F : \mathbb{R}^N \rightsquigarrow \mathbb{R}^N$, we associate the hamiltonian H_F defined by

$$\forall (x,p) \in \mathbb{R}^N \times \mathbb{R}^N, \qquad H_F(x,p) := \inf_{v\in F(x)} \langle v, p\rangle. \tag{3}$$

Note that the hamiltonian H_F is concave with respect to p.

Let us now recall two basic definitions of nonsmooth analysis.

If K is a closed subset of $\mathbb{R}^N$ and x belongs to K, the *contingent cone to K at x* is the set of vectors $v \in \mathbb{R}^N$ such that

$$\liminf_{h\to 0^+} d_K(x+hv)/h = 0$$

($d_K(y)$ denotes the distance from the point y to the set K). The contingent cone is a closed cone. It is denoted by $T_K(x)$.

We also denote by $T_K(x)^-$ the polar cone of $T_K(x)$, i.e.,

$$T_K(x)^- := \{p \in \mathbb{R}^N \mid \forall v \in T_K(x),\ \langle p, v\rangle \le 0\}.$$

The polar cone is a closed convex cone. The contingent cone plays the role of tangent half-space, while the polar cone plays the role of exterior normal for nonsmooth sets.

If K is a subset of $\mathbb{R}^N$ and x belongs to K, the Dubovitsky–Miljutin cone to K at x is the set of vectors $v \in \mathbb{R}^N$ for which there exists some $\alpha > 0$ such that

$$x +]0,\alpha](v + \alpha B) \subset K.$$

The Dubovitsky–Miljutin cone to K at x is denoted by $D_K(x)$. It is an open cone. If K is a closed subset of $\mathbb{R}^N$, then $D_{\mathbb{R}^N\setminus K}(x) = \mathbb{R}^N \setminus T_K(x)$ for $x \in K$ [25]. Moreover, $D_K(x) \subset \mathrm{Int}\,(T_K(x))$, but there is no equality in general even if the set K is a Lipschitz manifold.

Notation. Below, B_N denotes the closed unit ball of $\mathbb{R}^N$ (endowed with the euclidean topology). If there is no ambiguity, we write only B. In the same way, $\overset{o}{B}_N$ denotes the open unit ball of $\mathbb{R}^N$.

DEFINITION 1.1 (semipermeability). *A closed set M has a semipermeable boundary for the set-valued map $F : \mathbb{R}^N \rightsquigarrow \mathbb{R}^N$ (or enjoys the semipermeability property) in a neighborhood of $x_0 \in \partial M$ if there is some positive radius r such that*

$$\forall x \in M \cap (x_0 + rB),\ \forall p \in T_M(x)^-,\ H_F(x,p) = 0.$$

An equivalent definition of semipermeability in terms of trajectories is the following.

PROPOSITION 1.1 (semipermeability). *Assume that the set-valued map F satisfies the following conditions:*

$$\text{(4)} \qquad \begin{cases} \text{(a)} & F : \mathbb{R}^N \rightsquigarrow \mathbb{R}^N \text{ has convex compact values;} \\ \text{(b)} & F \text{ is } \ell\text{-Lipschitz, i.e.,} \\ & \quad \forall x, y \in \mathbb{R}^N \times \mathbb{R}^N,\ F(y) \subset F(x) + \ell B. \end{cases}$$

Then a closed set $M \subset \mathbb{R}^N$ has a semipermeable boundary in a neighborhood of $x_0 \in \partial M$ if and only if there are open subsets O and O' of $\mathbb{R}^N$ with $x_0 \in O \subset O'$ and a time $T > 0$ such that

(i) *$\forall x \in M \cap O$, there is at least one solution $x(\cdot) \in \mathcal{S}_F(x)$ which remains in $M \cap O'$ on $[0,T]$;*

(ii) *$\forall x \in M \cap O'$, any solution of the differential inclusion for $-F$ remains in M on $[0,T]$;*

(iii) *$\forall x \in \partial M \cap O'$, any solution of the differential inclusion for F remains in $\widehat{M} := \overline{\mathbb{R}^N \backslash M}$ on $[0,T]$.*

The notations of this definition are kept throughout this paper.

Before proving that result, let us point out an important consequence.

COROLLARY 1.1 (barrier solutions). *Assume that F and M are as in Proposition* 1.1.

If M enjoys the semipermeability property, then any solution $x(\cdot)$ of the differential inclusion for F starting from $\partial M \cap O$ which remains in M on $[0,T]$ remains in ∂M on $[0,T]$.

Such a solution is "a barrier solution."

Remarks.

(1) Thanks to Proposition 1.1(i), at least one barrier solution starts from any initial position of $\partial M \cap O$.

(2) Combining Proposition 1.1 and Corollary 1.1, we recover the definition of semipermeability given at the beginning of this paper. Property (α) holds true thanks to (iii), and Corollary 1.1 is exactly the same as property (β).

Proof. The solution $x(\cdot)$ remains in $\widehat{M}$ from Proposition 1.1(iii), and in M from assumption. Since $M \cap \widehat{M} = \partial M$, the corollary holds true. □

Proof of Proposition 1.1. Assume that the set M satisfies the described property in a neighborhood of a point x_0. Then M is (locally) viable[2] for F in O, so that the

[2]If $K \subset \mathbb{R}^N$ is locally compact, the *viability theorem* [3], [4], [17] gives the equivalence among the following statements.
(i) K is a viability domain for F; i.e.,$\forall x \in K,\ F(x) \cap T_K(x) \neq \emptyset$.
(ii) K is viable for F; i.e., $\forall x \in K,\ \exists x(\cdot) \in \mathcal{S}_F(x)$ and $t > 0$ such that $x(s) \in K\ \forall s \in [0,t]$.
(iii) K satisfies the following: $\forall x \in K,\ \forall p \in T_K(x)^-,\ H_F(x,p) \leq 0$.

viability theorem, applied to the locally compact set $M \cap O$, states that $H_F(x,p) \leq 0$ for $x \in O \cap M$ and $p \in T_M(x)^-$. Moreover, M is locally invariant[3] for $-F$, so that the invariance theorem, applied to the locally compact set $M \cap O$ again, states that $H_{-F}(x,-p) \geq 0$ for $x \in O \cap M$ and $p \in T_M(x)^-$. Since

$$H_{-F}(x,-p) = \inf_{v \in F(x)} \langle -v, -p \rangle = H_F(x,p),$$

we have finally proved that $H_F(x,p) = 0$ for any $x \in O \cap M$ and $p \in T_M(x)^-$.

Conversely, if the boundary of M is semipermeable, there is some radius $r > 0$ such that

$$\forall x \in (x_0 + rB) \cap M,\ \forall p \in T_M(x)^-,\ H_F(x,p) = 0.$$

Set $\rho := \max_{x \in (x_0+rB)} \max_{v \in F(x)} \|v\|$ and $T := \frac{r}{4\rho}$. Define, for any $i = 1, \ldots, 4$,

$$O_i := x_0 + \frac{ir}{4} \overset{o}{B}.$$

Note that any solution of the differential inclusion for F (or for $-F$) starting from O_i $(i = 1, \ldots, 3)$ remains in O_{i+1} on $[0,T]$.

Since the tangential condition

$$\forall x \in (x_0 + rB) \cap M,\ \forall p \in T_M(x)^-,\ H_F(x,p) \leq 0$$

is satisfied, the viability theorem states that for any initial position $x \in O_1 \cap M$, there is an $x(\cdot) \in S_F(x)$ such that $x(t) \in M$ as long as $x(t) \in x_0 + rB$, i.e., at least on $[0,T]$. In particular, such a solution remains in $M \cap O_2$ on $[0,T]$. Thus (i) holds true with $O := O_1$ and $O' := O_2$.

Since $H_F(x,p) = 0$ implies that $H_{-F}(x,-p) \geq 0$, the tangential condition

$$\forall x \in (x_0 + rB) \cap M,\ \forall p \in T_M(x)^-,\ H_{-F}(x,-p) \geq 0$$

is fulfilled. Thus M is locally invariant for $-F$, and any solution of the differential inclusion for $-F$ starting from $M \cap O_2$ (and also from $M \cap O_3$) remains in M as long as it remains in $x_0 + rB$, and in particular, on $[0,T]$. Thus (ii) holds true.

Assume, contrary to our claim, that (iii) is false. There is a solution $x(\cdot)$ of the differential inclusion for F starting from $\partial M \cap O_2$ which does not remain in $\widehat{M}$ on $[0,T]$. We already know that $x(\cdot)$ remains in O_3 on $[0,T]$. There is some time $t \in]0,T]$ such that $x(t)$ belongs to the interior of M and to O_3. From Filippov's theorem [12], the set-valued map $x \rightsquigarrow \mathcal{S}_F(x)$, endowed with the uniform topology, is continuous. Thus there is a neighborhood W of $x(0)$ such that, from any initial position $y \in W$, at least one solution $y(\cdot) \in \mathcal{S}_F(y)$ sufficiently close to $x(\cdot)$ on $[0,T]$ starts so that $y(t)$ belongs to the interior of M and to O_3. Since x belongs to ∂M, there is some $\bar{y} \in W$ which does not belong to M. Let us denote by $\bar{y}(\cdot)$ the associated solution. We now consider the function $z(\cdot)$ defined by $z(s) := \bar{y}(t-s)$ for $s \in [0,T]$. Then $z(\cdot)$ is a solution of the differential inclusion for $-F$ starting from $M \cap O_3$, which leaves M before T. This is in contradiction to the proof of (ii). So (iii) holds true. □

[3]The *invariance theorem* [3] states that, for F satisfying (4) and for $K \subset \mathbb{R}^N$ locally compact, there is an equivalence among the following statements.
(i) K is an invariance domain for F, i.e.,$\forall x \in K,\ F(x) \subset T_K(x)$.
(ii) K is invariant for F, i.e., $\forall x \in K,\ \exists t > 0$ such that $\forall x(\cdot) \in \mathcal{S}_F(x),\ \forall s \in [0,t],\ x(s) \in K$.
(iii) K satisfies the following: $\forall x \in K,\ \forall p \in T_K(x)^-,\ H_F(x,-p) \geq 0$.

1.2. Semipermeable boundaries are Lipschitz manifolds.

PROPOSITION 1.2. *Assume that the boundary of M is semipermeable in a neighborhood of x_0 for a set-valued map $F : \mathbb{R}^N \rightsquigarrow \mathbb{R}^N$ satisfying* (4). *If the values of F have a nonempty interior, then ∂M is a Lipschitz manifold in a neighborhood of x_0.*

To prove Proposition 1.2, let us recall a sufficient condition for a closed set to be a Lipschitz manifold (see [16, Thm. 1.2.2.2, p. 12]).

LEMMA 1.1. *Let K be a closed subset of $\mathbb{R}^N$ and let x belong to ∂K. Assume that there exist some open set C, some $\rho > 0$, and some neighborhood U of x such that*

$$\forall y \in \partial M \cap U, \ \begin{cases} y + [0,\rho]C \subset K, \\ y - [0,\rho]C \subset \widehat{K}, \end{cases}$$

where $\widehat{K} := \overline{\mathbb{R}^N \backslash K}$.

Then ∂K is a Lipschitz manifold in a neighborhood of x.

Proof of Proposition 1.2. It is enough to combine Lemma 1.1 with the following lemma.

LEMMA 1.2. *Let M be as in Proposition* 1.1 *and let F satisfy* (4). *Also let x belong to $\partial M \cap O$ and $v \in \mathrm{Int}(F(x))$. Then*

$$\begin{aligned} x + [0,t](v + \tfrac{a}{2}B) &\subset \widehat{M}, \\ x + [0,t](-v + \tfrac{a}{2}B) &\subset M, \end{aligned}$$

where $a := d_{\partial F(x)}(v)$ and $t := \min\{T, \frac{a}{\ell(2\|v\|+a)}\}$.

Proof of Lemma 1.2. Note that $v + aB \subset F(x)$.

LEMMA 1.3. *If C_1, C_2, and C_3 are compact convex subsets of $\mathbb{R}^N$,*

$$[C_1 + C_3 \subset C_2 + C_3] \ \Rightarrow \ [C_1 \subset C_2].$$

Since the set-valued map F is ℓ-Lipschitz, use Lemma 1.3 to obtain

$$\forall y \in x + \frac{a}{2\ell}B, \qquad v + \frac{a}{2}B \subset F(y).$$

The map $s \to x + s(v + \frac{a}{2}u)$ is a solution of the differential inclusion for F on $[0,t]$ (for any $u \in B$) because $\|s(v + \frac{a}{2\ell}u)\| \le \frac{a}{2\ell}$ for $s \in [0,t]$. Since $x \in \partial M \cap O$ and M is semipermeable, any solution of the differential inclusion for F remains in $\widehat{M}$ (see Proposition 1.1(iii)). Thus

$$x + [0,t]\left(v + \frac{a}{2}B\right) \subset \widehat{M}.$$

We can prove in a similar way (using the fact that M is locally invariant by $-F$ from Proposition 1.1(ii)) that

$$x + [0,t]\left(-v + \frac{a}{2}B\right) \subset M. \qquad \square$$

To complete the proof of Proposition 1.2, let v belong to the interior of $F(x)$ and set $a := d_{\partial F(x)}(v)$. Then

$$\forall y \in x + \frac{a}{2\ell}, \qquad v + \frac{a}{2}B \subset F(y).$$

In particular, v belongs to the interior of $F(y)$ and $d_{\partial F(y)}(v) \geq \frac{a}{2}$. Thus, from Lemma 1.2,

$$\forall y \in x + \frac{a}{2\ell}, \begin{cases} y + [0,t](-v + \frac{a}{4}\overset{o}{B}) \subset M, \\ y + [0,t](v + \frac{a}{4}\overset{o}{B}) \subset \widehat{M}, \end{cases}$$

where $t := \min\{T, \frac{a}{\ell(4\|v\|+a)}\}$. □

Recall that equality $\mathbb{R}^N \backslash T_M(x) = D_{\widehat{M}}(x)$ is always fulfilled. Moreover, we have the following corollary.

COROLLARY 1.2. *Under the assumptions and notations of Proposition* 1.2*, we have, for any* $x \in \partial M \cap O$*,*

$$D_{\widehat{M}}(x) = D_{\mathbb{R}^N \backslash M}(x) \quad \textit{and} \quad \mathbb{R}^N \backslash T_{\widehat{M}}(x) = D_M(x) = D_{\mathbb{R}^N \backslash \widehat{M}}(x).$$

Note, moreover, that

$$\mathrm{Int}(F(x)) \cap T_M(x) = \emptyset \quad \textit{and} \quad -\mathrm{Int}(F(x)) \cap T_{\widehat{M}}(x) = \emptyset.$$

Proof. We prove only the first equality, the proof of the second one being essentially the same. Since $\mathbb{R}^N \backslash T_M(x) = D_{\mathbb{R}^N \backslash M}(x) \subset D_{\widehat{M}}(x)$, it remains to prove that

$$D_{\widehat{M}}(x) \subset D_{\mathbb{R}^N \backslash M}(x).$$

Let v belong to $D_{\widehat{M}}(x)$. Since ∂M is a Lipschitz manifold, there is a Lipschitz function $\phi : W \to \mathbb{R}$ ($W \subset \mathbb{R}^{N-1}$ open) and an open neighborhood $W' \subset \mathbb{R}^N$ of x such that

$$\partial M \cap W' = \{(y, \phi(y)) \mid y \in W\}.$$

We can assume, without loss of generality, that M is the epigraph of ϕ, while $\widehat{M}$ is the hypograph of ϕ and

$$\left(\mathbb{R}^N \backslash M\right) \cap W' = \{(y,t) \mid t < \phi(y) \ \&\ y \in W\}.$$

Set $x := (x_y, x_t)$ and $v := (v_y, v_t)$ (where x_y and v_y belong to $\mathbb{R}^{N-1}$ and x_t and v_t belong to $\mathbb{R}$). There is some $\alpha > 0$ such that

$$(x_y, x_t) +]0, \alpha] \left((v_y, v_t) + \alpha B_N\right) \subset \widehat{M}.$$

Thus, for any $u := (u_y, u_t) \in B_N$ and for any $\theta \in]0, \alpha]$,

$$\phi(x_y + \theta\,(v_y + \alpha u_y)) \geq x_t + \theta(v_t + \alpha u_t).$$

In particular, for any $u := (u_y, u_t) \in B_N$ and for any $\theta \in]0, \alpha/2]$,

$$\phi\left(x_y + \theta\left(v_y + \frac{\alpha}{2}u_y\right)\right) \geq x_t + \theta\left(v_t + \frac{\alpha}{\sqrt{2}}\right) > x_t + \theta\left(v_t + \frac{\alpha}{2}u_t\right),$$

because $(\frac{1}{2}u_y, \frac{\sqrt{2}}{2}) \in B_N$. This actually means that

$$(x_y, x_t) + \left]0, \frac{\alpha}{2}\right] \left((v_y, v_t) + \frac{\alpha}{2}B_N\right) \subset \mathbb{R}^N \backslash M,$$

so that v belongs to $D_{\mathbb{R}^N \backslash M}(x)$. □

2. Regularity of barrier solutions. We show here that, with any barrier solution, we can associate a Lipschitzian function $p(\cdot) : [0,T] \to \mathbb{R}^N$ such that $\|p(t)\| = 1$ and

$$0 = \langle x'(t), p(t)\rangle = H(x(t), p(t)) \text{ for almost every } t \in [0,T].$$

Moreover, $p(t)$ is an exterior normal to M at $x(t)$:

$$\forall t \in]0,T[, \qquad T_M(x(t)) = (p(t))^- .$$

The function $p(\cdot)$ is called the adjoint of $x(\cdot)$. In the case when the hamiltonian H is derivable, this adjoint coincides with the usual adjoint up to a multiplicative coefficient, and $(x(\cdot), p(\cdot))$ satisfies the Pontryagin principle.

2.1. Two preliminary lemmas. We first estimate the variations of the contingent cone to closed sets with semipermeable boundaries along the barrier solutions.

LEMMA 2.1. *Let M be as in Proposition* 1.1 *and let F satisfy* (4). *Assume that x belongs to $\partial M \cap O$ and that $x(\cdot) \in \mathcal{S}_F(x)$ is a barrier solution* (*i.e., it remains in $\partial M \cap O'$ on $[0,T]$*).

There is a constant C, which only depends on T and on the Lipschitz constant ℓ of F, such that

$$\forall 0 \le t \le T,\ \forall v \in T_M(x(t)),\ d_{T_M(x)}(v) \le Ct\|v\|. \tag{5}$$

Moreover,

$$\forall 0 \le s \le t \le T,\ \forall v \in T_{\widehat{M}}(x(s)),\ d_{T_{\widehat{M}}(x(t))}(v) \le C(t-s)\|v\|. \tag{6}$$

Proof of Lemma 2.1. We prove only (5), since the proof of (6) is essentially the same. If v belongs to $T_M(x(t))$, there exist $h_n \to 0^+$, $v_n \to v$ such that $x(t) + h_n v_n$ belongs to M for any n. The Filippov theorem [12] provides the existence of solutions $y_n(\cdot) \in \mathcal{S}_{-F}(x(t) + h_n v_n)$ such that

$$\|x'(t-s) + y_n'(s)\| \le \ell e^{\ell s} h_n \|v_n\| \text{ for almost every } s \in [0,t]. \tag{7}$$

The solutions $y_n(\cdot)$ remain in M on $[0,T]$ because M is (locally) invariant for $-F$ from Proposition 1.1.

Set $w_n := \frac{y_n(t)-x}{h_n}$. We now prove that the sequence $\{w_n\}$ converges, up to a subsequence, to some $w \in T_M(x)$ such that $\|v - w\| \le (e^{\ell t} - 1)\|v\|$. Indeed,

$$\begin{aligned} y_n(t) - x &= (y_n(t) - (x(t) + h_n v_n)) + (x(t) - x) + h_n v_n \\ &= \textstyle\int_0^t (y_n'(s) + x'(t-s))ds + h_n v_n. \end{aligned}$$

Combining this latter equality with (7) yields

$$\|w_n - v_n\| \le \frac{1}{h_n} \int_0^t \|y_n'(s) + x'(t-s)\| ds \le \|v_n\|(e^{\ell t} - 1).$$

Thus $\{w_n\}$ is bounded and converges, up to a subsequence, to some w which belongs to $T_M(x)$ and satisfies

$$\|w - v\| \le \|v\|(e^{\ell t} - 1).$$

So Lemma 2.1 is proved by setting $C := \sup_{t \in [0,T]} \frac{e^{\ell t} - 1}{t}$. ∎

We now compute $D_M(x)$ for some particular points x.

LEMMA 2.2. *Let $x(\cdot)$ be a barrier solution on $[0,T]$ and assume that condition* (4) *and the following condition are satisfied:*

$$\forall x \in O',\ \forall v \in \partial F(x),\ T_{F(x)}(v) \text{ is a half-space.} \tag{8}$$

For each $t \in]0,T[$ where the derivative $x'(t)$ exists,

$$-\mathrm{Int}[T_{F(x(t))}(x'(t))] \ = \ D_M(x(t)) \tag{9}$$

and

$$\mathrm{Int}[T_{F(x(t))}(x'(t))] \ = \ D_{\widehat{M}}(x(t)). \tag{10}$$

In particular, $D_M(x(t))$ and $D_{\widehat{M}}(x(t))$ are both equal to open half-spaces.

Assumption (8) plays a major role below. It is equivalent to

$$\begin{cases} \text{(i)} & \partial F(x) \text{ is a } \mathcal{C}^1 \text{ manifold for any } x \in O'. \\ \text{(ii)} & F(x) \text{ is convex with a nonempty interior for any } x \in O'. \end{cases}$$

Proof of Lemma 2.2. Let $t \in]0,T[$ be such that the derivative $v := x'(t)$ exists at time t. Recall that v belongs to $F(x(t)) \cap T_M(x(t))$. Let us first prove that

$$-\mathrm{Int}[T_{F(x(t))}(v)] \ \subset \ D_M(x(t)). \tag{11}$$

Let w belong to the interior of $-T_{F(x(t))}(v)$. Since $F(x)$ is convex, there are some $\lambda > 0$ and $a > 0$ such that $w + aB$ is contained in $\lambda(v - F(x(t)))$, i.e.,

$$v - \tau w + \tau a B \subset F(x(t)),$$

with $\tau := 1/\lambda$. Since F is ℓ-Lipschitz, Lemma 1.3 implies that

$$\forall y \in x(t) + \frac{a}{2\tau\ell}B, \qquad v - \tau w + \frac{a\tau}{2}B \subset F(y).$$

For $h > 0$ sufficiently small (say, $h \in [0,\epsilon]$ with $\epsilon > 0$) the solutions of the differential inclusion for $-F$ starting from $x(t+h)$ remain in $x(t) + \frac{a}{2\tau\ell}B$ on $[0,h]$. In particular, $s \to x(t+h) - s(v - \tau w + u)$ is a solution of the differential inclusion for $-F$ on $[0,h]$ if $\|u\| \le \frac{a\tau}{2}$. Since $x(t+h)$ belongs to M, the solutions of the differential inclusion for $-F$ starting from $x(t+h)$ remain in M on $[0,T]$. Thus

$$\forall s \in [0,h], \quad x(t+h) - s\left(v - \tau w + \frac{a\tau}{2}B\right) \subset M. \tag{12}$$

Since $v = x'(t)$, there is some $\epsilon' > 0$ such that, for $h \in [0,\epsilon']$,

$$\|x(t+h) - x(t) - hv\| \le \frac{ha\tau}{4}. \tag{13}$$

Combining (12) with $s = h$ with (13) yields, for any $h \in [0, \inf\{\epsilon, \epsilon'\}]$,

$$x(t) + h\tau w + \frac{ha\tau}{4}B \ \subset \ M.$$

Thus w belongs to $D_M(x(t))$.

We can prove in the same way that

$$\mathrm{Int}[T_{F(x(t))}(v)] \subset D_{\widehat{M}}(x(t))$$

because $\widehat{M}$ is (locally) invariant by F. Since $D_M(x(t)) \cap D_{\widehat{M}}(x(t)) = \emptyset$, and since both sets contain an open half-space and are open, $D_M(x(t))$ and $D_{\widehat{M}}(x(t))$ are, respectively, equal to the interiors of the half-spaces $-T_{F(x(t))}(x'(t))$ and $T_{F(x(t))}(x'(t))$. □

2.2. The adjoint of a barrier solution.

THEOREM 2.1 (definition of the adjoint). *Let M be a closed set with a semipermeable boundary and let x belong to $\partial M \cap O$ (cf. Proposition 1.1). Let $x(\cdot) \in \mathcal{S}_F(x)$ be a barrier solution on $[0,T]$ and C be the constant defined by Lemma 2.1. There is a $2C$-Lipschitzian function $p(\cdot) : [0,T] \to \mathbb{R}^N$ such that $\|p(t)\| = 1$ for any $t \in [0,T]$ and*

$$\forall t \in]0,T[, \quad T_M(x(t)) = (p(t))^-,$$

where $(p(t))^- = \{v \in \mathbb{R}^N \mid \langle p(t), v\rangle \le 0\}$. The function $p(\cdot)$ is called the adjoint of $x(\cdot)$ on $[0,T]$. Moreover, if $p(\cdot)$ is the adjoint of some barrier solution $x(\cdot)$, then

$$(p(0))^- \subset T_M(x(0)). \tag{14}$$

The adjoint $p(\cdot)$ is uniquely defined. Theorem 2.1 states that the contingent cone $T_M(x(t))$ is a half-space for $t > 0$ and $p(t)$ is the unique outward normal at $x(t)$. This means that, at $x(t)$, the closed set M is "smooth."

Proof of Theorem 2.1. *Existence.* If $x'(t)$ exists, Corollary 1.2 states that

$$D_M(x(t)) \subset T_M(x(t)) = \mathbb{R}^N \backslash D_{\mathbb{R}^N \backslash M}(x(t)) = \mathbb{R}^N \backslash D_{\widehat{M}}(x(t)).$$

From Lemma 2.2, the left- and right-hand sides of the inclusions are half-spaces. Thus $T_M(x(t))$ is a half-space and there is some $p(t)$ satisfying $\|p(t)\| = 1$ and $T_M(x(t)) = (p(t))^-$.

Now fix any $t \in]0,T[$. Since the solution $x(\cdot)$ is absolutely continuous, there are $t_n \to t^+$ and $s_n \to t^-$ such that the derivatives $x'(t_n)$ and $x'(s_n)$ exist. The sequences $(p(t_n))_{n\in\mathbb{N}}$ and $(p(s_n))_{n\in\mathbb{N}}$ converge, respectively, to p_1 and p_2 (up to a subsequence). Lemma 2.1 yields

$$(p_1)^- \subset \operatorname{Limsup} (p(t_n))^- \subset T_M(x(t))$$

(where Limsup denotes the Kuratowski upper limit [2]) and, for any $v \in T_M(x(t))$,

$$0 = \liminf_n d(v, T_M(x(s_n))) = \liminf_n \langle v, p(s_n)\rangle_+ = \langle v, p_2\rangle_+$$

(where $s_+ := \max\{s, 0\}$), so that $\langle v, p_2\rangle \le 0$.

Thus $(p_1)^- \subset T_M(x(t)) \subset (p_2)^-$ and $T_M(x(t))$ is equal to a half-space. Let us denote by $p(t)$ the common value $p_1 = p_2$. Then $T_M(x(t)) = (p(t))^-$. The function $p(\cdot)$ is defined on $]0,T[$.

Note that, for $t = 0$, the same proof shows that any upper limit p_1 of the functions $p(t_n)$ satisfies $(p_1)^- \subset T_M(x(0))$. Let us now prove that $p(\cdot)$ is Lipschitz continuous and so can be defined (uniquely) on $[0,T]$.

The adjoint is Lipschitzian. Let C be the constant of Lemma 2.1 and let $0 < s < t < T$. There are two cases.

(1) Either $\langle p(t), p(s)\rangle \ge 0$. Then we denote by v the projection of $p(s)$ onto $T_M(x(t))$, and by w the projection of v onto $T_M(x(s))$. From Lemma 2.1, the distance between v and w is smaller than or equal to $C(t-s)\|v\|$. Since

$$v = p(s) - \langle p(t), p(s)\rangle p(t) \quad \text{and} \quad w = v - \langle p(s), v\rangle p(s),$$

we have $\|v - w\| = 1 - \langle p(t), p(s)\rangle^2$.

Moreover, $\|v\|^2 = 1 - \langle p(t), p(s)\rangle^2 = \|v - w\|$. Combining this equation with $\|v - w\| \leq C(t-s)\|v\|$ yields

$$\|v - w\| \leq C^2(t-s)^2.$$

Note that

$$\|p(t) - p(s)\|^2 = 2 - 2\langle p(t), p(s)\rangle.$$

We conclude that

$$\|p(t) - p(s)\|^2 \leq 2 - 2\langle p(t), p(s)\rangle^2 \leq 2C^2(t-s)^2.$$

Thus

$$\|p(t) - p(s)\| \leq 2C|t-s|. \tag{15}$$

(2) Or $\langle p(t), p(s)\rangle \leq 0$. Then $p(s)$ belongs to $(p(t))^- = T_M(x(t))$. Thus Lemma 2.1 states that

$$d(p(s), (p(s))^-) \leq C(t-s).$$

Since the left-hand side is equal to 1, $\langle p(t), p(s)\rangle \leq 0$ cannot occur unless $t - s \geq 1/C$. In that case, equation (15) is fulfilled. Thus we have proved that $p(\cdot)$ is a $2C$-Lipschitz function. □

LEMMA 2.3 (characterization of the adjoint). *The adjoint $p(\cdot)$ of a barrier solution $x(\cdot)$ on $[0,T]$ is the unique continuous function satisfying*

$$\begin{cases} \text{(a)} & \forall t \in [0,T],\ \|p(t)\| = 1, \\ \text{(b)} & H_F(x(t), p(t)) = \langle x'(t), p(t)\rangle = 0 \text{ for almost every } t \in [0,T]. \end{cases} \tag{16}$$

COROLLARY 2.1. *Under the notations and the assumptions of Lemma* 2.3,

$$T_M(x(t)) = -T_{F(x(t))}(x'(t)) = (p(t))^-$$

for almost every $t \in]0,T[$.

COROLLARY 2.2. *Assume, moreover, that $F(x)$ is strictly convex for any $x \in O'$. Then any barrier solution is $\mathcal{C}^1$.*

Proof. Indeed, the set-valued map $t \rightsquigarrow \operatorname{Arg\,min}_{v \in F(x(t))} \langle v, p(t)\rangle$ is upper semicontinuous and, in fact, single-valued because $F(x)$ is strictly convex. So it is continuous. From Lemma 2.3, $x'(t)$ is almost everywhere equal to the continuous function $t \to \operatorname{Arg\,min}_{v \in F(x(t))} \langle v, p(t)\rangle$ and so is continuous. □

Proof of Lemma 2.3 Assume that $p(\cdot)$ is the adjoint of $x(\cdot)$ on $[0,T]$. We have to prove (16). Let $t \in]0,T[$. The set $F(x(t))$ is convex and has a nonempty interior from assumption (8). Let w belong to the interior of $F(x(t))$. From Corollary 1.2, $w \notin T_M(x(t))$. Thus $\langle w, p(t)\rangle > 0$ from Theorem 2.1. Since $F(x(t)) = \overline{\operatorname{Int}(F(x(t)))}$, we have proved $H_F(x(t), p(t)) \geq 0$. If $x'(t)$ exists, then $x'(t)$ belongs to $F(x(t))$ and so $\langle x'(t), p(t)\rangle \geq 0$. Moreover, $x'(t)$ belongs to $T_M(x(t)) = (p(t))^-$, and thus $\langle x'(t), p(t)\rangle \leq 0$. So (16(b)) holds true because $x(\cdot)$ is almost everywhere derivable.

Conversely, let us now assume that some continuous function $p(\cdot) : [0,T] \to \mathbb{R}^N$ satisfies (16). We have to prove that $p(\cdot)$ is the adjoint of $x(\cdot)$ on $[0,T]$. Fix any $t \in (0,T]$ where the derivative $x'(t)$ exists and where (16(b)) is fulfilled. Combining

Lemma 2.2 and Theorem 2.1 yields that $T_M(x(t))$ is the closure of the half-space $D_M(x(t))$. Thus

$$T_M(x(t)) = -T_{F(x(t))}(x'(t)) = \overline{\bigcup_{\lambda>0} \lambda(x'(t) - F(x(t)))}. \tag{17}$$

Thanks to (16), for any $\lambda \geq 0$ and any $w \in F(x(t))$, one has

$$\langle \lambda(x'(t) - w), p(t)\rangle \leq 0. \tag{18}$$

Combining (17) and (18) yields

$$\forall v \in T_M(x(t)),\ \langle v, p(t)\rangle \leq 0.$$

So $(p(t))^- = T_M(x(t))$ for almost every $t \in (0,T]$. In particular, $p(t)$ coincide almost everywhere with the adjoint of $x(\cdot)$, which is continuous. So $p(\cdot)$ is equal to the adjoint of $x(\cdot)$. □

We study here the regularity properties of the function which associates its adjoint to a solution.

PROPOSITION 2.1. *Assume that $x_n(\cdot)$ are barrier solutions starting from $x_n \in \partial M \cap O$ and converging to some $x(\cdot)$ barrier solution starting from $x \in \partial M \cap O$. If $p_n(\cdot)$ are the adjoint of $x_n(\cdot)$ on $[0,T]$, then the $p_n(\cdot)$ converge uniformly to the adjoint of $x(\cdot)$ on $[0,T]$.*

Proof of Proposition 2.1. Since the $(p_n(\cdot))$ are uniformly continuous, Ascoli's theorem states that $p_n(\cdot)$ converge uniformly to some continuous function $p(\cdot)$ (up to a subsequence). To prove that $p(\cdot)$ is the adjoint of $x(\cdot)$, it is sufficient to show that $p(\cdot)$ satisfies (16). For any $t \in [0,T]$, $\|p(t)\| = 1$. From Lemma 2.3, for almost every $t \in (0,T]$, $\langle x_n'(t), p_n(t)\rangle = 0$. Thus

$$\forall t \in [0,T],\ \int_0^t \langle x_n'(s), p_n(s)\rangle ds = 0.$$

The sequence of functions $p_n(\cdot)$ converges uniformly to $p(\cdot)$, and the sequence $x_n'(\cdot)$ converges weakly to $x(\cdot)$. Thus

$$\forall t \in [0,T],\ \int_0^t \langle x'(s), p(s)\rangle ds = 0,$$

which implies that $\langle x'(t), p(t)\rangle = 0$ for almost every $t \in]0,T[$.

Let $t \in (0,T]$ and v belong to $F(x(t))$. We are going to prove that $\langle v, p(t)\rangle \geq 0$. From Michael's theorem [22], a continuous function $\tilde{v}(\cdot) : \mathbb{R}^N \to V$ exists, such that

$$\forall x \in \mathbb{R}^N, \quad \tilde{v}(x) \in F(x) \quad \text{and} \quad \tilde{v}(x(t)) = v.$$

Note that $\tilde{v}(x_n(t))$ converge to $\tilde{v}(x(t))$, and that $\langle \tilde{v}(x_n(t)), p_n(t)\rangle \geq 0$. Letting $n \to +\infty$ yields $\langle v, p(t)\rangle \geq 0$ for any $v \in F(x(t))$ and any $t \in (0,T]$.

Thus $p(\cdot)$ satisfies (16) and is indeed the adjoint of $x(\cdot)$. Since any converging subsequence of the uniformly continuous sequence $(p_n(\cdot))$ converges to the adjoint of $x(\cdot)$, we have proved that the $p_n(\cdot)$ converge to the adjoint of $x(\cdot)$. □

2.3. Barrier solutions and the Pontryagin principle. We now prove that, under some assumptions of differentiability of the hamiltonian H, barrier solutions satisfy the Pontryagin principle.

In Isaacs' pioneering work on differential games [18], semipermeable hypersurfaces are constructed by using the method of characteristics, which is very close to the Pontryagin principle. We show here that this method of construction is a priori justified since the barrier solutions indeed satisfy the Pontryagin principle.

THEOREM 2.2. *Assume that the set-valued map F satisfies* (4), (8) *and, moreover, that its associated hamiltonian H is $\mathcal{C}^2$ on $\mathbb{R}^N \times [\mathbb{R}^N \backslash \{0\}]$.*

Let M be a closed set with a semipermeable boundary. If $x(\cdot)$ is a barrier solution and $p(\cdot)$ is its adjoint, and if $q(\cdot) : [0,T] \to \mathbb{R}^N$ is defined by

$$\forall t \in [0,T],\ q(t) := p(t) \exp\left(- \int_0^t \left\langle \frac{\partial H}{\partial x}(x(s),p(s)),p(s) \right\rangle ds \right),$$

then $(x(\cdot),q(\cdot))$ is a solution to the hamiltonian system

$$\left\{ \begin{array}{l} x'(t) = \frac{\partial H}{\partial p}(x(t),q(t)), \\ q'(t) = -\frac{\partial H}{\partial x}(x(t),q(t)), \\ q(0) := p(0). \end{array} \right. \tag{19}$$

COROLLARY 2.3. *Suppose that the assumptions of the previous theorem are satisfied. If $x(\cdot)$ is a barrier solution, then $x(\cdot)$ is $\mathcal{C}^1$ on $[0,T]$. Moreover, for any $q \neq 0$ such that*

$$\langle q, x'(0) \rangle = H(x(0),q) = 0,$$

the solution $(x(\cdot),q(\cdot))$ of (19) *with initial condition $(x(0),q)$ satisfies*

$$\forall t \in]0,T[, \quad T_M(x(t)) = (q(t))^-.$$

Proof of Theorem 2.2. Since, for any $t \in [0,T]$, $H(x(t),p(t)) = 0$, since $\langle x'(t),p(t) \rangle = 0$ for almost every $t \in [0,T]$, and since H is differentiable, one has

$$x'(t) = \mathrm{Arg} \min_{v \in F(x(t))} \langle v, p(t) \rangle = \frac{\partial H}{\partial p}(x(t),p(t))$$

for almost every $t \in [0,T]$. In fact, these equalities hold true everywhere because the right-hand side is continuous. Thus $x'(\cdot)$ is defined everywhere on $[0,T]$ and is continuous.

Let $t \in]0,T[$ be such that $p'(t)$ exists. Let us prove that

$$\forall v \perp p(t),\ \left\langle v, \left(\frac{\partial^2 H}{\partial x \partial p}(x(t),p(t)) \right)^* p(t) + p'(t) \right\rangle \geq 0. \tag{20}$$

Since $v \perp p(t)$, v belongs to the boundary of $T_M(x(t))$ from Theorem 2.1, and so to $T_{\partial M}(x(t))$. Thus there are $h_n \to 0^+$ and $v_n \to v$ with $x(t) + h_n v_n \in \partial M$. For any n, let us consider the solutions $x_n(\cdot)$ (with final conditions) to

$$\left\{ \begin{array}{l} x_n'(s) = \frac{\partial H}{\partial p}(x_n(s),p(s)) \text{ for } s \in [0,t], \\ x_n(t) := x(t) + h_n v_n. \end{array} \right.$$

Note that the $x_n(\cdot)$ converge to $x(\cdot)$ and that, moreover,

$$\forall s \in [t,T],\ \frac{x_n(s)-x(s)}{h_n} \to z(s), \tag{21}$$

where $z(\cdot)$ is the solution to

$$\begin{cases} z'(s) = \frac{\partial^2 H}{\partial x \partial p}(x(s),p(s))z(s), \\ z(t) = v. \end{cases}$$

Since $x_n(t) \in \partial M$, the functions $s \to x_n(t-s)$ are solutions of the differential inclusion for $-F$ starting from M, and thus remain in M from the semipermeability of M. From Theorem 2.1, $T_M(x(s)) = (p(s))^-$ for every s. From (21), $z(s)$ belongs to $T_M(x(s))$ so that $\langle z(s), p(s)\rangle \le 0$. In particular,

$$\forall s \in [0,t],\ \langle z(s),p(s)\rangle - \langle z(t),p(t)\rangle \le 0,$$

because $\langle z(t),p(t)\rangle = \langle v,p(t)\rangle = 0$. Dividing by $s-t$ and letting $s \to t^-$ gives

$$\langle z'(t),p(t)\rangle + \langle z(t),p'(t)\rangle \ge 0,$$

and thus inequality (20) holds true.

Since H is positively homogeneous, Euler's rule states that

$$\left(\frac{\partial^2 H}{\partial x \partial p}(x(t),p(t))\right)^* p(t) = \frac{\partial H}{\partial x}(x(t),p(t)).$$

From (20), for every t where $p'(t)$ exists, there is some $\lambda(t) \in \mathbb{R}$ such that

$$p'(t) = -\frac{\partial H}{\partial x} + \lambda(t)p(t).$$

(Note that one can compute $\lambda(t)$ explicitly because $\|p(t)\| = 1$ implies that $\langle p(t),p'(t)\rangle = 0$, and thus $\lambda(t) = \langle \frac{\partial H}{\partial x}, p(t)\rangle$.)

Now let $\mu(t) := \exp(-\int_0^t \lambda(s)ds)$ and $q(t) := \mu(t)p(t)$. Using the fact that H is positively homogeneous, it is easy shown that $(x(\cdot),q(\cdot))$ is a solution to the hamiltonian system. □

Acknowledgment. We would like to warmly thank Pr. T. Rzeżuchowski for his helpful suggestions and friendly advice.

REFERENCES

[1] L. Ambrosio, P. Cannarsa, and H.M. Soner, *On the propagation of singularities of semi-convex functions*, Ann. Scuola Sup. Pisa Ser. IV, XX (1993), pp. 597–616.
[2] J.-P. Aubin and H. Frankowska, *Set-Valued Analysis*, Birkhaüser, Boston, 1990.
[3] J.-P. Aubin, *Viability Theory*, Birkhaüser, Boston, 1991.
[4] J.-P. Aubin and H. Frankowska, *Partial differential inclusions governing feedback controls*, J. Convex Analysis, 2 (1995), pp. 19–40.
[5] J.-P. Aubin and H. Frankowska, *Viability kernels of control systems*, in Nonlinear Synthesis, C. Byrnes and A. Kurzhanski, eds., Birkäuser, Boston, pp. 12–33.
[6] M. Bardi, *A boundary value problem for the minimum time function*, SIAM J. Control Optim., 28 (1989), pp. 950–965.
[7] G. Barles and B. Perthame, *Exit time problems in optimal control and vanishing method*, SIAM J. Control Optim., 26 (1988), pp. 1123–1134.

[8] P. Cannarsa and H. Frankowska, *Some characterizations of the value function in control theory*, SIAM J. Control Optim., 29 (1991), pp. 1322–1347.
[9] P. Cannarsa and C. Sinestrari, *Convexity properties of the minimum time function*, Calc. Var., 3 (1995), pp. 273–298.
[10] P. Cardaliaguet, M. Quincampoix, and P. Saint-Pierre, *Optimal times for constrained non-linear control problems without local controllability*, Appl. Math. Optim., to appear.
[11] N. Caroff and H. Frankowska, *Conjugate points and shocks in nonlinear optimal control* Trans. Amer. Math. Soc., 348 (1996), pp. 3133–3153.
[12] A.F. Filippov, *Classical solutions of differential equations with multivalued right hand side*, SIAM J. Control, 5 (1967), pp. 609–621.
[13] H. Frankowska, *Local controllability and infinitesimal generators of semi-groups of set-valued maps*, SIAM J. Control Optim., 25 (1987), pp. 412–432.
[14] H. Frankowska, *Control of Nonlinear Systems and Differential Inclusions*, manuscript.
[15] H. Frankowska and M. Quincampoix, *Viability kernels of differential inclusions with constraints: Algorithm and applications*, Math. Systems Estimation Control, 1 (1992), pp. 371–388.
[16] P. Grisvard, *Elliptic Problems in Nonsmooth Domains*, Pitman, Boston, 1985.
[17] H.G. Guseinov, A.I. Subbotin, and V.N. Ushakov, *Derivatives for multivalued mappings with applications to the game-theorical problems of control*, Problems Control Inform. Theory, 14 (1985), pp. 295–298.
[18] R. Isaacs, *Differential Games*, John Wiley, New York, 1965.
[19] H. Ishii, *Uniqueness of unbounded viscosity solutions of Hamilton-Jacobi equations*, Indiana Univ. Math. J., 33 (1984), pp. 721–748.
[20] S.N. Kruzkov, *Generalized solutions of Hamilton-Jacobi equations of eikonal type* I, Math. USSR-Sb, 27 (1975), pp. 406–446.
[21] P.L. Lions, *Generalized Solutions of Hamilton-Jacobi equations*, Pitman, Boston, 1982.
[22] E. Michael, *Continuous selections*, Ann. Math., 63 (1956), pp. 361–381.
[23] N.N. Petrov, *Controllability of autonomous systems*, Differential Equations, 4 (1968), pp. 311–317.
[24] N.N. Petrov, *On the Bellman function for time-optimal process problem*, J. Appl. Math. Mech., 34 (1970), pp. 785–791.
[25] M. Quincampoix, *Differential inclusions and target problems*, SIAM J. Control Optim., 30 (1992), pp. 324–335.
[26] M. Quincampoix, *Enveloppes d'invariance pour des inclusions différentielles Lipschitziennes et applications aux problèmes de cibles*, C. R. Acad. Sci. Paris Sér. I Math., 314 (1992), pp. 343–347.

SIAM J. CONTROL OPTIM.
Vol. 35, No. 5, pp. 1653–1671, September 1997

012

ON THE REGULARITY OF SEMIPERMEABLE SURFACES IN CONTROL THEORY WITH APPLICATION TO THE OPTIMAL EXIT-TIME PROBLEM (PART II)*

PIERRE CARDALIAGUET†

Abstract. In control theory, a semipermeable surface is an (in general nonsmooth) oriented surface that, on one hand, contains solutions (the so-called barrier solutions) of the controlled system and, on the other hand, may be crossed by the solutions of this system in only one direction. Without making any assumption on the regularity of the boundary of the semipermeable surface, we show that the barrier solutions contained in this semipermeable surface satisfy the Pontryagin principle, that this surface is a Lipschitz manifold, and that it is, locally, the graph of a semiconcave function. Applying these results to the optimal exit-time function from a given open set yields, without any controllability assumption at the boundary of the open set, that this function is semiconcave on an open dense subset of its domain.

Key words. semipermeable surfaces, differential inclusion, viability theory, minimal time function

AMS subject classifications. 49J24, 49J52, 49N60

PII. S0363012996312155

Introduction. In this paper, we continue our investigation of the regularity of semipermeable surfaces for the differential inclusion

$$\begin{cases} x'(t) \in F(x(t)), \\ x(0) = x_0, \end{cases}$$

where F is given by $F(x) := \bigcup_{u \in U} f(x,u)$. (See Part I of this paper, published in the same issue.)

In this paper, we prove that, under suitable assumptions, semipermeable surfaces for f are locally the graph of some semiconcave function. We apply this result to the optimal exit-time function, which is proved to be locally semiconcave in an open dense subset of its domain.

1. Regularity of semipermeable boundaries. In this section, we show that the contingent cone and the Dubovitsky–Miljutin cone at a point x to a closed set M with semipermeable boundary are determined by the "regulation map" [3], which is the set-valued map $\mathcal{R} : M \rightsquigarrow \mathbb{R}^N$ defined by

$$\forall x \in M,\ \mathcal{R}(x) := F(x) \cap T_M(x).$$

We also show that, if the boundary of M is semipermeable in a neighborhood of $x \in \partial M$, then $T_M(x)$ is a union of closed half-spaces and

$$D_M(x) = \mathrm{Int}\left(T_M(x)\right).$$

Moreover, we give a formula relating $T_M(x)$ and $\mathcal{R}(x)$.

We finally prove that the restriction of $T_M(\cdot)$ to $\partial M \cap O$ has a closed graph.

*Received by the editors June 5, 1996; accepted for publication (in revised form) July 1, 1996.
http://www.siam.org/journals/sicon/35-5/31215.html

†CEREMADE, URA CNRS 749, Université Paris-Dauphine, Place du Maréchal de Lattre de Tassigny, 75775 Paris cedex 16, France (cardaliaguet@ceremade.dauphine.fr).

1.1. Characterization formula for the computation of the contingent cone.

PROPOSITION 1.1. *Assume that F satisfies*

$$\text{(1)}\qquad \begin{cases} \text{(i)} & F : \mathbb{R}^N \rightsquigarrow \mathbb{R}^N \text{ has convex compact values,} \\ \text{(ii)} & F \text{ is } \ell\text{-Lipschitz, i.e.,} \\ & \quad \forall x, y \in \mathbb{R}^N \times \mathbb{R}^N,\ F(y) \subset F(x) + \ell B, \end{cases}$$

and that the boundary of M is semipermeable in a neighborhood of $x \in \partial M$. Then we have the following estimate of $T_M(x)$:

$$\text{(2)}\qquad \bigcup_{v \in \mathcal{R}(x)} -T_{F(x)}(v) \ \subset\ T_M(x).$$

Proof. Let v belong to $F(x) \cap T_M(x)$. There exist sequences $v_n \to v$ and $h_n \to 0^+$ such that, for any n, $x + h_n v_n$ belongs to M. Any solution of the differential inclusion for $-F$ remains in M on $[0, T]$. So for $t \in [0, T]$,

$$\mathcal{A}_{-F}(x + h_n v_n)(t) \subset M,$$

where $\mathcal{A}_G(y)(s)$ denotes the reachable set from y at time s for the set-valued map G. Then[1]

$$x + h_n v_n - tF(x + h_n v_n) \subset \mathcal{A}_{-F}(x + h_n v_n)(t) + o(t)B,$$

where $o(t)/t \to 0^+$ if $t \to 0^+$. (Note that $o(t)$ does not depend on n.) Thus we have

$$x + h_n v_n - tF(x + h_n v_n) \subset M + o(t)B.$$

If we set $t = h_n$, we deduce

$$x + h_n(v - F(x)) \subset M + o(h_n)B,$$

which proves that $v - F(x)$ is contained in $T_M(x)$. Since $T_M(x)$ is a closed cone, this completes the proof of (2) because

$$-T_{F(x)}(v) = \overline{\bigcup_{\lambda > 0} \lambda(v - F(x))}. \qquad \square$$

Unfortunately, inclusion (2) may be strict. Moreover, it may happen that the contingent cone to a closed set with a semipermeable boundary is not completely determined by the regulation map.

Example. Let $F(x, y, z) := [0, 1] \times [0, 1] \times [2, 3]$ be the dynamics. Set $K_1 := \{(x, y, z) \mid x \le 0\}$ and $K_2 := \{(x, y, z) \mid x \le 0 \text{ or } x + y \le 0\}$. Then ∂K_1 and ∂K_2 are semipermeable barriers for F. Moreover,

$$F(0, 0, z) \cap T_{K_1}(0, 0, z) = F(0, 0, z) \cap T_{K_2}(0, 0, z) = \{0\} \times [0, 1] \times [2, 3]$$

[1] If $G : \mathbb{R}^N \rightsquigarrow \mathbb{R}^N$ is an ℓ-Lipschitzian set-valued map, with compact values; then

$$y + tG(y) \subset \mathcal{A}_G(y)(t) + o(t)B,$$

where $o(t)/t \to 0^+$ if $t \to 0^+$, and $o(t)$ depend only on ℓ (see Frankowska [7]).

for any $z \in \mathbb{R}$. But

$$T_{K_1}(0,0,z) \neq T_{K_2}(0,0,z). \qquad \square$$

However, under some additional assumptions on F, there is an equality in equation (2).

THEOREM 1.1 (characterization formula for the contingent cone). *Let F satisfy* (1) *and*

$$\forall x \in O',\ \forall v \in \partial F(x),\ T_{F(x)}(v) \text{ is a half-space.} \tag{3}$$

Let M be a closed set with a semipermeable boundary in a neighborhood of $x \in \partial M$. Then

$$\bigcup_{v \in \mathcal{R}(x)} -T_{F(x)}(v) \ = \ T_M(x). \tag{4}$$

Thus, in general, $T_M(x)$ is a union of half-spaces.

To prove Theorem 1.1, we need two remarks.

LEMMA 1.1. *Let F satisfy* (1). *If $x_n \to x$, $h_n \to 0^+$, and, for any n, $x_n(\cdot) \in \mathcal{S}_F(x_n)$, then $(\frac{x_n(h_n)-x_n}{h_n})$ converges, up to a subsequence, to some element of $F(x)$.*

LEMMA 1.2. *Let C be a convex compact subset of $\mathbb{R}^N$, $v_n \in C$ converge to v, and $l_n \to +\infty$. Then*

$$-T_C(v) \subset \ \operatorname*{Limsup}_{n\to\infty}\ l_n(v_n - C).$$

The proof of Lemma 1.1 is straightforward, so we give only the proof of Lemma 1.2. If w belongs to C, then

$$[m < l] \ \Rightarrow \ [m(w - C) \subset l(w - C)], \tag{5}$$

because $(w - C)$ is a convex set and contains 0. If we can find a subsequence of the (v_n) such that $v_{n_k} = v$ for any k, then the result is an obvious consequence of equation (5) with $w = v$. If it is not the case, set

$$m_n = \inf\{\sqrt{l_n}, \|v - v_n\|^{-1/2}\}.$$

Then $l_n(v_n - C)$ contains $m_n(v_n - C)$, because v_n belongs to C and $m_n \leq l_n$. Moreover,

$$m_n(v_n - C) = m_n(v - C) + m_n(v_n - v).$$

Since the sequence $m_n(v_n - v)$ converges to 0 and the upper limit of $m_n(v - C)$ is equal to $-T_C(v)$ because $m_n \to +\infty$, we have proved Lemma 1.2. $\square$

Proof of Theorem 1.1. Let w belong to $T_M(x)$. There exist $h_n \to 0^+$ and $w_n \to w$ such that, for any n, $x + h_n w_n$ belongs to M. For any $n \in \mathbb{N}$, there exists $x_n(\cdot) \in S_F(x + h_n w_n)$, which remains in M on $[0, T]$.

From Lemma 1.1, for any fixed $l > 0$, the sequence $v_l^n := \frac{x_n(lh_n)-(x+h_n w_n)}{lh_n}$ converges (up to a subsequence) to a limit v_l, which belongs to $F(x)$.

With the same ideas used in the proof of Proposition 1.1, we have

$$x_n(lh_n) - tF(x_n(lh_n)) \ \subset \ M + o(t)B$$

so that, by setting $t = mh_n$, with $m > 0$, we can deduce

$$x + h_n w_n + l h_n v_l^n - m h_n F(x) \ \subset \ M + o(mh_n)B.$$

Thus

$$w + l v_l - m F(x) \ \subset \ T_M(x). \tag{6}$$

Dividing by l and letting $l \to +\infty$ yield

$$\operatorname*{Limsup}_{l \to +\infty} \ v_l \subset T_M(x).$$

This upper limit (in the Kuratowski sense) is nonempty because the v_l belong to $F(x)$, which is compact. Let v belong to this upper limit. Note that v belongs to $\mathcal{R}(x)$ because $F(x)$ and $T_M(x)$ are closed. There exists l_n converging to $+\infty$ such that the limit of v_{l_n} is equal to v. For simplicity, we set $v_n := v_{l_n}$.

Inclusion (6) holds true for any $m \geq 0$. So, with $m := l_n$,

$$w + l_n[v_n - F(x)] \subset T_M(x).$$

Since $T_M(x)$ is closed, Lemma 1.2 yields

$$w - T_{F(x)}(v) \subset T_M(x).$$

Assume for a while that w does not belong to $-T_{F(x)}(v)$. From assumption (3), $-T_{F(x)}(v)$ is a half-space. So the cone spanned by $w - T_{F(x)}(v)$ is equal to $\mathbb{R}^N$ and is contained in $T_M(x)$. This is impossible, since $T_M(x)$ cannot be the full space, as we showed in Corollary 1.2 in Part I of this paper.

1.2. The Dubovitsky–Miljutin cone. We now compute the Dubovitsky–Miljutin cone in terms of the regulation map. For that purpose, we shall need a result concerning the existence of particular barrier solutions.

Throughout this section, we assume that the set-valued map F satisfies (1) and (3).

PROPOSITION 1.2. *Let M be a closed set with a semipermeable boundary in a neighborhood of $x \in \partial M$. Assume that w belongs to $\mathcal{R}(x)$. Then there is some barrier solution $x(\cdot)$ starting from x such that the adjoint $p(\cdot)$ of $x(\cdot)$ satisfies*

$$\langle w, p(0) \rangle = 0.$$

Let us point out a particular case of that proposition.

COROLLARY 1.1. *Assume, moreover, that $F(x)$ is strictly convex. Then there is a barrier solution $x(\cdot)$ starting from x such that $x'(0) = w$.*

The proof of this corollary is a straightforward consequence of Proposition 1.2 and of the following lemma.

LEMMA 1.3. *Let $x(\cdot)$ be a barrier solution and $p(\cdot)$ be its adjoint. Set*

$$F_p := \{v \in F(x) \mid \langle v, p(0) \rangle = 0\}.$$

Then

$$\operatorname*{Limsup}_{t \to 0^+} \ \frac{x(t) - x(0)}{t} \subset F_p,$$

where Limsup *denotes the Kuratowski upper limit* [2].

Proof. The upper limit of $\frac{x(t)-x}{t}$ is contained in $F(x)$. It remains to prove that, if z belongs to this upper limit, then $\langle z, p(0)\rangle = 0$. But

$$\begin{aligned} |\langle \tfrac{x(t)-x}{t}, p(0)\rangle| &\leq \tfrac{1}{t}\int_0^t |\langle x'(s), p(0)\rangle| ds \\ &\leq \tfrac{1}{t}\int_0^t |\langle x'(s), p(s)\rangle| ds + \tfrac{1}{t}\int_0^t \|x'(s)\| 2Cs ds \\ &\leq \rho C t, \end{aligned}$$

where $\rho := \max_{t\in[0,T]} \|x'(t)\|$ and the constant C comes from Theorem 2.1 of Part I. Letting $t \to 0^+$ proves our claim. □

Proof of Proposition 1.2. We use the same kind of ideas as in the proof of Theorem 1.1. Since w belongs to $T_{\partial M}(x)$, there exists $h_n \to 0^+$, $w_n \to w$ such that $x + h_n w_n$ belongs to ∂M. There also exist barrier solutions $x_n(\cdot)$ starting from $x + h_n w_n$. We denote by $p_n(\cdot)$ their adjoint. From Theorem 5.3.1 of [2], the $x_n(\cdot)$ converge, up to a subsequence, to some solution $x(\cdot)$ starting from x while the $p_n(\cdot)$ converge to some function $p(\cdot)$ which is the adjoint of $x(\cdot)$ (see Proposition 2.1 of Part I).

As in Theorem 1.1, for any $\lambda > 0$, there exists $v_\lambda \in F(x)$ such that

$$\frac{x_n(\lambda h_n) - (x + h_n w_n)}{\lambda h_n} \to v_\lambda$$

up to a subsequence. Moreover, a subsequence of the (v_λ) converge to some $v \in F(x) \cap T_M(x)$. As in the proof of Theorem 1.1, we can prove that

$$w \in -T_{F(x)}(v). \tag{7}$$

We want to prove that $\langle w, p(0)\rangle = 0$. For that purpose, let us first show that $\langle p(0), v\rangle = 0$. We denote by ρ the following bound:

$$\rho := \sup\{\|F(x_n(t))\| \mid n \in \mathbb{N} \text{ and } t \in [0,T]\}.$$

For any $n \in \mathbb{N}$,

$$\begin{aligned} |\langle p_n(0)\ , &\ x_n(\lambda h_n) - (x + h_n w_n)\rangle| \\ &\leq \textstyle\int_0^{\lambda h_n} |\langle p_n(0), x_n'(s)\rangle| ds \\ &\leq \textstyle\int_0^{\lambda h_n} |\langle p_n(s), x_n'(s)\rangle| ds + \int_0^{\lambda h_n} 2Cs\|x_n'(s)\| ds \\ &\leq \rho C \lambda^2 h_n^2 \end{aligned}$$

because $p_n(\cdot)$ are $2C$-Lipschitz. Dividing these inequalities by λh_n and letting $n \to +\infty$ yields $\langle p(0), v_\lambda\rangle = 0$. Finally, letting $\lambda \to +\infty$ gives the desired formula: $\langle p(0), v\rangle = 0$.

Since $p(\cdot)$ is an adjoint,

$$0 = \inf_{z\in F(x)} \langle z, p(0)\rangle = \langle v, p(0)\rangle.$$

To complete the proof, we need the following remark.

LEMMA 1.4. *Assume that F satisfies* (3). *If $p \in \mathbb{R}^N\backslash\{0\}$ and $v \in F(x)$ are such that $H(x,p) = \langle v, p\rangle$, then*

$$T_{F(x)}(v) = (p)^+,$$

where $(p)^+ := \{w \in \mathbb{R}^N \mid \langle w, p\rangle \geq 0\ \}$.

Thus Lemma 1.4 states that

$$(p(0))^- = -T_{F(x)}(v).$$

Since, from (7), w belongs to the right-hand side of the equality, $\langle w, p(0)\rangle \le 0$. But w belongs to $F(x)$. So $\langle w, p(0)\rangle \ge 0$ from Theorem 2.1 of Part I. So $\langle w, p(0)\rangle = 0$, and Proposition 1.2 is proved. □

THEOREM 1.2 (the Dubovitsky–Miljutin cone). *Let M be a closed set with a semipermeable boundary in a neighborhood of $x \in \partial M$. The Dubovitsky–Miljutin cone is determined by the regulation map*

$$D_M(x) \;=\; \bigcup_{v\in\mathcal{R}(x)} \operatorname{Int}\left(-T_{F(x)}(v)\right) \;=\; \operatorname{Int} T_M(x).$$

COROLLARY 1.2. *Under the notations and assumptions of the previous theorem,*

$$T_{\widehat{M}}(x) = \bigcap_{v\in\mathcal{R}(x)} T_{F(x)}(v) \quad \textit{and} \quad D_{\widehat{M}}(x) = \bigcap_{v\in\mathcal{R}(x)} \operatorname{Int}\left(T_{F(x)}(v)\right).$$

Proof. Combining Corollary 1.2 of Part I with Theorem 1.2 gives the first formula. The second formula comes from Theorem 1.1.

Proof of Theorem 1.2.

(1) Let us first prove that

$$\bigcup_{v\in\mathcal{R}(x)} \operatorname{Int}\left(-T_{F(x)}(v)\right) \;=\; \operatorname{Int} T_M(x). \tag{8}$$

The left-hand side of the equality is contained in the right-hand side of Theorem 1.1. Let us prove the converse inclusion. For that purpose, we denote by A the left-hand side of the equality.

Let w belong to the interior of $T_M(x)$. Let us set

$$\mathcal{P} := \{p \in \mathbb{R}^N,\ \|p\| \le 1 \text{ and } \exists v \in \mathcal{R}(x) \text{ with } (p)^+ = -T_{F(x)}(v)\}.$$

Note that $\mathcal{P}$ is compact since $\mathcal{R}(x)$ is compact. Since w belongs to $\operatorname{Int}(T_M(x))$, Theorem 1.1 states that there is some positive α such that, for any $b \in B$, there is some $v \in \mathcal{R}(x)$ such that $w + \alpha b \in -T_{F(x)}(v)$. Then

$$\inf_{b\in B} \sup_{p\in\mathcal{P}} \langle w + \alpha b, p\rangle \ge 0$$

so that

$$\sup_{p\in\overline{Co}(\mathcal{P})} \inf_{b\in B} \langle w + \alpha b, p\rangle \ge 0.$$

In particular,

$$\sup_{p\in\overline{Co}(\mathcal{P})} \langle w, p\rangle \;\ge\; \alpha \inf_{p\in\overline{Co}(\mathcal{P})} \|p\|.$$

Note that the right-hand side of the inequality is positive since, otherwise, there would exist $p_i \in \mathcal{P}$, $\lambda_i \in [0,1]$, $\Sigma\lambda_i = 1$, such that $\sum_i \lambda_i p_i = 0$. Then $\mathbb{R}^N = \bigcup_i (p_i)^+ \subset$

$T_M(x)$, which is in contradiction with Corollary 1.2 of Part I. So we have proved that there is some $p \in \mathcal{P}$ such that $\langle w, p\rangle > 0$, and so w belongs to the interior of $-T_{F(x)}(v)$ for some $v \in \mathcal{R}(x)$. Thus equality (8) is proved.

(2) Since $D_M(x)$ is open and contained in $T_M(x)$, $D_M(x)$ is contained in the interior of $T_M(x)$. So, from equality (8), it remains to prove that

$$\bigcup_{v\in\mathcal{R}(x)} \operatorname{Int}\left(-T_{F(x)}(v)\right) \subset D_M(x).$$

For that purpose, let v belong to $\mathcal{R}(x)$.

From Proposition 1.2, there is some solution $x(\cdot)$ which remains on ∂M and some adjoint $p(\cdot)$ to $x(\cdot)$ on $[0,T]$ such that $\langle p(0), v\rangle = 0$. Let us denote

$$F_p := \{z \in F(x) \mid \langle z, p(0)\rangle = 0\}.$$

Lemma 1.3 states that

$$\operatorname*{Limsup}_{t\to 0^+} \frac{x(t)-x}{t} \subset F_p. \tag{9}$$

Now let w belong to the interior of $-T_{F(x)}(v)$. We have to prove that w belongs to $D_M(x)$. For any $z \in F_p$,

$$\langle z, p(0)\rangle = \inf_{v'\in F(x)} \langle v', p(0)\rangle = 0,$$

and so, Lemma 1.4 states that

$$-T_{F(x)}(z) = (p(0))^- = -T_{F(x)}(v).$$

Since, for any $z \in F_p$,

$$\operatorname{Int}\left(-T_{F(x)}(z)\right) = \bigcup_{\lambda>0} \lambda(z - \operatorname{Int}(F(x))),$$

there is some $\lambda_z > 0$ such that w is contained in the interior of $\lambda_z(z - F(x))$. Since F_p is compact, the λ_z are bounded on F_p by some $\lambda > 0$ and there is some $\alpha > 0$ such that $w + \alpha B$ is contained in $\lambda(z - F(x))$ for any $z \in F_p$. Since we want to prove that $w \in D_M(x)$ and that $D_M(x)$ is a cone, we can assume, without loss of generality, that $\lambda = 1$. So, $z - w + \alpha B$ is contained in $F(x)$ for any $z \in F_p$.

Since F is Lipschitz,

$$\forall y \in x + \frac{\alpha}{2\ell}B,\ \forall z \in F_p, \quad z - w + \frac{\alpha}{2}B \subset F(y).$$

For t sufficiently small (say, $t \le \epsilon$ with $\epsilon > 0$), the reachable set for $-F$ from $x(t)$ at time t is contained in $x + \frac{\alpha}{2\ell}B$. Thus

$$\forall z \in F_p,\ \ x(t) - t\left(z - w + \frac{\alpha}{2}B\right) \subset \mathcal{A}_{-F}(x(t))(t). \tag{10}$$

From (9), there is some $\epsilon' > 0$ such that $\frac{x(t)-x}{t}$ is contained in $F_p + \frac{\alpha}{4}B$ for $t \in]0, \epsilon']$. Thus, for any $t \le \inf\{\epsilon, \epsilon'\}$,

$$\begin{aligned} x + t(w + \tfrac{\alpha}{4}B) &\subset x - t[z_t - \tfrac{x(t)-x}{t} - w + \tfrac{\alpha}{2}B] \\ &= x(t) - t[z_t - w + \tfrac{\alpha}{2}B], \end{aligned} \tag{11}$$

where $z_t \in F_p$ is the projection onto F_p of $\frac{x(t)-x}{t}$. Since $\mathcal{A}_{-F}(x(t))(t)$ is contained in M, combining (10) and (11) yields that $x + t(w + \frac{\alpha}{4}B)$ is contained in M for any $t \in \inf\{\epsilon, \epsilon'\}$. Thus w belongs to $D_M(x)$, which completes the proof of Theorem 1.2. □

1.3. Upper semicontinuity of the contingent cone. We are now ready to state the main theorem of this section.

THEOREM 1.3 (regularity of the contingent cone). *Assume that F satisfies* (1) *and* (3) *and that the boundary of M is semipermeable in a neighborhood of $x \in \partial M$. Then*

$$\operatorname*{Limsup}_{x' \to x,\ x' \in \partial M} T_M(x') \subset T_M(x).$$

In other words, the set-valued map $x \rightsquigarrow T_M(x) \cap B$ from $\partial M \cap O$ to $\mathbb{R}^N$ is upper semicontinuous.

Remarks.

(1) In some problems, we already know that M is convex. Then the set-valued map $x \rightsquigarrow T_M(x)$ is lower semicontinuous with convex values [2]. Thus Theorem 1.3 yields that the set-valued map $x \rightsquigarrow T_M(x)$ is continuous on ∂M, and $T_M(x)$ is always equal to a half-space. This means that ∂M is a $\mathcal{C}^1$ manifold. This is, in particular, the case when

• K is convex and F has a convex graph[2]; then the viability kernel $\mathrm{Viab}_F(K)$ is convex [4]. Thus the boundary of $\mathrm{Viab}_F(K)$, which is semipermeable in the interior of K, is a smooth manifold in the interior of K.

• F has a convex graph and 0 belongs to $\mathrm{Int}(F(x_0))$; then the reachable set from x_0 is convex. Thus its boundary, which is semipermeable, is a smooth manifold.

(2) From Corollary 1.2, $T_{\widehat{M}}(x) = \overline{\mathbb{R}^N \backslash T_M(x)}$ for $x \in \partial M$. Thus the set-valued map $x \rightsquigarrow T_{\widehat{M}}(x)$ is lower semicontinuous; i.e., $\widehat{M}$ is sleek [2].

Proof of Theorem 1.3. Let w_n belong to $T_M(x_n)$, with $x_n \to x$, $w_n \to w$, x_n and x belonging to $\partial M \cap O$. We have to prove that w belongs to $T_M(x)$.

Theorem 1.1 yields the existence of some $v_n \in \mathcal{R}(x_n)$ such that w_n belong to $-T_{F(x_n)}(v_n)$. From Proposition 1.2, for any $n \in \mathbb{N}$, there exists some barrier solution $x_n(\cdot)$ starting from x_n, with associated adjoint denoted by $p_n(\cdot)$ and satisfying $\langle v_n, p_n(0)\rangle = 0$.

From Theorem 5.3.1 in [2], a subsequence of the $x_n(\cdot)$ converges to some barrier solution $x(\cdot)$ starting from x. Moreover, the adjoint maps $p_n(\cdot)$ converge to the adjoint function $p(\cdot)$ of $x(\cdot)$ from Proposition 2.1 of Part I.

Since, $H_F(x_n, p_n(0)) = \langle v_n, p_n(0)\rangle = 0$, Lemma 1.4 states that

$$-T_{F(x_n)}(v_n) = (p_n(0))^-.$$

Thus, for any $n \in \mathbb{N}$, $\langle w_n, p_n(0)\rangle \le 0$, and so $\langle w, p(0)\rangle \le 0$. Moreover, $p(\cdot)$ is the adjoint of the solution $x(\cdot)$, so that Theorem 2.1 of Part I states that

$$(p(0))^- \subset T_M(x(0)).$$

In particular, w belongs to $T_M(x)$. □

[2]This assumption is satisfied in particular if the controlled system is affine in x.

2. Semiconcavity property. The aim of this section is to show that, under suitable assumptions, semipermeable surfaces are locally the graphs of semiconcave functions.

2.1. A new assumption on the controlled system. We now introduce a new assumption:

$$\begin{aligned} &\text{(a)}\quad \text{The values of } F \text{ are convex and have nonempty interior.}\\ &\text{(b)}\quad \exists \gamma > 0 \text{ such that}\\ &\qquad \forall x \in O',\ \forall v \in \partial F(x) \text{ and } \forall w \in T_{F(x)}(v),\\ &\qquad\qquad d_{F(x)}(v+w) \le \gamma \|w\|^2. \end{aligned} \tag{12}$$

It can be proved that condition (12) is stronger than (3). Condition (12) is related to the boundedness of the curvature of the convex sets $F(x)$ for $x \in O'$. It is also related to some property of "uniform strict concavity" of the hamiltonian H. Namely, we have the following proposition.

PROPOSITION 2.1. *Let F satisfy* (1) *and $H : \mathbb{R}^N \times \mathbb{R}^N \to \mathbb{R}$ be the hamiltonian associated with F. The following assertions are equivalent.*

(i) *F satisfies* (12).

(ii)

$$\begin{aligned} &\textit{There is some } \epsilon > 0 \textit{ such that}\\ &\forall x \in O',\ p \to H(x,p) + \epsilon\|p\| \textit{ is concave.} \end{aligned} \tag{13}$$

(iii) *There is some convex set-valued map $F_0 : O' \rightsquigarrow \mathbb{R}^N$ and some $\epsilon > 0$ such that $F(x) = F_0(x) + \epsilon B$.*

(iv) *F has $\mathcal{C}^1$ convex values with a nonempty interior, and the normal map*[3] *$n_x : \partial F(x) \to \partial B_N$ is λ-Lipschitz, with λ independent of $x \in O'$.*

Proposition 2.1 can be proved by using classical arguments of convex analysis, and thus we do not give its proof.

Remark. If F is Lipschitz, then the set-valued map F_0 defined in (iii) is also Lipschitz.

Let us now give an example of a set-valued map satisfying (12).

PROPOSITION 2.2. *Assume that $f : \mathbb{R}^N \to \mathbb{R}^N$ is Lipschitz and that $A : \mathbb{R}^N \to \mathcal{L}(\mathbb{R}^N, \mathbb{R}^N)$ and $A^{-1} : \mathbb{R}^N \to \mathcal{L}(\mathbb{R}^N, \mathbb{R}^N)$ are Lipschitz. Then the set-valued map*

$$F(x) := f(x) + \bigcup_{u \in B} A(x)u$$

satisfies (1) *and* (12) *on bounded sets.*

Proof. The hamiltonian associated with F is

$$H(x,p) = \langle f(x), p\rangle - \|A^*(x)p\|,$$

and it satisfies (13) thanks to the following lemma.

LEMMA 2.1. *Assume that Q is a positive definite quadratic form with largest eigenvalue a and smallest eigenvalue b. Then, for any $\alpha \in [0, b/a^{\frac{1}{2}}]$, the map*

$$p \to (\langle p, Qp\rangle)^{\frac{1}{2}} - \alpha\|p\|$$

is convex. □

[3] That is, the map such that $\|n_x(v)\| = 1$ for any $v \in \partial F(x)$ and $n_x(v)$ is an outward normal to $F(x)$ at v.

2.2. Further regularity.

THEOREM 2.1. *Assume that the boundary of M is semipermeable and that $F : \mathbb{R}^N \rightsquigarrow \mathbb{R}^N$ satisfies* (1) *and* (12). *There is some constant k such that, for any $x \in \partial M \cap O$, for any $w \in T_M(x)$ with $\|w\| \leq T^2$,*

$$d_M(x+w) \leq k\|w\|^{\frac{3}{2}}. \tag{14}$$

The notations O and T are defined in Proposition 1.1 of Part I.

The exponent $\frac{3}{2}$ is rather surprising: one would expect an exponent 2, as for F. We have an example proving that this exponent is optimal at least when F is Lipschitz. For smooth hamiltonians, the problem is still open.

Proof. If the vector w belongs to $T_M(x)$ with $x \in \partial M$, Proposition 1.2 states that there is some barrier solution $x(\cdot) \in \mathcal{S}_F(x)$ which remains in ∂M on $[0,T]$ and some adjoint $p(\cdot)$ of $x(\cdot)$ on $[0,T]$ such that $\langle p(0), w\rangle \leq 0$.

Set $\tau := \|w\|^{\frac{1}{2}}$ and let $w(t)$ be the projection of w onto $T_M(x(t))$. Let us set, for any $t \leq \tau$,

$$\alpha(t) := d_{F(x(t))}\left(x'(t) - \frac{w(t)}{\tau}\right).$$

Let $y(\cdot)$ be defined on $[0,\tau]$ by

$$\begin{cases} y'(t) := -x'(\tau - t) + \frac{1}{\tau} w(\tau - t), \\ y(0) := x(\tau). \end{cases}$$

(Recall that $\|w\| \leq T^2$, so that $\tau \leq T$.) The Filippov theorem yields the existence of a solution $z(\cdot) \in \mathcal{S}_{-F}(x(\tau))$ such that

$$\forall t \in [0,\tau],\ \|z(t) - y(t)\| \leq e^{\ell t} \int_0^t \alpha(s)ds. \tag{15}$$

Let us prove that

$$\int_0^\tau \alpha(s)ds \leq \gamma\|w\|^{\frac{3}{2}} \tag{16}$$

and that

$$\|y(\tau) - (x+w)\| \leq C\|w\|^{\frac{3}{2}}. \tag{17}$$

Proof of (16). From Corollary 2.1 of Part I, for almost every $t \in]0,T[$, $T_M(x(t)) = -T_{F(x(t))}(x'(t))$. From its very definition, $w(t)$ belongs to $T_M(x(t))$, so $-w(t)$ belongs to $T_{F(x(t))}(x'(t))$. Thus assumption (12) yields that

$$\alpha(t) := d_{F(x(t))}\left(x'(t) - \frac{w(t)}{\tau}\right) \leq \gamma\frac{\|w(t)\|^2}{\tau^2} \leq \gamma\|w\|$$

(note that $\|w(t)\| \leq \|w\|$ because $w(t)$ is the projection of w onto a half-space). After integration, we obtain (16).

Proof of (17). From Corollary 2.1 of Part I, for almost every $t \in]0,T[$, $T_M(x(t)) = (p(t))^-$. Thus

$$w(t) = w - \langle p(t), w\rangle_+ p(t),$$

where x_+ denotes $\sup\{0, x\}$. Recall that $p(\cdot)$ is $2C$-Lipschitz. (See Theorem 2.1 of Part I.) Thus

$$\langle w, p(t)\rangle_+ \leq [\langle w, p(0)\rangle + 2Ct\|w\|]_+ \leq 2Ct\|w\|,$$

which implies

$$\|w(t) - w\| \leq 2Ct\|w\|.$$

After integration, we obtain

$$\left\| \int_0^\tau w(\tau - s)ds - \tau w \right\| \leq C\|w\|\tau^2.$$

Since $y(\tau) := x + \frac{1}{\tau}\int_0^\tau w(\tau - s)ds$, we have obtained (17).

Combining (15), (16), and (17), we conclude that

$$\|z(\tau) - (x + w)\| \leq (C + e^{\ell T^2}\gamma)\|w\|^{\frac{3}{2}}.$$

Let us now recall that $z(\cdot)$ is a solution of the differential inclusion for $-F$ starting from $x(\tau)$ which belongs to $M \cap O'$. Thus $z(t)$ belongs to M for any $t \leq \tau$ and

$$d_M(x + w) \leq \|z(\tau) - (x + w)\| \leq k\|w\|^{\frac{3}{2}},$$

which is the desired conclusion. □

2.3. A criteria of semiconcavity. We show here that the result described in Theorem 2.1 is related to semiconcavity of a function. Let us first recall what a semiconcave function is [8].

DEFINITION 2.1. *Let Ω be a open convex subset of $\mathbb{R}^N$ and $\phi : \Omega \to \mathbb{R}$. The function ϕ is semiconcave on Ω if, $\forall x, y \in \Omega$, $\forall \lambda \in [0, 1]$,*

$$\lambda\phi(x) + (1 - \lambda)\phi(y) \leq \phi\left(\lambda x + (1 - \lambda)y\right) + \lambda(1 - \lambda)\omega(\|y - x\|),$$

where $\omega : \mathbb{R}^+ \to \mathbb{R}^+$ is the modulus of semiconcavity, $\omega(t) \to 0^+$ if $t \to 0^+$.

We now provide a criterion for semiconcavity.

PROPOSITION 2.3. *Let Ω be a convex open subset of $\mathbb{R}^N$ and $\phi : \Omega \to \mathbb{R}$. Assume that ϕ satisfies the following conditions:*

(i) *ϕ is ρ-Lipschitz;*
(ii) *$\forall x \in \Omega$, $-T_{Epi(\phi)}(x, \phi(x)) \bigcup T_{Epi(\phi)}(x, \phi(x)) = \mathbb{R}^{N+1}$;*
(iii) *There exists a constant k such that, $\forall x \in \Omega$, $\forall (v_x, v_t) \in T_{Epi(\phi)}(x, \phi(x))$ such that $x + v_x \in \Omega, d_{Epi(\phi)}\left((x, \phi(x)) + (v_x, v_t)\right) \leq k\|(v_x, v_t)\|^{\frac{3}{2}}$.*

Then ϕ is semiconcave with a modulus $\omega(t) := k(\rho^2 + 1)^2 t^{\frac{3}{2}}$.

Let us point out that condition (ii) is fulfilled in particular if $T_{Epi(\phi)}(x, \phi(x))$ is a union of half-spaces, as in Theorem 1.1. Condition (iii) is exactly the same as (14).

Proof of Proposition 2.3. *First step.* We claim that condition (ii) implies that, if (v_x, v_t) belongs to $\partial T_{Epi(\phi)}(x, \phi(x))$, then $-(v_x, v_t)$ belongs to $T_{Epi(\phi)}(x, \phi(x))$. Indeed, since (v_x, v_t) belongs to $\partial T_{Epi(\phi)}(x, \phi(x))$, there exists a sequence $(v_x^n, v_t^n) \notin T_{Epi(\phi)}(x, \phi(x))$ converging to (v_x, v_t). From assumption (ii), $-(v_x^n, v_t^n)$ belongs to $T_{Epi(\phi)}(x, \phi(x))$ for any n. Letting $n \to +\infty$ yields $-(v_x, v_t) \in T_{Epi(\phi)}(x, \phi(x))$, which is the desired conclusion.

Second step. We now prove that, if (v_x, v_t) belongs to $T_{Epi(\phi)}(z, \phi(z))$, then, for any $h \in [0,1]$ such that $z + hv_x \in \Omega$, one has

$$\phi(z + hv_x) - (\phi(z) + hv_t) \leq k(\rho^2+1)^{\frac{1}{2}} \left(h\|(v_x, v_t)\|\right)^{\frac{3}{2}}.$$

Let (y_h, ρ_h) belong to the projection of $(z + hv_x, \phi(z) + hv_t)$ onto $\mathrm{Epi}(\phi)$. Then

$$\begin{aligned}
&\phi(z + hv_x) - (\phi(z) + hv_t)\\
&\quad = (\phi(z + hv_x) - \phi(y_h)) + (\phi(y_h) - \rho_h) + (\rho_h - (\phi(z) + hv_t))\\
&\quad \leq \rho\|y_h - (z + hv_x)\| + 0 + |\rho_h - (\phi(z) + hv_t)|\\
&\quad \leq (\rho^2+1)^{\frac{1}{2}} \|(y_h, \rho_h) - (z + hv_x, \phi(z) + hv_t)\|\\
&\quad \leq k(\rho^2+1)^{\frac{1}{2}} (h\|(v_x, v_t)\|)^{\frac{3}{2}}
\end{aligned}$$

from assumption (iii).

Third step. Let x and y belong to Ω and $\lambda \in (0,1)$. Set $z := \lambda x + (1-\lambda)y$. Let us define τ by

$$\tau := \min\{t \mid (y - x, t) \in T_{Epi(\phi)}(z, \phi(z))\}.$$

Note that $\tau > -\infty$ because ϕ is ρ-Lipschitz. In fact, it is easily seen that $|\tau| \leq \rho\|y - x\|$. Note also that $(y - x, \tau)$ belongs to the boundary of $T_{Epi(\phi)}(z, \phi(z))$. Thus, from the first step, $-(y - x, \tau)$ belongs to $T_{Epi(\phi)}(z, \phi(z))$. Applying the second step with $v_x := -(y - x)$, $v_t := -\tau$, and $h := 1 - \lambda$ yields

$$\phi(x) - (\phi(z) - (1-\lambda)\tau) \leq k(\rho^2+1)^{\frac{1}{2}} ((1-\lambda)\|(y - x, \tau)\|)^{\frac{3}{2}}, \tag{18}$$

while applying the second step with $h := \lambda$ and $(v_x, v_t) := (y - x, \tau)$ yields

$$\phi(y) - (\phi(z) + \lambda\tau) \leq k(\rho^2+1)^{\frac{1}{2}} (\lambda\|(y - x, \tau)\|)^{\frac{3}{2}}. \tag{19}$$

Summing (18) multiplied by λ and (19) multiplied by $(1 - \lambda)$ gives

$$\begin{aligned}
&\lambda\phi(x) + (1-\lambda)\phi(y) - \phi(\lambda x + (1-\lambda)y)\\
&\leq k(\rho^2+1)^{\frac{1}{2}} [\lambda((1-\lambda)\|(y - x, \tau)\|)^{\frac{3}{2}} + (1-\lambda)(\lambda\|(y - x, \tau)\|)^{\frac{3}{2}}]\\
&\leq k(\rho^2+1)^{\frac{1}{2}} \lambda(1-\lambda)\|(y - x, \tau)\|^{\frac{3}{2}}.
\end{aligned}$$

Since $|\tau| \leq \rho\|y - x\|$, $\|(y - x, \tau)\| \leq (\rho^2+1)^{\frac{1}{2}}\|y - x\|$. Thus we have finally proved that

$$\begin{aligned}
&\lambda\phi(x) + (1-\lambda)\phi(y) - \phi(\lambda x + (1-\lambda)y)\\
&\qquad \leq k(\rho^2+1)^2 \lambda(1-\lambda)\|y - x\|^{\frac{3}{2}},
\end{aligned}$$

which is the desired conclusion. □

COROLLARY 2.1. *Suppose that the assumptions of Theorem* 2.1 *are fulfilled. Assume that the boundary of M is semipermeable. Then M is locally the epigraph of a semiconcave function with a modulus of semiconcavity of the form $\omega(t) := kt^{\frac{3}{2}}$.*

Semiconcave functions enjoy nice regularity properties. For instance, singularities of semiconcave functions (i.e., the points at which this function is not differentiable) propagate [1].

Corollary 2.1 is an application of Proposition 1.2 of Part I, Theorem 2.1, and Proposition 2.3.

3. Regularity of the optimal exit-time map. An important application of the previous results is the study of the optimal exit-time problem. For an open set Ω, we study the map

$$\theta_\Omega(x) := \inf\{t \geq 0 \mid \exists x(\cdot) \in \mathcal{S}_F(x) \text{ with } x(t) \notin \Omega\}. \tag{20}$$

We set $\theta_\Omega(x) := +\infty$ if no solution reaches the complement of Ω. We refer to the introduction of the first part for bibliographical comments on the regularity of θ_Ω. In Cardaliaguet, Quincampoix, and Saint-Pierre [5], it is proved that the epigraph of θ_Ω, denoted by $\mathrm{Epi}(\theta_\Omega)$, is the viability kernel of $\mathcal{H} := \overline{\Omega} \times \mathbb{R}^+$ for the set-valued map $\Phi : \mathbb{R}^{N+1} \rightsquigarrow \mathbb{R}^{N+1}$ defined by

$$\forall (x,t) \in \mathbb{R}^{N+1},\ \Phi(x,t) := \begin{cases} F(x) \times \{-1\} & \text{if } x \in \Omega, \\ \overline{Co}\,(\{0\} \times \{0\} \bigcup F(x) \times \{-1\}) & \text{otherwise.} \end{cases}$$

The set-valued map Φ obviously satisfies condition (1) on $\Omega \times \mathbb{R}$.

Since, for any $x \in \Omega$, $\theta_\Omega(x) > 0$, $(x, \theta_\Omega(x))$ belongs to the boundary of $\mathrm{Viab}_\Phi(\mathcal{H})$ and to the interior of $\mathcal{H}$. Thus the Quincampoix theorem [9] states that the boundary of $\mathrm{Epi}(\theta_\Omega)$ is semipermeable for Φ in a neighborhood of any point $(x, \phi(x))$ for $x \in \Omega$.

Unfortunately, the set-valued map Φ never satisfies condition (3) (because $\Phi(x)$ always has an empty interior), and we cannot straightforwardly apply the regularity results previously obtained. For this reason, we first prove the existence of a set-valued map Ψ satisfying (12) (provided that F satisfies (12)) such that the boundary of $\mathrm{Viab}_\Phi(\mathcal{H})$ is still semipermeable for Ψ. Then we apply Theorems 1.3 and 2.1 to the particular case of the map θ_Ω. We complete this paper by showing that θ_Ω is smooth along some optimal trajectories.

Notation. Below, the variable in $\mathbb{R}^{N+1}$ is divided into (x,t), with x belonging to the state space $\mathbb{R}^N$ and t to the time space $\mathbb{R}$.

3.1. Reduction of the problem. The hamiltonian associated with map Φ is

$$H_\Phi(x,t,p_x,p_t) := -p_t + H(x,p_x),$$

where H is the hamiltonian associated with F.

PROPOSITION 3.1. *Assume that the set-valued map $F : \mathbb{R}^N \rightsquigarrow \mathbb{R}^N$ satisfies* (1) *and* (12). *Then there is a set-valued map $\Psi : \mathbb{R}^{N+1} \rightsquigarrow \mathbb{R}^{N+1}$ satisfying* (1) *and* (12) *such that*

$$\begin{aligned} &\forall (x,t) \in \mathbb{R}^{N+1},\ \forall (p_x,p_t) \in \mathbb{R}^{N+1}, \\ &\quad H_\Psi(x,t,p_x,p_t) = 0 \iff H_\Phi(x,t,p_x,p_t) = 0, \end{aligned} \tag{21}$$

and moreover, $\Phi(x,t) \subset \Psi(x,t)$ for any $(x,t) \in \mathbb{R}^{N+1}$.

In that case, Proposition 1.1 of Part I states that it is equivalent for a closed set M to enjoy the semipermeability property for Φ and to enjoy the semipermeability property for Ψ.

Proof of Proposition 3.1. *Construction of* Ψ. From Proposition 2.1, there is some set-valued map $F_0 : \mathbb{R}^N \rightsquigarrow \mathbb{R}^N$ such that $F(x) = F_0(x) + \epsilon B$ and has associated hamiltonian $h(x,p) := H(x,p) + \epsilon\|p\|$. Since F is Lipschitz, F_0 is Lipschitz.

Let us define Ψ by

$$\Psi(x,t) := \overline{Co}\left(\bigcup_{\|u_x\|^2 + |u_t|^2 \leq 2} (2 + u_t)[F_0(x) \times \{-1\}] + \epsilon\{u_x\} \times \{0\} \right).$$

The set-valued map Ψ satisfies the regularity condition (1). Setting $u_t = -1$, it is easily seen that $\Phi(x,t) \subset \Psi(x,t)$. So it remains to show that the hamiltonian H_Ψ of Ψ satisfies (13) and (21).

H_Ψ satisfies (13) *and* (21). A little computation gives

$$H_\Psi(x,t,p_x,p_t) = \min_{v\in F_0(x)} \left[-2p_t + 2\langle v,p_x\rangle - \sqrt{2}\left((-p_t+\langle v,p_x\rangle)^2 + \epsilon^2\|p_x\|^2\right)^{\frac12}\right].$$

Note that equivalence (21) is fulfilled. The eigenvalues of the quadratic form Q_v defined by $(p_x,p_t) \to (-p_t+\langle v,p_x\rangle)^2 + \epsilon^2\|p_x\|^2$ are

$$\begin{cases} \lambda_1 = \epsilon^2 & \text{(eigenvector } (w,0) \text{ where } w \perp v\text{)},\\ \lambda_2 = \frac{\epsilon^2+\|v\|^2+1+\left((\epsilon^2+\|v\|^2+1)^2-4\epsilon^2\right)^{\frac12}}{2} & \text{(eigenvector } (v,\epsilon^2+\|v\|^2-\lambda_2)\text{)},\\ \lambda_3 = \frac{\epsilon^2+\|v\|^2+1-\left((\epsilon^2+\|v\|^2+1)^2-4\epsilon^2\right)^{\frac12}}{2} & \text{(eigenvector } (v,\epsilon^2+\|v\|^2-\lambda_3)\text{)}. \end{cases}$$

Thus Q_v is positive definite and Lemma 2.1 states that, for

$$\alpha := \min_{x\in O'} \min_{v\in F_0(x)} \frac{\epsilon^2}{(\epsilon^2+\|v\|^2+1)^{\frac32}},$$

the map $(p_x,p_t) \to \left[(-p_t+\langle v,p_x\rangle)^2 + \epsilon^2\|p_x\|^2\right]^{\frac12} - \alpha\|(p_x,p_t)\|$ is convex. So $(p_x,p_t) \to H_\Psi(p_x,p_t) + \alpha\sqrt{2}\|(p_x,p_t)\|$ is concave, and Proposition 3.1 is proved since $\alpha > 0$. □

3.2. The regularity results. We now apply Theorems 1.3 and 2.1 in the case when M is the epigraph of θ_Ω.

THEOREM 3.1 (regularity of θ_Ω). *Let $F : \mathbb{R}^N \rightsquigarrow \mathbb{R}^N$ be a set-valued map satisfying* (1) *and* (12). *Then*

(i) *θ_Ω is Lipschitz continuous in a open subset $\mathcal{L}$ of $\mathbb{R}^N$ which is dense in* $\mathrm{dom}(\theta_\Omega)$. *Moreover, θ_Ω is also semiconcave in $\mathcal{L}$ with a modulus of semiconcavity of the form $\omega(t) = kt^{\frac32}$.*

(ii) *A point x belongs to* $\mathrm{dom}(\theta_\Omega)\backslash\mathcal{L}$ *if and only if*

$$\liminf_{x'\to x} \frac{\theta_\Omega(x') - \theta_\Omega(x)}{\|x'-x\|} = -\infty.$$

Remarks.

(1) We do not assume in Theorem 3.1 that the set Ω is smooth or that the hamiltonian is differentiable. In this situation, assumption (12) is crucial. For instance, for the dynamic $F(x,y) := [-1,1]\times[-1,1]$ (which clearly does not satisfy (12)), there are open sets Ω such that the optimal exit-time θ_Ω for F is not continuous in any open subset of Ω.

(2) Theorem 3.1 still holds partially true for constrained optimal exit-time problems (for the formulation of this problem, see [5]). Then $\mathcal{L}$ is an open dense subset of $\mathrm{dom}(\theta_\Omega) \cap \mathrm{Int}(K)$, where K is the constraint and θ_Ω is (locally) Lipschitz and semiconcave in $\mathcal{L}$.

(3) In the proof of Theorem 3.1, we never need the fact that the function we are studying is the optimal exit-time function. So our results are still valid in any case when the boundary of the epigraph of the lower semicontinuous function is semipermeable for a set-valued map Ψ satisfying (1) and (12).

Before beginning the proof of Theorem 3.1, let us recall some basic facts concerning the epigraph of a lower semicontinuous function.

LEMMA 3.1. *Let $\phi : \mathbb{R}^N \to \mathbb{R} \cup \{+\infty\}$ be a lower semicontinuous function. We set* $\mathrm{dom}(\phi) := \{x \mid \phi(x) < +\infty\}$ *and* $M := \mathrm{Epi}(\phi)$. *Let $x \in \mathrm{dom}(\phi)$.*

(α) If (v_x, v_t) belongs to $D_M(x, \phi(x))$, then v_x belongs to $D_{dom(\phi)}(x)$.

(β) If ϕ is not continuous at x, then there is some $t > 0$ such that $(x, \phi(x) + t)$ belongs to $\partial\mathrm{Epi}(\phi)$.

(γ) If x belongs to $\mathrm{dom}(\phi)$, *then, for any $t > 0$, the vector $(0, -1)$ belongs to $T_{Epi(\phi)}(x, \phi(x) + t)$.*

Proof of Theorem 3.1. Thanks to the construction of Proposition 3.1, $M := \mathrm{Epi}(\theta_\Omega)$ satisfies the conclusion of Theorems 1.3 and 2.1.

Set

$$A := \{x \in \mathrm{dom}(\theta_\Omega) \mid (0, -1) \in T_M(x, \theta_\Omega(x))\}$$

and $\mathcal{L} := \mathrm{dom}(\theta_\Omega) \backslash A$.

Let us show that A is closed in $\mathrm{dom}(\theta_\Omega)$ and has an empty interior, and that $\mathcal{L}$ is open. Then we prove that θ_Ω is locally Lipschitz on $\mathcal{L}$. Note that, from the very definition of the contingent cone, (b) is satisfied.

• *A is closed.* Let (x_n) be a sequence of A converging to $x \in \mathrm{dom}(\theta_\Omega)$. We have to prove that x belongs to A. There are two cases.

(1) Either

$$\liminf_n \theta_\Omega(x_n) = \theta_\Omega(x).$$

Then a subsequence $(x_{n'}, \theta_\Omega(x_{n'}))$ converges to $(x, \theta_\Omega(x))$ and so

$$\mathrm{Limsup}\, T_M(x_{n'}, \theta_\Omega(x_{n'})) \subset T_M(x, \theta_\Omega(x))$$

from Theorem 1.3. In particular, $(0, -1)$ belongs to $T_M(x, \theta_\Omega(x_n))$ and thus belongs to $T_M(x, \theta_\Omega(x))$. So x belongs to A.

(2) Or θ_Ω is not continuous at x. Then Lemma 3.1 (β) states that there is some $\tau > 0$ such that $(x, \tau + \theta_\Omega(x))$ belongs to ∂M. So, for any $t \in (0, \tau]$, the point $(x, \theta_\Omega(x) + t)$ belongs to the boundary of M and $(0, -1)$ belongs to $T_M(x, \theta_\Omega(x) + t)$ (Lemma 3.1 (γ)). Letting $t \to \theta_\Omega(x)$ means that $(0, -1)$ belongs to $T_M(x, \theta_\Omega(x))$.

So we have proved that, in both cases, x belongs to A. Thus A is closed in $\mathrm{dom}(\theta_\Omega)$.

• *A has an empty interior.* Assume to the contrary that $x + rB$ is contained in A. Then, if (y, ρ) belongs to $[(x + rB) \times \mathbb{R}] \cap M$, $(0, -1)$ belongs to $T_M(y, \rho)$ because either $\rho = \theta_\Omega(y)$ (and so it is true from the very definition of A) or $\rho > \theta_\Omega(y)$ (and then it is true from Lemma 3.1). So $[(x + r\overset{o}{B}) \times \mathbb{R}] \cap M$ is a locally compact viability domain for the map $(y, \rho) \to (0, -1)$. The viability theorem states that there is some $\tau > 0$ such that $(x, \theta_\Omega(x) - t)$ remains in M on $[0, \tau]$. This is impossible, because M is the epigraph of θ_Ω.

• *$\mathcal{L} := dom(\theta_\Omega) \backslash A$ is open.* Let x belong to $\mathrm{dom}(\theta_\Omega)$, but not to A. From Theorem 1.1, $T_M(x, \theta_\Omega(x))$ is a union of half-spaces, so that there is some $(p_x, p_t) \in \mathbb{R}^{N+1}$ such that $(p_x, p_t)^- \subset T_M(x, \theta_\Omega(x))$. Since $(0, -1)$ does not belong to $T_M(x, \theta_\Omega(x))$, $p_t < 0$. Moreover, Theorem 1.2 states that

$$\mathrm{Int}[(p_x, p_t)]^- \subset D_M(x, \theta_\Omega(x)).$$

Since, for any $v \in \mathbb{R}^N$, one can find some ρ such that $\langle p_x, v\rangle + p_t\rho < 0$ because $p_t \neq 0$, v belongs to $D_{dom(\theta_\Omega)}(x)$ from Lemma 3.1 (α). Thus $\mathbb{R}^N\backslash\{0\}$ is contained in $D_{dom(\theta_\Omega)}(x)$, which means that x belongs to the interior of $\mathrm{dom}(\theta_\Omega)$.

• *θ_Ω is locally Lipschitz on $\mathcal{L}$.* Let $x + rB$ be contained in $\mathcal{L}$. In a first step, we prove that θ_Ω is continuous on $x + rB$. Then we prove that there is some $\gamma > 0$ such that

$$\forall y \in x + rB,\ \forall v \in \mathbb{R}^N,\ \left(\{v\} \times \left[-\frac{\|v\|}{\gamma}, \frac{\|v\|}{\gamma}\right]\right) \cap T_{\partial M}(x, \phi(x)) \neq \emptyset. \tag{22}$$

We finally show that θ_Ω is $1/\gamma$-Lipschitz on $x + rB$.

First step. From Lemma 3.1, if θ_Ω is not continuous at a point $y \in \mathrm{dom}(\theta_\Omega)$, there is some $\tau > 0$ such that $(y, \theta_\Omega(y)+\tau)$ belongs to ∂M. From Lemma 3.1 again, $(0, -1)$ belongs then to $T_M(y, \theta_\Omega(y) + \tau)$. With the same arguments as we used previously, we can show that $(0, -1)$ belongs to $T_M(x, \theta_\Omega(x))$, which contradicts $y \in \mathcal{L}$. So θ_Ω is continuous on $x + rB$.

Second step. The graph of the restriction of θ_Ω to $x+rB$ is compact because θ_Ω is continuous on $x + rB$. From Theorem 1.3, the set-valued map $T_M(\cdot)$ on the compact $\mathrm{Graph}(\theta_\Omega)$ restricted to $x + rB$ has a closed graph. Let $\epsilon > 0$ be the distance from $(0, -1)$ to this closed graph. Set $\gamma := \epsilon/2$.

Let $y \in x + rB$ and $v \in \mathbb{R}^N$. Since

$$\left(v, -\frac{\|v\|}{\gamma}\right) \in \bigcup_{\lambda\geq 0} \lambda\left((0, -1) + \epsilon\, \overset{o}{B}_{N+1}\right),$$

$(v, -\frac{\|v\|}{\gamma})$ does not belong to $T_M(y, \theta_\Omega(y))$. From Theorem 1.1, $T_M(y, \theta_\Omega(y))$ is a union of half-spaces. Thus $(-v, \frac{\|v\|}{\gamma})$ belongs to $T_M(y, \phi(y))$. So we have obtained, $\forall w \in \mathbb{R}^N$,

$$\left(w, \frac{\|w\|}{\gamma}\right) \in T_M(y, \theta_\Omega(y)) \quad \text{and} \quad \left(w, -\frac{\|w\|}{\gamma}\right) \notin T_M(y, \theta_\Omega(y)).$$

Then the Quincampoix lemma [9] gives (22).

Third step. Thanks to (22), the set $\partial M \cap ((x+r\ \overset{o}{B}) \times \mathbb{R})$ is a locally viable domain for the constant set-valued map G_v defined by $y \rightsquigarrow \{v\} \times [-\|v\|\gamma^{-1}, \|v\|\gamma^{-1}]$ and for any $v \in \mathbb{R}^N$.

Let y and z belong to $x + rB$. For $v := z - y$, the viability theorem states that there is a solution $(y(\cdot), \rho(\cdot))$ of the differential inclusion for G_v starting from $(y, \theta_\Omega(y))$ which remains in ∂M until it leaves $(x + rB) \times \mathbb{R}$. For $t = 1$ the solution still satisfies $(y(1), \rho(1)) \in \partial M$. Thus $\rho(1)$ is equal to $\theta_\Omega(z)$. From the very definition of the set-valued map G_v, $|\rho(1) - \rho(0)| \leq \frac{\|z-y\|}{\gamma}$, i.e.,

$$|\theta_\Omega(z) - \theta_\Omega(y)| \leq \frac{\|z - y\|}{\gamma}.$$

So θ_Ω is Lipschitz. Thanks to Proposition 2.3, θ_Ω is semiconcave with a modulus of semiconcavity of the form $\omega(t) = kt^{\frac{3}{2}}$. This completes the proof of Theorem 3.1. □

3.3. Regularity along optimal trajectories. With any initial position $x \in \Omega$, one can associate at least one optimal solution $x(\cdot) \in \mathcal{S}_F(x)$ with the optimal exit-time problem, i.e., a solution such that

$$\theta_\Omega(x) = \inf\{t \geq 0 \mid x(t) \notin \Omega\}.$$

The aim of this section is to show that, for any optimal trajectory $x(\cdot)$, either the optimal exit-time θ_Ω is differentiable at $x(t)$ for any $t \in]0, \theta_\Omega(x(0))[$ or it is not differentiable at any point $x(t)$ for any $t \in]0, \theta_\Omega(x(0))[$. A similar phenomenon is observed in [6] for the Bolza problem, where any optimal solution enters immediately into the differentiability domain of the (Lipschitz continuous) value function.

Throughout this section, we assume that the set-valued map F satisfies (1) and (12).

LEMMA 3.2. *If a solution $x(\cdot) \in \mathcal{S}_F(x)$ is optimal for the optimal exit-time problem, then the map $(x(\cdot), \theta_\Omega(x(\cdot)))$ is a barrier solution for the set-valued map Ψ and the closed set* $\mathrm{Epi}(\theta_\Omega)$.

Proof. Note first that $\theta_\Omega(x(t)) = \theta_\Omega(x) - t$. Thus $(x(\cdot)), \theta_\Omega(x(\cdot)))$ is a solution for Φ and so for Ψ (because $\Phi(x,t) \subset \Psi(x,t)$ from Proposition 3.1). This solution remains on the boundary of $\mathrm{Epi}(\theta_\Omega)$ from the very definition of the epigraph of a function. Thus this solution is a barrier solution. ∎

PROPOSITION 3.2. *Let $x \in \Omega$ and $x(\cdot) \in \mathcal{S}_F(x)$ be an optimal solution of the optimal exit-time problem.*

If x belongs to $\mathcal{L}$, then $x(t)$ belongs to $\mathcal{L}$ for any $t \in]0, \theta_\Omega(x)[$.

If x does not belong to $\mathcal{L}$ and if the adjoint $(p_x(\cdot), p_t(\cdot))$ of the barrier solution $(x(\cdot), \theta_\Omega(x(\cdot)))$ satisfies $p_t(0) = 0$, then $x(t)$ does not belong to $\mathcal{L}$ for $t \in [0, \theta_\Omega(x)[$.

COROLLARY 3.1. *Let $x \in \Omega$ and $x(\cdot) \in \mathcal{S}_F(x)$ be an optimal solution of the optimal exit-time problem.*

- *If x belongs to $\mathcal{L}$, then θ_Ω is differentiable at $x(t)$ for $t \in]0, \theta_\Omega(x)[$.*
- *If θ_Ω is not differentiable at some point $x(t)$ with $t \in]0, \theta_\Omega(x)[$, then θ_Ω is not differentiable at any point $x(s)$ for $s \in [0, \theta_\Omega(x)[$. Moreover, $x(s)$ does not belong to $\mathcal{L}$ for $s \in]0, \theta_\Omega(x)[$.*

Proof of Corollary 3.1. We set $M := \mathrm{Epi}(\theta_\Omega)$. From Proposition 3.2, $x(t)$ belongs to $\mathcal{L}$ for $t \in [0, \theta_\Omega(x)[$. From Lemma 3.2, $(x(\cdot), \theta_\Omega(x(\cdot)))$ is a barrier solution. Thus Theorem 2.1 of Part I states that there is a function $(p_x(\cdot), p_t(\cdot))$ adjoint of $(x(\cdot), \theta_\Omega(x(\cdot)))$ on $]0, \theta_\Omega(x)[$. Moreover,

$$T_{Epi(\theta_\Omega)}(x(t), \theta_\Omega(x(t))) = (p_x(t), p_t(t))^-.$$

Since $x(t)$ belongs to $\mathcal{L}$ for $t \in [0, \theta_\Omega(x)[$, $p_t(t) \neq 0$ on $[0, \theta_\Omega(x)[$. In particular,

$$\nabla \theta_\Omega(x(t)) = -\frac{p_x(t)}{p_t(t)},$$

and θ_Ω is derivable on $[0, \theta_\Omega(x)[$.

For proving the second point, note first that $x(t) \notin \mathcal{L}$. Indeed, since $t > 0$, $T_M(x(t), \theta_\Omega(x(t)))$ is a half-space (Theorem 2.1 of Part I). Thus θ_Ω is derivable at $x(t)$ unless the half-space $T_M(x(t), \theta_\Omega(x(t)))$ is vertical, i.e., $x(t) \notin \mathcal{L}$.

If $0 \leq s < t$, then $x(s)$ does not belong to $\mathcal{L}$ from Proposition 3.2, and thus θ_Ω is not derivable at $x(s)$. If we denote by $(p_x(\cdot), p_t(\cdot))$ the adjoint of $(x(\cdot), \theta_\Omega(x(\cdot)))$, Theorem 2.1 of Part I states that

$$\forall s \in]0, \theta_\Omega(x)[,\ T_M(x(s), \theta_\Omega(x(s))) = (p_x(\cdot), p_t(\cdot))^-.$$

Since $x(t) \notin \mathcal{L}$, $p_t(t) = 0$. Thus Proposition 3.2 states that $x(s) \notin \mathcal{L}$ for $s \in [t, \theta_\Omega(x)[$. This completes the proof of the corollary. ∎

Proof of Proposition 3.2.

- Let us assume that $x(t)$ does not belong to $\mathcal{L}$ for some $t \in]0, \theta_\Omega(x)[$. Thus $(0, -1)$ belongs to $T_M(x(t), \theta_\Omega(x(t)))$ and there are sequences $h_n \to 0^+$, $v_n \to 0$, and

$\tau_n \to -1$ such that, for any n, $(x(t)+h_nv_n, \theta_\Omega(x(t))+h_n\tau_n)$ belongs to M, i.e.,

$$\theta_\Omega(x(t)+h_nv_n) \le \theta_\Omega(x(t))+h_n\tau_n.$$

From Filippov's theorem applied to $-F$, there is a sequence of solutions $y_n(\cdot) \in \mathcal{S}_{-F}(x+h_nv_n)$, such that

$$\forall s \in [0,t],\ \|y_n(s)-x(t-s)\|_{[0,t]} \le e^{\ell t}h_n\|v_n\|.$$

Since $y_n(\cdot)$ are solutions to the differential inclusion for $-F$, $(y_n(\cdot), \theta_\Omega(x(t))+h_n\tau_n+\cdot)$ is a solution for $-\Phi$ and so for $-\Psi$ starting from M and remaining in M from semipermeability of M. Let us recall that $\theta_\Omega(x(t)) = \theta_\Omega(x)-t$. Thus

$$(y_n(t), \theta_\Omega(x)+h_n\tau_n) \in M.$$

Since $\|y_n(t)-x\| \le h_n\|v_n\|$ from the very definition of $y_n(\cdot)$, we have finally proved that $(0,-1)$ belongs to $T_M(x,\theta_\Omega(x))$.

• Let us now assume that x does not belong to $\mathcal{L}$, that $x(\cdot)$ is an optimal solution starting from x, and that the adjoint $(p_x(\cdot),p_t(\cdot))$ of $(x(\cdot),\theta_\Omega(x(\cdot)))$ satisfies $p_t(0)=0$.

In a first step, we show that, if $0<s<t<\theta_\Omega(x)$ and if (v,τ) belongs to $T_{\widehat{M}}(x(s),\theta_\Omega(x(s)))$, then there is some w with $(w,\tau) \in T_{\widehat{M}}(x(t),\theta_\Omega(x(t)))$ and $\|w-v\| \le e^{\ell t}\|v\|$.

Since (v,τ) belongs to $T_{\widehat{M}}(x(s),\theta_\Omega(x(s)))$, there are sequences $h_n \to 0^+$, $v_n \to v$, and $\tau_n \to \tau$ such that $(x(s)+h_nv_n, \theta_\Omega(x(s))+h_n\tau_n)$ belongs to $\widehat{M}$. From Filippov's theorem, there are solutions $x_n(\cdot)$ of the differential inclusion for F starting from $x(s)+h_nv_n$ at time s such that

$$\|x_n(\cdot)-x(\cdot)\|_{[s,t]} \le e^{\ell t}h_n\|v_n\|.$$

Then $\frac{x_n(t)-x(t)}{h_n}$ converges, up to a subsequence, to some w with $\|w-v\| \le e^{\ell t}\|v\|$. Moreover, the solutions $(x_n(\cdot), \theta_\Omega(x(s))+h_n\tau_n-\cdot)$ start from $\widehat{M}$, so that, for semipermeability, $(x_n(t), \theta_\Omega(x(t))+h_n\tau_n)$ belongs to $\widehat{M}$. So (w,τ) belongs to $T_{\widehat{M}}(x(t), \theta_\Omega(x(t)))$ with $\|w-v\| \le e^{\ell t}\|v\|$, and our claim is proved.

Let us now prove that, for any $t \in]0,\theta_\Omega(x)[$, $p_t(t)=0$. Recall that, from Theorem 2.1 of Part I, the adjoint $(p_x(\cdot),p_t(\cdot))$ is Lipschitz. In particular, $p_t(s) \to 0$ when $s \to 0^+$. Fix $t \in]0,\theta_\Omega(x)[$ and let $0<s<t$. Let (v_s,τ_s) be the projection of $(0,1)$ onto $T_{\widehat{M}}(x(s),\theta_\Omega(x(s))) = (p_x(s),p_t(s))^+$. Note that (v_s,τ_s) converges to $(0,1)$ when $s \to 0^+$ since $(p_x(s),p_t(s)) \to (p_x(0),0)$. We previously proved that there is some w_s such that (w_s,τ_s) belongs to $T_{\widehat{M}}(x(t),\theta_\Omega(x(t)))$ and $\|w_s-v_s\| \le e^{\ell t}\|v_s\|$. Since $v_s \to 0$, $w_s \to 0$ when $s \to 0^+$. The contingent cone $T_{\widehat{M}}(x(t),\theta_\Omega(x(t)))$ is closed, so that $(0,1)$, which is the limit of (w_s,τ_s), belongs to $T_{\widehat{M}}(x(t),\theta_\Omega(x(t)))$. Thus we have proved that, for any $t \in]0,\theta_\Omega(x)[$, $(0,1)$ belongs to $T_{\widehat{M}}(x(t),\theta_\Omega(x(t)))$, which is equal to $(p_x(t),p_t(t))^+$; thus $p_t(t) \ge 0$. Since M is an epigraph, $p_t(t)$ is nonpositive, and so $p_t=0$ for any $t \in]0,\theta_\Omega(x)[$.

To complete the proof, let us recall that $T_M(x(t),\theta_\Omega(x(t)))$ is equal to $(p_x(t),p_t(t))^-$ for any $t \in]0,\theta_\Omega(x)[$. Since $p_t(t)=0$, the vector $(0,-1)$ belongs to $T_M(x(t),\theta_\Omega(x(t)))$, which means that $x(t)$ does not belong to $\mathcal{L}$ for any $t \in]0,\theta_\Omega(x)]$. □

REFERENCES

[1] L. Ambrosio, P. Cannarsa, and H. M. Soner, *On the propagation of singularities of semi-convex functions*, Ann. Scuola Sup. Ser. IV., XX (1993), pp. 597–616.
[2] J.-P. Aubin and H. Frankowska, *Set-Valued Analysis*, Birkhäuser, Boston, 1990.
[3] J.-P. Aubin, *Viability Theory*, Birkhäuser, Boston, 1991.
[4] J.-P. Aubin and H. Frankowska, *Viability kernels of control systems*, in Nonlinear Synthesis, C. Byrnes and A. Kurzhanski, eds., Birkhäuser, Boston, 1991, pp. 12–33.
[5] P. Cardaliaguet, M. Quincampoix, and P. Saint-Pierre, *Optimal times for constrained non-linear control problems without local controllability*, Appl. Math. Optim., to appear.
[6] N. Caroff and H. Frankowska, *Conjugate points and shocks in nonlinear optimal control*, Trans. Amer. Math. Soc., 348 (1996), pp. 3133–3153.
[7] H. Frankowska, *Local controllability and infinitesimal generators of semi-groups of set-valued maps*, SIAM J. Control Optim., 25 (1987), pp. 412–432.
[8] S. N. Kruzkov, *Generalized solutions of Hamilton-Jacobi equations of eikonal type* I, Math. USSR-Sb, 27 (1975), pp. 406–446.
[9] M. Quincampoix, *Differential inclusions and target problems*, SIAM J. Control Optim., 30 (1992), pp. 324–335.

SIAM J. CONTROL OPTIM.
Vol. 35, No. 5, pp. 1672–1689, September 1997

013

MIXED OBJECTIVE CONTROL SYNTHESIS: OPTIMAL $\ell_1/\mathcal{H}_2$ CONTROL*

MURTI V. SALAPAKA†, MOHAMMED DAHLEH†, AND PETROS VOULGARIS‡

Abstract. In this paper we consider the problem of minimizing the ℓ_1 norm of the transfer function from the exogenous input to the regulated output over all internally stabilizing controllers while keeping its $\mathcal{H}_2$ norm under a specified level. The problem is analyzed for the discrete-time, single-input single-output (SISO), linear-time invariant case. It is shown that an optimal solution always exists. Duality theory is employed to show that any optimal solution is a finite impulse response sequence, and an a priori bound is given on its length. Thus, the problem can be reduced to a finite-dimensional convex optimization problem with an a priori determined dimension. Finally, it is shown that, in the region of interest of the $\mathcal{H}_2$ constraint level, the optimal is unique and continuous with respect to changes in the constraint level.

Key words. robust control, duality, ℓ_1 optimization, discrete time

AMS subject classifications. 49N05, 49N10, 49N15, 49N35, 93C55

PII. S0363012995286666

1. Notation.

The following notation is employed in this paper:

$\text{int}(X)$	The interior of a set X.
$\|x\|_1$	The 1-norm of the vector $x \in R^n$.
$\|x\|_2$	The 2-norm of the vector $x \in R^n$.
$\hat{x}(\lambda)$	The λ transform of a right-sided real sequence $x = (x(k))_{k=0}^{\infty}$ defined as $\hat{x}(\lambda) := \sum_{k=0}^{\infty} x(k)\lambda^k$.
ℓ_1	The Banach space of right-sided absolutely summable real sequences with the norm given by $\| x \|_1 := \sum_{k=0}^{\infty} \|x(k)\|$.
ℓ_∞	The Banach space of right-sided, bounded sequences with the norm given by $\| x \|_\infty := \sup_k \|x(k)\|$.
c_0	The subspace of ℓ_∞ with elements x that satisfy $\lim_{k\to\infty} x(k) = 0$.
ℓ_2	The Banach space of right-sided square summable sequences with the norm given by $\| x \|_2 := \left[\sum_{k=0}^{\infty} x(k)^2\right]^{\frac{1}{2}}$.
$\mathcal{H}_2$	The isometric isomorphic space of l_2 under the λ transform $\hat{x}(\lambda)$ with the norm given by $\| \hat{x}(\lambda) \|_2 = \| x \|_2$.
X^*	The dual space of the Banach space X. $\langle x, x^* \rangle$ denotes the value of the bounded linear functional x^* at $x \in X$.
$W(X^*, X)$	The weak star topology on X^* induced by X. In this topology, $x_n^* \to x^*$ if and only if $\langle x, x_n^* \rangle \to \langle x, x^* \rangle$ for all $x \in X$.
T^*	The adjoint operator of $T : X \to Y$ which maps Y^* to X^*.

The following identities also hold (see, e.g., [7]): $(\ell_1)^* = \ell_\infty,\ (c_0)^* = \ell_1,\ (\ell_2)^* = \ell_2$.

2. Introduction.

Consider the finite-dimensional linear time invariant system depicted in Figure 2.1, where P denotes the plant and K denotes the controller. The

*Received by the editors May 26, 1995; accepted for publication (in revised form) July 2, 1996. This research was supported by National Science Foundation grants ECS-9204309, ECS-9216690, and ECS-9308481.

http://www.siam.org/journals/sicon/35-5/28666.html

†Mechanical and Environmental Engineering Department, University of California at Santa Barbara, Santa Barbara, CA 93106 (vasu@engineering.ucsb.edu, dahleh@engineering.ucsb.edu).

‡Coordinated Science Laboratory, University of Illinois at Urbana–Champaign, 1308 W. Main St., Urbana, IL 61801 (petros@ktisivios.csl.uiuc.edu).

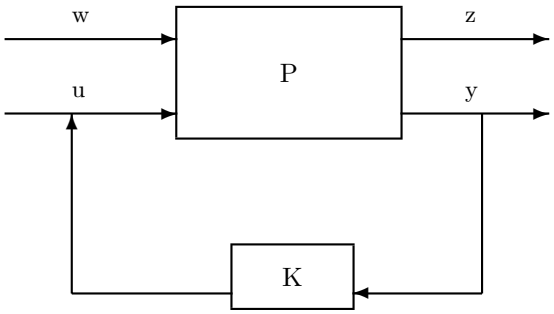


FIG. 2.1. *Plant controller configuration.*

signal w is the exogenous input and z is the regulated output. The signals u and y denote the control input and the measured output, respectively. Let T_{zw} be the closed-loop transfer function which maps w to z.

Many important control problems can be reduced to the above setup, where the objective is to minimize a suitably defined measure of T_{zw}. In the standard ℓ_1 problem the design of an internally stabilizing controller such that the ℓ_∞ norm of the regulated output z due to the worst-case magnitude bounded disturbance w is addressed. It is shown in [3] that this problem reduces to solving finite-dimensional linear programs. The analogous problem, with the signal measures being the ℓ_2 norm, is the standard $\mathcal{H}_\infty$ problem. The standard $\mathcal{H}_2$ problem is concerned with the minimization of the energy contained in the pulse response of the closed loop, T_{zw}. This can be viewed as minimizing the variance of the regulated output z due to a white noise input w. Both problems are discussed in [4].

It is well known (see, e.g., [2]) that optimization with respect to a particular norm may not necessarily yield good performance with respect to another. Thus, if enhanced performance is required with respect to multiple measures, then it is necessary to include all these measures directly into the design process. As a logical step, the design of controllers to satisfy mixed performance criteria has recently been the focus of researchers. Several state-space results on the interplay between the $\mathcal{H}_2$ and $\mathcal{H}_\infty$ are available (see, e.g., [5]). In [1] it is shown that a wide variety of mixed control problems reduce to convex optimization problems, and it is argued that the present technology makes it possible to deem the problem solved if it can be reduced to a convex optimization problem. In this light it is appropriate to exploit as much structure in the problem as possible, so that the standard software available becomes computationally efficient. In [6] the problem of minimizing the ℓ_1 norm of the closed loop under linear inequality constraints is addressed. Every such problem is equivalent to a linear programming problem which has a canonical dual problem associated with it. Contrary to the finite-dimensional case, it is not true that every infinite-dimensional linear program has the same optimal value as its dual. However, it was shown by the authors that under some conditions this "duality gap" does not exist between the primal and the dual, which is advantageous from a computational point of view. The problem of minimizing the ℓ_1 norm of the closed loop while keeping the $\mathcal{H}_\infty$ norm under a prescribed level falls under the above category. In [8] the problem of minimizing the ℓ_1 norm of a single-input single-output (SISO) transfer function while keeping the $\mathcal{H}_\infty$ norm of the closed-loop system under a specified value is reduced to solving a sequence of finite-dimensional convex optimization problems and an unconstrained $\mathcal{H}_\infty$ problem. In [9] it is shown that the $\mathcal{H}_2/\ell_1$ problem, the

problem of minimizing the $\mathcal{H}_2$ norm of the closed loop while maintaining the ℓ_1 norm below a prescribed value, reduces to a finite-dimensional convex optimization problem. However, no a priori bound on its dimension is furnished, which substantially degrades the efficiency of the solution procedure since it may require a large number of iterations.

In this paper, the $\ell_1/\mathcal{H}_2$ problem, which is the problem of minimizing the ℓ_1 norm of T_{zw} while keeping the $\mathcal{H}_2$ norm below a prescribed level, is considered. This problem not only complements the one studied in [9], but it also turns out that much stronger results can be obtained which make this problem considerably more attractive to solve. In particular, it is shown that the problem reduces to a convex finite-dimensional one, and an a priori bound on its dimension is established. The latter feature is important in reducing the computational burden. Furthermore, the developments in this paper are substantially different than those in [9] and are more far reaching.

The paper is organized as follows. In section 3 relevant duality theory results are given. In section 4 the problem statement is made precise. In section 5 it is shown that an optimal solution always exists, and that it is a finite impulse response sequence. An a priori bound is given on its length. In section 6 the region of interest of the constraint level on the $\mathcal{H}_2$ norm is determined. It is shown that in this region the optimal is unique and is continuous with respect to changes in the constraint level. In section 7 an example is given to demonstrate the theory developed.

3. Mathematical preliminaries. In this section we present a Lagrange duality theorem that applies to the minimization of a convex functional subject to both equality and inequality constraints. A sensitivity result which follows directly from the Lagrange duality theorem is presented. We employ the terminology used in [7], which is standard. First, we need the following definitions.

DEFINITION 3.1. *Let P be a convex positive cone in a vector space X. We write $x \geq y$ if $x-y \in P$. We write $x > 0$ if $x \in \text{int}(P)$. Similarly, $x \leq y$ if $x-y \in -P := N$ and $x < 0$ if $x \in \text{int}(N)$. Given a vector space X with positive cone P the positive cone in X^*, $P^{\oplus}$ is defined as*

$$P^{\oplus} := \{x^* \in X^* : \langle x, x^* \rangle \geq 0 \text{ for all } x \in P\}.$$

DEFINITION 3.2. *Let X be a vector space and Z be a vector space with positive cone P. A mapping $G : X \to Z$ is convex if $G(tx+(1-t)y) \leq tG(x)+(1-t)G(y)$ for all x, y in X and t with $0 \leq t \leq 1$ and is strictly convex if $G(tx+(1-t)y) < tG(x)+(1-t)G(y)$ for all $x \neq y$ in X and t with $0 < t < 1$.*

The following is a Lagrange duality theorem.

THEOREM 3.3. *Let X be a Banach space, Ω be a convex subset of X, Y be a finite-dimensional normed space, and Z be a normed space with positive cone P. Let Z^* denote the dual space of Z with a positive cone $P^{\oplus}$. Let $f : \Omega \to R$ be a real valued convex functional, $g : X \to Z$ be a convex mapping, $H : X \to Y$ be an affine linear map, and $0 \in \text{int}[\text{range(H)}]$. Define*

$$\mu_0 := \inf\{f(x) : \ g(x) \leq 0, \ H(x) = 0, \ x \in \Omega\}. \tag{3.1}$$

Suppose that there exists $x_1 \in \Omega$ such that $g(x_1) < 0$ and $H(x_1) = 0$, and suppose that μ_0 is finite. Then

$$\mu_0 = \max\{\varphi(z^*, y) : z^* \geq 0, \ z^* \in Z^*, \ y \in Y\}, \tag{3.2}$$

where $\varphi(z^*,\ y) := \inf\{f(x) + \langle g(x), z^*\rangle + \langle H(x), y\rangle : x \in \Omega\ \}$, *and the maximum is achieved for some* $z_0^* \geq 0,\ z_0^* \in Z^*,\ y_0 \in Y$.

Furthermore, if the infimum in (3.1) *is achieved by some* $x_0 \in \Omega$, *then*

$$\langle g(x_0), z_0^*\rangle + \langle H(x_0), y_0\rangle = 0, \tag{3.3}$$

and

$$x_0 \text{ minimizes } f(x) + \langle g(x), z_0^*\rangle + \langle H(x), y_0\rangle \text{ over all } x \in \Omega. \tag{3.4}$$

Proof. Given any $z^* \geq 0,\ y \in Y$, we have $\inf\{f(x)+\langle g(x), z^*\rangle+\langle H(x), y\rangle : x \in \Omega\}$ $\leq \inf\{f(x) + \langle g(x), z^*\rangle + \langle H(x), y\rangle : x \in \Omega,\ g(x) \leq 0,\ H(x) = 0\} \leq \inf\{f(x) : x \in \Omega,\ g(x) \leq 0,\ H(x) = 0\} = \mu_0$. Therefore, it follows that $\max\{\varphi(z^*, y) : z^* \geq 0,\ y \in Y\} \leq \mu_0$. From Problem 7 of [7, Chap. 8] (see Lemma 9.1 in the Appendix for problem statement and proof), we know that there exists $z_0^* \in Z^*, z_0^* \geq 0,\ y_0 \in Y$ such that $\mu_0 = \varphi(z_0^*,\ y_0)$. This proves (3.2).

Suppose there exists $x_0 \in \Omega, H(x_0) = 0,\ g(x_0) \leq 0$, and $\mu_0 = f(x_0)$; then $\mu_0 = \varphi(z_0^*,\ y_0) \leq f(x_0) + \langle g(x_0), z_0^*\rangle + \langle H(x_0), y_0\rangle \leq f(x_0) = \mu_0$. Therefore, we have $\langle g(x_0), z_0^*\rangle + \langle H(x_0), y_0\rangle = 0$ and $\mu_0 = f(x_0) + \langle g(x_0), z_0^*\rangle + \langle H(x_0), y_0\rangle$. This proves the theorem. ∎

We refer to (3.1) as the *primal* problem and (3.2) as the *dual* problem.

COROLLARY 3.4. *Let* X, Y, Z, f, H, g, Ω *be as in Theorem* 3.3. *Let* x_0 *be the solution to the problem*

$$\begin{gathered} \text{minimize } f(x) \\ \text{subject to } x \in \Omega,\ H(x) = 0,\ g(x) \leq z_0, \end{gathered}$$

with (z_0^*, y_0) *as the dual solution. Let* x_1 *be the solution to the problem*

$$\begin{gathered} \text{minimize } f(x) \\ \text{subject to } x \in \Omega,\ H(x) = 0,\ g(x) \leq z_1, \end{gathered}$$

with (z_1^*, y_1) *as the dual solution. Then*

$$\langle z_1 - z_0, z_1^*\rangle\ \leq f(x_0) - f(x_1)\ \leq \langle z_1 - z_0, z_0^*\rangle. \tag{3.5}$$

Proof. From Theorem 3.3 we know that for any $x \in \Omega$,

$$\begin{aligned} &f(x_0) + \langle g(x_0) - z_0, z_0^*\rangle + \langle H(x_0), y_0\rangle \\ &\leq\ f(x) + \langle g(x) - z_0, z_0^*\rangle + \langle H(x), y_0\rangle. \end{aligned}$$

In particular, we have

$$\begin{aligned} &f(x_0) + \langle g(x_0) - z_0, z_0^*\rangle + \langle H(x_0), y_0\rangle \\ &\leq\ f(x_1) + \langle g(x_1) - z_0, z_0^*\rangle + \langle H(x_1), y_0\rangle. \end{aligned}$$

From Theorem 3.3 we know that $\langle g(x_0) - z_0, z_0^*\rangle + \langle H(x_0), y_0\rangle = 0$ and $H(x_1) = 0$. This implies

$$f(x_0) - f(x_1)\ \leq \langle g(x_1) - z_0, z_0^*\rangle\ \leq \langle z_1 - z_0, z_0^*\rangle.$$

A similar argument gives the other inequality. This proves the corollary. ∎

4. Problem formulation. Consider the standard feedback problem represented in Figure 2.1, where P and K are the plant and the controller, respectively. Let w represent the exogenous input, z represent the output of interest, y be the measured output, and u be the control input where z and w are assumed scalar. Let ϕ be the closed-loop map which maps $w \to z$. From the Youla parametrization (see, e.g., [2]) it is known that all achievable closed-loop maps under stabilizing controllers are given by $\phi = h - u * q$ ($*$ denotes convolution), where $h, u, q \in \ell_1$, h, u depend only on the plant P, and q is a free parameter in ℓ_1. Throughout the paper we make the following assumption.

ASSUMPTION 1. All the zeros of $\hat{u}$ (the λ transform of u) inside the unit disc are real and distinct. Also, $\hat{u}$ has no zeros on the unit circle.

The assumption that all zeros of $\hat{u}$ which are inside the open unit disc are real and distinct is not restrictive and is made to streamline the presentation of the paper. Let the zeros of u which are inside the unit disc be given by $z_1, z_2, \ldots, z_n$. Let

$$\Theta := \{\phi : \text{ there exists } q \in \ell_1 \text{ with } \phi = h - u * q\}.$$

Θ is the set of all achievable closed-loop maps under stabilizing controllers. Let $A : \ell_1 \to R^n$ be given by

$$A = \begin{pmatrix} 1 & z_1 & z_1^2 & z_1^3 & \ldots \\ 1 & z_2 & z_2^2 & z_2^3 & \ldots \\ \vdots & \vdots & \vdots & \vdots & \vdots \\ 1 & z_n & z_n^2 & z_n^3 & \ldots \end{pmatrix},$$

and $b \in R^n$ be given by

$$b = \begin{pmatrix} \hat{h}(z_1) \\ \hat{h}(z_2) \\ \vdots \\ \hat{h}(z_n) \end{pmatrix}.$$

THEOREM 4.1. *The following is true:*

$$\begin{aligned} \Theta &= \{\phi \in \ell_1 : \hat{\phi}(z_i) = \hat{h}(z_i) \text{ for all } i = 1, \ldots, n\} \\ &= \{\phi \in \ell_1 : A\phi = b\}. \end{aligned}$$

Proof. The proof is given in [2, pp. 104–105]. ∎

The problem

$$\begin{aligned} \nu_\infty &:= \inf\{\| h - u * q \|_1 : q \in \ell_1\} \\ &= \inf\{\| \phi \|_1 : \phi \in \ell_1 \text{ and } A\phi = b\} \end{aligned} \tag{4.1}$$

is the standard ℓ_1 problem. In [3] it is shown that this problem has a solution which is possibly nonunique. Optimal solutions are shown to be finite impulse response sequences. Let

$$\begin{aligned} \mu_\infty &:= \inf\{\| h - u * q \|_2{}^2 : q \in \ell_1\}, \\ &= \inf\{\| \phi \|_2{}^2 : \phi \in \ell_1 \text{ and } A\phi = b\}, \end{aligned} \tag{4.2}$$

which is the standard $\mathcal{H}_2$ problem. The solution to this problem is unique, and the solution is an infinite impulse response sequence. Define

$$m_1 := \inf_{A\phi=b, \|\phi\|^2 \le \mu_\infty} \| \phi \|_1, \tag{4.3}$$

which is the ℓ_1 norm of the unique optimal solution of the standard $\mathcal{H}_2$ problem. Let

$$m_2 := \inf_{A\phi=b, \|\phi\| \le \nu_\infty} \| \phi \|_2{}^2, \tag{4.4}$$

which is the infimum over the ℓ_2 norms of the optimal solutions of the standard ℓ_1 problem.

The problem of interest is as follows: *given a positive constant $\gamma > \mu_\infty$ obtain a solution to the following mixed objective problem:*

$$\begin{aligned} \nu_\gamma &:= \inf\{\| h - u * q \|_1 : q \in \ell_1 \text{ and} \langle h - u * q, h - u * q \rangle \le \gamma\} \\ &= \inf\{\| \phi \|_1 : \phi \in \ell_1 \ , A\phi = b \text{ and } \langle \phi, \phi \rangle \le \gamma\}. \end{aligned} \tag{4.5}$$

Note that $\langle ., . \rangle$ is the inner product associated with ℓ_2. In the following sections we will study this problem from the point of view of existence, structure, continuity, and computation of the optimal solutions.

5. Analysis of optimal solutions and their properties. In the first part of this section we show that (4.5) always has a solution. In the second part we show that any solution to (4.5) is of finite length, and in the third, we give an a priori bound on the length.

5.1. Existence of a solution. Here we show that a solution to (4.5) always exists. We use the following lemma (see, e.g., [7]) to prove the main result of this subsection.

LEMMA 5.1 (Banach–Alaoglu). *Let X be a Banach space with X^* as its dual. Then the set $\{x^* : x^* \in X^*, \| x^* \| \le M\}$ is $W(X^*, X)$ compact for any $M \in R$.*

THEOREM 5.2. *There exists $\phi_0 \in \Phi$ such that*

$$\| \phi_0 \|_1 = \inf_{\phi \in \Phi} \{\| \phi \|_1\},$$

where Φ:=$\{\phi \in \ell_1 : A\phi = b$ and$\langle \phi, \phi \rangle \le \gamma\}$ with $\gamma > \mu_\infty$. Therefore, the infimum in (4.5) *is a minimum.*

Proof. We denote the feasible set of our problem by Φ:=$\{\phi \in \ell_1 : A\phi = b$ and$\langle \phi, \phi \rangle \le \gamma\}$. $\nu_\gamma < \infty$ because $\gamma > \mu_\infty$, and therefore the feasible set is not empty. Let $B := \{\phi \in \ell_1 : \| \phi \|_1 \le \nu_\gamma + 1\}$. It is clear that

$$\nu_\gamma = \inf_{\phi \in \Phi \cap B} \{\| \phi \|_1\}.$$

Therefore, given $i > 0$, there exists $\phi_i \in \Phi \cap B$ such that $\| \phi_i \|_1 \le \nu_\gamma + \frac{1}{i}$. B is a bounded set in $\ell_1 = c_0^*$. It follows from the Banach–Alaoglu lemma that B is $W(c_0^*, c_0)$ compact. Using the fact that c_0 is separable, we know that there exists a subsequence $\{\phi_{i_k}\}$ of $\{\phi_i\}$ and $\phi_0 \in \Phi \cap B$ such that $\phi_{i_k} \to \phi_0$ in the $W(c_0^*, c_0)$ sense; that is, for all v in c_0,

$$\langle v, \phi_{i_k} \rangle \to \langle v, \phi_0 \rangle \text{ as } k \to \infty. \tag{5.1}$$

Let the jth row of A be denoted by a_j and the jth element of b be given by b_j. Then, as $a_j \in c_0$ we have

$$\langle a_j, \phi_{i_k} \rangle \to \langle a_j, \phi_0 \rangle \text{ as } k \to \infty \text{ for all } j = 1, 2, \ldots, n. \tag{5.2}$$

As $A(\phi_{i_k}) = b$ we have $\langle a_j, \phi_{i_k} \rangle = b_j$ for all k and for all j which implies $\langle a_j, \phi_0 \rangle = b_j$ for all j. Therefore, we have $A(\phi_0) = b$. As $l_2 \subset c_0$ we have from (5.1) that for all v in l_2,

$$\langle v, \phi_{i_k} \rangle \to \langle v, \phi_0 \rangle \text{ as } k \to \infty, \tag{5.3}$$

which shows that $\phi_{i_k} \to \phi_0$ in $W(l_2^*, l_2)$. Also, from the construction of ϕ_{i_k}, we know that $\| \phi_{i_k} \|_2 \leq \sqrt{\gamma}$. From Lemma 5.1 we conclude that $\langle \phi_0, \phi_0 \rangle \leq \gamma$, and therefore we have shown that $\phi_0 \in \Phi$. Recall that ϕ_{i_k} were chosen so that $\| \phi_{i_k} \|_1 \leq \nu_\gamma + \frac{1}{i_k}$. From Lemma 5.1 we have that $\| \phi_0 \|_1 \leq \nu_\gamma + \frac{1}{i_k}$ for all k. Therefore $\| \phi_0 \|_1 \leq \nu_\gamma$. As $\phi_0 \in \Phi$ (which is the feasible set) we have $\| \phi_0 \|_1 = \nu_\gamma$. This proves the theorem. ∎

5.2. Structure of optimal solutions. In this subsection we use Lagrange duality results to show that every optimal solution is of finite length. The following two lemmas establish the dual problem.

LEMMA 5.3.

$$\nu_\gamma = \max\{\varphi(y_1, y_2) : y_1 \geq 0 \text{ and } y_2 \in R^n\}, \tag{5.4}$$

where

$$\varphi(y_1, y_2) := \inf_{\phi \in \ell_1} \{\| \phi \|_1 + y_1(\langle \phi, \phi \rangle - \gamma) + \langle b - A\phi, y_2 \rangle\}.$$

Proof. We will apply Theorem 3.3 to get the result. Let X, Ω, Y, Z in Theorem 3.3 correspond to ℓ_1, ℓ_1, R^n, and R, respectively. Let $g(\phi) := \langle \phi, \phi \rangle - \gamma$, $H(\phi) := b - A\phi$. With this notation, we have $Z^* = R$.

A has full range, which implies $0 \in \text{int}[\text{range}(H)]$. $\gamma > \mu_\infty$, and therefore there exists ϕ_1 such that $\langle \phi_1, \phi_1 \rangle - \gamma < 0$ and $H(\phi_1) = 0$. Therefore, all the conditions of Theorem 3.3 are satisfied. From Theorem 3.3 we have

$$\nu_\gamma = \max_{y_1 \geq 0, y_2 \in R^n} \inf_{\phi \in \ell_1} \{\| \phi \|_1 + y_1(\langle \phi, \phi \rangle - \gamma) + \langle b - A\phi, y_2 \rangle\}.$$

This proves the lemma. ∎

The right-hand side of (5.4) is the *dual problem.*

LEMMA 5.4. *The dual problem is given by*

$$\max\{\varphi(y_1, y_2) : y_1 \geq 0 \text{ and } y_2 \in R^n\}, \tag{5.5}$$

where

$$\varphi(y_1, y_2) := \inf_{\phi \in \ell_1, \phi(i)v(i) \geq 0} \{\| \phi \|_1 + y_1(\langle \phi, \phi \rangle - \gamma) + \langle b, y_2 \rangle - \langle \phi, v \rangle\}.$$

$v(i)$ *is defined by* $v(i) := A^* y_2(i)$.

Proof. Let $y_1 \geq 0,\ y_2 \in R^n$. It is clear that

$$\begin{aligned} &\inf_{\phi \in \ell_1} \{\| \phi \|_1 + y_1(\langle \phi, \phi \rangle - \gamma) + \langle b - A\phi, y_2 \rangle\} \\ &= \inf_{\phi \in \ell_1} \{\| \phi \|_1 + y_1(\langle \phi, \phi \rangle - \gamma) + \langle b, y_2 \rangle - \langle \phi, v \rangle\}. \end{aligned}$$

Suppose $\phi \in \ell_1$ and there exists i such that $\phi(i)v(i) < 0$; then define $\phi^1 \in \ell_1$ such that $\phi^1(j) = \phi(j)$ for all $j \neq i$ and $\phi^1(i) = 0$. Therefore, we have $\| \phi \|_1 + y_1(\langle \phi, \phi \rangle - \gamma) + \langle b, y_2 \rangle - \langle \phi, v \rangle \geq \| \phi^1 \|_1 + y_1(\langle \phi^1, \phi^1 \rangle - \gamma) + \langle b, y_2 \rangle - \langle \phi^1, v \rangle$. This shows that we can restrict ϕ in the infimization to satisfy $\phi(i)v(i) \geq 0$. This proves the lemma. ∎

The following theorem is the main result of this subsection. It shows that any solution of (4.5) is a finite impulse response sequence.

THEOREM 5.5. *Define* $\mathcal{T}$:=$\{\phi \in \ell_1 : \text{there exists } L^* \text{ with } \phi(i) = 0 \text{ if } i \geq L^*\}$. *The dual of the problem is given by*

$$\max\{\varphi(y_1, y_2) : y_1 \geq 0, y_2 \in R^n\}, \tag{5.6}$$

where

$$\varphi(y_1, y_2) := \inf_{\phi \in \mathcal{T}, \phi(i)v(i) \geq 0} \{\| \phi \|_1 + y_1(\langle \phi, \phi \rangle - \gamma) + \langle b, y_2 \rangle - \langle \phi, v \rangle\}.$$

$v(i)$ *defined by* $v(i) = A^* y_2(i)$. *Also, any optimal solution* ϕ_0 *of* (4.5) *belongs to* $\mathcal{T}$.

Proof. Let $y_1^\gamma \geq 0,\ y_2^\gamma \in R^n$ be the solution to

$$\max_{y_1 \geq 0, y_2 \in R^n} \inf_{\phi \in \ell_1, \phi(i)v(i) \geq 0} \{\| \phi \|_1 + y_1(\langle \phi, \phi \rangle - \gamma) + \langle b - A\phi, y_2 \rangle\}.$$

It is easy to show that there exists L^* such that $v^\gamma(i) := (A^* y_2^\gamma)(i)$ satisfies $|v^\gamma(i)| < 1$ if $i \geq L^*$. If $\phi(i)v^\gamma(i) \geq 0$ for all i, then,

$$\begin{aligned}
&\| \phi \|_1 + y_1^\gamma(\langle \phi, \phi \rangle - \gamma) + \langle b, y_2^\gamma \rangle - \langle \phi, v^\gamma \rangle \\
&= \sum_{i=0}^{\infty} \{|\phi(i)| + y_1^\gamma(\phi(i))^2 - \phi(i)v^\gamma(i)\} - y_1^\gamma \gamma + \langle y_2^\gamma, b \rangle \\
&= \sum_{i=0}^{\infty} \{\phi(i)(\operatorname{sgn}(v^\gamma(i)) - v^\gamma(i)) + y_1^\gamma(\phi(i))^2\} - y_1^\gamma \gamma + \langle y_2^\gamma, b \rangle \\
&= \sum_{i=0}^{L^*} \{\phi(i)(\operatorname{sgn}(v^\gamma(i)) - v^\gamma(i)) + y_1^\gamma(\phi(i))^2\} \\
&\quad + \sum_{i=L^*+1}^{\infty} \{\phi(i)(\operatorname{sgn}(v^\gamma(i)) - v^\gamma(i)) + y_1^\gamma(\phi(i))^2\} - y_1^\gamma \gamma + \langle y_2^\gamma, b \rangle.
\end{aligned}$$

Suppose $|v^\gamma(i)| < 1$. Then we have

$$\phi(i)(\operatorname{sgn}(v^\gamma(i)) - v^\gamma(i)) + y_1^\gamma(\phi(i))^2 \geq 0$$

and equal to zero only if $\phi(i) = 0$. Therefore, in the infimization, we can restrict $\phi(i) = 0$ whenever $|v^\gamma(i)| < 1$. As $|v^\gamma(i)| < 1$ for all $i \geq L^*$ it follows that we can restrict ϕ to $\mathcal{T}$ in the infimization. In Theorem 5.2 we showed that there exists a solution ϕ_0 to the primal. From Theorem 3.3 we have that ϕ_0 minimizes

$$\| \phi \|_1 + y_1^\gamma(\langle \phi, \phi \rangle - \gamma) + \langle b, y_2^\gamma \rangle - \langle \phi, v^\gamma \rangle,$$

over all ϕ in ℓ_1. From the previous discussion it follows that $\phi_0 \in \mathcal{T}$. This proves the theorem. ∎

5.3. An a priori bound on the length of any optimal solution. In this subsection we give an a priori bound on the length of any solution to (4.5). First we establish the following three lemmas.

LEMMA 5.6. *Let* $\gamma > \mu_\infty$, $m_1 := \inf_{A\phi=b, \langle\phi,\phi\rangle\leq\mu_\infty} \| \phi \|_1$, *and* $\nu_\gamma := \inf_{A\phi=b, \langle\phi,\phi\rangle\leq\gamma} \| \phi \|_1$. *Let* y_1^γ, y_2^γ *represent a dual solution as obtained in* (5.5). *Then* $y_1^\gamma \leq M_\gamma$ *where* $M_\gamma := \frac{m_1}{\gamma-\mu_\infty}$.

Proof. Let $\gamma > \gamma_1 > \mu_\infty$ and $\nu_{\gamma_1} := \inf_{A\phi=b, <\phi,\phi>\leq\gamma_1} \| \phi \|_1$. Let y_1^γ, y_2^γ represent a dual solution as obtained in (5.5). From Corollary 3.4 we have

$$\langle \gamma - \gamma_1, y_1^\gamma \rangle \leq \nu_{\gamma_1} - \nu_\gamma \leq \nu_{\gamma_1} \leq m_1,$$

which implies that $y_1^\gamma \leq \frac{m_1}{\gamma-\gamma_1}$. This holds for all $\gamma > \gamma_1 > \mu_\infty$. Therefore, $M_\gamma := \frac{m_1}{\gamma-\mu_\infty}$ is an a priori bound on y_1^γ. This proves the lemma. ∎

LEMMA 5.7. *Let* ϕ_0 *be a solution of the primal* (4.5). *Let* y_1^γ, y_2^γ *represent the corresponding dual solution as obtained in* (5.5). *Let* $v^\gamma := A^* y_2^\gamma$, *then*

$$\begin{aligned} y_1^\gamma \phi_0(i) &= \frac{v^\gamma(i) - 1}{2} \; if \; v^\gamma(i) > 1 \\ &= \frac{v^\gamma(i) + 1}{2} \; if \; v^\gamma(i) < -1 \\ &= \quad 0 \qquad if \; |v^\gamma(i)| \leq 1. \end{aligned}$$

Also, $\| v^\gamma \|_\infty \leq \alpha_\gamma$ *where* $\alpha_\gamma = \frac{2m_1\sqrt{\gamma}}{\gamma-\mu_\infty} + 1$.

Proof. Let

$$L(\phi) := \sum_{i=0}^{\infty} \{\phi(i)(\mathrm{sgn}(v^\gamma(i)) - v^\gamma(i)) + y_1^\gamma(\phi(i))^2\} - \gamma y_1^\gamma + \langle b, y_2^\gamma \rangle.$$

Suppose $|v^\gamma(i)| = 1$. Now, if $y_1^\gamma = 0$, then it is clear that $y_1^\gamma \phi_0(i) = 0$. If $y_1^\gamma > 0$, then as ϕ_0 minimizes $L(\phi)$ we have $\phi_0(i) = 0$. We have already shown that if $|v^\gamma(i)| < 1$, then $\phi_0(i) = 0$. Therefore, $y_1^\gamma \phi_0(i) = 0$ if $|v^\gamma(i)| \leq 1$.

Suppose $v^\gamma(i) > 1$. Then it is easy to show that there exists $\phi(i)$ such that $\phi(i) \geq 0$ and $\phi(i)(\mathrm{sgn}(v^\gamma(i)) - v^\gamma(i)) + y_1^\gamma(\phi(i))^2 < 0$. As any optimal minimizes $L(\phi)$, we know that $\phi_0(i)(\mathrm{sgn}(v^\gamma(i)) - v^\gamma(i)) + y_1^\gamma(\phi_0(i))^2 < 0$, which implies $\phi_0(i) > 0$ and therefore $1 - v^\gamma(i) + 2y_1^\gamma\phi_0(i) = 0$. This implies that $y_1^\gamma\phi_0(i) = \frac{v^\gamma(i)-1}{2}$. Similarly, the result follows when $v^\gamma(i) < -1$. Therefore, $\| v^\gamma \|_\infty \leq 2M_\gamma \| \phi_0 \|_\infty + 1 \leq \frac{2m_1}{\gamma-\mu_\infty} \| \phi_0 \|_\infty + 1 \leq \frac{2m_1\sqrt{\gamma}}{\gamma-\mu_\infty} + 1$. The last inequality follows from the fact that $\langle \phi_0, \phi_0 \rangle \leq \gamma$. This implies that $\alpha_\gamma := \frac{2m_1\sqrt{\gamma}}{\gamma-\mu_\infty} + 1$ is an a priori upper bound on $\| v^\gamma \|_\infty$. This proves the lemma. ∎

LEMMA 5.8 (see [2]). *If* $y_2 \in R^n$ *is such that* $\| A^* y_2 \|_\infty \leq \alpha_\gamma$, *then there exists a positive integer* L^* *independent of* y_2 *such that* $|(A^* y_2)(i)| < 1$ *for all* $i \geq L^*$.

Proof. Define

$$A_L^* = \begin{pmatrix} 1 & 1 & 1 & \dots & 1 \\ z_1 & z_2 & z_3 & \dots & z_n \\ \vdots & \vdots & \vdots & \vdots & \vdots \\ z_1^L & z_2^L & z_3^L & \dots & z_n^L \end{pmatrix},$$

$A_L^* : R^n \to R^{L+1}$. With this definition we have $A_\infty^* = A^*$. Let $y_2 \in R^n$ be such that $\| A^* y_2 \|_\infty \leq \alpha_\gamma$. Choose any L such that $L \geq (n-1)$. As $z_i, i = 1, \ldots, n$, are distinct, A_L^* has full column rank. A_L^* can be regarded as a linear map taking $(R^n, \| \, . \, \|_1) \to (R^{L+1}, \| \, . \, \|_\infty)$. As A_L^* has full column rank, we can define the left inverse of A_L^*, $(A_L^*)^{-l}$, which takes $(R^{L+1}, \| \, . \, \|_\infty) \to (R^n, \| \, . \, \|_1)$. Let the induced norm of $(A_L^*)^{-l}$ be given by $\| \, (A_L^*)^{-l} \, \|_{\infty,1}$. $y_2 \in R^n$ is such that $\| A^* y_2 \|_\infty \leq \alpha_\gamma$, and therefore $\| A_L^* y_2 \|_\infty \leq \alpha_\gamma$. It follows that

$$\| y_2 \|_1 \leq \| \, (A_L^*)^{-l} \, \|_{\infty,1} \, \| A_L^* y_2 \|_\infty \leq \| \, (A_L^*)^{-l} \, \|_{\infty,1} \, \alpha_\gamma. \tag{5.7}$$

Choose L^* such that

$$\max_{k=1,\ldots,n} |z_k|^{L^*} \ \| \, (A_L^*)^{-l} \, \|_{\infty,1} \, \alpha_\gamma < 1. \tag{5.8}$$

There always exists such an L^* because $|z_k| < 1$ for all $k = 1, \ldots, n$. Note that L^* does not depend on y_2. For any $i \geq L^*$ we have

$$\begin{aligned} |(A^* y_2)(i)| = \left| \sum_{k=1}^{k=n} z_k^i y_2(k) \right| &\leq \max_{k=1,\ldots,n} |z_k|^i \| \, y_2 \, \|_1 \\ &\leq \max_{k=1,\ldots,n} |z_k|^i \ \| \, (A_L^*)^{-l} \, \|_{\infty,1} \, \alpha_\gamma \\ &\leq \max_{k=1,\ldots,n} |z_k|^{L^*} \ \| \, (A_L^*)^{-l} \, \|_{\infty,1} \, \alpha_\gamma. \end{aligned}$$

The second inequality follows from (5.7). From (5.8) we have $|(A^* y_2)(i)| < 1$ if $i \geq L^*$. This proves the lemma. □

The following theorem is the main result of the section.

THEOREM 5.9. *Every solution ϕ_0 of the primal* (4.5) *is such that $\phi(i) = 0$ if $i \geq L^*$, where L^* given in Lemma* 5.8 *can be determined a priori. Furthermore, the upper bound on lengths of the optimal solutions is nonincreasing as a function of γ.*

Proof. Using Lemma 5.7 we can bound on $\| \, v^\gamma \, \|_\infty$ by α_γ. Applying Lemma 5.8, we conclude that there exists L_γ^* (which can be determined a priori) such that $|v^\gamma(i)| < 1$ if $i \geq L_\gamma^*$. Using the fact that $\phi_0(i) = 0$ if $|v^\gamma(i)| < 1$, we conclude that $\phi_0 = 0$ if $i \geq L_\gamma^*$. L_γ^*; was chosen to satisfy

$$\max_{k=1,\ldots,n} |z_k|^{L^*} \ \| \, (A_L^*)^{-l} \, \|_{\infty,1} \, \alpha_\gamma < 1.$$

α_γ is nonincreasing as a function of γ. Therefore L_γ^* is nonincreasing as a function of γ. This proves the theorem. □

Note that as $\alpha_\gamma = \frac{2m_1\sqrt{\gamma}}{\gamma - \mu_\infty} + 1$, we have that the upper bound on lengths of the solutions increases to infinity as γ decreases to μ_∞. This is commensurate with the fact that the optimal solution for the standard $\mathcal{H}_2$ problem (4.2) is an infinite impulse response sequence.

The above theorem shows that the problem at hand is a finite-dimensional convex problem of a priori determined dimension. In particular, in view of Theorem 5.9 the problem that needs to be solved is as follows:

$$\nu_\gamma = \min_{A_{L^*}\phi = b, \langle \phi, \phi \rangle \leq \gamma} \sum_{k=0}^{L^*} |\phi(k)|, \tag{5.9}$$

where

$$A_{L^*} = \begin{pmatrix} 1 & z_1 & z_1^2 & \dots & z_1^{L^*} \\ 1 & z_2 & z_2^2 & \dots & z_2^{L^*} \\ \vdots & \vdots & \vdots & \vdots & \vdots \\ 1 & z_n & z_n^2 & \dots & z_n^{L^*} \end{pmatrix},$$

and L^* is given in Lemma 5.8. An alternative representation can be given, as the following lemma suggests.

LEMMA 5.10. *The primal is given by*

$$\begin{aligned} & \textit{minimize} \sum_{k=0}^{L^*} \phi^+(k) + \phi^-(k) \\ \textit{subject to} \quad & A_{L^*}(\phi^+ - \phi^-) = b \\ & \langle \phi^+ - \phi^-, \phi^+ - \phi^- \rangle \le \gamma \\ & \phi^+, \phi^- \textit{ in } R^{L^*} \textit{ with } \phi^+, \phi^- \ge 0. \end{aligned} \tag{5.10}$$

Proof. Note that in the above theorem the ordering is componentwise for the inequalities. We will show that (5.9) is equivalent to (5.10). Let p_0 denote the value attained by the objective functional in (5.10). Suppose ϕ^+, ϕ^- satisfy the constraints of (5.10). Let $\phi := \phi^+ - \phi^-$. Then it is clear that ϕ satisfies the constraints of (5.9). Also, for each k, $|\phi(k)| = |\phi^+ - \phi^-| \le |\phi^+| + |\phi^-| = \phi^+(k) + \phi^-(k)$. This implies that $\nu_\gamma \le p_0$.

Suppose that ϕ satisfies the constraints of (5.9). Define ϕ^+ such that $\phi^+(k) = \phi(k)$ if $\phi(k) \ge 0$, and 0 otherwise. Similarly, define ϕ^- such that $\phi^-(k) = -\phi(k)$ if $\phi(k) \le 0$, and 0 otherwise. It is clear that $\phi = \phi^+ - \phi^-$ and that ϕ^+, ϕ^- satisfy the constraints of (5.10). Also, $|\phi(k)| = \phi^+(k) + \phi^-(k)$. This proves that $\nu_\gamma \ge p_0$. Therefore, $\nu_\gamma = p_0$. It is easy to show that if ϕ_0^+, ϕ_0^- is optimal for (5.10) then $\phi_0 := \phi_0^+ - \phi_0^-$ is optimal for (5.9). This proves the lemma. ∎

This type of convex problem can be solved efficiently using standard methods [1].

6. Uniqueness and continuity of the optimal solution. In this section we address the issue of uniqueness and continuity of solutions to the primal problem with respect to changes in the constraint level on the $\mathcal{H}_2$ norm of the closed-loop map. In the first part we address the issue of uniqueness, and in the second part, we show that the optimal solution is continuous in the region where it is unique.

6.1. Uniqueness of the optimal solution. The following three lemmas are established before the main result of this subsection.

LEMMA 6.1. *Let $y_1^\gamma \ge 0$, $y_2^\gamma \in R^n$ be a solution to* (5.5). *If $y_1^\gamma = 0$, then $\nu_\gamma = \nu_\infty$. This implies that* (4.5) *reduces to solving a standard ℓ_1 problem.*

Proof. Let $v := A^* y_2$ and ϕ^1 be such that $A\phi^1 = b$. If $y_1^\gamma = 0$, then the dual (5.5) is given by

$$\begin{aligned} & \max_{y_2 \in R^n} \inf_{\phi(i)v(i) \ge 0} \{ \| \phi \|_1 + \langle b - A\phi, y_2 \rangle \} \\ = & \max_{y_2 \in R^n} \inf_{\phi(i)v(i) \ge 0} \sum_{i=0}^{\infty} \{ \phi(i)(\operatorname{sgn}(v(i)) - v(i)) \} + \langle \phi^1, v \rangle \\ = & \max_{v \in Range(A^*)} \inf_{\phi(i)v(i) \ge 0} \sum_{i=0}^{\infty} \{ \phi(i)(\operatorname{sgn}(v(i)) - v(i)) \} + \langle \phi^1, v \rangle. \end{aligned}$$

Suppose $\| v \|_\infty > 1$; then there exists j such that $|v(j)| > 1$. Thus we can choose $\phi(j)$ with $\phi(j)v(j) \geq 0$ such that $\phi(j)(\text{sgn}(v(j)) - v(j)) < M$ for any M. This implies that

$$\inf_{\phi(i)v(i)\geq 0} \sum_{i=0}^{\infty} \{\phi(i)(\text{sgn}(v(i)) - v(i))\} + \langle \phi^1, v\rangle = -\infty.$$

Therefore, we can restrict v in the maximization to satisfy $\| v \|_\infty \leq 1$. From arguments similar to that of the proof of Theorem 5.9, $\phi(i) = 0$ whenever $|v(i)| < 1$. Therefore, the infimum term is zero whenever $\| v \|_\infty \leq 1$. This implies that the dual problem reduces to

$$\max_{v\in Range(A^*), \|v\|_\infty \leq 1} \langle \phi^1, v\rangle,$$

which is the same as the dual of the standard ℓ_1 problem as given in (4.1) [2]. This proves the lemma. □

LEMMA 6.2. *Let Ω be a convex subset of a Banach space X and $f : \Omega \to R$ be strictly convex. If f achieves its minimum on Ω then the minimizer is unique.*

Proof. Let $m := \min_{x\in\Omega} f(x)$. Let $x_1, x_2 \in \Omega$ be such that $f(x_1) = f(x_2) = m$. Let $0 < \lambda < 1$. From convexity of Ω we have $\lambda x_1 + (1-\lambda)x_2 \in \Omega$. From strict convexity of f we have that if $x_1 \neq x_2$ then $f(\lambda x_1 + (1-\lambda)x_2) < \lambda f(x_1) + (1-\lambda)f(x_2) = m$, which is a contradiction. Therefore $x_1 = x_2$. This proves the lemma. □

LEMMA 6.3. *Let $y_1^\gamma \geq 0$, $y_2^\gamma \in R^n$ be a solution in* (5.5). *If $y_1^\gamma > 0$ then the solution ϕ_0 of* (4.5) *is unique.*

Proof. Let $L(\phi) := \| \phi \|_1 + y_1^\gamma(\langle \phi, \phi\rangle - \gamma) + \langle b - A\phi, y_2^\gamma\rangle$. From Theorem 3.3 we know that ϕ_0 minimizes $L(\phi)$, $\phi \in \ell_1$. If $y_1^\gamma > 0$ then it is easy to show that $L(\phi)$ is strictly convex in ℓ_1. From the previous lemma it follows that ϕ_0 is unique. This proves the lemma. □

The main result of this subsection is now presented.

THEOREM 6.4. *Define $S := \{\phi : A\phi = b$ and $\| \phi \|_1 = \nu_\infty\}$, $m_2 := \inf_{\phi\in S}\langle \phi, \phi\rangle$. The following is true.*

(1) *If $\gamma \geq m_2$, then problem* (4.5) *is equivalent to the standard ℓ_1 problem whose solution is possibly nonunique.*

(2) *If $\mu_\infty < \gamma < m_2$, then the solution to* (4.5) *is unique.*

Proof. Suppose $m_2 < \gamma$. Then there exists $\phi_1 \in \ell_1$ such that $A\phi_1 = b$, $\| \phi_1 \|_1 = \nu_\infty$, and $\langle \phi_1, \phi_1\rangle \leq \gamma$. This implies that $\nu_\gamma = \inf_{A\phi=b, \langle\phi,\phi\rangle\leq\gamma} \| \phi \|_1 \leq \nu_\infty$. The other inequality is obvious. This proves (1).

Let $\mu_\infty < \gamma < m_2$ and suppose $y_1^\gamma = 0$; then we have shown in Lemma 6.1 that $\nu_\gamma = \nu_\infty$. Therefore, there exists $\phi_1 \in \ell_1$ such that $\| \phi_1 \|_1 = \nu_\infty$, $A\phi_1 = b$, and $\langle \phi_1, \phi_1\rangle \leq \gamma < m_2$. This implies that $\phi_1 \in S$ and $\langle \phi_1, \phi_1\rangle < m_2$, which is a contradiction. Therefore $y_1^\gamma > 0$. From Lemma 6.3 we know that ϕ_0 is unique. This proves (2). □

The above theorem shows that in the region where the constraint level on the $\mathcal{H}_2$ is essentially of interest (i.e., active) the optimal solution is unique.

6.2. Continuity of the optimal solution. Following is a theorem which shows that the ℓ_1 norm of the optimal solution is continuous with respect to changes in the constraint level γ.

THEOREM 6.5. *Let $\nu_\gamma := \inf_{A\phi=b, \langle\phi,\phi\rangle\leq\gamma} \| \phi \|_1$. Then ν_γ is a continuous function of γ on (μ_∞, ∞).*

Proof. If $\gamma \in (\mu_\infty, \infty)$, then it is obvious that $\gamma \in \text{int}\{\text{dom}(\nu_\gamma)\}$, where $\text{dom}(\nu_\gamma) := \{\gamma : -\infty < \nu_\gamma < \infty\}$ is the domain of ν_γ. From Proposition 1 of [7, §8.3] we know that ν_γ is a convex function of γ. The theorem follows from the fact that every convex function is continuous in the interior of its domain. ∎

Now we prove that the optimal solution is continuous with respect to changes in the constraint level in the region where the optimal is unique.

THEOREM 6.6. *Let $\mu_\infty < \gamma < m_2$. Let ϕ_γ represent the solution of $\nu_\gamma = \min_{A\phi=b, \langle\phi,\phi\rangle \le \gamma} \| \phi \|_1$. Then $\phi_{\gamma_k} \to \phi_\gamma$ in the norm topology if $\gamma_k \to \gamma$.*

Proof. Let $m_1 := \min_{A\phi=b, \langle\phi,\phi\rangle \le \mu_\infty} \| \phi \|_1$. Then it is obvious that $\| \phi_\gamma \|_1 = \nu_\gamma \le m_1$. Without loss of generality, assume that $\gamma_k \ge \gamma/2$. Let L^* represent the upper bound on the length of $\phi_{\frac{\gamma}{2}}$. Then, as the upper bound is nonincreasing (see Theorem 5.9) we can assume that $\phi_{\gamma_k} \in R^{L^*}$. Let $B := \{x : x \in R^{L^*} : \| x \|_1 \le m_1\}$; then we have $\phi_{\gamma_k} \in B$. Therefore, there exists a subsequence ϕ_{k_i} of ϕ_{γ_k} and ϕ_1 such that

$$\phi_{k_i} \to \phi_1 \text{ as } i \to \infty \text{ in } (R^{L^*}, \| \, . \, \|_1). \tag{6.1}$$

It is clear, as in the proof of Theorem 5.2, that $A\phi_1 = b$ as $A\phi_{k_i} = b$ for all i. Also,

$$\| \phi_1 \|_2{}^2 \le \| \phi_1 - \phi_{k_i} \|_2{}^2 + \| \phi_{k_i} \|_2{}^2 \le \| \phi_1 - \phi_{k_i} \|_2{}^2 + \gamma_{k_i}.$$

Taking limits on both sides as $i \to \infty$ we get $\langle \phi_1, \phi_1 \rangle \le \gamma$. This implies that ϕ_1 is a feasible element in the problem of ν_γ. From Theorem 6.5 we have $\| \phi_{k_i} \|_1 \to \nu_\gamma$. From (6.1) we have $\| \phi_1 \|_1 = \nu_\gamma$. From uniqueness of the optimal solution we have $\phi_1 = \phi_\gamma$. From uniqueness of the optimal solution, it also follows that $\phi_{\gamma_k} \to \phi_\gamma$. This proves the theorem. ∎

7. An example. In this section we illustrate the theory developed in the previous sections with an example. Consider the SISO plant,

$$\hat{P}(\lambda) = \lambda - \frac{1}{2}, \tag{7.1}$$

where we are interested in the sensitivity map $\phi := (I - PK)^{-1}$. Using Youla parametrization, we get that all achievable transfer functions are given by $\hat{\phi} = (I - \hat{P}\hat{K})^{-1} = 1 - (\lambda - \frac{1}{2})\hat{q}$ where $\hat{q}$ is a stable map. The matrices A and b are given by

$$A = \left(1, \frac{1}{2}, \frac{1}{2^2}, \dots\right), \quad b = 1.$$

It is easy to check that for this problem

$$\mu_\infty := \inf\{\| \phi \|_2{}^2 : \phi \in \ell_1 \text{ and } A\phi = b\} = 0.75$$

and

$$m_1 := \inf_{A\phi=b, \|\phi\|_2{}^2 \le \mu_\infty} \| \phi \|_1 = 1.5,$$

with the optimal solution ϕ_2 given by

$$\hat{\phi}_2(\lambda) = \sum_{t=0}^{\infty} \frac{0.75}{2^t} \lambda^t.$$

Performing a standard ℓ_1 optimization [2] we obtain

$$\nu_\infty := \inf\{\| \phi \|_1 : \phi \in \ell_1 \text{ and } A\phi = b\} = 1$$

and

$$m_2 := \inf_{A\phi=b, \|\phi\|_1 \le \nu_\infty} \| \phi \|_2{}^2 = 1,$$

with the optimal solution $\phi_1 = 1$. We choose the constraint level to be 0.95. Therefore, $\alpha_\gamma = \frac{2m_1\sqrt{\gamma}}{\gamma - \mu_\infty} + 1 = 15.62$. For this example $n = 1$ and $z_1 = \frac{1}{2}$. L^*, the a priori bound on the length of the optimal, is chosen to satisfy

$$\max_{k=1,\ldots,n} |z_k|^{L^*} \ \| (A_L^*)^{-l} \|_{\infty,1} \ \alpha_\gamma < 1, \tag{7.2}$$

where L is any positive integer such that $L \ge (n-1)$. We choose $L = 0$, and therefore $A_L = 1$ and $\| (A_L^*)^{-l} \|_{\infty,1} = 1$. We choose $L^* = 4$, which satisfies (7.2). Therefore, the optimal solution ϕ_0 satisfies $\phi_0(i) = 0$ if $i \ge 4$. The problem reduces to the following finite-dimensional convex optimization problem:

$$\nu_\gamma = \min_{A_{L^*}\phi=1, \|\phi\|_2{}^2 \le 0.95} \left\{ \sum_{k=0}^{3} |\phi(k)| : \phi \in R^4 \right\},$$

where $A_{L^*} = (1, \frac{1}{2}, \frac{1}{2^2}, \frac{1}{2^3})$. We obtain (using Matlab Optimization Toolbox) the optimal solution ϕ_0 to be

$$\hat{\phi}_0(\lambda) = 0.9732 + 0.0535\lambda.$$

Therefore, we have $\| \phi_0 \|_1 = 1.02670$ and $\| \phi_0 \|_2{}^2 \cong 0.95$. The same computation was carried out for various values of the constraint level, $\gamma \in [0.75,\ 1]$. The tradeoff curve between the ℓ_1 and the $\mathcal{H}_2$ norms of the optimal solution is given in Figure 7.1. For all values of γ in the chosen range, the square of the $\mathcal{H}_2$ norm of the optimal closed loop was equal to the constraint level γ. Although, when the constraint level γ equals 0.75, the optimal closed-loop map is an infinite impulse response sequence, the optimal closed-loop map has very few nonzero terms in its impulse response even for values of γ very close to 0.75. For example, with $\gamma = 0.755$ the optimal closed-loop map is given by

$$\hat{\phi}_{0.755} = 0.7708 + 0.3632\lambda + 0.1596\lambda^2 + 0.0578\lambda^3 + 0.0065\lambda^4.$$

As a final remark, we can use the structure of this example to illustrate that the optimal unconstrained $\mathcal{H}_2$ solution can have $\mathcal{H}_2$ norm much smaller than the $\mathcal{H}_2$ norm of the optimal ℓ_1 (unconstrained) solution. Hence, minimizing only the ℓ_1 norm, which is an upper bound on the $\mathcal{H}_2$ norm, may require substantial sacrifices in terms of $\mathcal{H}_2$ performance. Indeed, instead of the P used in the example before, consider the plant $\hat{P}_a(\lambda) = \lambda - a$, where now a is a zero in the unit disk (i.e., $|a| < 1$) and very close to the unit circle (i.e., $|a| \cong 1$). Then the optimal unconstrained $\mathcal{H}_2$ norm given by

$$(b_a(A_aA_a^*)^{-1}b_a)^{1/2} = (1 - |a|^2)^{1/2},$$

where $A_a = (1, a, a^2, \ldots)$, $b_a = 1$ [2], is close to 0. On the other hand, for the optimal ℓ_1 unconstrained solution $\phi_{a,1}$, we have $\phi_{a,1} = 1$, which has $\mathcal{H}_2$ norm equal to 1. Therefore, minimizing only with respect to ℓ_1 may lead to undesirable $\mathcal{H}_2$ performance.

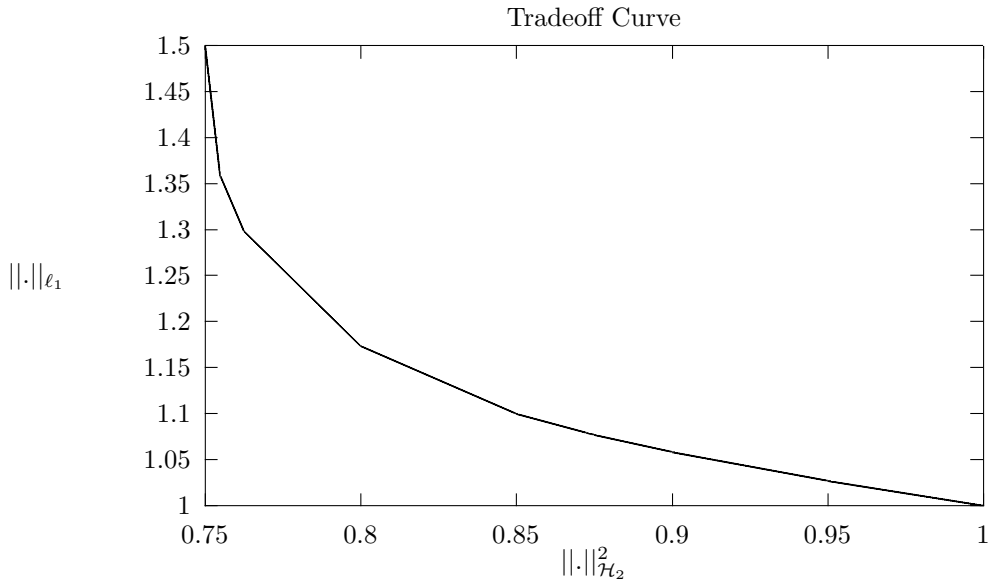


FIG. 7.1. *Here the ℓ_1 and the $\mathcal{H}_2$ norms of the optimal closed loop for various values of γ are plotted. The x axis can be read as the square of $\mathcal{H}_2$ norm or the value of γ.*

8. Conclusions. In this paper the mixed problem of $\ell_1/\mathcal{H}_2$ for the SISO discrete-time case is solved. The problem was reduced to a finite-dimensional convex optimization problem with an a priori determined dimension. The region of the constraint level in which the optimal is unique was determined, and it was shown that in this region the optimal solution is continuous with respect to changes in the constraint level of the $\mathcal{H}_2$ norm. A Lagrange duality theorem and a sensitivity result were used. The techniques used in obtaining the results of this paper are general enough to be adapted to analyze other mixed objective problems. Also, several of the results seem to have natural extensions for the multiple-input, multiple-output case, but a full investigation of this case is beyond the scope of this paper.

9. Appendix. Here we give the lemma needed in the proof of Theorem 3.3.

LEMMA 9.1. *Let X be a Banach space, Ω be a convex subset of X, Y be a finite-dimensional normed space, and Z be a normed space with positive cone P. Let Z^* denote the dual space of Z with a positive cone $P^{\oplus}$. Let $f : \Omega \to R$ be a real valued convex functional, $g : X \to Z$ be a convex mapping, $H : X \to Y$ be an affine linear map, and $0 \in \text{int}[\{y \in Y : \ H(x) = y \text{ for some } x \in \Omega\}]$. Define*

$$\mu_0 := \inf\{f(x) : \ g(x) \leq 0, \ H(x) = 0, \ x \in \Omega\}. \tag{9.1}$$

Suppose there exists $x_1 \in \Omega$ such that $g(x_1) < 0$ and $H(x_1) = 0$ and suppose μ_0 is finite. Then, there exist $z_0^* \geq 0$ and y_0^* such that

$$\mu_0 = \inf\{f(x) + \langle g(x), z_0^* \rangle + \langle H(x), y_0^* \rangle : x \in \Omega\}. \tag{9.2}$$

Proof. Let

$$\Omega_1 := \{x : \ x \in \Omega, \ H(x) = 0\}.$$

Applying Theorem 8.3.1 of [7, p. 217] to Ω_1 we know that there exists $z_0^* \in P^{\oplus}$ such that

$$\mu_0 = \inf\{f(x) + \langle g(x), z_0^* \rangle : x \in \Omega_1\}. \tag{9.3}$$

Consider the convex subset

$$H(\Omega) := \{y \in Y : \ y = H(x) \text{ for some } x \in \Omega\}$$

of Y. For $y \in H(\Omega)$, define

$$k(y) := \inf\{f(x) + \langle g(x), z_0^* \rangle : \ x \in \Omega, \ H(x) = y\}.$$

We now show that k is convex. Suppose $y, y' \in H(\Omega)$ and x, x' are such that $H(x) = y$ and $H(x') = y'$. Suppose $0 < \lambda < 1$. We have

$$\begin{aligned} \lambda(f(x) + \langle g(x), z_0^* \rangle) + (1-\lambda)(f(x') + \langle g(x'), z_0^* \rangle) \quad &\geq \quad f(\lambda x + (1-\lambda)x') \\ &\quad +\langle g(\lambda x + (1-\lambda)x'), z_0^* \rangle \\ &\geq \quad k(\lambda y + (1-\lambda)y'). \end{aligned}$$

(The first inequality follows from the convexity of f and g. The second inequality is true because $H(\lambda x + (1-\lambda)x') = (\lambda y + (1-\lambda)y')$. Taking the infimum on the left-hand side, we obtain $\lambda k(y) + (1-\lambda)k(y') \geq k(\lambda y + (1-\lambda)y')$. This proves that k is a convex function.

We now show that $k : H(\Omega) \to R$ (i.e., we show that $k(y) > -\infty$ for all $y \in H(\Omega)$). As $0 \in \text{int}[H(\Omega)]$, we know that there exists an $\epsilon > 0$ such that if $||y|| \leq \epsilon$, then $y \in H(\Omega)$. Take any $y \in H(\Omega)$ such that $y \neq 0$. Choose λ, y' such that

$$\lambda = \frac{\epsilon}{2||y||} \text{ and } y' = -\lambda y.$$

This implies that $y' \in H(\Omega)$. Let $\beta = \dfrac{\lambda}{\lambda + 1}$. We have

$$(1-\beta)y' + \beta y = 0.$$

Therefore, from convexity of the function k, we have

$$\beta k(y) + (1-\beta)k(y') \geq k(0) = \mu_0.$$

But $\mu_0 > -\infty$ by assumption. Therefore, $k(y) > -\infty$. Note that for all $y \in H(\Omega)$, $k(y) < \infty$. This proves that k is a well-defined function.

Let $[k, H(\Omega)]$ be defined as given below:

$$[k, H(\Omega)] := \{(r, y) \in R \times Y : y \in H(\Omega), \ k(y) \leq r\}.$$

We first show that $[k, H(\Omega)]$ has a nonempty interior. As k is a well-defined convex function on the finite-dimensional convex set $H[\Omega]$ and $0 \in \text{int}[H(\Omega)]$, we have from Corollary 7.9.1 of [7, p. 194] that k is continuous at 0. Let $r_0 = k(0) + 2$ and choose ϵ' such that $0 < \epsilon' < 1$. As k is continuous at 0 we know that there exists $\delta > 0$ such that $y \in H(\Omega)$, and $||y|| \leq \delta$ implies that

$$|k(y) - k(0)| < \epsilon'.$$

This means that if $y \in H(\Omega)$ and $||y|| \leq \delta$, then

$$k(y) < k(0) + \epsilon' < k(0) + 1 < r_0 - \frac{1}{2}.$$

Therefore, for all $y \in H(\Omega)$ with $||y|| \leq \delta$, we have $k(y) < r_0 - \frac{1}{2}$. This implies that for all $(r, y) \in R \times Y$ such that $|r - r_0| < \frac{1}{4}, y \in H(\Omega)$, and $||y|| \leq \delta$, we have $k(y) < r$. This proves that $(r_0, 0) \in \text{int}([k, H(\Omega)])$.

It is clear that $(k(0), 0) \in R \times Y$ is not in the interior of $[k, H(\Omega)]$. Using Theorem 5.12.2 of [7, p. 133] we know that there exists $(s, y^*) \neq (0, 0) \in R \times Y^*$ such that for all $(r, y) \in [k, H(\Omega)]$ the following is true:

$$\langle y, y^* \rangle + rs \geq \langle 0, y^* \rangle + k(0)s = s\mu_0. \tag{9.4}$$

In particular, $rs \geq s\mu_0$ for all $r \geq \mu_0$ (note that $(r, 0) \in [k, H(\Omega)]$ for all $r \geq \mu_0$). This means that $s \geq 0$.

Suppose $s = 0$. We have from (9.4) that $\langle y, y^* \rangle \geq 0$ for all $y \in H(\Omega)$. As $0 \in \text{int}[H(\Omega)]$, it follows that there exists an $\epsilon \in R$ such that $||y|| \leq \epsilon$ implies that $\langle y, y^* \rangle \geq 0$ and $\langle -y, y^* \rangle \geq 0$. This implies that if $||y|| \leq \epsilon$, then $\langle y, y^* \rangle = 0$. But then, for any $y \in Y$, one can choose a positive constant α such that $||\alpha y|| \leq \epsilon$, and therefore $\langle \alpha y, y^* \rangle = 0$. This implies that $(s, y^*) = (0, 0)$, which is not possible. Therefore, we conclude that $s > 0$.

Let $y_0^* = y^*/s$. From (9.4) we have

$$\langle y, y_0^* \rangle + r \geq \mu_0 \quad \text{for all } (r, y) \in [k, H(\Omega)]. \tag{9.5}$$

This implies that for all $y \in H(\Omega)$,

$$\langle y, y_0^* \rangle + k(y) \geq \mu_0. \tag{9.6}$$

(This is because $(k(y), y) \in [k, H(\Omega)]$.) Therefore, for all $x \in \Omega$,

$$\langle H(x), y_0^* \rangle + f(x) + \langle g(x), z_0^* \rangle \geq \mu_0, \tag{9.7}$$

which implies that

$$\inf\{f(x) + \langle g(x), z_0^* \rangle + \langle H(x), y_0^* \rangle : \ x \in \Omega\} \geq \mu_0. \tag{9.8}$$

But if $x \in \Omega$ is such that $H(x) = 0$, then

$$f(x) + \langle g(x), z_0^* \rangle = f(x) + \langle g(x), z_0^* \rangle + \langle H(x), y_0^* \rangle \tag{9.9}$$

$$\geq \inf\{f(x) + \langle g(x), z_0^* \rangle + \langle H(x), y_0^* \rangle : \ x \in \Omega\} \geq \mu_0. \tag{9.10}$$

Taking the infimum on the left-hand side of the above inequality over all $x \in \Omega$ which satisfy $H(x) = 0$ (that is infimum over all $x \in \Omega_1$), we have

$$\mu_0 = \inf\{f(x) + \langle g(x), z_0^* \rangle + \langle H(x), y_0^* \rangle : \ x \in \Omega\}. \tag{9.11}$$

This proves the lemma. □

REFERENCES

[1] S. P. Boyd and C. H. Barratt, *Linear Controller Design: Limits Of Performance*, Prentice-Hall, Englewood Cliffs, NJ, 1991.

[2] M. A. Dahleh and I. J. Diaz-Bobillo, *Control of Uncertain Systems: A Linear Programming Approach*, Prentice-Hall, Englewood Cliffs, NJ, 1995.

[3] M. A. Dahleh and J.B. Pearson, ℓ_1 *Optimal feedback controllers for MIMO discrete-time systems*, IEEE Trans. Automat. Control, 32 (1987), pp. 314–322.

[4] J. C. Doyle, K. Glover, P. Khargonekar, and B. A. Francis, *State space solutions to standard $\mathcal{H}_2$ and $\mathcal{H}_\infty$ control problems*, IEEE Trans. Automat. Control, 34 (1989), pp. 831–847.

[5] J. C. Doyle, K. Zhou, and B. Bodenheimer, *Optimal control with mixed $\mathcal{H}_2$ and $\mathcal{H}_\infty$ performance objectives*, in Proceedings of the American Control Conference, Vol. 3, Pittsburgh, PA, 1989, pp. 2065–2070.

[6] N. Elia, M. A. Dahleh, and I. J. Diaz-Bobillo, *Controller design via infinite dimensional linear programming*, in Proceedings of the American Control Conference, Vol. 3, San Francisco, CA, 1993, pp. 2165–2169.

[7] D. G. Luenberger, *Optimization by Vector Space Methods*, John Wiley, New York, 1969.

[8] M. Sznaier, *Mixed $\ell_1/\mathcal{H}_\infty$ controllers for SISO discrete time systems*, in Proceedings of the IEEE Conference on Decision and Control, Vol. 1, San Antonio, TX, 1993, pp. 34–38.

[9] P. Voulgaris, *Optimal $\mathcal{H}_2/\ell_1$ control: The siso case*, in Proceedings of the IEEE Conference on Decision and Control, Vol. 4, Orlando, FL, December 1994, pp. 3181–3186.

SIAM J. CONTROL OPTIM.
Vol. 35, No. 5, pp. 1690–1710, September 1997

014

NUMERICALLY RELIABLE COMPUTATION OF OPTIMAL PERFORMANCE IN SINGULAR H_∞ CONTROL*

PASCAL GAHINET† AND ALAN J. LAUB‡

Abstract. A numerically stable algorithm is described for the computation of the optimal H_∞ performance $\gamma_{\rm opt}$ when the feedthrough matrices D_{12} or D_{21} are rank-deficient. Using only orthogonal transformations, the singular H_∞ problem is reduced to a regular subproblem and a standard Riccati-based γ-iteration is applied to this subproblem to compute $\gamma_{\rm opt}$. Various interpretations of this scheme are given in terms of the infinite zero structure of the plant and the deflation of Hamiltonian pencils. The implementation of this algorithm is straightforward and its performance and reliability are confirmed by extensive numerical testing.

Key words. H_∞ optimal control, singular problem, numerical computation, algebraic Riccati equation, linear matrix inequality

AMS subject classifications. 93C05, 93C35, 93C60, 93C45, 93B40, 49B99

PII. S0363012994269958

1. Introduction.

It is well known that the solvability of regular H_∞ problems can be characterized in terms of the stabilizing solutions X_∞ and Y_∞ of two indefinite Riccati equations [9]. In addition, a particular solution, called the central controller, is given by explicit formulas in terms of X_∞ and Y_∞ [18, 9, 20]. Extensions of this Riccati-based characterization to singular problems where D_{12} or D_{21} are rank-deficient have been proposed in [32, 34, 5, 27]. The suboptimality conditions of [32, 34] involve reduced-order algebraic Riccati equations (AREs) that are extracted via coordinate transformations on the plant matrices. Suboptimal controllers are then obtained by solving an almost disturbance decoupling problem on the transformed plant. Unfortunately, this approach involves nonorthogonal similarity transformations and is therefore prone to numerical instability.

A numerically more appealing approach is proposed in [5, 6]. There the authors consider ϵ-regularizations of singular H_∞ problems and characterize suboptimality in terms of the limits as ϵ goes to zero of the regularized Riccati solutions $X_\infty(\epsilon)$ and $Y_\infty(\epsilon)$. It is further argued that these limits can be directly computed via standard pencil-based algorithms. However, the possible singularity of the pencils considered in [5, 6] may also be a source of numerical difficulties as shown in subsection 5.1 of this paper. Moreover, the controller formulas in [5] are obtained as the limit of the entropy-maximizing central controller of [9]. As discussed previously in [33], this limit may be improper and even ill defined in terms of transfer functions (see also [5]). The zero-compensation scheme of [6] is also likely to produce improper controllers.

The main contribution of the present paper is a new algorithm for computing the optimal H_∞ performance in the singular case. While relying on the suboptimality conditions of [5, 6], this algorithm proceeds by reducing these conditions to standard, reduced-order Riccati equations. In contrast to [32, 34], this reduction involves only numerically stable singular value decompositions (SVDs) and is straightforward to

*Received by the editors June 22, 1994; accepted for publication (in revised form) July 3, 1996. http://www.siam.org/journals/sicon/35-5/26995.html

†The MathWorks Inc., 24 Prime Park Way, Natick, MA 01760-1500 (pascal@mathworks.com).

‡College of Engineering, University of California, Davis, 1050 Engineering II, Davis, CA 95616-5294 (laub@ucdavis.edu).

implement. And even though our algorithm amounts to a unitary deflation of the pencil W_{12_∞} considered in [5], it does not attempt to explicitly compute the stable eigenspace of this potentially singular pencil. This makes it an efficient and reliable alternative to previously available schemes.

The paper is organized as follows. Section 2 gives some background on matrix pencils, in particular on singular pencils. Section 3 recalls the problem statement and the linear matrix inequality (LMI)–based suboptimality conditions for general H_∞ problems. Section 4 explores the connection between the LMI- and Riccati-based characterizations of suboptimality for singular H_∞ problems. While leading to Riccati-based conditions similar to those of [5, 6], this analysis lays the ground for the subsequent derivation of our algorithm. Section 5 contains the main results on the computation of the Riccati "solutions" X_∞ and Y_∞ in the singular case. An iterative scheme to extract a regular subproblem is presented, and it is shown that the resulting subproblem fully determines X_∞ and Y_∞. The overall algorithm for the computation of γ_{opt} is presented in section 6, and section 7 discusses numerical stability issues. Finally, section 8 reports the results of numerical testing.

2. Background on singular matrix pencils. This section recalls basic notions on matrix pencils that are useful in the subsequent analysis. A more complete discussion is found in [17, 37]. A matrix pencil is any pair (A, B) of matrices of the same dimensions and is usually denoted by $A-\lambda B$. Note that A, B need not be square. The pencil $A-\lambda B$ is called *regular* if A is square and $\det(A-\lambda B) \neq 0$ for some $\lambda \in \mathbb{C}$, and is called *singular* otherwise. Two pencils $A_1-\lambda B_1$ and $A_2-\lambda B_2$ are said to be *equivalent* if there exist two invertible matrices P, Q such that $A_2-\lambda B_2 = P(A_1-\lambda B_1)Q$.

Kronecker's theory of matrix pencils [17] shows that any pencil $A-\lambda B$ can be brought by equivalence transformation to the canonical form

$$P(A-\lambda B)Q = \text{Diag}\left(R_{p_1}, \ldots, R_{p_r}, C_{q_1}, \ldots, C_{q_s}, I-\lambda N, M-\lambda I\right), \tag{2.1}$$

where

- R_{p_i} is a $p_i \times (p_i+1)$ bidiagonal matrix of the form $\begin{pmatrix} -\lambda & 1 & & \\ & \ddots & \ddots & \\ & & -\lambda & 1 \end{pmatrix}$;
- C_{q_j} is a $(q_j+1) \times q_j$ bidiagonal matrix of the form $\begin{pmatrix} 1 & & \\ -\lambda & \ddots & \\ & \ddots & 1 \\ & & -\lambda \end{pmatrix}$;
- $N \in \mathbb{R}^{m\times m}$ is a nilpotent Jordan matrix and $M \in \mathbb{R}^{n\times n}$.

The rectangular R and C blocks determine the Kronecker row and column structure, respectively, and constitute the singular part of the pencil. The two regular pencils $M-\lambda I$ and $I-\lambda N$ constitute the regular part of the pencil and determine its finite and infinite eigenstructure, respectively. Indeed, all generalized eigenvalues of $I-\lambda N$ are at infinity since N is nilpotent, while all eigenvalues of $M-\lambda I$ are finite.

While computing the decomposition (2.1) is an ill-conditioned problem, the singular/regular decomposition and the pencil eigenstructure can be computed in a backward stable manner using Van Dooren's deflation algorithm [37]. Using orthogonal transformations P and Q, this algorithm deflates any pencil $A-\lambda B$ to the staircase

form

$$P(A-\lambda B)Q = \begin{pmatrix} A_r-\lambda B_r & \star & \star & \star \\ 0 & A_f-\lambda B_f & \star & \star \\ 0 & 0 & A_i-\lambda B_i & \star \\ 0 & 0 & 0 & A_c-\lambda B_c \end{pmatrix}, \tag{2.2}$$

where

- $A_f-\lambda B_f$ and $A_i-\lambda B_i$ are square regular pencils associated with the finite and infinite generalized eigenvalues, respectively;
- $A_r-\lambda B_r$ and $A_c-\lambda B_c$ are singular rectangular pencils, A_r having full row rank and A_c having full column rank.

Note that backward stability does not suppress the intrinsic sensitivity of this decomposition to rounding errors. In particular, the dimensions of each characteristic subpencil can be drastically altered by small perturbations of the data.

Note finally that the finite generalized eigenvalues of $A-\lambda B$ have the following simple rank characterization.

LEMMA 2.1. *If*

$$\nu := \max_{\lambda} \ \mathrm{Rank}(A-\lambda B)$$

denotes the normal rank of the pencil $A-\lambda B$, the finite generalized eigenvalues of $A-\lambda B$ are the complex values λ_k such that

$$\mathrm{Rank}\,(A-\lambda_k B) < \nu\ . \tag{2.3}$$

Proof. The proof is essentially in [17] and is included for completeness. The rank of $A-\lambda B$ is readily assessed from the canonical decomposition (2.1). Observing that each R_{p_i} is of rank p_i and each C_{q_j} is of rank q_j regardless of λ, it follows that

$$\mathrm{Rank}(A-\lambda B) = \sum_{i=1}^{r} p_i + \sum_{j=1}^{s} q_j + \mathrm{Rank}(I-\lambda N) + \mathrm{Rank}(M-\lambda I).$$

Now, $I-\lambda N$ remains invertible for all λ since N is a nilpotent Jordan matrix. Thus, the overall rank can change only when $\mathrm{Rank}(M-\lambda I)$ drops, that is, when λ is an eigenvalue of M and hence a finite generalized eigenvalue of $A-\lambda B$. ∎

3. Singular H_∞ control.

3.1. Problem statement. Consider a linear time-invariant plant $P(s)$ with state-space equations

$$\left\{ \begin{array}{rcl} \dot{x} &=& Ax + B_1 w + B_2 u, \\ z &=& C_1 x + D_{11} w + D_{12} u, \\ y &=& C_2 x + D_{21} w + D_{22} u, \end{array} \right. \tag{3.1}$$

where the vectors w, u, z, and y denote the exogenous inputs, control inputs, controlled outputs, and measured outputs, respectively. The plant dimensions are summarized by

$$A \in \mathbb{R}^{n\times n}, \qquad D_{11} \in \mathbb{R}^{p_1\times m_1}, \qquad D_{22} \in \mathbb{R}^{p_2\times m_2}\ .$$

Let $T_{wz}(s)$ denote the closed-loop transfer function from w to z under dynamic output-feedback $u = K(s)y$.

Given some performance level $\gamma > 0$, the suboptimal H_∞ control problem consists of finding an internally stabilizing controller $K(s)$ such that

$$\|T_{wz}(s)\|_\infty < \gamma.$$

Solutions of this problem (if any) will be called γ-*suboptimal* controllers. The optimal H_∞ gain $\gamma_{\rm opt}$ is defined as the smallest achievable performance γ. The following assumptions on the plant matrices are made throughout the paper:

(A1) (A, B_2, C_2) is stabilizable and detectable;
(A2) $D_{22} = 0$;
(A3) the matrix pencils $\left(\begin{smallmatrix} A-\lambda I & B_2 \\ C_1 & D_{12} \end{smallmatrix}\right)$ and $\left(\begin{smallmatrix} A-\lambda I & B_1 \\ C_2 & D_{21} \end{smallmatrix}\right)$ have no finite generalized eigenvalue on the imaginary axis.

Note that (A2) incurs no loss of generality and merely amounts to redefining the measured outputs as $y - D_{22}u$. For simplicity, most results and proofs are stated for the case $D_{11} = 0$. See section 6 for the general case $D_{11} \neq 0$.

Of particular interest in this paper are plants where D_{12} and/or D_{21} are rank-deficient. The corresponding H_∞ problem is traditionally referred to as *singular*, following the LQG terminology. Singular problems are more difficult since they cannot be handled by the standard Riccati-based approach to H_∞ control [9]. Note, however, that such problems are all but ill-posed from both theoretical and control viewpoints. On the contrary, most singular H_∞ problems have meaningful solutions, especially suboptimal ones. Singularity means only that *Riccati-based* central controllers [9] are not well defined [5]. Note finally that singular problems are by no means exotic and do frequently arise in applications such as loop-shaping, μ-synthesis, etc.

3.2. LMI-based conditions for solvability. The derivation of the algorithm presented in this paper makes extensive use of the LMI-based characterization of H_∞ suboptimality [12, 24]. This result is recapped in the next theorem and applies to any plant, whether it be regular or singular.

THEOREM 3.1 (solvability of the suboptimal H_∞ problem). *Consider a proper continuous-time plant $P(s)$ of order n and minimal realization* (3.1)*, and assume* (A1)–(A2) *and $D_{11} = 0$. Given bases $\mathcal{N}_{12}$ and $\mathcal{N}_{21}$ of the null spaces of $(B_2^T,\ D_{12}^T)$ and (C_2, D_{21}), respectively, the suboptimal H_∞ problem with performance γ is solvable if and only if there exist pairs of symmetric matrices (R, S) in $\mathbb{R}^{n\times n}$ such that*

$$\mathcal{N}_{12}^T \begin{pmatrix} AR + RA^T + \gamma^{-2} B_1 B_1^T & RC_1^T \\ C_1 R & -I \end{pmatrix} \mathcal{N}_{12} < 0, \tag{3.2}$$

$$\mathcal{N}_{21}^T \begin{pmatrix} A^T S + SA + \gamma^{-2} C_1^T C_1 & SB_1 \\ B_1^T S & -I \end{pmatrix} \mathcal{N}_{21} < 0, \tag{3.3}$$

$$R > 0, \qquad S > 0, \qquad \lambda_{\min}(RS) \geq \gamma^{-2}. \tag{3.4}$$ ∎

Theorem 3.1 shows that the performance γ is achievable if and only if the following feasibility problem is solvable:

$$\text{find } R \in \mathcal{R} \text{ and } S \in \mathcal{S} \text{ such that } \ \lambda_{\min}(RS) \geq \gamma^{-2}, \tag{3.5}$$

where

$$\mathcal{R} := \{R > 0 : \ \ (3.2) \text{ holds}\}, \qquad \mathcal{S} := \{S > 0 : \ \ (3.3) \text{ holds}\}. \tag{3.6}$$

Since (3.2)–(3.4) are LMI constraints on the matrices R and S [12], this feasibility problem could be tackled directly by convex optimization techniques. In particular, it

falls within the scope of efficient LMI solvers such as those described in [3, 26, 36, 25]. However, solving LMIs is of higher complexity than solving Riccati equations. For this reason, solvability conditions involving only standard linear algebra techniques remain computationally appealing.

4. Riccati-based conditions for singular problems. Henceforth, the discussion is implicitly specialized to the singular case (D_{12} or D_{21}^T column-rank deficient), most results becoming trivial in the regular case. Although the Riccati-based characterization of [9] is not applicable to singular problems, it can be generalized along the lines of [5]. This section sheds new light on the results of [5], in particular on the connection between the LMI- and Riccati-based characterizations of suboptimality. This analysis is the foundation of the new algorithm proposed in section 5.

4.1. From LMIs to Riccati equations. We first outline the procedure for turning the LMI conditions of Theorem 3.1 into Riccati equations. This procedure is adapted from [31, 12] and relies on the following two technical lemmas.

LEMMA 4.1 (Finsler's lemma [35, 28]). *Given a symmetric matrix M, an unstructured matrix P, and any matrix W_P whose columns form a basis for the null space of P, there exists $\alpha > 0$ such that $M - \alpha\, P^T P < 0$ if and only if $W_P^T\, M\, W_P < 0$.*

LEMMA 4.2. *Given any matrices $A \in \mathbb{R}^{n\times n}$, $F = F^T \in \mathbb{R}^{n\times n}$, $C \in \mathbb{R}^{m\times n}$ such that (C, A) has no unobservable mode on the imaginary axis, the set*

$$\mathcal{R} := \{R > 0 : AR + RA^T + RC^T CR + F < 0\} \tag{4.1}$$

is nonempty if and only if the Riccati equation $A^T X + XA + XFX + C^T C = 0$ has a stabilizing solution $X_{\rm st} \geq 0$. (Here "stabilizing" means that the closed-loop matrix $A + FX$ has all its eigenvalues in the open left-half plane.)

Moreover, $X_{\rm st}$ is a lower limit point for the set $\mathcal{R}_{\rm inv} := \{R^{-1} : R \in \mathcal{R}\}$ in such cases. That is,

- *$X_{\rm st} < R^{-1}$ for all $R \in \mathcal{R}$;*
- *there exists a sequence $\{R_n\}$ of elements of $\mathcal{R}$ such that $\lim_{n\to\infty} R_n^{-1} = X_{\rm st}$. In other words, $X_{\rm st}$ lies on the boundary of $\mathcal{R}_{\rm inv}$.*

Proof. See Appendix A. ∎

A straightforward application of these two lemmas leads to the following ARE-based condition for the existence of *positive definite* solutions to the LMI (3.2).

LEMMA 4.3. *Assume* (A3) *and let*

$$\pi_{12} := I - D_{12}^+ D_{12}, \quad \hat{B}_2 := B_2 D_{12}^+, \quad \hat{A} := A - \hat{B}_2 C_1, \quad \hat{C}_1 := (I - D_{12} D_{12}^+) C_1. \tag{4.2}$$

The set $\mathcal{R}$ defined in (3.6) *is nonempty if and only if the Riccati equation*

$$\hat{A}^T X + X\hat{A} + X(\gamma^{-2} B_1 B_1^T - \hat{B}_2 \hat{B}_2^T) X + \hat{C}_1^T \hat{C}_1 \;-\; \alpha\, X B_2 \pi_{12} B_2^T X = 0 \tag{4.3}$$

has a stabilizing solution $X(\alpha) \geq 0$ for $\alpha > 0$ large enough.

Proof. Following [12], (3.2) is equivalent to

$$W_{12}^T \left\{ \hat{A}R + R\hat{A}^T + R\hat{C}_1^T \hat{C}_1 R + \gamma^{-2} B_1 B_1^T - \hat{B}_2 \hat{B}_2^T \right\} W_{12} < 0, \tag{4.4}$$

where W_{12} denotes any full-rank matrix whose columns span the null space of $\pi_{12} B_2^T$. Invoking Lemma 4.1, this inequality is feasible if and only if

$$\hat{A}R + R\hat{A}^T + R\hat{C}_1^T \hat{C}_1 R + \gamma^{-2} B_1 B_1^T - \hat{B}_2 \hat{B}_2^T \;-\; \alpha\, B_2 \pi_{12} B_2^T < 0. \tag{4.5}$$

is feasible for some $\alpha > 0$. Now, $(\hat{C}_1, \hat{A})$ has no unobservable mode on the imaginary axis by virtue of (A3). By Lemma 4.2, it follows that (4.5) has positive definite solutions if and only if the ARE (4.3) has a stabilizing solution $X(\alpha) \geq 0$ for $\alpha > 0$ large enough. □

Remark 4.4. The Riccati equation (4.3) can be viewed as the standard H_∞ ARE for some ϵ-regularization of the plant $P(s)$. Specifically, define $\epsilon := \sqrt{1/\alpha}$; introduce an orthogonal transformation $V = (V_1, V_2)$ such that $D_{12}(V_1, V_2) = (\Delta, 0)$, where Δ has full column rank; and consider the plant obtained by replacing C_1, D_{11}, and D_{12} with

$$\begin{pmatrix} C_1 \\ 0 \end{pmatrix}, \qquad \begin{pmatrix} D_{11} \\ 0 \end{pmatrix}, \qquad \begin{pmatrix} \Delta & 0 \\ 0 & \epsilon I \end{pmatrix} V^T,$$

respectively. The resulting D_{12} matrix has full column rank, and it is readily verified from the definition (4.2) of $\hat{A}$, $\hat{B}_2$, and $\hat{C}_1$ that the corresponding controller ARE for X_∞ coincides with (4.3).

4.2. Riccati-based characterization of suboptimality. In the regular case (full-rank D_{12}), (4.3) reduces to the usual ARE for X_∞ since $\pi_{12} = 0$. In the singular case, however, (4.3) defines a family of AREs parametrized by α, and testing the existence of a stabilizing solution for large enough α becomes nontrivial. To remove the dependence on α, note that $X(\alpha)$ has the following properties:

(i) there exists $\alpha_0 > 0$ such that (4.3) has a stabilizing solution $X(\alpha)$ for all $\alpha > \alpha_0$;

(ii) $X(\alpha)$ is a monotonically decreasing function of α on the half-line $(\alpha_0, +\infty)$ [39].

When $\mathcal{R}$ is nonempty, (i) is an immediate consequence of Lemma 4.3. More generally, observe that the dependence on α and the dependence on γ are of the same nature. As a result, (i) follows from the analysis in [30, 11] (see, e.g., Corollary 5.2 in [11]).

Since $X(\alpha)$ is a decreasing function of $\alpha > \alpha_0$, it has a (not necessarily finite) limit as α goes to $+\infty$. This suggests introducing the following extension of the usual Riccati solution X_∞.

DEFINITION 4.5 (asymptotic stabilizing solution). *The family of Riccati equations* (4.3) *is said to have an asymptotic stabilizing solution X_∞ if and only if*

(1) *these equations have a stabilizing solution $X(\alpha)$ on some interval $(\alpha_0, +\infty)$;*

(2) *$X(\alpha)$ has a finite limit X_∞ as α goes to $+\infty$:*

$$X_\infty := \lim_{\alpha \to +\infty} X(\alpha). \tag{4.6}$$

Note that X_∞ coincides with the limit of the Riccati solutions $X(\epsilon)$ considered in [5] (see Remark 4.1 above). Lemma 4.3 has a simple reformulation in terms of asymptotic stabilizing solution.

LEMMA 4.6. *The set $\mathcal{R}$ defined in* (3.6) *is nonempty if and only if the family of α-parametrized Riccati equations* (4.3) *has an asymptotic stabilizing solution $X_\infty \geq 0$. Moreover, X_∞ is a lower limit point for the set $\mathcal{R}_{\text{inv}} := \{R^{-1} : R \in \mathcal{R}\}$.*

Proof. Sufficiency is immediate from Lemma 4.3, Definition 4.5, and the fact that $X_\infty \leq X(\alpha)$ since $X(\alpha)$ is monotonically decreasing in α. To prove necessity, recall from Lemma 4.3 that $\mathcal{R} \neq \emptyset$ implies that $X(\alpha) \geq 0$ for α large enough. Hence, $X(\alpha)$ is bounded from below as $\alpha \to +\infty$, and its limit X_∞ is finite and satisfies $X_\infty \geq 0$.

Finally, recall from Lemma 4.2 that, for α large enough, $X(\alpha)$ is a lower limit point for the set $\{R^{-1} : R > 0$ and R solves (4.5)$\}$. Now, as α goes to $+\infty$, this set

tends to $\{R^{-1} : R > 0 \text{ and } R \text{ solves (4.4)}\}$. Since (4.4) and (3.2) are equivalent [12], we conclude that X_∞ is a lower limit point for $\mathcal{R}_{\text{inv}}$. □

By duality, similar properties hold for the stabilizing solutions $Y(\alpha)$ of

$$\tilde{A}Y + Y\tilde{A}^T + Y(\gamma^{-2}C_1^TC_1 - \tilde{C}_2^T\tilde{C}_2)Y + \tilde{B}_1\tilde{B}_1^T - \alpha Y C_2^T(I - D_{21}D_{21}^+)C_2Y = 0, \tag{4.7}$$

where $\tilde{C}_2 = D_{21}^+C_2$, $\tilde{A} = A - B_1\tilde{C}_2$, and $\tilde{B}_1 = B_1(I - D_{21}^+D_{21})$. Defining Y_∞ as

$$Y_\infty := \lim_{\alpha \to +\infty} Y(\alpha), \tag{4.8}$$

the solvability of the suboptimal singular H_∞ problem can be characterized as follows (see also Theorem 4 in [5]).

THEOREM 4.7. *The singular H_∞ problem with performance γ is solvable if and only if both*

(a) *the α-dependent Riccati equations* (4.3) *and* (4.7) *have asymptotic stabilizing solutions X_∞ and Y_∞;*

(b) *X_∞ and Y_∞ further satisfy*

$$X_\infty \geq 0, \qquad Y_\infty \geq 0, \qquad \rho(X_\infty Y_\infty) \leq \gamma^2. \tag{4.9}$$

Proof. Necessity is immediate from the previous discussion. Note that the condition $\rho(X_\infty Y_\infty) \leq \gamma^2$ follows from the existence of $(R, S) \in \mathcal{R} \times \mathcal{S}$ such that $\lambda_{\min}(RS) \geq \gamma^{-2}$, and from the fact that $0 \leq X_\infty < R^{-1}$ and $0 \leq Y_\infty < S^{-1}$ for all $(R, S) \in \mathcal{R} \times \mathcal{S}$. As for sufficiency, (a) together with Lemma 4.3 and Definition 4.5 ensure that $\mathcal{R}$ and $\mathcal{S}$ are nonempty. Moreover, there exist sequences $R_n \in \mathcal{R}$ and $S_n \in \mathcal{S}$ such that $R_n^{-1} \to X_\infty$ and $S_n^{-1} \to Y_\infty$. Since $\rho(X_\infty Y_\infty) < \gamma^2$, we deduce that $\rho(R_n^{-1}S_n^{-1}) < \gamma^2$ for n large enough, or equivalently that $\lambda_{\min}(R_nS_n) > \gamma^{-2}$. Hence, there exist $R \in \mathcal{R}$ and $S \in \mathcal{S}$ such that $\lambda_{\min}(RS) \geq \gamma^{-2}$, which ensures that the H_∞ performance γ is achievable by virtue of Theorem 3.1. □

Remark 4.8. From Lemma 4.6, the asymptotic stabilizing solution X_∞ coincides with the strict lower limit point $P(\mu)$ in Theorem 13 of [32]. The matrix X_∞ is also a particular solution of the quadratic matrix inequality (QMI)

$$\begin{pmatrix} A^TX + XA + \gamma^{-2}XB_1B_1^TX + C_1^TC_1 & XB_2 + C_1^TD_{12} \\ B_2^TX + D_{12}^TC_1 & D_{12}^TD_{12} \end{pmatrix} \geq 0 \tag{4.10}$$

considered in [34]. To see this, let $V = (V_1, V_2)$ be any orthogonal matrix such that $D_{12}(V_1, V_2) = (\Delta, 0)$ with Δ full-column-rank. Postmultiplying and premultiplying by $\text{diag}(I, V)$, the QMI (4.10) reads

$$\begin{pmatrix} A^TX + XA + \gamma^{-2}XB_1B_1^TX + C_1^TC_1 & XB_2V_1 + C_1^T\Delta & XB_2V_2 \\ V_1^TB_2^TX + \Delta^TC_1 & \Delta^T\Delta & 0 \\ V_2^TB_2^TX & 0 & 0 \end{pmatrix} \geq 0,$$

which is equivalent to

$$XB_2V_2 = 0, \qquad \hat{A}^TX + X\hat{A} + X(\gamma^{-2}B_1B_1^T - \hat{B}_2\hat{B}_2^T)X + \hat{C}_1^T\hat{C}_1 \geq 0. \tag{4.11}$$

To show that X_∞ satisfies these constraints, recall that

$$\hat{A}^TX(\alpha) + X(\alpha)\hat{A} + X(\alpha)(\gamma^{-2}B_1B_1^T - \hat{B}_2\hat{B}_2^T)X(\alpha) + \hat{C}_1^T\hat{C}_1 = \alpha X(\alpha)B_2\pi_{12}B_2^TX(\alpha) \geq 0.$$

Taking the limit as $\alpha \to +\infty$ shows that X_∞ satisfies the inequality in (4.11) and that $0 = X_\infty B_2\pi_{12}B_2^TX_\infty = X_\infty B_2V_2V_2^TB_2^TX_\infty$, whence $X_\infty B_2V_2 = 0$.

5. Computation of X_∞ and Y_∞ in the singular case. From Theorem 4.7, the main task when testing the feasibility of some H_∞ performance γ is the computation of the asymptotic stabilizing solutions X_∞ and Y_∞ defined by (4.6) and (4.8). This section contains the main results of the paper. It begins with a discussion of the pencil-based algorithm proposed in [5] and brings out some numerical shortcomings of that approach. A new and numerically stable algorithm to compute X_∞ and Y_∞ is then presented. Using only SVDs, this algorithm proceeds by iterative row/column compressions of the plant matrices A, B_1, B_2, C_1, D_{12} until a reduced-order, regular subproblem has been extracted. That is, until obtaining a standard reduced-order Riccati equation that is equivalent to the LMI feasibility condition (3.2). A formal description and a justification of this algorithm are given in subsections 5.2 and 5.3. Finally, subsection 5.4 brings out insightful connections with the deflation of the Hamiltonian pencil considered in [5].

5.1. Pencil-based computation of X_∞. In principle, to determine whether a given performance γ is achievable, we could approximate X_∞ and Y_∞ by $X(\alpha)$ and $Y(\alpha)$ for α large enough and iteratively increase α until (4.3) and (4.7) have stabilizing solutions that satisfy

$$X(\alpha) \geq 0, \qquad Y(\alpha) \geq 0, \qquad \rho(X(\alpha)Y(\alpha)) \leq \gamma^2.$$

In practice, however, this scheme will run into two difficulties. First, the convergence toward X_∞ or Y_∞ may be very slow, thus requiring a large number of iterations. Worse, the computation of the Riccati solutions $X(\alpha)$ and $Y(\alpha)$ becomes ill conditioned for large values of α. For these reasons, a direct computation of X_∞ and Y_∞ is numerically desirable.

To this end, Copeland and Safonov [5] recall that the stabilizing solution $X(\alpha)$ is related to the stable eigenspace $\mathbf{S}(\alpha)$ of the matrix pencil

$$M(\alpha)-\lambda N = \begin{pmatrix} \hat{A} & \gamma^{-2}B_1B_1^T - \hat{B}_2\hat{B}_2^T & B_2\pi_{12} \\ -\hat{C}_1^T\hat{C}_1 & -\hat{A}^T & 0 \\ 0 & \pi_{12}B_2^T & \alpha^{-1}I \end{pmatrix} - \lambda \begin{pmatrix} I & 0 & 0 \\ 0 & I & 0 \\ 0 & 0 & 0 \end{pmatrix} \tag{5.1}$$

("stable eigenspace" refers to the invariant subspace associated with the *finite and stable* generalized eigenvalues) [22, 38]. Specifically, given any basis

$$\begin{pmatrix} P_\alpha \\ Q_\alpha \\ Z_\alpha \end{pmatrix}$$

of $\mathbf{S}(\alpha)$, (4.3) has a stabilizing solution if and only if P_α is square invertible, in which case $X(\alpha)$ is given by $X(\alpha) = Q_\alpha P_\alpha^{-1}$. As a result, computing X_∞ amounts to computing the limit of $\mathbf{S}(\alpha)$ as α goes to infinity. Introducing $M_\infty := \lim_{\alpha\to+\infty} M(\alpha)$, this limit is explicitly characterized in [5] as

$$\lim_{\alpha\to+\infty} \mathbf{S}(\alpha) = \text{Span} \begin{pmatrix} P_1 \\ Q_1 \\ Z_1 \end{pmatrix} \oplus \text{Span} \begin{pmatrix} P_2 \\ 0 \\ Z_2 \end{pmatrix}, \tag{5.2}$$

where the first subspace is the stable eigenspace $\mathbf{S}_\infty$ of the limit pencil $M_\infty - \lambda N$, and the columns of $\left(\begin{smallmatrix} P_2 \\ Z_2 \end{smallmatrix}\right)$ span the lower infinite eigenspace of $\left(\begin{smallmatrix} A-\lambda I & B_2 \\ C_1 & D_{12} \end{smallmatrix}\right)$ (see [5] for details). Consequently, the asymptotic stabilizing solution X_∞ of (4.3) exists if and

only if (P_1, P_2) is square and invertible, in which case it is given by

$$X_\infty = (Q_1,\ 0)\ (P_1,\ P_2)^{-1}.$$

In fact, [19] further shows that X_∞ depends only on P_1 and Q_1.

The computation of $\mathbf{S}_\infty$ requires deflation of the pencil $M_\infty - \lambda N$. To enhance numerical stability, [5] advises using the deflation algorithm 3.6 of [37] rather than the QZ algorithm because of possible generalized eigenvalues at infinity. In fact, the most pressing reason for ruling out the QZ algorithm is the possible singularity of the pencil $M_\infty - \lambda N$ (or equivalently, of the pencil $W_{12_\infty}(0, s)$ considered in [5]). Such singularities have been overlooked in [5] and complicate the computation of the required eigenspace. First, a combination of Algorithms 4.1 and 4.5 in [37] should be used rather than Algorithm 3.6. Second, the computation of the finite spectrum of a singular pencil is a badly conditioned problem [37]. Finally, the algorithms of [37] will not produce, in general, the regular decomposition assumed in [5, equation (293), p. 387]. Instead, they will deflate $M_\infty - \lambda N$ to the staircase form (2.2).

To compute the stable eigenspace $\mathbf{S}_\infty$ in the presence of the singular component $M_r - \lambda N_r$, we must either swap the diagonal blocks $M_r - \lambda N_r$ and $M_f - \lambda N_f$ or block-diagonalize the subpencil $\left(\begin{smallmatrix} M_r - \lambda N_r & \star \\ 0 & M_f - \lambda N_f \end{smallmatrix}\right)$. Both operations involve nonorthogonal equivalence transformations [8]. Besides their possible ill-conditioning, these transformations are often nonunique, in which case the remaining degrees of freedom add dimensions to the stable eigenspace $\mathbf{S}_\infty$. These various difficulties are illustrated by the simple example below.

Example 5.1. Consider a plant with state-space matrices

$$A = \begin{pmatrix} 3 & -2 \\ -2 & 0 \end{pmatrix},\ B_1 = \begin{pmatrix} 1 \\ 1 \end{pmatrix},\ B_2 = \begin{pmatrix} 1 & 5 \\ -1 & 0 \end{pmatrix},\ C_1 = \begin{pmatrix} 2 & -1 \\ 0 & 1 \end{pmatrix},\ D_{11} = 0,\ D_{12} = \begin{pmatrix} 1 & 0 \\ 0 & 0 \end{pmatrix};$$

take $\gamma = 1$; and form the pencil $M_\infty - \lambda N$. Running Algorithms 4.1 and 4.5 of [37] on this pencil yields the following deflated form (2.2):

$$M_\infty - \lambda N \equiv \left(\begin{array}{cc|cc|c} 5 & 1-\lambda & 1 & 2 & 0 \\ \hline 0 & 0 & -1-\lambda & 0 & -2 \\ 0 & 0 & 1 & 1-\lambda & 1 \\ \hline 0 & 0 & 0 & 0 & 1+\lambda \\ 0 & 0 & 0 & 0 & 5 \end{array}\right). \tag{5.3}$$

This pencil is clearly singular with regular part $M_f - \lambda N_f = \left(\begin{smallmatrix} -1-\lambda & 0 \\ 1 & 1-\lambda \end{smallmatrix}\right)$. The only stable eigenvalue being $\lambda = -1$, the stable eigenspace is spanned by the nontrivial solutions of the linear system of equations

$$\left(\begin{array}{cc|cc} 5 & 2 & 1 & 2 \\ \hline 0 & 0 & 0 & 0 \\ 0 & 0 & 1 & 2 \end{array}\right) \begin{pmatrix} x_1 \\ x_2 \\ x_3 \\ x_4 \end{pmatrix} = 0. \tag{5.4}$$

The stable eigenspace is therefore of dimension 2, whereas the stable eigenspace of $M_f - \lambda N_f$ is of dimension 1. This extra dimension comes from the nonuniqueness of the diagonalizing equivalence transformations.

In the general case when $M_f - \lambda N_f$ has several stable eigenvalues, we need either to assemble and solve the system (5.4) for each eigenvalue or to diagonalize the pencil [8] while keeping a record of the remaining degrees of freedom. Both approaches involve nonorthogonal equivalence transformations, are difficult to implement, and

will produce nonorthonormal bases for the stable eigenspace in general. A more extensive discussion of these difficulties is found in [8].

Finally, the finite spectrum of singular pencils is highly sensitive to rounding errors. For instance, a perturbation $\epsilon = 1.0 \times 10^{-10}$ in the (2,2) entry of the pencil (5.3) is enough to make its finite spectrum vanish entirely! Computing the stable eigenspace of singular pencils may thus be numerically hazardous. □

Circumventing such numerical hazards was our main motivation for seeking an alternative algorithm to compute X_∞. The results of this research are presented next.

5.2. A new algorithm to compute X_∞. We begin with an overall description of the algorithm in terms of computational procedure and major operations performed.

ALGORITHM 5.2

Purpose: Test the existence of a nonnegative asymptotic stabilizing solution X_∞ to the α-parametrized Riccati equation (4.3), and compute it when it exists.

Initialization: Set $k := 0$, $Z := I$, and

$$\mathcal{A}_0 := A, \qquad \mathcal{G}_0 := B_1, \qquad \mathcal{B}_0 := B_2, \qquad \mathcal{C}_0 := C_1, \qquad \mathcal{D}_0 := D_{12}.$$

Regularization loop: Iteratively update the matrices $\mathcal{A}_k$, $\mathcal{G}_k$, $\mathcal{B}_k$, $\mathcal{C}_k$, and $\mathcal{D}_k$ at iteration k as follows:

1. Perform a column compression on $\mathcal{D}_k$. That is, compute (via SVD, say) an orthogonal matrix $V = (V_1, V_2)$ such that

$$\mathcal{D}_k\,(V_1, V_2) = \,(\Delta\,,\,0) \tag{5.5}$$

with Δ full column rank. Terminate if $\mathcal{D}_k$ has full column rank.

2. Perform a row compression on $\mathcal{B}_k V_2$. That is, compute (via SVD, say) an orthogonal matrix $W = (W_1, W_2)$ such that

$$\begin{pmatrix} W_1^T \\ W_2^T \end{pmatrix} \mathcal{B}_k V_2 = \begin{pmatrix} 0 \\ L \end{pmatrix} \tag{5.6}$$

with L full row rank. Terminate if $\mathcal{B}_k V_2$ is either full row rank or identically zero.

3. Define

$$\begin{aligned} &\mathcal{A}_{k+1} := W_1^T\,\mathcal{A}_k\,W_1\,, \qquad \mathcal{G}_{k+1} := W_1^T\,\mathcal{G}_k\,, \qquad \mathcal{C}_{k+1} := \mathcal{C}_k\,W_1 \\ &\mathcal{B}_{k+1} := W_1^T\,(\mathcal{A}_k W_2\,,\,\mathcal{B}_k V_1)\,, \qquad \mathcal{D}_{k+1} := (\mathcal{C}_k W_2\,,\,\mathcal{D}_k V_1). \end{aligned} \tag{5.7}$$

4. Overwrite Z by ZW_1, set $k := k + 1$, and return to Step 1.

Reduced-order Riccati equation: Let K denote the number of iterations performed in the regularization loop, and let $\hat{\mathcal{A}}_K$, $\hat{\mathcal{B}}_K$, and $\hat{\mathcal{C}}_K$ be the counterparts of $\hat{A}$, $\hat{B}_2$, and $\hat{C}_1$ in (4.2) when replacing A, B_2, C_1, and D_{12} with $\mathcal{A}_K$, $\mathcal{B}_K V_1$, $\mathcal{C}_K$, and $\mathcal{D}_K V_1$.

If $\mathcal{B}_K V_2$ has full row rank, $X_\infty = 0$. Otherwise, (4.3) has a nonnegative asymptotic stabilizing solution X_∞ if and only if the Riccati equation

$$\hat{\mathcal{A}}_K^T X_r + X_r \hat{\mathcal{A}}_K + X_r\,(\gamma^{-2}\mathcal{G}_K\mathcal{G}_K^T - \hat{\mathcal{B}}_K\hat{\mathcal{B}}_K^T)\,X_r + \hat{\mathcal{C}}_K^T\hat{\mathcal{C}}_K = 0 \tag{5.8}$$

has a stabilizing solution $X_r \geq 0$, and X_∞ is then given by

$$X_\infty \;=\; Z\,X_r\,Z^T. \tag{5.9}$$

□

In simple terms, this algorithm extracts a regular subproblem from the original singular H_∞ problem. Specifically, it iteratively reduces the original plant (A, B_1, B_2,

C_1, D_{12}) to the subproblem of state-space data ($\mathcal{A}_K$, $\mathcal{G}_K$, $\mathcal{B}_K V_1$, $\mathcal{C}_K$, $\mathcal{D}_K V_1$). This subproblem is then handled in the usual way by solving the reduced-order Riccati equation (5.8). Since the size of $\mathcal{A}_k$ is strictly decreasing, the regularization loop will terminate in a finite number of steps. Note that the row and column compressions involve only numerically stable operations (SVDs). In addition, the matrix Z relating X_r to X_∞ has orthonormal columns since each W_1 satisfies $W_1^T W_1 = I$. Thus, the overall regularization amounts to an orthonormal change of coordinates that isolates the nontrivial part of X_∞ (simply observe from (5.9) that the orthogonal complement of Span(Z) lies entirely in the null space of X_∞). Incidentally, this regularization is equivalent to an implicit unitary deflation of the pencil $W_{12_\infty}(0, s)$ considered in [5], as will be seen in subsection 5.4. Finally, note that (5.9) is closely related to the identity in Theorem 13 of [32], the main difference here being the orthogonality of the coordinate transformation Z.

5.3. Justification. To justify the regularization scheme of Algorithm 5.2, we show that each iteration of the regularization loop produces an equivalent problem of smaller dimension. This property is best seen by working with the LMI-based characterization of Theorem 3.1. For notational simplicity, start with the original problem data A, B_1, B_2, C_1, D_{12} and apply one iteration of Algorithm 5.2. Denote the resulting matrices by

$$\begin{aligned} \mathcal{A} &:= W_1^T A W_1, \quad \mathcal{G} := W_1^T B_1, \quad \mathcal{B} := W_1^T (A W_2, B_2 V_1), \\ \mathcal{C} &:= C_1 W_1, \quad \mathcal{D} := (C_1 W_2, \Delta), \end{aligned} \tag{5.10}$$

where $V = (V_1, V_2)$ and $W = (W_1, W_2)$ are orthogonal matrices such that

$$D_{12}\,(V_1, V_2) = (\Delta\,,\,0), \qquad (B_2 V_2)^T\,(W_1, W_2) = (0,\,L^T). \tag{5.11}$$

(Δ and L have full column and row rank, respectively.) The next theorem shows that X_∞ is fully determined by the reduced problem of data $(\mathcal{A}, \mathcal{G}, \mathcal{B}, \mathcal{C}, \mathcal{D})$.

THEOREM 5.3. *With the notation* (5.10), (5.11) *and the assumptions* (A1)–(A3), *there is equivalence between the following two statements:*

(a) *the α-parametrized Riccati equation* (4.3) *has an asymptotic stabilizing solution $X_\infty \geq 0$;*

(b) *the counterpart of* (4.3) *resulting from the substitutions*

$$(A, B_1, B_2, C_1, D_{12}) \to (\mathcal{A}, \mathcal{G}, \mathcal{B}, \mathcal{C}, \mathcal{D})$$

has an asymptotic stabilizing solution $X_1 \geq 0$.
Moreover, X_∞ and X_1 are related by

$$X_\infty = W_1 X_1 W_1^T. \tag{5.12}$$

Proof. See Appendix B. ∎

By applying Theorem 5.3 to each iteration of the regularization loop, we can characterize X_∞ in terms of the stabilizing solution of the final reduced-order Riccati equation (5.8). This corollary completes the justification of Algorithm 5.2.

COROLLARY 5.4. *For plants satisfying* (A1)–(A3), *there is equivalence between*

(a) (4.3) *has an asymptotic stabilizing solution $X_\infty \geq 0$;*

(b) *the reduced-order Riccati equation* (5.8) *has a stabilizing solution $X_r \geq 0$ (with the convention $X_r = 0$ when $\mathcal{B}_K V_2$ has full row rank upon termination).*

If either property holds, the asymptotic stabilizing solution X_∞ of (4.3) *is given by*

$$X_\infty = Z X_r Z^T, \tag{5.13}$$

where Z is the matrix accumulated during the iterative regularization of the plant. ∎

Summing up, the regularization loop clears the way to turning the LMI (3.2) into a well-defined Riccati equation that can be solved with standard Riccati solvers [21, 2] (see section 6 for more details).

5.4. Interpretation in terms of pencil deflation. This subsection gives two interpretations of Algorithm 5.2 in terms of the deflation of

(1) the system matrix

$$P_{12}(\lambda) = \begin{pmatrix} A - \lambda I & B_2 \\ C_1 & D_{12} \end{pmatrix}, \tag{5.14}$$

(2) the extended Hamiltonian pencil associated with the H_∞ controller ARE

$$\hat{A}^T X + X\hat{A} + X(\gamma^{-2} B_1 B_1^T - \hat{B}_2 \hat{B}_2^T)X + \hat{C}_1^T \hat{C}_1 = 0. \tag{5.15}$$

Algorithm 5.2 is shown to perform an orthogonal deflation of these pencils to extract their finite eigenstructures. However, this algorithm differs from the pencil algorithm discussed in [5] in two important ways:

- the deflation is performed implicitly in a highly efficient and numerically stable manner;
- no attempt is made to compute the stable eigenspace of the Hamiltonian pencil associated with (5.15) or to deal with related difficulties when this pencil is singular (see discussion in subsection 5.1). Instead, Algorithm 5.2 extracts a reduced Hamiltonian pencil which is regular, contains all information needed to compute X_∞, and can be deflated in a numerically stable way.

First consider the effect of one iteration of Algorithm 5.2 on the system matrix $P_{12}(\lambda)$. With the notation (5.10), (5.11), we have

$$\begin{pmatrix} W^T & 0 \\ 0 & I \end{pmatrix} P_{12}(\lambda) \begin{pmatrix} W & 0 \\ 0 & V \end{pmatrix} = \left(\begin{array}{cc|cc} \mathcal{A} - \lambda I & W_1^T A W_2 & W_1^T B_2 V_1 & 0 \\ \star & \star & \star & W_2^T B_2 V_2 \\ \hline \mathcal{C} & C_1 W_2 & \Delta & 0 \end{array}\right)$$

$$\equiv \left(\begin{array}{cc|cc} W_2^T B_2 V_2 & & \star & \star \\ \hline 0 & & \mathcal{A} - \lambda I & \mathcal{B} \\ 0 & & \mathcal{C} & \mathcal{D} \end{array}\right), \tag{5.16}$$

where $\equiv$ stands for "equivalent by row and column permutations." Recalling that $W_2^T B_2 V_2$ has full row rank by construction, it is immediate from Lemma 2.1 that $P_{12}(\lambda)$ and $\left(\begin{smallmatrix} \mathcal{A}-\lambda I & \mathcal{B} \\ \mathcal{C} & \mathcal{D} \end{smallmatrix}\right)$ have the same finite zeros. Repeating the argument for each iteration, we conclude that $P_{12}(\lambda)$ and

$$P_K(\lambda) = \begin{pmatrix} \mathcal{A}_K - \lambda I & \mathcal{B}_K V_1 \\ \mathcal{C}_K & \mathcal{D}_K V_1 \end{pmatrix} \tag{5.17}$$

share the same finite eigenvalues (finite zeros). Moreover, the deflated system matrix $P_K(\lambda)$ is associated with a regular problem since $\mathcal{D}_K V_1$ has full column rank, and subsection 5.3 shows that this problem contains all the information needed to compute X_∞. A more complete interpretation in terms of infinite zero structure can be found

in [32]. From [10, 11], note that the finite zeros of $P_K(\lambda)$ are exactly the unobservable modes of

$$(\hat{\mathcal{C}}_K, \hat{\mathcal{A}}_K) = \left((I - (\mathcal{D}_K V_1)(\mathcal{D}_K V_1)^+)\mathcal{C}_K\,,\ \mathcal{A}_K - \mathcal{B}_K V_1 (\mathcal{D}_K V_1)^+ \mathcal{C}_K\right).$$

The previous analysis has an immediate counterpart in terms of extended Hamiltonian pencils. Recall that the Hamiltonian matrix associated with (5.15) is

$$H_X := \begin{pmatrix} \hat{A} & \gamma^{-2} B_1 B_1^T - \hat{B}_2 \hat{B}_2^T \\ -\hat{C}_1^T \hat{C}_1 & -\hat{A}^T \end{pmatrix}. \tag{5.18}$$

Following [38, 22, 2], we can eliminate all pseudoinversions by replacing H_X with the extended Hamiltonian pencil

$$\begin{aligned} \mathcal{H} - \lambda\mathcal{E} &= \begin{pmatrix} H_{11} & H_{12} \\ H_{21} & H_{22} \end{pmatrix} - \lambda \begin{pmatrix} I & 0 \\ 0 & 0 \end{pmatrix} \\ &:= \left(\begin{array}{cc|cc} A & \gamma^{-2} B_1 B_1^T & 0 & B_2 \\ 0 & -A^T & -C_1^T & 0 \\ \hline C_1 & 0 & -I & D_{12} \\ 0 & -B_2^T & -D_{12}^T & 0 \end{array}\right) - \lambda \left(\begin{array}{cc|cc} I & 0 & 0 & 0 \\ 0 & I & 0 & 0 \\ \hline 0 & 0 & 0 & 0 \\ 0 & 0 & 0 & 0 \end{array}\right). \end{aligned} \tag{5.19}$$

In the regular case, H_{22} is invertible since D_{12} has full column rank. Hence, $\mathcal{H} - \lambda\mathcal{E}$ is a regular pencil and given any basis

$$\begin{pmatrix} P \\ Q \\ \star \\ \star \end{pmatrix}$$

of its stable eigenspace, it is easily verified from the identity $H_X = H_{11} - H_{12} H_{22}^{-1} H_{21}$ that the columns of $\binom{P}{Q}$ span the stable eigenspace of the Hamiltonian matrix H_X. Consequently, $X = QP^{-1}$ is the stabilizing solution of (5.15) whenever P is invertible [2]. Note that the pencil $\mathcal{H} - \lambda\mathcal{E}$ is an extended version of the pencil $W_{12_\infty}(0, s)$ considered in [5].

When D_{12} is rank-deficient, by contrast, the pencil $\mathcal{H} - \lambda\mathcal{E}$ may be singular and the connection with the (asymptotic) stabilizing ARE solution becomes fuzzier. To understand the deflating action of Algorithm 5.2, observe that $\mathcal{H} - \lambda\mathcal{E}$ is equivalent by row and column permutations to

$$\mathcal{H} - \lambda\mathcal{E} \equiv \left(\begin{array}{c|c} P_{12}(\lambda) & \mathrm{Diag}(\gamma^{-2} B_1 B_1^T, -I) \\ \hline 0 & -P_{12}^T(-\lambda) \end{array}\right). \tag{5.20}$$

From the previous discussion and (5.16), we readily infer that one iteration of Algorithm 5.2 deflates the pencil $\mathcal{H} - \lambda\mathcal{E}$, via orthogonal transformations, to the equivalent pencil

$$\left(\begin{array}{c|c|c} W_2^T B_2 V_2 & \star & \star \\ \hline 0 & \mathcal{H}_1 & \star \\ \hline 0 & 0 & -V_2^T B_2^T W_2 \end{array}\right) - \lambda \left(\begin{array}{c|c|c} 0 & \star & 0 \\ \hline 0 & \mathcal{E}_1 & \star \\ \hline 0 & 0 & 0 \end{array}\right), \tag{5.21}$$

where

$$\mathcal{H}_1 - \lambda\mathcal{E}_1 := \left(\begin{array}{cc|cc} \mathcal{A} & \gamma^{-2}\mathcal{G} & 0 & \mathcal{B} \\ 0 & -\mathcal{A}^T & -\mathcal{C}^T & 0 \\ \hline \mathcal{C} & 0 & -I & \mathcal{D} \\ 0 & -\mathcal{B}^T & -\mathcal{D}^T & 0 \end{array}\right) - \lambda \left(\begin{array}{cc|cc} I & 0 & 0 & 0 \\ 0 & I & 0 & 0 \\ \hline 0 & 0 & 0 & 0 \\ 0 & 0 & 0 & 0 \end{array}\right). \tag{5.22}$$

By inspection of (5.21), (5.22), it follows that

- one iteration of Algorithm 5.2 extracts a subpencil $\mathcal{H}_1 - \lambda\mathcal{E}_1$ with the same Hamiltonian structure;
- $\mathcal{H} - \lambda\mathcal{E}$ and $\mathcal{H}_1 - \lambda\mathcal{E}_1$ have the same finite generalized eigenvalues. This follows from (5.21) together with Lemma 2.1 and the fact that $W_2^T B_2 V_2$ has full row rank;
- upon termination, either $\mathcal{H}_K$ is empty (case $X_\infty = 0$) or $\mathcal{H}_K - \lambda\mathcal{E}_K$ is the regular extended Hamiltonian pencil associated with the reduced-order ARE (5.8).

Summing up, Algorithm 5.2 amounts to a particular orthogonal deflation of the Hamiltonian pencil $\mathcal{H} - \lambda\mathcal{E}$. Unlike in [5], the stable eigenspace of $\mathcal{H} - \lambda\mathcal{E}$ is not explicitly computed. Instead, X_∞ is computed from the stable eigenspace of the reduced pencil $\mathcal{H}_K - \lambda\mathcal{E}_K$.

6. Computation of $\gamma_{\mathbf{opt}}$. This section summarizes the overall procedure for computing the optimal H_∞ performance γ_{opt} in the singular case. The algorithm is given for the general case $D_{11} \neq 0$, which is a straightforward extension of the previous results. The regularization loop of Algorithm 5.2 remains unchanged in this general setup, and its SVD-based implementation is straightforward. Once the reduced-order ARE (5.8) has been extracted, it can be solved by standard Schur techniques. To enhance numerical stability, we recommend using the generalized eigenproblem implementation described in [2] as follows.

(1) Form the (regular) extended pencil

$$\mathcal{H}_K - \lambda\mathcal{E}_K = \left(\begin{array}{cc|ccc} \mathcal{A}_K & 0 & 0 & \gamma^{-1}\mathcal{G}_K & \mathcal{B}_K V_1 \\ 0 & -\mathcal{A}_K^T & -\mathcal{C}_K^T & 0 & 0 \\ \hline \mathcal{C}_K & 0 & -I & \gamma^{-1}D_{11} & \mathcal{D}_K V_1 \\ 0 & \gamma^{-1}\mathcal{G}_K^T & \gamma^{-1}D_{11}^T & -I & 0 \\ 0 & (\mathcal{B}_K V_1)^T & (\mathcal{D}_K V_1)^T & 0 & 0 \end{array}\right) - \lambda \left(\begin{array}{cc|ccc} I & 0 & 0 & 0 & 0 \\ 0 & I & 0 & 0 & 0 \\ \hline 0 & 0 & 0 & 0 & 0 \\ 0 & 0 & 0 & 0 & 0 \\ 0 & 0 & 0 & 0 & 0 \end{array}\right),$$

where $\mathcal{A}_K$, $\mathcal{G}_K$, $\mathcal{B}_K$, $\mathcal{C}_K$, $\mathcal{D}_K$, and V_1 are the matrices returned by Algorithm 5.2 upon termination.

(2) Eliminate its infinite eigenstructure by using Algorithm 3.6 of [37].

(3) Compute an orthonormal basis for its stable eigenspace by using the QZ algorithm with reordering of the stable eigenvalues [38].

Algorithm 5.2 is readily dualized to test the existence of Y_∞ and compute it: simply replace $(A, B_1, B_2, C_1, D_{11}, D_{12})$ with $(A^T, C_1^T, C_2^T, B_1^T, D_{11}^T, D_{21}^T)$ in the initialization step.

Combining this regularization algorithm with the characterization of Theorem 4.7, the optimal H_∞ performance γ_{opt} is estimated to the desired accuracy via the usual γ-iteration scheme. Note that the trick proposed in [29] to avoid numerical difficulties when testing $X_\infty \geq 0$ or $Y_\infty \geq 0$ is also applicable to our context.

7. Numerical stability and related issues. This section discusses the numerical reliability of the algorithm outlined in section 6. Recall that the main difficulty is the computation of X_∞ and Y_∞. Since Algorithm 5.2 relies on SVDs and since the SVD is a backward stable operation, the computed γ_{opt} is the exact value for a nearby problem (i.e., for slightly perturbed values of the plant matrices). Yet this is not enough to guarantee that this computed value is meaningful from a control viewpoint. Indeed, it was shown in [13] that the optimal H_∞ gain γ_{opt} may be discontinuous near singular problems. In other words, arbitrarily small perturbations of

the plant data may result in large variations of $\gamma_{\rm opt}$. To understand how this might affect Algorithm 5.2, we first recall a few important facts about such discontinuities.

7.1. Discontinuities of $\gamma_{\rm opt}$. As shown in [13], $\gamma_{\rm opt}$ is an upper semicontinuous function of the plant data that may be discontinuous near singular plants. Consider a singular plant $P(s)$ with optimal performance $\gamma^* := \gamma_{\rm opt}(P)$, and suppose that $\gamma_{\rm opt}$ is discontinuous at P with a gap $\delta > 0$ between its upper and lower limits. Then there exist arbitrarily small perturbations P_ϵ of P such that

$$\gamma_{\rm opt}(P_\epsilon) < \gamma^* - \delta/2 \; .$$

While the value $\gamma_{\rm opt}(P_\epsilon)$ is mathematically correct, it is clearly not meaningful from a control standpoint. Indeed, the H_∞ performance should be robust to small perturbations in the state-space data, while here $\gamma_{\rm opt}(P_\epsilon)$ jumps up by at least $\delta/2$ (a possibly large amount) when P_ϵ is perturbed to P.

For these reasons, we claim that only the upper limit γ^* is meaningful in the vicinity of the singular plant P. This value is typically insensitive to small perturbations of the data and can be achieved with reasonable control effort. In contrast, the lower limit(s) are only achievable via high gain or marginal closed-loop stability.

7.2. Implementation tips. Based on the previous discussion, we recommend the following "singularity-preferring" policy: if the plant is nearly singular, make it singular and compute the optimal H_∞ performance for the resulting singular plant. This amounts to computing the upper limit of $\gamma_{\rm opt}$ in the vicinity of the plant. Now, singularity has to do with the rank of the matrices $\mathcal{D}_k$ and $\mathcal{B}_k V_2$ in Algorithm 5.2. (The regularization is complete when these matrices have full rank.) Thus, "nearly singular" is equivalent to "nearly rank-deficient," which in turn has to do with small singular values. Due to the finite precision of computer arithmetic, singular values should be considered zero when they fall below some adequate relative tolerance TOL:

SVD truncation rule: Given the SVD of a matrix M, zero all singular values smaller than $\text{TOL} \times \sigma_{\max}(M)$ where $\sigma_{\max}(M)$ denotes the largest singular value of M.

To implement our singularity-preferring policy, it suffices to set TOL to some conservative value, e.g., the square root of the relative machine precision. This will automatically turn nearly singular problems into truly singular ones. Note that steering clear of nearly rank-deficient matrices also improves the numerical conditioning of the reduced Riccati equation (5.8).

8. Numerical experiments. The following simple example illustrates the performance of Algorithm 5.2.

Example 8.1. Consider the plant data

$$A = \begin{pmatrix} 0 & -1 & 2 \\ 1 & -2 & 3 \\ 0 & 1 & 0 \end{pmatrix}, \quad B_1 = \begin{pmatrix} 1 \\ -1 \\ 0 \end{pmatrix}, \quad B_2 = \begin{pmatrix} 1 & 0 \\ 0 & 0 \\ 0 & -1 \end{pmatrix},$$

$$C_1 = \begin{pmatrix} 1 & 0 & 1 \\ 1 & 0 & 1 \end{pmatrix}, \quad D_{12} = \begin{pmatrix} 0 & 1 \\ 0 & 0 \end{pmatrix}.$$

Clearly D_{12} is rank-deficient and the standard Riccati equation for X_∞ is ill-defined for this problem.

With this data and $\gamma = 1$, Algorithm 5.2 performs one regularization iteration and returns the asymptotic stabilizing solution

$$X_\infty = \begin{pmatrix} 0 & 0 & 0 \\ 0 & 0.06857 & 0.18733 \\ 0 & 0.18733 & 0.51180 \end{pmatrix}.$$

The eigenvalues of X_∞ are $0.58037, 0, 0$.

By comparison, the eigenvalues of the computed solution $X(\alpha)$ of (4.3) for $\alpha = 10^{10}$ are $0.5804, 1.12 \times 10^{-5}$, and 7.73×10^{-12}. This illustrates the relatively slow convergence of $X(\alpha)$ to X_∞. Note that $\alpha = 10^{10}$ corresponds to the following ϵ-regularization of the true plant data:

$$C_1 \to \begin{pmatrix} 1 & 0 & 1 \\ 1 & 0 & 1 \\ 0 & 0 & 0 \end{pmatrix}, \qquad D_{12} \to \begin{pmatrix} 0 & 1 \\ 0 & 0 \\ 10^{-5} & 0 \end{pmatrix}. \qquad \square$$

Algorithm 5.2 has been implemented in MATLAB's *LMI Control Toolbox* [16] developed by the authors and Nemirovski. To solve the reduced-order Riccati equations we used the Riccati solvers developed by Laub [23]. The algorithm was tested on a large number of randomly generated singular problems, and the computed γ_{opt} was compared with (1) the optimal value obtained from LMI-based optimization using Nestorov and Nemirovski's projective method [14, 25, 16] and (2) the optimal value obtained with a standard γ-iteration algorithm after ϵ-regularization of the plant. Algorithm 5.2 proved extremely reliable and always returned an optimal value within 1% of the LMI-based optimum. In contrast, the ϵ-regularization scheme proved very sensitive to the choice of ϵ. Algorithm 5.2 therefore appears as a fast and reliable alternative to ϵ-regularization for singular problems.

9. Concluding remarks. We have presented a new and numerically stable algorithm to compute the optimal H_∞ gain γ_{opt} when D_{12} or D_{21} are rank-deficient. Its efficiency and reliability have been confirmed by extensive numerical testing.

Given γ_{opt}, there remains the issue of computing (sub)optimal controllers that achieve some performance $\gamma \geq \gamma_{\text{opt}}$. As pointed out in [5], the usual central controller formulas applied to the asymptotic stabilizing solutions X_∞ and Y_∞ do not always yield a proper or well-posed controller. In fact, there does not seem to be any obvious way of deriving adequate controllers from knowledge of X_∞ and Y_∞ (see [15] for insight drawn from the LMI approach). Consequently, the ϵ-regularization of the plant data remains useful for the computation of the H_∞ controller. Even so, exact knowledge of the optimal performance prior to any regularization proves valuable for two reasons. First, it eliminates the dependence of the computed optimal performance on the chosen regularization level ϵ. Here the true value is obtained at once and without the possible numerical instability introduced by ϵ. Second, increasing ϵ tends to improve numerical conditioning and to yield better-behaved controllers. Once some performance $\gamma > 0$ has been unambiguously diagnosed as feasible with Algorithm 5.2, it is easy to maximize ϵ subject to γ remaining feasible.

Appendix A.

Proof of Lemma 4.2. The proof is adapted from [31]. Temporarily assuming that $(C, -A)$ is detectable, Theorem 2.23 of [31] shows that the set

$$\mathcal{G} := \{R = R^T : AR + RA^T + RC^TCR + F < 0\}$$

is nonempty if and only if the Riccati equation

$$-AR - RA^T - RC^TCR - F = 0 \tag{A.1}$$

has a stabilizing solution $R_{\rm st}$. Moreover, it is shown that $R_{\rm st}$ is then an upper limit point of $\mathcal{G}$.

If $\mathcal{R} \neq \emptyset$, the set $\mathcal{G}$ contains a positive definite element R_0 and consequently $R_{\rm st} > R_0$ is positive definite and invertible. Defining $X_{\rm st} := R_{\rm st}^{-1}$, it is readily verified that $X_{\rm st}$ is a positive definite and stabilizing solution of $A^TX+XA+XFX+C^TC = 0$. In addition, from $0 < R < R_{\rm st}$ for all $R \in \mathcal{R}$, we deduce that $0 < X_{\rm st} < R^{-1}$ for all $R \in \mathcal{R}$. This, together with the upper limiting property of $R_{\rm st}$, guarantees that $X_{\rm st}$ is a lower limit point of $\mathcal{R}_{\rm inv}$.

Conversely, if $A^TX + XA + XFX + C^TC = 0$ has a stabilizing solution $X_{\rm st} > 0$, it retains a positive definite stabilizing solution $X_{\rm st}(\epsilon)$ when C^TC is perturbed to $C^TC + \epsilon I$ with $\epsilon > 0$ small enough [7]. Hence, $\mathcal{R}$ is nonempty since $X_{\rm st}^{-1}(\epsilon) \in \mathcal{R}$.

Finally, the case where (C, A) has stable unobservable modes can be handled by a limiting argument (see, e.g., Lemma 8.1 in [12]). Note that the stable unobservable subspace of (C, A) determines the null space of $X_{\rm st}$. ∎

Appendix B. We begin with a useful technical lemma.

LEMMA B.1 (Projection lemma) [12, 4]. *Consider a symmetric matrix M and matrices P, Q of compatible dimensions, and let W_P and W_Q be any matrices whose columns form bases for the null spaces of P and Q, respectively.*

With this notation, the matrix inequality $M + Q^TXP + P^TX^TQ < 0$ is solvable for X if and only if the inequalities $W_P^T\,M\,W_P < 0$ and $W_Q^T\,M\,W_Q < 0$ hold.

The regularizing effect of Algorithm 5.2 is best analyzed in the LMI framework, that is, using the LMI characterization of Theorem 3.1. The implications in terms of asymptotic stabilizing Riccati solutions are then easily deduced from the results of section 4. The next theorem shows that one regularization iteration reduces the characteristic LMI (3.2) to a smaller LMI with the same structure and where A, B_1, B_2, C_1, and D_{12} are replaced by $\mathcal{A}$, $\mathcal{G}$, $\mathcal{B}$, $\mathcal{C}$, and $\mathcal{D}$, respectively.

THEOREM B.2. *Assume that D_{12} is column-rank deficient. With the notation* (5.10), (5.11) *there exists a symmetric matrix R satisfying the LMI*

$$\mathcal{N}_{12}^T \begin{pmatrix} AR + RA^T + \gamma^{-2}B_1B_1^T & RC_1^T \\ C_1R & -I \end{pmatrix} \mathcal{N}_{12} < 0 \tag{B.1}$$

if and only if there exists a symmetric matrix R_1 satisfying the reduced-order LMI

$$\tilde{\mathcal{N}}^T \begin{pmatrix} \mathcal{A}R_1 + R_1\mathcal{A}^T + \gamma^{-2}\mathcal{G}\mathcal{G}^T & R_1\mathcal{C}^T \\ \mathcal{C}R_1 & -I \end{pmatrix} \tilde{\mathcal{N}} < 0, \tag{B.2}$$

where $\tilde{\mathcal{N}}$ is any orthonormal basis for the null space of $(\mathcal{B}^T, \mathcal{D}^T)$. Moreover, all solutions of (B.1) *are of the form*

$$R = W \begin{pmatrix} R_1 & \star \\ (\star)^T & \Psi \end{pmatrix} W^T, \tag{B.3}$$

where R_1 solves (B.2) *and Ψ is an arbitrary symmetric matrix.*

Proof. Applying Lemma 4.1 with $P = (B_2^T, D_{12}^T)$, the LMI (B.1) is feasible if and only if there exists $R = R^T$ such that

$$\begin{pmatrix} AR + RA^T + \gamma^{-2}B_1B_1^T & RC_1^T \\ C_1R & -I \end{pmatrix} - \alpha \begin{pmatrix} B_2 \\ D_{12} \end{pmatrix} \begin{pmatrix} B_2 \\ D_{12} \end{pmatrix}^T < 0\,, \qquad \alpha > 0\,. \tag{B.4}$$

Using the decomposition $\left(\begin{smallmatrix} B_2 \\ D_{12} \end{smallmatrix}\right)V = \left(\begin{smallmatrix} B_2V_1 & B_2V_2 \\ \Delta & 0 \end{smallmatrix}\right)$, (B.4) can also be written as

$$\begin{pmatrix} AR + RA^T + \gamma^{-2}B_1B_1^T & RC_1^T \\ C_1R & -I \end{pmatrix} - \alpha \begin{pmatrix} B_2V_1 \\ \Delta \end{pmatrix}\begin{pmatrix} B_2V_1 \\ \Delta \end{pmatrix}^T$$
$$-\alpha \begin{pmatrix} B_2V_2 \\ 0 \end{pmatrix}\begin{pmatrix} B_2V_2 \\ 0 \end{pmatrix}^T < 0.$$

Recalling that W_1 is a basis for the null space of $V_2^TB_2^T$ by construction, we can eliminate the last term by invoking Lemma 4.1 once again, this time with $P = (V_2^TB_2^T, 0)$ and $W_P = \left(\begin{smallmatrix} W_1 & 0 \\ 0 & I \end{smallmatrix}\right)$. It follows that (B.4) is feasible if and only if

$$\begin{pmatrix} W_1^T(AR + RA^T + \gamma^{-2}B_1B_1^T)W_1 & W_1^TRC_1^T \\ C_1RW_1 & -I \end{pmatrix} - \alpha \begin{pmatrix} W_1^TB_2V_1 \\ \Delta \end{pmatrix}\begin{pmatrix} W_1^TB_2V_1 \\ \Delta \end{pmatrix}^T < 0 \tag{B.5}$$

holds for some $\alpha > 0$. With the notation (5.10), (5.11) and

$$(R_1, R_2) := W_1^TR\,(W_1, W_2), \qquad \Omega_1 := \begin{pmatrix} W_1^TAW_2 \\ C_1W_2 \end{pmatrix}, \qquad \Omega_2 := \begin{pmatrix} W_1^TB_2V_1 \\ \Delta \end{pmatrix},$$

condition (B.5) is equivalent to the existence of $\alpha > 0$ and of matrices $R_1 = R_1^T$ and R_2 such that

$$\left\{\begin{pmatrix} \mathcal{A}R_1 + R_1\mathcal{A}^T + \gamma^{-2}\mathcal{G}\mathcal{G}^T & R_1\mathcal{C}^T \\ \mathcal{C}R_1 & -I \end{pmatrix} - \alpha\,\Omega_2\Omega_2^T\right\} + \begin{pmatrix} I \\ 0 \end{pmatrix} R_2\,\Omega_1^T + \Omega_1R_2^T\,(I\,,\,0) < 0. \tag{B.6}$$

To obtain the reduced LMI (B.2), it now suffices to eliminate the variable R_2 by invoking Lemma B.1 with $P = \Omega_1^T$ and $Q = (I\,,\,0)$. Observing that the projected inequality $W_Q^T\,M\,W_Q < 0$ is trivial here, we conclude that (B.4) is feasible if and only if the other projected inequality $W_P^T\,M\,W_P < 0$ holds or, equivalently from Lemma 4.1, if and only if the LMI

$$\left\{\begin{pmatrix} \mathcal{A}R_1 + R_1\mathcal{A}^T + \gamma^{-2}\mathcal{G}\mathcal{G}^T & R_1\mathcal{C}^T \\ \mathcal{C}R_1 & -I \end{pmatrix} - \alpha\,\Omega_2\Omega_2^T\right\} - \beta\,\Omega_1\Omega_1^T < 0$$

holds for some symmetric R_1 and positive scalars α, β. Without loss of generality, we may take $\alpha = \beta$ and group the last two terms into the single term

$$-\alpha(\Omega_1, \Omega_2)\begin{pmatrix} \Omega_1^T \\ \Omega_2^T \end{pmatrix}.$$

Then the LMI condition (B.2) readily follows by applying Lemma 4.1 with $P = (\Omega_1, \Omega_2)^T = (\mathcal{B}^T, \mathcal{D}^T)$ and $W_P = \tilde{\mathcal{N}}$.

By construction, the solutions R of (B.1) are related to the solutions R_1 of (B.2) by

$$R = W\begin{pmatrix} R_1 & R_2 \\ R_2^T & \Psi \end{pmatrix}W^T,$$

where R_2 satisfies (B.6) and $\Psi = \Psi^T$ is arbitrary: the proof is complete. □

To prove Theorem 5.3, it now suffices to reinterpret this result in terms of asymptotic stabilizing solutions of the corresponding AREs.

Proof of Theorem 5.3. First observe that Theorem B.2 can be strengthened to positive definite solutions of the LMIs (B.1), (B.2). Specifically, if $R > 0$ solves (B.1), then $R_1 = W_1^T R W_1$ is a positive definite solution of (B.2) from the proof of Theorem B.2. Conversely, given a solution $R_1 > 0$ of (3.3) and any R_2 solving (B.6), the matrix

$$R := W \begin{pmatrix} R_1 & R_2 \\ R_2^T & \Psi \end{pmatrix} W^T \tag{B.7}$$

solves (B.1) for any symmetric Ψ, and Ψ can always be chosen to make R positive definite.

Now, from Lemma 4.3 together with (A3), the LMI (B.1) has positive definite solutions if and only if (a) holds. Similarly, (B.2) has positive definite solutions if and only if (b) holds since the pencil

$$P(\lambda) = \begin{pmatrix} \mathcal{A} - \lambda I & \mathcal{B} \\ \mathcal{C} & \mathcal{D} \end{pmatrix}$$

has no finite eigenvalue on the imaginary axis. Indeed, $P_{12}(\lambda)$ and $P(\lambda)$ share the same finite spectra as established in subsection 5.4. Consequently, (a) and (b) are equivalent.

This leaves us with proving the identity $X_\infty = W_1 X_1 W_1^T$. Schematically, this identity stems from the correspondence (B.7) between solutions of the LMIs (B.1) and (B.2) as well the fact that X_∞ and X_1 are lower limit points for these LMIs (see Lemma 4.2). Specifically, $X_\infty \le R^{-1}$ for all solutions $R > 0$ of (B.1) while $X_1 \le R_1^{-1}$ for all solutions $R_1 > 0$ of (B.2).

Let $R > 0$ solve (B.1), and partition R as in (B.7). By the standard inversion formula for 2×2 block matrices, we have

$$\begin{aligned} W^T R^{-1} W &= \begin{pmatrix} R_1^{-1} & 0 \\ 0 & 0 \end{pmatrix} + \begin{pmatrix} R_1^{-1} R_2 \\ -I \end{pmatrix} (\Psi - R_2^T R_1^{-1} R_2)^{-1} \begin{pmatrix} R_1^{-1} R_2 \\ -I \end{pmatrix}^T \\ &\ge \begin{pmatrix} R_1^{-1} & 0 \\ 0 & 0 \end{pmatrix}. \end{aligned}$$

Since R_1 solves (B.2), it follows that $R^{-1} \ge W_1 R_1^{-1} W_1^T \ge W_1 X_1 W_1^T$. Now this holds for any solution $R > 0$ of (B.1), whence $X_\infty \ge W_1 X_1 W_1^T$.

Conversely, for any $R_1 > 0$ solving (B.2), construct a solution $R > 0$ of (B.1) as indicated above and choose $\Psi = \alpha I$ with $\alpha > 0$ large enough. Then

$$W \begin{pmatrix} R_1 & R_2 \\ R_2^T & \alpha I \end{pmatrix}^{-1} W^T = R^{-1} \ge X_\infty,$$

which, by letting α go to $+\infty$, ensures that $W_1 R_1^{-1} W_1^T \ge X_\infty$. Since this must hold for any $R_1 > 0$ solving (B.2), we infer that $W_1 X_1 W_1^T \ge X_\infty$, and the proof is complete. ∎

By applying these results to each iteration, it becomes clear that Algorithm 5.2 generates a sequence of LMIs that are equivalent to (B.1) and of decreasing order (i.e., the size of the matrix variable R is strictly decreasing). From the algorithm description, the regularization loop terminates with either of the following two situations.

Termination type (A): $\mathcal{B}_K V_2$ has full row rank. Then W_1 is "empty," (B.5) is trivially satisfied, and any $R > 0$ solves (B.1). From Lemma 4.6, this case yields $X_\infty = 0$.

Termination Type (B): $\mathcal{D}_K$ has full column rank or $\mathcal{B}_K V_2 = 0$. In both cases, the resulting data $\mathcal{A}_K, \mathcal{G}_K, \mathcal{B}_K, \mathcal{C}_K, \mathcal{D}_K$ define a *regular* problem, and the final "reduced" LMI is equivalent to the Riccati inequality (with the notation of Algorithm 5.2):

$$\hat{\mathcal{A}}_K R_r + R_r \hat{\mathcal{A}}_K^T + R_r\, \hat{\mathcal{C}}_K^T \hat{\mathcal{C}}_K\, R_r + \gamma^{-2} \mathcal{G}_K \mathcal{G}_K^T - \hat{\mathcal{B}}_K \hat{\mathcal{B}}_K^T < 0. \tag{B.8}$$

When $\mathcal{D}_K$ has full column rank, this last equivalence readily follows from subsection 3.3. When $\mathcal{B}_K V_2 = 0$ instead, the identity

$$\begin{pmatrix} \mathcal{B}_K \\ \mathcal{D}_K \end{pmatrix} V_2 = 0$$

shows that the input directions along V_2 are disconnected from both the plant dynamics and the output z; hence, they can be discarded to obtain a regular problem.

REFERENCES

[1] B.D.O. Anderson, D.J. Clements, A.J. Laub, and J.B. Matson, *Spectral factorizations with imaginary axis zeros*, in final preparation.

[2] W.F. Arnold and A.J. Laub, *Generalized eigenproblem algorithms and software for algebraic Riccati equations*, Proc. IEEE, 72 (1984), pp. 1746–1754.

[3] S.P. Boyd, and L. El Ghaoui, *Method of centers for minimizing generalized eigenvalues*, Linear Algebra Appl., 188 (1993), pp. 63–111.

[4] S. Boyd, L. El Ghaoui, E. Feron, and V. Balakrishnan, *Linear Matrix Inequalities in Systems and Control Theory*, SIAM, Philadelphia, 1994.

[5] B.R. Copeland and M.G. Safonov, *A generalized eigenproblem solution for singular H^2 and H^∞ problems*, in Control and Dynamic Systems, vol. 50, C.T. Leondes, ed., Academic Press, San Diego, 1992.

[6] B.R. Copeland and M.G. Safonov, *A zero compensation approach to singular H_2 and H_∞ problems*, Internat. J. Robust Nonlinear Control, 5 (1995), pp. 71–106.

[7] D.F. Delchamps, *A note on the analyticity of the Riccati metric*, in Algebraic and Geometric Methods in Linear Systems Theory, Lecture Notes in Applied Mathematics, 18, American Mathematical Society, Providence, RI, 1980, pp. 37–41.

[8] J.W. Demmel and B. Kågström, *Computing stable eigendecompositions of matrix pencils*, Linear Algebra Appl., 88/89 (1987), pp. 139–186.

[9] J.C. Doyle, K. Glover, P. Khargonekar, and B. Francis, *State-space solutions to standard H_2 and H_∞ control problems*, IEEE Trans. Automat. Control, 34 (1989), pp. 831–847.

[10] A. Emami-Naeini and P.M. Van Dooren, *On computation of transmission zeros and transfer functions*, in Proc. Control and Decision Conf., 1982, pp. 51–55.

[11] P. Gahinet, *On the game Riccati equations arising in H_∞ control problems*, SIAM J. Control Optim., 32 (1994), pp. 635–647.

[12] P. Gahinet and P. Apkarian, *A linear matrix inequality approach to H_∞ control*, Internat. J. Robust and Nonlinear Control, 4 (1994), pp. 421–448.

[13] P. Gahinet and A. Stoorvogel, *Continuity properties of the H_∞ optimal gain*, in Proc. European Control Conf., 1993, pp. 1171–1175.

[14] P. Gahinet and A. Nemirovski, *General-purpose LMI solvers with benchmarks*, in Proc. Control and Decision Conf., 1993, pp. 3162–3165.

[15] P. Gahinet, *Explicit controller formulas for LMI-based H_∞ synthesis*, Automatica J. IFAC, 32 (1996), pp. 1007–1014.

[16] P. Gahinet, A. Nemirovski, A.J. Laub, and M. Chilali, *LMI Control Toolbox*, The MathWorks Inc., Natick, MA, 1995.

[17] F.R. Gantmacher, *Theory of Matrices*, Vols. I and II, Chelsea, New York, 1959.

[18] K. Glover and J.C. Doyle, *State-space formulae for all stabilizing controllers that satisfy an H_∞-norm bound and relations to risk sensitivity*, Systems Control Lett., 11 (1988), pp. 167–172.

[19] K.C. Goh and M.G. Safonov, *The Extended $j\omega$-axis Eigenstructure of a Hamiltonian Matrix Pencil*, in Proc. Control and Decision Conf., 1992, pp. 1897–1902.

[20] M. Green, K. Glover, D. Limebeer, and J. Doyle, *A J-spectral factorization approach to H_∞ optimization*, SIAM J. Control Optim., 28 (1990), pp. 1350–1371.

[21] A.J. Laub, *A Schur method for solving algebraic Riccati equations*, IEEE Trans. Automat. Control, AC-24 (1979), pp. 913–921.
[22] A.J. Laub, *Schur techniques in invariant imbedding methods for solving two-point boundary value problems*, in Proc. Control and Decision Conf., 1982, pp. 56–61.
[23] A.J. Laub and P. Gahinet, *Numerical improvements for solving Riccati equations*, IEEE Trans. Automat. Control, to appear.
[24] T. Iwasaki and R.E. Skelton, *All controllers for the general H_∞ control problem: LMI existence conditions and state-space formulas*, Automatica J. IFAC, 30 (1994), pp. 1307–1317.
[25] P. Gahinet and A. Nemirovski, *The projective method for solving linear matrix inequalities*, Math. Programming Series B, 77 (1997), pp. 163–190.
[26] Yu. Nesterov and A. Nemirovski, *Interior Point Polynomial Methods in Convex Programming: Theory and Applications*, SIAM, Philadelphia, 1994.
[27] H.H. Niemann and J. Stoustrup, *A CACSD package for H_∞ and LTR design*, in Proc. Amer. Control Conf., 1993, pp. 2327–2331.
[28] I.R. Petersen, *Stabilization of an uncertain linear system in which uncertain parameters enter into the input matrix*, SIAM J. Control Optim., 26 (1988), pp. 1257–1264.
[29] M.G. Safonov, D.J. Limebeer, and R.Y. Chiang, *Simplifying the H_∞ theory via loop-shifting, matrix-pencil and descriptor concepts*, Internat. J. Control, 50 (1989), pp. 2467–2488.
[30] C. Scherer, *H_∞-control by state-feedback and fast algorithms for the computation of optimal H_∞-norms*, IEEE Trans. Automat. Control, 35 (1990), pp. 1090–1099.
[31] C. Scherer, *The Riccati Inequality and State-Space H_∞-Optimal Control*, Ph.D. thesis, University of Wurzburg, Wurzburg, Germany, 1990.
[32] C. Scherer, *H_∞ optimization without assumptions on finite or infinite zeros*, SIAM J. Control Optim., 30 (1992), pp. 143–166.
[33] A.A. Stoorvogel, *The singular minimum entropy H_∞ control problem*, Systems Control Lett., 16 (1991), pp. 411–422.
[34] A.A. Stoorvogel, *The singular H_∞ control problem with dynamic measurement feedback*, SIAM J. Control Optim., 29 (1991), pp. 160–184.
[35] F. Uhlig, *A recurring theorem about pairs of quadratic forms and extensions*, Linear Algebra Appl., 25 (1979), pp. 219–237.
[36] L. Vandenberghe and S. Boyd, *Primal-Dual potential reduction method for problems involving matrix inequalities*, Math. Programming Series B, 69 (1995), pp. 205–236.
[37] P. Van Dooren, *The computation of Kronecker's canonical form of a singular pencil*, Linear Algebra Appl., 27 (1979), pp. 103–140.
[38] P. Van Dooren, *A generalized eigenvalue approach for solving Riccati equations*, SIAM J. Sci. Statist. Comput., 2 (1981), pp. 121–135.
[39] H.K. Wimmer, *Monotonicity of maximal solutions of algebraic Riccati equations*, Systems Control Lett., 5 (1985), pp. 317–319.

SIAM J. CONTROL OPTIM.
Vol. 35, No. 5, pp. 1711–1738, September 1997

015

THE RELATIONSHIP BETWEEN THE MAXIMUM PRINCIPLE AND DYNAMIC PROGRAMMING FOR THE CONTROL OF PARABOLIC VARIATIONAL INEQUALITIES*

CĂTĂLIN POPA†

Abstract. We show that the well-known relationship between the dual extremal arc in the maximum principle and the optimal value function (of dynamic programming), calculated on the optimal trajectory, is valid for the control of parabolic variational inequalities. It follows that every optimal control is given by a feedback law. In the case when the functions defining the performance index are convex also with respect to the state variable, a more specific result is obtained.

Key words. optimal control, dual extremal arc, optimal value function, parabolic variational inequality, optimality conditions, dynamic programming principle, feedback law

AMS subject classifications. 49L10, 49K20, 35K85

PII. S0363012994198585

1. Introduction. The aim of this paper is to establish the expected relationship between the maximum principle and dynamic programming for optimal control problems governed by a certain class of variational inequalities of parabolic type, including the parabolic obstacle problem and semilinear parabolic equations.

Let Ω be an open and bounded subset of $\mathbf{R}^N$ having a sufficiently smooth boundary. The control system we deal with is described by an abstract version of the following nonlinear parabolic equation (with boundary and initial conditions):

$$\begin{cases} \dfrac{\partial y}{\partial t} + A_0 y + \beta(y) \ni Bu + f \text{ a.e. in } Q = (0,T)\times\Omega, \\ \alpha_1 y + \alpha_2 \dfrac{\partial y}{\partial \nu} = 0 \text{ on } (0,T)\times\partial\Omega, \\ y(0,x) = y_0(x) \text{ in } \Omega, \end{cases} \tag{1.1}$$

where $-A_0$ is a strictly elliptic differential operator in Ω having its principal part in divergence form, β is a maximal monotone graph in $\mathbf{R}^2$, and B is a linear continuous operator from a given Hilbert space $\mathcal{U}$ (the control space) to $L^2(\Omega)$ (the state space). Of course, the cases of semilinear parabolic equations and parabolic obstacle problem are covered by (1.1).

In the present paper we shall be concerned with a certain class of free final state optimal control problems of Bolza type governed by (1.1), in which the involved integrand is convex, lower semicontinuous and possibly $+\infty$ with respect to the control variable u, and Lipschitz continuous (on bounded subsets of $L^2(\Omega)$) with respect to the state variable y. For such problems, V. Barbu has obtained in [2] (see also [1] for the convex case) first-order necessary conditions of optimality (the maximum principle) by an approach involving their regularization and penalization. For systems governed by the parabolic obstacle problem or by semilinear parabolic equations, these conditions (in fact the adjoint equations) take a more explicit form.

Now let $(t,y) \mapsto V(t,y)$ be the optimal value function of dynamic programming associated with one of the optimal control problems of the considered class. Our main

*Received by the editors November 22, 1994; accepted for publication (in revised form) July 15, 1996.

http://www.siam.org/journals/sicon/35-5/19858.html

†Facultatea de Matematică, Universitatea "Al. I. Cuza," Bdul. Copou 11, 6600 Iaşi, Romania.

result states that, for every optimal pair (u^*, y^*) of the fixed problem, we can select a dual extremal arc p which besides the optimality conditions (in Barbu's form) also satisfies the following inclusion:

$$-p(t) \in \partial_y V(t, y^*(t)) \text{ a.e. } t \in (0, T). \tag{1.2}$$

This is the above-mentioned connection between the maximum principle and dynamic programming. Moreover, we shall prove that it is possible to choose p such that the following additional inclusion is simultaneously verified:

$$-p(0) \in \partial_y V(0, y_0). \tag{1.3}$$

However, when the domain of the elliptic operator $y \mapsto A_0 y + \beta(y)$ is not dense in $L^2(\Omega)$ but its closure is a convex cone (this is the case of parabolic obstacle problem), a less precise result is established: the inclusions (1.2) and (1.3) must be adjusted by adding the normal cones to the closure of the domain at its vertex and at y_0, respectively.

In the finite-dimensional case, the inclusions (1.2), (1.3) have been proved for a very large class of nonsmooth free endpoint problems by F. H. Clarke and R. B. Vinter in [7]. Let us describe in a few words the ingenious idea behind their approach. According to the dynamic programming principle, they introduce an auxiliary control variable v (in our notation) into the state equation, corresponding to the new term

$$-\int_0^T \inf\{(p, v(t)) : p \in \partial_y V(t, y), |y - y^*(t)| \le \delta\}\, dt$$

in the performance index. So the optimal control problem obtained in this way is solved by the same optimal pair as the original one. The part of the maximum principle which refers to the additional control variable is just the relation (1.2), as $\delta \to 0$.

This paper shows that the above idea works also in our infinite-dimensional framework. But here we shall combine it with Barbu's approach to the optimality conditions for the control of parabolic variational inequalities. However, some specific difficulties arise when we try to adapt the approach in [7] to our case. First, the way used there to prove that the given optimal pair solves also the auxiliary optimal control problem involves a certain approximation procedure for the new state equation. Expressing this approximation in the form of an integral equation, one easily shows that it converges uniformly on $[0, T]$. But this does not work here mainly due to the fact that the operator $y \mapsto A_0 y + \beta(y)$ in (1.1) is not Lipschitz continuous in $L^2(\Omega)$ norm. (Generally, differential operators are not continuous.) How could we surpass this difficulty? The key idea is based on the following simple observation: the approximation proposed in [7] for the new state equation (applied to our situation) can be regarded as a Trotter product formula approximation in $L^2(\Omega)$; its convergence will follow by applying an extension due to Y. Kobayashi of a well-known result of Brézis and Pazy. The reduction of the convergence to this result is rather technical and requires a previous approximation of both original and auxiliary controls by step functions followed by a refined partition of the time interval. (It is the author's opinion that, for infinite-dimensional control systems, Trotter-type product formulas represent just the adequate tool to prove the convergence of such an approximation scheme as that proposed in [7] for the new state equation.) On the other hand, the adjoint equation of (1.1) arising in the optimality conditions is a much more complicated object than

in the finite-dimensional case. So the procedure of passing to the limit as $\delta \to 0$ used in [7] must be modified according to Barbu's approach to optimality conditions. In a few words, we can express the new situation in the following way: since a new parameter η arises in the regularization process, we have to take limits when both δ and η tend to zero.

For optimal control problems involving parabolic infinite-dimensional systems, the inclusion (1.2) was proved by Barbu for two special situations by using two types of arguments. The first refers to convex optimal control problems governed by linear parabolic equations (see [3, p. 319]). In this case every dual extremal arc satisfies (1.2). This happens because the necessary optimality conditions here are also sufficient, and so any optimal pair of the considered problem solves also all the problems defining $V(t, y^*(t))$ for $t \in [0, T]$. The second situation concerns optimal control of semilinear parabolic equations, but under two very restrictive additional hypotheses: the integrand in the performance index is Gâteaux differentiable with respect to the control variable u (therefore the subdifferential of the integrand with respect to u is single valued) and B^* is injective (see [2, Cor. 5.1, p. 209]).

But here we work under natural hypotheses, i.e., those under which the necessary conditions of optimality are obtained in [2]. However, as in the finite-dimensional case, additional assumptions lead to more specific results. So we shall show that if the two functions defining the performance index are convex also with respect to the state variable, then (1.2) holds not only *almost everywhere* but even *everywhere* (see [7, Prop. 5.5] for a related finite-dimensional result). As an intermediate stage we shall prove the convexity of the optimal value function with respect to the space variable. This involves a Trotter-type product formula for the corresponding dynamic programming equation.

Finally, let us mention that, as a consequence of the main results (Theorems 2.1 and 2.2), we find that every optimal control of our considered problems is given by a feedback law (Theorems 2.3 and 2.4).

2. The framework and the main results. Let $\mathcal{U}$ be a real Hilbert space, and set $\mathcal{H} = L^2(\Omega)$, where Ω is a fixed open and bounded subset of $\mathbf{R}^N$ having a sufficiently smooth boundary. The scalar products and norms of $\mathcal{U}$ and $\mathcal{H}$ are denoted by the same symbols: $(\cdot, \cdot)$ and $|\cdot|$, respectively. Also, let $\mathcal{V}$ be a real Hilbert space continuously and densely imbedded in $\mathcal{H}$ with $\mathcal{V}'$ its dual space. Identifying $\mathcal{H}$ with its own dual, we assume

$$\mathcal{V} \subset \mathcal{H} \subset \mathcal{V}'.$$

Denote by $(\cdot, \cdot)$ the pairing between $\mathcal{V}$ and $\mathcal{V}'$, and by $|\cdot|_{\mathcal{V}}$ the norm of $\mathcal{V}$.

Consider the following optimal control problem:

(P) Minimize

$$\int_0^T (h(u(t)) + g(y(t)))\, dt + \ell(y(T)) \tag{2.1}$$

over all $u \in L^2(0, T; \mathcal{U})$, where $y \in W^{1,2}([0, T]; \mathcal{H})$ satisfies the state equation

$$y'(t) + Ay(t) + \beta(y(t) - \psi) \ni Bu(t) + f(t) \text{ a.e. } t \in (0, T) \tag{2.2}$$

and the initial condition

$$y(0) = y_0. \tag{2.3}$$

We impose on the data of (P) the following hypotheses:
(H1) The inclusion $\mathcal{V} \subset \mathcal{H}$ is compact.
(H2) $A : \mathcal{V} \to \mathcal{V}'$ is a linear continuous and symmetric operator which, for some $\omega > 0$ and $\alpha \in \mathbf{R}$, satisfies

$$(Ay, y) \geq \omega |y|_{\mathcal{V}}^2 - \alpha |y|^2 \text{ for all } y \in \mathcal{V}. \tag{2.4}$$

Moreover, there exists $c \geq 0$ such that, for every nondecreasing function $\xi \in C^1(\mathbf{R})$ with $\xi(0) = 0$ and $\xi' \leq 1$,

$$(Ay, \xi(y)) \geq -c(1 + |\xi(y)|)(1 + |y|) \text{ for all } y \in D(A) = \{z \in \mathcal{V} : Az \in \mathcal{H}\}. \tag{2.5}$$

(H3) β is a maximal monotone graph in $\mathbf{R} \times \mathbf{R}$ with $0 \in D(\beta)$ and ψ is a given function in $\mathcal{H}$ such that for some $c \geq 0$,

$$(Ay, \beta_\eta(y - \psi)) \geq -c(1 + |\beta_\eta(y - \psi)|)(1 + |y|) \text{ for all } y \in D(A) \text{ and } \eta > 0, \tag{2.6}$$

where $\beta_\eta(r) = \eta^{-1}(r - (I + \eta\beta)^{-1}r)$ for all $r \in \mathbf{R}, \eta > 0$.
(H4) $B : \mathcal{U} \to \mathcal{H}$ is a linear continuous operator.
(H5) $f \in L^2(0, T; \mathcal{H})$.
(H6) y_0 is a given function in $\mathcal{V}$ such that

$$\int_\Omega j(y_0(x) - \psi(x))\, dx < +\infty,$$

where $j : \mathbf{R} \to (-\infty, +\infty]$ is a convex function whose subdifferential is β.
(H7) $h : \mathcal{U} \to (-\infty, +\infty]$ is convex, lower semicontinuous, and not identically $+\infty$ and, for some $c_1 > 0$ and $c_2 \in \mathbf{R}$, satisfies

$$h(u) \geq c_1 |u|^2 - c_2 \text{ for all } u \in \mathcal{U}. \tag{2.7}$$

(H8) $g, \ell : \mathcal{H} \to \mathbf{R}$ are Lipschitz continuous on bounded subsets and bounded from below by affine functions.
By means of j (specified in (H6)), we define the following convex function $\phi : \mathcal{H} \to (-\infty, +\infty]$:

$$\phi(y) = \int_\Omega j(y(x) - \psi(x))\, dx. \tag{2.8}$$

We have (see [2, Prop. 1.9, p. 24])

$$\partial\phi(y) = \{w \in \mathcal{H} : w(x) \in \beta(y(x) - \psi(x)) \text{ a.e. } x \in \Omega\}$$

so that we may write (2.2) like this:

$$y' + Ay + \partial\phi(y) \ni Bu + f \text{ a.e. } t \in (0, T). \tag{2.9}$$

Notice that (H6) can be rewritten as
(H6)′ $y_0 \in \mathcal{V} \cap D(\phi)$.

A standard existence result (see [2, Thm. 4.3, p. 131]) states that, under the hypotheses (H2)–(H6) (except (2.5)), for every $u \in L^2(0, T; \mathcal{U})$, the problem (2.9), (2.3) has a unique solution $y \in W^{1,2}([0, T]; \mathcal{H}) \cap L^2(0, T; D(A)) \cap C([0, T]; \mathcal{V})$. If, instead of (H6), $y_0 \in \overline{\mathcal{V} \cap D(\phi)} = \overline{D(\phi)}$, then (2.9), (2.3) has a unique solution

$y \in C([0,T];\mathcal{H})$ such that $\sqrt{t}y' \in L^2(0,T;\mathcal{H})$ and $\sqrt{t}y \in L^2(0,T;D(A))$. Let us mention that (2.6) ensures the maximal monotonicity of $A+\partial\phi$ in $\mathcal{H}\times\mathcal{H}$.

One also knows that, under the hypotheses (H2)–(H5) (without (2.5)), (H7), and (H8), the control problem (P) admits at least one optimal pair for every $y_0 \in \overline{D(\phi)}$ (see [2, Prop. 5.1, p. 173]). (Note only that (2.7) and the boundedness from below by affine functions of g and ℓ ensure here that the functional (2.1) tends to $+\infty$ as $\int_0^T |u|^2\,dt \to \infty$.)

We associate with the optimal control problem (P) the corresponding optimal value function $V : [0,T]\times\overline{D(\phi)} \to \mathbf{R}$, i.e.,

(2.10)
$$V(t,y) = \inf\left\{ \int_t^T (h(u(s)) + g(z(s)))\,ds + \ell(z(T)) : \right.$$
$$\left. z' + Az + \beta(z-\psi) \ni Bu + f \text{ a.e. } s \in (t,T), z(t) = y, u \in L^2(t,T;\mathcal{U}) \right\}.$$

The following result is well known, but we shall give its proof for the reader's convenience.

PROPOSITION 2.1. *Under the hypotheses* (H2)–(H5) (*except* (2.5)), (H7), *and* (H8), *the function* $\overline{D(\phi)} \ni y \mapsto V(t,y)$ *is Lipschitz continuous on bounded subsets, uniformly with respect to* $t \in [0,T]$.

(Note that only here and in the proof of the existence of an optimal pair the condition (2.7) and the boundedness from below by affine functions of g and ℓ are needed.)

Proof. Let $y_1, y_2 \in \overline{D(\phi)}$, arbitrary, such that $|y_1|, |y_2| \le r$. By (2.10), we have

$$V(t,y_2) - V(t,y_1) \le \int_t^T (g(z_2(s)) - g(z_1(s)))\,ds + \ell(z_2(T)) - \ell(z_1(T)),$$

where z_2 and z_1 are the solutions of (2.2) corresponding to the initial states y_2 and y_1 at the time t, and the infimum defining $V(t,y_1)$ is attained. Taking (2.7) into account, we obtain

$$\int_t^T |u(s)|^2\,ds \le \text{ const.},$$

where the above constant depends only on r. (It is independent of t.) So $|z_1(s)|$ and $|z_2(s)|$ are bounded by a constant which depends only on r. Thus we may use the Lipschitz continuity on bounded subsets of g and ℓ to deduce the same thing for $V(t,\cdot)$ uniformly with respect to $t \in [0,T]$. The proof is finished.

Now to give a sense to the gradient of V with respect to y also at those points of $\overline{D(\phi)}$ which are not interior points for $\overline{D(\phi)}$, we consider the extension of V to the whole space $\mathcal{H}$ denoted by $\tilde{V}$ and defined as follows:

$$\tilde{V}(t,y) = V(t,P_K y) \text{ for } (t,y) \in [0,T]\times\mathcal{H},$$

where P_K is the projection operator of $\mathcal{H}$ onto $K = \overline{D(\phi)}$. Obviously, by virtue of Proposition 2.1, the function $\mathcal{H} \ni y \mapsto \tilde{V}(t,y)$ is also Lipschitz continuous on bounded subsets uniformly with respect to $t \in [0,T]$. So we may define

$$\partial_y V(t,y) = \partial_y \tilde{V}(t,y) \text{ for } (t,y) \in [0,T]\times\overline{D(\phi)},$$

where $\partial_y \tilde{V}(t,y)$ is the generalized gradient (in Clarke's sense) of $y \mapsto \tilde{V}(t,y)$.

Finally, let $\mathcal{Y} = H^s(\Omega) \cap \mathcal{V}$, where $s > N/2$, and set $Q = (0,T) \times \Omega$. We are now prepared to state our main results.

THEOREM 2.1. *Let (u^*, y^*) be an arbitrary optimal pair of the control problem* (P) *where $\beta : \mathbf{R} \to \mathbf{R}$ is Lipschitz continuous on bounded subsets. Suppose that* (H1)–(H8) *are satisfied. Then there exist $p \in BV([0,T];\mathcal{Y}') \cap L^\infty(0,T;\mathcal{H}) \cap L^2(0,T;\mathcal{V})$ and $\mu \in (L^\infty(Q))'$ such that*

$$p' - Ap - \mu \in \partial g(y^*) \text{ a.e. in } Q, \quad p' - Ap - \mu \in L^\infty(0,T;\mathcal{H}), \tag{2.11}$$

$$\mu_a(t,x) \in p(t,x)\partial\beta(y^*(t,x) - \psi(x)) \text{ a.e. in } Q, \tag{2.12}$$

$$-p(T) \in \partial\ell(y^*(T)), \tag{2.13}$$

$$B^*p(t) \in \partial h(u^*(t)) \text{ a.e. } t \in (0,T), \tag{2.14}$$

and

$$-p(t) \in \partial_y V(t, y^*(t)) \text{ a.e. } t \in (0,T). \tag{2.15}$$

If the following additional hypothesis on β holds:

(H9) *ϕ (given by* (2.8)*) is bounded on bounded subsets of $\mathcal{V} \cap D(\phi)$ in the $\mathcal{V}$ norm, then p (together with μ) can be chosen such that besides* (2.11)–(2.15) *it also satisfies*

$$-p(0) \in \partial_y V(0, y_0). \tag{2.16}$$

Moreover, if, for some $c \geq 0$, β satisfies

$$\beta'(r) \leq c(|\beta(r)| + |r| + 1) \text{ a.e. } r \in \mathbf{R}, \tag{2.17}$$

then $p \in AC([0,T];\mathcal{Y}') \cap C_w([0,T];\mathcal{H})$ and $\mu = \mu_a \in L^1(Q)$.

Here p' is the derivative of p in the sense of $\mathcal{V}'$-valued distribution and $\mu_a \in L^1(Q)$ is the absolutely continuous part of μ; also, $\partial\beta$, ∂g, and $\partial\ell$ are the generalized gradients of β, g, and ℓ, and ∂h is the subdifferential of h. Finally, $C_w([0,T];\mathcal{H})$ is the space of weakly continuous functions from $[0,T]$ to $\mathcal{H}$.

Now set $K = \{y \in L^2(\Omega) : y \geq \psi \text{ a.e. in } \Omega\}$.

THEOREM 2.2. *Let (u^*, y^*) be an arbitrary optimal pair for the control problem* (P) *where β is defined by*

$$\beta(r) = \begin{cases} 0 & if\ r > 0, \\ (-\infty, 0] & if\ r = 0, \\ \emptyset & if\ r < 0. \end{cases} \tag{2.18}$$

Suppose that (H1)–(H8) *hold. Then there exists a function $p \in BV([0,T];\mathcal{Y}') \cap L^\infty(0,T;\mathcal{H}) \cap L^2(0,T;\mathcal{V})$, which with the necessary conditions of optimality*

$$(p' - Ap)_a \in \partial g(y^*) \text{ a.e. in } \{(t,x) \in Q : y^*(t,x) > \psi(x)\}\ (p' - Ap \in (L^\infty(Q))'), \tag{2.19}$$

$$p(f + Bu^* - Ay^* - y^{*\prime}) = 0 \text{ a.e. in } Q, \tag{2.20}$$

$$-p(T) \in \partial\ell(y^*(T)), \tag{2.21}$$

$$B^*p(t) \in \partial h(u^*(t)) \text{ a.e. } t \in (0,T) \tag{2.22}$$

also satisfies

$$-p(t) \in \partial_y V(t, y^*(t)) + N_K(\psi) \text{ a.e. } t \in (0,T), \tag{2.23}$$

$$-p(0) \in \partial_y V(0, y_0) - N_K(y_0). \tag{2.24}$$

Here $N_K(\psi)$ and $N_K(y_0)$ denote the normal cones to K at ψ and y_0, respectively; for example, $N_K(\psi) = \{w \in \mathcal{H} : (w, y - \psi) \leq 0 \text{ for all } y \in K\}$. Note also that $K = \overline{D(\phi)} = D(\phi)$, where ϕ (defined by (2.8)) corresponds to β given by (2.18).

Remark 2.1. While for semilinear parabolic equations (in Theorem 2.1), $\overline{D(\phi)} = \mathcal{H}$ and therefore all points of $\overline{D(\phi)}$ are interior (consequently, in (2.15), (2.16), $\partial_y V(t,\cdot)$ is just Clarke's generalized gradient), in the case of the parabolic obstacle problem (of Theorem 2.2), $\overline{D(\phi)} = \{y \in L^2(\Omega) : y \geq \psi \text{ a.e. in } \Omega\}$, and no point of $\overline{D(\phi)}$ is interior.

Remark 2.2. It would be interesting to see if similar results (to those above) can be obtained if we use other (possibly more intrinsic) notions of gradient of V (with respect to y) at the noninterior points of $\overline{D(\phi)}$. An alternative way to define $\partial_y V(t,y)$ is to consider as extension $\tilde{V}$ the function

$$\tilde{V}(t,y) = \begin{cases} V(t,y) & \text{if } y \in K, \\ +\infty & \text{if } y \notin K, \end{cases}$$

and as $\partial_y \tilde{V}(t,y)$, the subdifferential of $\tilde{V}$ with respect to y at (t,y) defined in [11] with the aid of the upper subderivative $h \mapsto \tilde{V}^{\uparrow}(t,y;h)$. It would also be useful to compare the results corresponding to various notions of gradient.

Let us give now some examples of parabolic variational inequalities covered by the abstract formulation (2.2) and satisfying (H1)–(H3). Let A_0 be the second-order elliptic differential operator defined by

$$A_0 y = -\sum_{i,j=1}^{N} \frac{\partial}{\partial x_i}\left(a_{ij}(x)\frac{\partial y}{\partial x_i}\right) + a_0(x)y,$$

where $a_{ij} \in C^1(\Omega), a_0 \in L^\infty(\Omega), a_{ij} = a_{ji}$ for all $i,j, a_0(x) \geq 0$ a.e. $x \in \Omega$ and, for some $\omega > 0$,

$$\sum_{i,j=1}^{N} a_{ij}(x)\xi_i\xi_j \geq \omega \sum_{i=1}^{N} |\xi_i|^2 \text{ for all } (\xi_1, \ldots, \xi_N) \in \mathbf{R}^N \text{ a.e. } x \in \Omega.$$

Consider the following mixed boundary value problem:

$$\begin{cases} \dfrac{\partial y}{\partial t}(t,x) + A_0 y(t,x) + \beta(y(t,x) - \psi(x)) \ni Bu(t) + f(t,x) \text{ a.e. } (t,x) \in Q, \\ \alpha_1 y + \alpha_2 \dfrac{\partial y}{\partial \nu} = 0 \text{ on } (0,T) \times \partial\Omega, \\ y(0,x) = y_0(x), \ x \in \Omega, \end{cases} \tag{2.25}$$

where β, B, f, and y_0 satisfy (H3)–(H6), $\psi \in H^2(\Omega)$, and $\alpha_1, \alpha_2 \geq 0$ with $\alpha_1 + \alpha_2 > 0$. By shifting the range of β, we may assume, without any loss of generality, that

$0 \in \beta(0)$. To interpret (2.25) as an evolution equation of the form (2.2), we have to specify the space $\mathcal{V}$ and the operator A: If $\alpha_2 \neq 0$, we take $\mathcal{V} = H^1(\Omega)$ and we define $A : \mathcal{V} \to \mathcal{V}'$ by

$$(Ay, z) = \sum_{i,j=1}^{N} \int_\Omega a_{ij} \frac{\partial y}{\partial x_i} \frac{\partial z}{\partial x_j}\, dx + \int_\Omega a_0 yz\, dx + \frac{\alpha_1}{\alpha_2} \int_{\partial\Omega} yz d\sigma \text{ for all } y, z \in \mathcal{V}. \tag{2.26}$$

If $\alpha_2 = 0$, take $\mathcal{V} = H_0^1(\Omega)$ and define $A : H_0^1(\Omega) \to H^{-1}(\Omega)$ by

$$(Ay, z) = \sum_{i,j=1}^{N} \int_\Omega a_{ij} \frac{\partial y}{\partial x_i} \frac{\partial z}{\partial x_j}\, dx + \int_\Omega a_0 yz\, dx \text{ for all } y, z \in H_0^1(\Omega). \tag{2.27}$$

We have $(Ay)(x) = (A_0 y)(x)$ a.e. $x \in \Omega$ for all $y \in D(A)$, where $D(A) = \{y \in H^2(\Omega) : \alpha_1 y + \alpha_2 \partial y/\partial\nu = 0 \text{ a.e. on } \partial\Omega\}$. In particular, $D(A) = H_0^1(\Omega) \cap H^2(\Omega)$ if $\alpha_2 = 0$. Obviously, for $\mathcal{V}$ and A defined as above, (H1) and (H2) hold. It remains to examine (2.6). It is easy to show (see [2, p. 137]) that, in the case when $\alpha_2 \neq 0$, the condition (2.6) is satisfied if one assumes in addition that

$$\left(\alpha_1 \psi + \alpha_2 \frac{\partial \psi}{\partial \nu}\right) \beta_\eta(y - \psi) \geq 0 \text{ a.e. on } \partial\Omega \text{ for all } y \in D(A) \text{ and } \eta > 0. \tag{2.28}$$

If $\alpha_2 = 0$, the condition (2.6) is satisfied if one has

$$\beta_\eta(-\psi) = 0 \text{ a.e. on } \partial\Omega \text{ for all } \eta > 0. \tag{2.29}$$

In the case of the parabolic obstacle problem (in which β is given by (2.18)), $\beta_\eta(r) = \eta^{-1} \min(r, 0)$, and, consequently, the conditions (2.28), (2.29) are satisfied if

$$\alpha_1 \psi + \alpha_2 \frac{\partial \psi}{\partial \nu} \leq 0 \text{ a.e. on } \partial\Omega,$$

respectively,

$$\psi \leq 0 \text{ a.e. on } \partial\Omega.$$

Remark 2.3. The additional hypothesis (H9), which assures the validity of (2.16) in Theorem 2.1, is automatically verified for β given by (2.18). Indeed, in this case $\phi = I_K$, where I_K is the indicator function of $K = \{y \in L^2(\Omega) : y \geq \psi \text{ a.e. in } \Omega\}$; therefore $\phi = 0$ on $D(\phi) = K$. As regards Theorem 2.1, if, for some $c_1 \geq 0$, $c_2 \in \mathbf{R}$, β satisfies also the growth condition

$$|\beta(r)| \leq c_1 |r| + c_2 \text{ for all } r \in \mathbf{R}$$

(in particular, if β is Lipschitz continuous on $\mathbf{R}$), then (H9) clearly holds. But this happens even for a sharper growth of β at ∞. Here is an example. Take (2.2) given by (2.25) as state equation with $N \leq 4$, $\beta(r) = r^3$, and $\alpha_2 = 0$. We have $\mathcal{V} = H_0^1(\Omega), D(\phi) = L^4(\Omega)$, and

$$\phi(y) = \frac{1}{4} \int_\Omega y^4(x)\, dx.$$

By Sobolev imbedding theorem, $H^1(\Omega) \subset L^4(\Omega)$ for $N \leq 4$, whence (H9) immediately follows.

Finally, combining (2.14) and (2.22) (after a previous use of [3, Prop. 2.1, p. 103]) with (2.15) and (2.23), respectively, we obtain the following direct consequences of Theorems 2.1 and 2.2, which state that every optimal control for the problem (P) is also a feedback optimal control.

THEOREM 2.3. *Let (u^*, y^*) be an arbitrary optimal pair of the control problem* (P) *where, as in Theorem* 2.1, *$\beta : \mathbf{R} \to \mathbf{R}$ is Lipschitz continuous on bounded subsets. Suppose that* (H1)–(H8) *hold. Then*

$$u^*(t) \in \partial h^*(-B^* \partial_y V(t, y^*(t))) \text{ a.e. } t \in (0,T).$$

THEOREM 2.4. *Under the assumptions of Theorem* 2.2, *if (u^*, y^*) is an arbitrary optimal pair of the problem* (P), *then*

$$u^*(t) \in \partial h^*(-B^*(\partial_y V(t, y^*(t)) + N_K(\psi))) \text{ a.e. } t \in (0,T).$$

Under the additional assumption (2.17), Theorem 2.3 was proved by V. Barbu [2, Thm. 5.6, p. 208].

3. Proofs of Theorems 2.1 and 2.2. Likewise as in [7] we shall construct an auxiliary optimal control problem by introducing a new control variable v into the control system (2.2). So we consider the following perturbed system:

$$\begin{cases} y' + Ay + \beta(y - \psi) \ni Bu + f + v, \\ y(0) = y^0. \end{cases} \tag{3.1}$$

For every $\delta > 0$, define the function $k_\delta : [0,T] \times \mathcal{H} \to \mathbf{R}$ by

$$k_\delta(t, v) = \sup\{(p, v) : p \in \partial_y V(t, y), |y - y^*(t)| \le \delta\}. \tag{3.2}$$

It is easy to see that $v \mapsto k_\delta(t, v)$ is convex on $\mathcal{H}$ and Lipschitz continuous on $\mathcal{H}$ uniformly with respect to $t \in [0,T]$ (by virtue of Proposition 2.1). So $\partial_v k_\delta(t, v)$ is well defined. Also, for every $v \in L^2(0,T;\mathcal{H})$, the function $t \mapsto k_\delta(t, v(t))$ is measurable (see [7, Lem. 8.1] and [6]); moreover, it belongs to $L^2(0,T)$.

Now let $0 < \delta' < \delta$. The following lemma is an expression of the dynamic programming principle, and it is the key of the proof (see [7, Lem. 8.4] for the finite-dimensional version).

LEMMA 3.1. *Suppose that* (H2)–(H5) (*except* (2.5)), (H7), *and* (H8) *hold and, in addition, $K = \overline{D(\phi)}$ is a* (*closed*) *convex cone in $\mathcal{H}$ with the vertex ψ. Let $u \in L^2(0,T;\mathcal{U})$, $v \in L^2(0,T;\mathcal{H})$, and $y^0 \in K$ such that $h(u) \in L^1(0,T)$, $v(t) \in K - \psi$ a.e. $t \in (0,T)$ and $|y(t) - y^*(t)| \le \delta'$ for all $t \in [0,T]$, where y is the solution of* (3.1) *corresponding to the given u, v and y^0. Then*

$$\int_0^T (h(u(t)) + g(y(t)))\, dt + \ell(y(T)) + \int_0^T k_\delta(t, -v(t))\, dt - V(0, y^0) \ge 0. \tag{3.3}$$

Proof. We shall prove Lemma 3.1 in several steps.

1. *Reduction to the case in which u and v are step functions.* We assert that it suffices to prove Lemma 3.1 for every triple of step functions $u : [0,T] \to \mathcal{U}$, $v : [0,T] \to \mathcal{H}$, $f : [0,T] \to \mathcal{H}$, which are constant on the same subintervals of the form $(k\theta, (k+1)\theta]$, $k = 0, 1, \dots, p-1$, such that $h(u) \in L^1(0,T)$, $v(t) \in K - \psi$, and $|y(t) - y^*(t)| \le \delta''$ for all $t \in [0,T]$, where y is the solution of (3.1) and $\delta' < \delta'' < \delta$.

Indeed, let us assume that we have proved Lemma 3.1 in this case. Then the general case will immediately follow if, for every u, v, y^0 as in the statement of Lemma 3.1, we construct two sequences of step functions $u_m : [0,T] \to \mathcal{U}, v_m : [0,T] \to \mathcal{H}$ taking constant values on $(k\theta_m, (k+1)\theta_m], k = 0, 1, \dots, p_m - 1$, such that $h(u_m) \in L^1(0,T), v_m(t) \in K - \psi$ for all $t \in [0,T]$, and

$$u_m \to u \text{ strongly in } L^2(0,T;\mathcal{U}), \tag{3.4}$$

$$\int_0^T h(u_m(t))\,dt \to \int_0^T h(u(t))\,dt, \tag{3.5}$$

$$v_m \to v \text{ strongly in } L^2(0,T;\mathcal{H}). \tag{3.6}$$

To see this, let y_m satisfy

$$\begin{cases} y_m' + Ay_m + \beta(y_m - \psi) \ni Bu_m + f_m + v_m \text{ a.e. in } (0,T), \\ y_m(0) = y^0, \end{cases}$$

where $f_m \to f$ strongly in $L^2(0,T;\mathcal{H})$. We have

$$y_m \to y \text{ strongly in } C([0,T];\mathcal{H}),$$

and, by uniform Lipschitz continuity of $v \mapsto k_\delta(t,v)$,

$$\int_0^T k_\delta(t, -v_m(t))\,dt \to \int_0^T k_\delta(t, -v(t))\,dt.$$

But for sufficiently large m, $|y_m(t) - y^*(t)| \le \delta''$ for all $t \in [0,T]$, whence, by virtue of our assumption,

$$\int_0^T (h(u_m(t)) + g(y_m(t)))\,dt + \ell(y_m(T)) + \int_0^T k_\delta(t, -v_m(t))\,dt - V(0,y^0) \ge 0.$$

Letting $m \to \infty$, we obtain the statement of Lemma 3.1 in the general case.

It remains only to indicate the construction of $\{u_m\}$, $\{v_m\}$. Clearly, there exist two sequences of step functions $\tilde{u}_m : [0,T] \to \mathcal{U}$, $\tilde{v}_m : [0,T] \to \mathcal{H}$, which are constant on $(k\theta_m, (k+1)\theta_m], k = 0, 1, \dots, p_m - 1$, such that $\tilde{u}_m \to u$ strongly in $L^2(0,T;\mathcal{U})$, $\tilde{v}_m \to v$ strongly in $L^2(0,T;\mathcal{H})$. Define the function u_m, v_m by

$$u_m(t) = (I + \lambda_m \partial h)^{-1} \tilde{u}_m(t) \text{ with } \lambda_m = \left(\int_0^T |\tilde{u}_m(s) - u(s)|^2\,ds \right)^{1/2},$$

$$v_m(t) = \left(I + \frac{1}{m} \partial I_{K-\psi} \right)^{-1} \tilde{v}_m(t)$$

for all $t \in [0,T]$, where, of course, $I_{K-\psi}$ is the indicator function of $K - \psi$. Obviously, u_m, v_m are step functions, $v_m(t) \in K - \psi$ for all $t \in [0,T]$, and by assumptions on $\tilde{u}_m, \tilde{v}_m$ we obtain (3.4) and (3.6). For the proof of (3.5) we refer to [9, Lem. 1].

2. *Approximation of the perturbed system.* Let u, v, and f be a triple of step functions as above, which are constant on $(k\theta, (k+1)\theta], k = 0, 1, \dots, p-1$ (do not forget that $v(t) \in K - \psi$ for all $t \in [0,T]$), and let y be the solution of (3.1) corresponding to these u, v, and y^0. By virtue of Proposition 7.1 from [7] (note that $t \mapsto k_\delta(t, -v(t))$ is

an essentially bounded function), we can choose $\tau \in (0,T)$ and a subsequence $\{N_j\}$ of the positive integers such that T/τ is an irrational number and

$$\sum_{i=m_j}^{n_j} \frac{T}{N_j} k_\delta \left(\tau + i\frac{T}{N_j}, -v\left(\tau + i\frac{T}{N_j}\right)\right) \to \int_0^T k_\delta(t, -v(t))\,dt \text{ as } j \to \infty, \tag{3.7}$$

where

$$m_j = \min\left\{i : \tau + i\frac{T}{N_j} > 0\right\}, \qquad n_j = \max\left\{i : \tau + i\frac{T}{N_j} < T\right\}.$$

Clearly, since T/τ is irrational, $\tau + iT/N_j$ cannot be either 0 or T.

Define $y_j : [0,T] \to \mathcal{H}$ as being the solution of

$$\begin{cases} y_j' + Ay_j + \beta(y_j - \psi) \ni Bu + f, \\ y_j(0) = y^0 \end{cases} \tag{3.8}$$

on the subinterval $[0, m_jT/N_j + \tau)$; the solution of

$$\begin{cases} y_j' + Ay_j + \beta(y_j - \psi) \ni Bu + f, \\ y_j(iT/N_j + \tau) = y_j(iT/N_j + \tau - 0) + \dfrac{T}{N_j} v(iT/N_j + \tau) \end{cases} \tag{3.9}$$

on the subinterval $[iT/N_j + \tau, (i+1)T/N_j + \tau), i = m_j, \ldots, n_j - 1$; and the solution of (3.9) but with $i = n_j$ on the subinterval $[n_jT/N_j + \tau, T]$. We shall show that for every $t \in [0,T]$

$$y_j(t) \to y(t) \text{ strongly in } \mathcal{H}. \tag{3.10}$$

Moreover, an uniform estimate holds at the points $t = iT/N_j + \tau$: For any $\eta > 0$, we can find a positive integer j_η such that if $j \geq j_\eta$, then

$$|y_j(iT/N_j + \tau) - y(iT/N_j + \tau)| < \eta \text{ for } i = m_j, \ldots, n_j. \tag{3.11}$$

To prove these, for every $j = 1, 2, \ldots$, consider the moments $t_j^{(0)}, t_j^{(1)}, \ldots, t_j^{(p)}$ and the positive integers $\nu_j^{(1)}, \nu_j^{(2)}, \ldots, \nu_j^{(p)}$ defined by

$$\begin{aligned} &t_j^{(k)} = t_j^{(k-1)} + \frac{T}{N_j} + \nu_j^{(k)} \frac{T}{N_j}, \qquad k = 1, 2, \ldots, p, \\ &t_j^{(0)} = (m_j - 1)\frac{T}{N_j} + \tau, \\ &t_j^{(k)} < k\theta < t_j^{(k)} + \frac{T}{N_j}, \qquad k = 1, 2, \ldots, p. \end{aligned}$$

Since T/τ is irrational, $t_j^{(k)}$ are well defined (because $k\theta$ cannot take the values $iT/N_j + \tau$). Obviously, $t_j^{(p)} = n_jT/N_j + \tau$. Next, for $k = 1, 2, \ldots, p$, define the operators $\mathcal{A}_k, \mathcal{B}_k$ by

$$\begin{aligned} &\mathcal{A}_k y = Ay + \beta(y - \psi) - Bu(k\theta) - f(k\theta), \qquad y \in D(A) \cap D(\partial\phi), \\ &\mathcal{B}_k y = -v(k\theta) \in -(K - \psi), y \in \mathcal{H}. \end{aligned}$$

Clearly, since K is a cone, if $y \in K$, $e^{-\mathcal{B}_k t}y = y + tv(k\theta)$ also belongs to K for $t \geq 0$, and so the expression $e^{-\mathcal{A}_k t}e^{-\mathcal{B}_k t}y$ makes sense. Therefore we may define the operators $T_k(t)$ with $t \geq 0$ like this:

$$T_k(t)y = e^{-\mathcal{A}_k t}e^{-\mathcal{B}_k t}y \text{ for } y \in K, k = 1, 2, \ldots, p.$$

Finally, denote by $S_k(t)$ the nonlinear semigroup generated by the operator $y \mapsto Ay + \beta(y - \psi) - Bu(k\theta) - f(k\theta) - v(k\theta)$.

Let $t \in (0, T]$ arbitrary but fixed. Let us denote by k_0 that integer for which $t \in (k_0\theta, (k_0 + 1)\theta]$. Define t_j and the positive integer ν_j by

$$t_j = t_j^{(k_0)} + \frac{T}{N_j} + \nu_j \frac{T}{N_j}, \quad t_j \leq t < t_j + \frac{T}{N_j}.$$

We set now the following notations:

$$\varepsilon_j = \frac{T}{N_j}, \quad \varepsilon_j^{(k)} = t_j^{(k-1)} + \frac{T}{N_j} - (k-1)\theta, \quad \overline{\varepsilon}_j^{(k)} = k\theta - t_j^{(k)}, \quad \theta_j^{(k)} = t_j^{(k)} - t_j^{(k-1)} - \frac{T}{N_j}.$$

We can write y_j (given by (3.8), (3.9)) as

$$y_j(t) = \begin{cases} e^{-\mathcal{A}_{k_0+1}(t-t_j)}e^{-\mathcal{B}_{k_0+1}\varepsilon_j}T_{k_0+1}\left(\dfrac{t_j - t_j^{(k_0)} - \varepsilon_j}{\nu_j}\right)^{\nu_j} e^{-\mathcal{A}_{k_0+1}\varepsilon_j^{(k_0+1)}} \\ \qquad \times \left(\displaystyle\prod_{k=1}^{k_0} e^{-\mathcal{A}_k\overline{\varepsilon}_j^{(k)}}e^{-\mathcal{B}_k\varepsilon_j}T_k\left(\frac{\theta_j^{(k)}}{\nu_j^{(k)}}\right)^{\nu_j^{(k)}} e^{-\mathcal{A}_k\varepsilon_j^{(k)}}\right) y^0 \text{ for } t \in (\theta, T], \\ e^{-\mathcal{A}_1(t-t_j)}e^{-\mathcal{B}_1\varepsilon_j}T_1\left(\dfrac{t_j - t_j^{(0)} - \varepsilon_j}{\nu_j}\right)^{\nu_j} e^{-\mathcal{A}_1\varepsilon_j^{(1)}} \text{ for } t \in (0, \theta]. \end{cases}$$

For simplicity we set

$$Y_j^{(k)} = e^{-\mathcal{A}_k\overline{\varepsilon}_j^{(k)}}e^{-\mathcal{B}_k\varepsilon_j}T_k\left(\frac{\theta_j^{(k)}}{\nu_j^{(k)}}\right)^{\nu_j^{(k)}} e^{-\mathcal{A}_k\varepsilon_j^{(k)}}, \qquad k = 1, 2, \ldots, p,$$

$$Y_j^{(k_0+1)}(t - k_0\theta) = e^{-\mathcal{A}_{k_0+1}(t-t_j)}e^{-\mathcal{B}_{k_0+1}\varepsilon_j}T_{k_0+1}\left(\frac{t_j - t_j^{(k_0)} - \varepsilon_j}{\nu_j}\right)^{\nu_j} e^{-\mathcal{A}_{k_0+1}\varepsilon_j^{(k_0+1)}},$$

$$Y_j^{(k+1)}(t - k\theta) = Y_j^{(k_0+1)}(t - k_0\theta)Y_j^{(k_0)} \cdots Y_j^{(k+1)}, \qquad k = 1, 2, \ldots, k_0 - 1.$$

Obviously, all the above operators are contractions. In order to estimate $|y_j(t) - y(t)|$, we shall write

(3.12)

$$
\begin{aligned}
&y_j(t)-y(t)\\
&=\sum_{k=1}^{k_0}\Bigg(Y_j^{(k+1)}(t-k\theta)Y_j^{(k)}y((k-1)\theta)\\
&\quad-\ Y_j^{(k+1)}(t-k\theta)e^{-\mathcal{A}_k\bar{\varepsilon}_j^{(k)}}e^{-\mathcal{B}_k\varepsilon_j}T_k\left(\frac{\theta_j^{(k)}}{\nu_j^{(k)}}\right)^{\nu_j^{(k)}}y((k-1)\theta)\\
&\quad+\ Y_j^{(k+1)}(t-k\theta)e^{-\mathcal{A}_k\bar{\varepsilon}_j^{(k)}}e^{-\mathcal{B}_k\varepsilon_j}T_k\left(\frac{\theta_j^{(k)}}{\nu_j^{(k)}}\right)^{\nu_j^{(k)}}y((k-1)\theta)\\
&\qquad\qquad-Y_j^{(k+1)}(t-k\theta)e^{-\mathcal{A}_k\bar{\varepsilon}_j^{(k)}}e^{-\mathcal{B}_k\varepsilon_j}S_k(\theta_j^{(k)})y((k-1)\theta)\\
&\quad+\ Y_j^{(k+1)}(t-k\theta)e^{-\mathcal{A}_k\bar{\varepsilon}_j^{(k)}}e^{-\mathcal{B}_k\varepsilon_j}y((k-1)\theta+\theta_j^{(k)})-Y_j^{(k+1)}(t-k\theta)e^{-\mathcal{A}_k\bar{\varepsilon}_j^{(k)}}e^{-\mathcal{B}_k\varepsilon_j}y(k\theta)\\
&\quad+\ Y_j^{(k+1)}(t-k\theta)e^{-\mathcal{A}_k\bar{\varepsilon}_j^{(k)}}e^{-\mathcal{B}_k\varepsilon_j}y(k\theta)-Y_j^{(k+1)}(t-k\theta)y(k\theta)\Bigg)\\
&\quad+\ Y_j^{(k_0+1)}(t-k_0\theta)y(k_0\theta)-e^{-\mathcal{A}_{k_0+1}(t-t_j)}e^{-\mathcal{B}_{k_0+1}\varepsilon_j}T_{k_0+1}\left(\frac{t_j-t_j^{(k_0)}-\varepsilon_j}{\nu_j}\right)^{\nu_j}y(k_0\theta)\\
&\quad+\ e^{-\mathcal{A}_{k_0+1}(t-t_j)}e^{-\mathcal{B}_{k_0+1}\varepsilon_j}T_{k_0+1}\left(\frac{t_j-t_j^{(k_0)}-\varepsilon_j}{\nu_j}\right)^{\nu_j}y(k_0\theta)\\
&\qquad\qquad-e^{-\mathcal{A}_{k_0+1}(t-t_j)}e^{-\mathcal{B}_{k_0+1}\varepsilon_j}S_{k_0+1}(t_j-t_j^{(k_0)}-\varepsilon_j)y(k_0\theta)\\
&\quad+\ e^{-\mathcal{A}_{k_0+1}(t-t_j)}e^{-\mathcal{B}_{k_0+1}\varepsilon_j}y(k_0\theta+t_j-t_j^{(k_0)}-\varepsilon_j)-e^{-\mathcal{A}_{k_0+1}(t-t_j)}e^{-\mathcal{B}_{k_0+1}\varepsilon_j}y(t)\\
&\quad+\ e^{-\mathcal{A}_{k_0+1}(t-t_j)}e^{-\mathcal{B}_{k_0+1}\varepsilon_j}y(t)-y(t)\ \text{for } t\in(\theta,T].
\end{aligned}
$$

If $t\in(0,\theta]$ (therefore $k_0=0$), then the corresponding expression for $y_j(t)-y(t)$ consists only of the last five lines of (3.12). Take $j\to\infty$ in (3.12). Since $\varepsilon_j,\varepsilon_j^{(k)},\bar{\varepsilon}_j^{(k)}$ tend to 0, $t_j^{(k)}\to k\theta, t_j\to t, \nu_j^{(k)}$ and ν_j tend to $+\infty$, and $\theta_j^{(k)}\to\theta$, we obtain (3.10). Here we have used the continuity at 0 of the functions $s\mapsto e^{-\mathcal{A}_k s}y((k-1)\theta)$ for $k=1,2,\dots,k_0+1, s\mapsto e^{-\mathcal{A}_k s}y(k\theta)$, and $s\mapsto e^{-\mathcal{B}_k s}y(k\theta)$ for $k=1,2,\dots,k_0$, $s\mapsto e^{-\mathcal{A}_{k_0+1}s}y(t)$, and $s\mapsto e^{-\mathcal{B}_{k_0+1}s}y(t)$; the uniform continuity of y; and a Trotter-type product formula due to Brézis and Pazy (see [4, Prop. 4.4]), and proved for multivalued operators (our case) by Y. Kobayashi in [8]; that is,

$$
\begin{aligned}
&T_k\left(\frac{s}{\nu_j^{(k)}}\right)^{\nu_j^{(k)}}y((k-1)\theta)\to S_k(s)y((k-1)\theta)\ \text{as}\ \nu_j^{(k)}\to\infty,\quad k=1,2,\dots,k_0,\\
&T_{k_0+1}\left(\frac{s}{\nu_j}\right)^{\nu_j}y(k_0\theta)\to S_{k_0+1}(s)y(k_0\theta)\ \text{as}\ \nu_j\to\infty,
\end{aligned}
$$

both uniformly in $s\in[0,T]$.

Now let $t=iT/N_j+\tau$. In this case $t_j=t$. According to (3.12), we shall obtain (3.11) as soon as we show that for any $\eta'>0$ there exits $j_{\eta'}>0$ such that, if $j\le j_{\eta'}$,

we have

$$\left| T_{k_0+1}\left(\frac{t_j - t_j^{(k_0)} - \varepsilon_j}{\nu_j}\right)^{\nu_j} y(k_0\theta) - S_{k_0+1}(t_j - t_j^{(k_0)} - \varepsilon_j)y(k_0\theta)\right| < \eta'$$
$$\text{for } \nu_j = 1, 2, \ldots, \theta_j^{(k_0+1)}/\varepsilon_j, \quad k_0 = 0, 1, \ldots, p-1.$$

(Note that $e^{-\mathcal{B}_{k_0+1}\varepsilon_j} y(t) - y(t) = \varepsilon_j v((k_0+1)\theta) \to 0$ uniformly in t as $j \to \infty$.) To this end, first, by the above-mentioned Trotter product formula, we choose a positive integer $\nu_{\eta'}$ such that if $\nu > \nu_{\eta'}$, then

$$|T_{k_0+1}(\varepsilon_j)^\nu y(k_0\theta) - S_{k_0+1}(\nu\varepsilon_j)y(k_0\theta)| < \eta', \qquad k_0 = 0, 1, \ldots, p-1. \tag{3.13}$$

Next choose $j_{\eta'}$ such that for $j > j_{\eta'}$

$$|T_{k_0+1}(\varepsilon_j)y(k_0\theta) - y(k_0\theta)| < \frac{\eta'}{2\nu_{\eta'}},$$

$$|S_{k_0+1}(\varepsilon_j)y(k_0\theta) - y(k_0\theta)| < \frac{\eta'}{2\nu_{\eta'}},$$

$k_0 = 0, 1, \ldots, p-1$. (Obviously, $j_{\eta'}$ depends on η' also via $\nu_{\eta'}$.) Hence, if $j > j_{\eta'}$, (3.13) holds also for $\nu \le \nu_{\eta'}$, therefore for all ν.

3. *The case when u and v are step functions.* The rest of the proof is similar to the proof of Lemma 8.4 from [7]. Using a variant of the dynamic programming principle (see [7, Lem. 8.3]) applied to the problem (P), we have

$$\int_0^{m_jT/N_j+\tau} (h(u(t)) + g(y_j(t)))\,dt + V(m_jT/N_j + \tau, y_j(m_jT/N_j + \tau - 0)) - V(0, y^0) \geq 0,$$

$$\int_{iT/N_j+\tau}^{(i+1)T/N_j+\tau} (h(u(t)) + g(y_j(t)))\,dt + V((i+1)T/N_j + \tau, y_j((i+1)T/N_j + \tau - 0)) - V(iT/N_j + \tau, y_j(iT/N_j + \tau)) \geq 0, \qquad i = m_j, \quad m_j + 1, \ldots, n_j - 1,$$

$$\int_{n_jT/N_j+\tau}^{T} (h(u(t)) + g(y_j(t)))\,dt + V(T, y_j(T)) - V(n_jT/N_j + \tau, y_j(n_jT/N_j + \tau)) \geq 0.$$

We can write the sum of all these inequalities as follows:

$$\int_0^T (h(u(t)) + g(y_j(t)))\,dt + V(T, y_j(T)) - V(0, y^0) - \sum_{i=m_j}^{n_j} (V(iT/N_j + \tau, y_j(iT/N_j + \tau)) - V(iT/N_j + \tau, y_j(iT/N_j + \tau - 0))) \geq 0. \tag{3.14}$$

But by the mean value theorem for generalized gradients (see [2, Cor. 1.2]), we have when $v(iT/N_j + \tau) \neq 0$

$$-\left(V\left(iT/N_j+\tau, y_j(iT/N_j+\tau-0)+\frac{T}{N_j}v(iT/N_j+\tau)\right)\right.$$
$$\left.-\ V(iT/N_j+\tau, y_j(iT/N_j+\tau-0))\right)$$
$$=\frac{T}{N_j}(p_j(iT/N_j+\tau), -v(iT/N_j+\tau)),$$

where

$$p_j(iT/N_j+\tau)\in\partial_y V(iT/N_j+\tau, z_j(iT/N_j+\tau)),$$
$$z_j(iT/N_j+\tau)=\lambda y_j(iT/N_j+\tau-0)+(1-\lambda)y_j(iT/N_j+\tau) \text{ for some } \lambda\in(0,1).$$

Since $|y(t)-y^*(t)|\le\delta''<\delta$ for all $t\in[0,T]$, by using (3.11), we obtain for j large enough

$$|y_j(iT/N_j+\tau-0)-y^*(iT/N_j+\tau)|\le\delta \text{ and } |y_j(iT/N_j+\tau)-y^*(iT/N_j+\tau)|\le\delta,$$

whence

$$|z_j(iT/N_j+\tau)-y^*(iT/N_j+\tau)|\le\delta, \qquad i=m_j,\dots,n_j.$$

Returning to (3.14), we have for sufficiently large j

$$\int_0^T (h(u(t))+g(y_j(t)))\,dt+\ell(y_j(T))-V(0,y^0)+\sum_{i=m_j}^{n_j}\frac{T}{N_j}k_\delta(iT/N_j+\tau,-v(iT/N_j+\tau))\ge 0.$$

Letting now $j\to\infty$, by Lebesgue dominated convergence theorem, (3.10) and (3.7) one obtains (3.3), which completes the proof of Lemma 3.1.

Remark 3.1. Obviously, in both our situations, $\overline{D(\phi)}$ is a (closed) convex cone in $\mathcal{H}$.

In other words, Lemma 3.1 states that $u^*, v^*\equiv 0$ and y_0 solve the following optimal control problem (remind that $0<\delta'<\delta$):

($\mathrm{P}_{\delta,\delta'}$) Minimize

$$\int_0^T (h(u(t))+g(y(t)))\,dt+\ell(y(T))+\int_0^T k_\delta(t,-v(t))\,dt-V(0,y^0)$$

over all $u\in L^2(0,T;\mathcal{U}), v\in L^2(0,T;\mathcal{H})$ with $v(t)\in K-\psi$ a.e. $t\in(0,T), y^0\in\mathcal{V}\cap D(\phi)$, where $y\in W^{1,2}([0,T];\mathcal{H})$ satisfies

$$\begin{cases} y'+Ay+\beta(y-\psi)\ni Bu+f+v \text{ a.e. in } (0,T),\\ y(0)=y^0, \end{cases}$$

and

$$|y(t)-y^*(t)|\le\delta' \text{ for all } t\in[0,T]. \tag{3.15}$$

But we can introduce the constraint on v and the state constraint (3.15) into the performance index. Indeed, for $\delta>0$, consider the function $I_\delta:[0,T]\times\mathcal{H}\to(-\infty,+\infty]$ defined by

$$I_\delta(t,y)=\begin{cases} 0 & \text{if } |y-y^*(t)|\le\dfrac{\delta}{2},\\ +\infty & \text{otherwise.}\end{cases}$$

Then, for $\delta' = \delta/2$, the problem $(P_{\delta,\delta'})$ is clearly equivalent with the following one:

(P_δ) Minimize

$$\begin{aligned}\int_0^T (h(u(t)) + g(y(t)))\,dt + \ell(y(T)) + \int_0^T k_\delta(t, -v(t))\,dt - V(0, y^0)\\ + \int_0^T I_{K-\psi}(v(t))\,dt + \int_0^T I_\delta(t, y(t))\,dt\end{aligned} \tag{3.16}$$

over all $u \in L^2(0,T;\mathcal{U}), v \in L^2(0,T;\mathcal{H}), y^0 \in \mathcal{V} \cap D(\phi)$, where $y \in W^{1,2}([0,T];\mathcal{H})$ satisfies

$$\begin{cases} y' + Ay + \beta(y - \psi) \ni Bu + f + v \text{ a.e. in } (0,T),\\ y(0) = y^0. \end{cases} \tag{3.17}$$

Suppose for the rest of the proof that the additional hypothesis (H9) holds (see Remark 2.3). If this is not the case, we impose on the initial state y^0 of the problem (P_δ) to be fixed, i.e., $y^0 = y_0$, the modifications which must be made being obvious.

As we have already mentioned in the introduction, the inclusions (2.15), (2.16) and (2.23), (2.24) will follow from a set of optimality conditions for (P_δ) as $\delta \to 0$. To this end, we associate with (P_δ) the following family of smooth (and penalized) problems:

$(P_{\delta,\eta})$ Minimize

$$\begin{aligned}\int_0^T (h(u(t)) + g_\eta(y(t)))\,dt + \ell_\eta(y(T)) + \int_0^T k_\delta(t, -v(t))\,dt - V_\eta(0, y^0)\\ + \int_0^T I_{K-\psi}(v(t))\,dt + \int_0^T I_{\delta,\eta}(t, y(t))\,dt\\ + \frac{1}{2}\int_0^T |u(t) - u^*(t)|^2\,dt + \frac{1}{2}\int_0^T |v(t)|^2\,dt\\ + \frac{1}{2}|V_\eta(0, y^0) - V_\eta(0, y_0)|^2 + \frac{1}{2}|y^0 - y_0|^2_{\mathcal{V}}\end{aligned} \tag{3.18}$$

over all $u \in L^2(0,T;\mathcal{U}), v \in L^2(0,T;\mathcal{H}), y^0 \in \mathcal{V} \cap D(\phi)$, where $y \in W^{1,2}([0,T];\mathcal{H})$ satisfies

$$\begin{cases} y' + Ay + \beta^\eta(y - \psi) = Bu + f + v \text{ a.e. in } (0,T),\\ y(0) = y^0. \end{cases} \tag{3.19}$$

Here $g_\eta, \ell_\eta, V_\eta$ are regularizations of g, ℓ, V, respectively (see [2, p. 28]), $I_{\delta,\eta}(t,\cdot)$ is the convex regularization of $I_\delta(t,\cdot)$, i.e., $I_{\delta,\eta}(t,y) = \inf\{|z-y|^2/2\eta : |z - y^*(t)| \le \delta/2\}$, and β^η is a certain regularization of β_η (see [2, p. 75]).

LEMMA 3.2. *Each of the problems* $(P_{\delta,\eta})$ *has at least one solution* $(u_{\delta,\eta}, v_{\delta,\eta}, y^0_{\delta,\eta}) \in L^2(0,T;\mathcal{U}) \times L^2(0,T;\mathcal{H}) \times (\mathcal{V} \cap D(\phi))$.

Proof. Let $\{(u_{\delta,\eta,m}, v_{\delta,\eta,m}, y^0_{\delta,\eta,m})\} \subset L^2(0,T;\mathcal{U}) \times L^2(0,T;\mathcal{H}) \times (\mathcal{V} \cap D(\phi))$ be a minimizing sequence for $(P_{\delta,\eta})$. For simplicity, we set $u_m = u_{\delta,\eta,m}, v_m = v_{\delta,\eta,m}$ and $y^0_m = y^0_{\delta,\eta,m}$. It is easy to see that

$$k_\delta(t,v) \ge -L|v|, (t,v) \in [0,T] \times \mathcal{H}, \tag{3.20}$$

where L is a nonnegative constant which depends only on the bounds of $t \mapsto |y^*(t)|$. Using (2.7), the boundedness from below by affine functions of g and ℓ, (3.20), and

the fact that $I_{K-\psi}$ and $I_{\delta,\eta}$ are nonnegative, we easily deduce that $\{u_m\}, \{v_m\}$, and $\{y_m^0\}$ are bounded in $L^2(0,T;\mathcal{U}), L^2(0,T;\mathcal{H})$, and $\mathcal{V}$, respectively. Hence, we find $(u_{\delta,\eta}, v_{\delta,\eta}, y_{\delta,\eta}^0) \in L^2(0,T;\mathcal{U}) \times L^2(0,T;\mathcal{H}) \times \mathcal{V}$ such that, on a subsequence of $\{m\}$, again denoted $\{m\}$, $u_m \to u_{\delta,\eta}$ weakly in $L^2(0,T;\mathcal{U})$, $v_m \to v_{\delta,\eta}$ weakly in $L^2(0,T;\mathcal{H})$, and $y_m^0 \to y_{\delta,\eta}^0$ weakly in $\mathcal{V}$ and strongly in $\mathcal{H}$ (because the inclusion $\mathcal{V} \subset \mathcal{H}$ is compact). Moreover, by hypothesis (H9) (which is verified also for β given by (2.18)) and the lower semicontinuity of ϕ, we see that $\phi(y_{\delta,\eta}^0) < +\infty$, that is, $y_{\delta,\eta}^0 \in D(\phi)$.

Now let y_m be the solution of (3.19) corresponding to $u = u_m, v = v_m$, and $y^0 = y_m^0$. We can rewrite (3.19) as

$$\begin{cases} y_m' + Ay_m + \beta_\eta(y_m - \psi) = Bu_m + f + v_m + \beta_\eta(y_m - \psi) - \beta^\eta(y_m - \psi), \\ y_m(0) = y_m^0. \end{cases} \tag{3.21}$$

Note that, by definition of β^η, $|\beta^\eta(r) - \beta_\eta(r)| \le 2\eta$ for all $r \in \mathbf{R}$ (see [2, p. 75]). Define

$$\phi_\eta(y) = \int_\Omega \int_0^{y(x)-\psi(x)} \beta_\eta(r)\, dr\, dx \text{ for all } y \in \mathcal{H}.$$

It is easy to see that ϕ_η is just the convex regularization of ϕ, i.e., $\phi_\eta(y) = \inf\{|z - y|^2/2\eta + \phi(z) : z \in \mathcal{H}\}$. Moreover, for every $y \in \mathcal{H}$, we have

$$\nabla \phi_\eta(y) = \beta_\eta(y - \psi) \text{ a.e. in } \Omega.$$

Fix $y^1 \in \mathcal{V} \cap D(\partial\phi)$. We multiply (scalarly in $\mathcal{H}$) (3.21) first by $y_m - y^1$ and then by y_m', and next we integrate on $[0,t]$. After some calculation (do not forget (2.4)), we obtain for every $t \in [0,T]$:

$$\begin{aligned} |y_m(t)| \le \Bigg(|y_m^0| + \int_0^t (|Bu_m(s)| + |v_m(s)| + |f(s)|)\, ds \\ + \ t(|Ay^1| + |(\partial\phi)^0(y^1)| + 2\eta(\mathrm{meas}\Omega)^{1/2}) \Bigg) e^{\alpha t} + (1 + e^{\alpha t})|y^1| \end{aligned}$$

and

$$\begin{aligned} &\frac{1}{2}\int_0^t |y_m'(s)|^2\, ds + \frac{1}{2}\omega|y_m(t)|_{\mathcal{V}}^2 + \phi_\eta(y_m(t)) \le \frac{1}{2}(Ay_m^0, y_m^0) + \phi_\eta(y_m^0) \\ &+\frac{1}{2}\alpha|y_m(t)|^2 + 2\int_0^t (|Bu_m(s)|^2 + |v_m(s)|^2 + |f(s)|^2)\, ds + 8t\eta^2 \mathrm{meas}\Omega. \end{aligned}$$

Here $|(\partial\phi)^0(y^1)| = \inf\{|z| : z \in \partial\phi(y^1)\}$. Using the boundedness in $\mathcal{V}$ of y_m^0 and the hypothesis (H9), we see that (Ay_m^0, y_m^0) and $\phi_\eta(y_m^0)$ ($\phi_\eta(y_m^0) \le \phi(y_m^0)$) are bounded. Thus, in view of the above two estimates, we may apply the Arzelà–Ascoli theorem (do not forget that the inclusion $\mathcal{V} \subset \mathcal{H}$ is compact) to conclude that, on a subsequence,

$$y_{\delta,\eta,m} \to \overline{y}_{\delta,\eta} \text{ strongly in } C([0,T];\mathcal{H}) \text{ as } m \to \infty. \tag{3.22}$$

Denote by $y_{\delta,\eta}$ the solution of the equation (with initial condition)

$$\begin{cases} y_{\delta,\eta}' + Ay_{\delta,\eta} + \beta^\eta(y_{\delta,\eta} - \psi) = Bu_{\delta,\eta} + f + v_{\delta,\eta}, \\ y_{\delta,\eta}(0) = y_{\delta,\eta}^0. \end{cases} \tag{3.23}$$

Subtracting (3.23) from (3.21) and multiplying the difference by $y_{\delta,\eta,m}-y_{\delta,\eta}$, we have (after an integration on $[0,t]$)

$$\frac{1}{2}|y_{\delta,\eta,m}(t)-y_{\delta,\eta}(t)|^2 \le \frac{1}{2}|y^0_{\delta,\eta,m}-y^0_{\delta,\eta}|^2+\alpha\int_0^t |y_{\delta,\eta,m}(s)-y_{\delta,\eta}(s)|^2\,ds$$
$$+\int_0^t (B(u_{\delta,\eta,m}(s)-u_{\delta,\eta}(s))+v_{\delta,\eta,m}(s)-v_{\delta,\eta}(s), y_{\delta,\eta,m}(s)-y_{\delta,\eta}(s))\,ds.$$

Letting $m\to\infty$ and taking (3.22) into account, we obtain

$$y_{\delta,\eta,m}\to y_{\delta,\eta} \text{ strongly in } C([0,T];\mathcal{H}).$$

Finally, by standard arguments, it follows that $(u_{\delta,\eta}, v_{\delta,\eta}, y^0_{\delta,\eta})$ solves the problem $(\mathrm{P}_{\delta,\eta})$, thereby completing the proof.

Fix now $\delta>0$. The following result is an effect of the penalization terms.

LEMMA 3.3. *Let $(u_{\delta,\eta}, v_{\delta,\eta}, y^0_{\delta,\eta})\in L^2(0,T;\mathcal{U})\times L^2(0,T;\mathcal{H})\times(\mathcal{V}\cap D(\phi))$ be a solution for the problem $(\mathrm{P}_{\delta,\eta})$. Then we have as $\eta\to 0$*

$$u_{\delta,\eta}\to u^* \textit{ strongly in } L^2(0,T;\mathcal{U}),$$
$$v_{\delta,\eta}\to v^*\equiv 0 \textit{ strongly in } L^2(0,T;\mathcal{H}),$$
$$y^0_{\delta,\eta}\to y_0 \textit{ strongly in } \mathcal{V}.$$

Proof. Let us denote by $J_{\delta,\eta}(u,v,y^0)$ the functional (3.18) and by $J_\delta(u,v,y^0)$ the functional (3.16). Also, we denote by $\tilde{J}_{\delta,\eta}(u,v,y^0)$ the functional obtained from (3.18) by eliminating the penalization terms $\frac{1}{2}\int_0^T|u(t)-u^*(t)|^2\,dt$, $\frac{1}{2}\int_0^T|v(t)|^2\,dt$, $\frac{1}{2}|V_\eta(0,y^0)-V_\eta(0,y_0)|^2$, and $\frac{1}{2}|y^0-y_0|^2_{\mathcal{V}}$.

Let $y^*_{\delta,\eta}$ be the solution of (3.19) corresponding to $u=u^*$, $v=v^*\equiv 0$, and $y^0=y_0$. Arguing as below, one shows that $y^*_{\delta,\eta}\to y^*$ strongly in $C([0,T];\mathcal{H})$ as $\eta\to 0$. At the same time one observes that $I_{\delta,\eta}(t,y^*_{\delta,\eta}(t))=0$ when $|y^*_{\delta,\eta}(t)-y^*(t)|\le\delta/2$. Taking these into account, for each $\delta>0$, we find $\eta_\delta>0$ such that if $\eta\le\eta_\delta$,

$$J_{\delta,\eta}(u_{\delta,\eta},v_{\delta,\eta},y^0_{\delta,\eta})\le J_{\delta,\eta}(u^*,0,y^0)\le 1. \tag{3.24}$$

Now, as in the proof of the preceding lemma, there exists $M>0$ (which can be chosen independent of δ) such that if $\eta\le\eta_\delta$,

$$\left(\int_0^T|u_{\delta,\eta}(t)|^2\,dt\right)^{1/2}\le M,\qquad \left(\int_0^T|v_{\delta,\eta}(t)|^2\,dt\right)^{1/2}\le M,\qquad |y^0_{\delta,\eta}|_{\mathcal{V}}\le M.$$

Consequently, there exists $(u_\delta,v_\delta,y^0_\delta)\in L^2(0,T;\mathcal{U})\times L^2(0,T;\mathcal{H})\times(\mathcal{V}\cap D(\phi))$ such that, on a subsequence of $\{\eta\}$,

$$u_{\delta,\eta}\to u_\delta \text{ weakly in } L^2(0,T;\mathcal{U}),$$
$$v_{\delta,\eta}\to v_\delta \text{ weakly in } L^2(0,T;\mathcal{H}),$$
$$y^0_{\delta,\eta}\to y^0_\delta \text{ weakly in } \mathcal{V}.$$

Next, let $y_{\delta,\eta}$ be the solution of (3.23) (where, of course, $u_{\delta,\eta}$, $v_{\delta,\eta}$, and $y^0_{\delta,\eta}$ are those from the statement of this lemma). We shall show that if the above convergences hold, then $y_{\delta,\eta}$ converges (possibly on a subsequence) to the solution of (3.17)

corresponding to $u = u_\delta$, $v = v_\delta$, $y^0 = y^0_\delta$. As in the proof of Lemma 3.2, one obtains

$$|y_{\delta,\eta}(t)| \ \leq \ \text{const. and} \ \int_0^T |y'_{\delta,\eta}(s)|^2 \, ds + |y_{\delta,\eta}(t)|^2_{\mathcal{V}} \ \leq \ \text{const. for } t \in [0,T],$$

where the constants are independent of η (and even of δ). Multiplying (3.23) by $\nabla\phi_\eta(y_{\delta,\eta}(t))$, we also obtain (after an integration on $[0,T]$)

$$\int_0^T |\nabla\phi_\eta(y_{\delta,\eta}(s))|^2 \, ds \leq \ \text{const.},$$

with the constant as above. Thus we have as $\eta \to 0$

$$y_{\delta,\eta} \to y_\delta \text{ strongly in } C([0,T];\mathcal{H}) \text{ and weakly in } W^{1,2}([0,T];\mathcal{H})$$

$$(\text{therefore } y'_{\delta,\eta} \to y'_\delta \text{ weakly in } L^2(0,T;\mathcal{H})),$$

$$\nabla\phi_\eta(y_{\delta,\eta}) \to \xi_\delta \text{ weakly in } L^2(0,T;\mathcal{H}).$$

Hence,

$$Ay_{\delta,\eta} \to Bu_\delta + f + v_\delta - y'_\delta - \xi_\delta \text{ weakly in } L^2(0,T;\mathcal{H}).$$

But since $Ay_{\delta,\eta} \to Ay_\delta$ weakly in $L^2(0,T;\mathcal{V}')$ (because $y_{\delta,\eta} \to y_\delta$ weakly in $L^2(0,T;\mathcal{V})$), we have

$$y'_\delta + Ay_\delta + \xi_\delta = Bu_\delta + f + v_\delta \text{ a.e. in } (0,T).$$

Now we define the lower semicontinuous convex function $\Phi : L^2(0,T;\mathcal{H}) \to (-\infty, +\infty]$ by

$$\Phi(y) = \int_0^T \phi(y(t)) \, dt.$$

We have

$$\partial\Phi(y) = \{w \in L^2(0,T;\mathcal{H}) : w(t) \in \partial\phi(y(t)) \text{ a.e. } t \in (0,T)\}$$

and, if we denote by Φ_η the convex regularization of Φ,

$$\nabla\Phi_\eta(y_{\delta,\eta}) = \nabla\phi_\eta(y_{\delta,\eta}) \text{ a.e. in } (0,T).$$

Since $y_{\delta,\eta} \to y_\delta$ strongly in $L^2(0,T;\mathcal{H})$ and $\nabla\Phi_\eta(y_{\delta,\eta}) \to \xi_\delta$ weakly in $L^2(0,T;\mathcal{H})$, by a well-known result (see [2, Thm. 1.2]) $\xi_\delta \in \partial\Phi(y_\delta)$, whence, as we have stated above,

$$\begin{cases} y'_\delta + Ay_\delta + \partial\phi(y_\delta) \ni Bu_\delta + f + v_\delta \text{ a.e. in } (0,T), \\ y_\delta(0) = y^0_\delta. \end{cases}$$

We still need the following semicontinuity property:

$$\liminf_{\eta \to 0} \int_0^T I_{\delta,\eta}(t, y_{\delta,\eta}(t)) \, dt \geq \int_0^T I_\delta(t, y_\delta(t)) \, dt. \tag{3.25}$$

To see this, let us observe that, by virtue of (3.24), $\int_0^T I_{\delta,\eta}(t, y_{\delta,\eta}(t))\,dt$ is bounded with respect to η. Hence, by a well-known expression of the convex regularization (see [3, p. 121]), it readily follows that

$$(I + \eta\partial I_\delta(t,\cdot))^{-1} y_{\delta,\eta}(t) - y_{\delta,\eta}(t) \to 0 \text{ strongly in } L^2(0,T;\mathcal{H}) \text{ as } \eta \to 0,$$

whence, by the lower semicontinuity of $y \mapsto I_\delta(t,y)$, we derive (3.25).

Now, using the convergence of $\{u_{\delta,\eta}\}, \{v_{\delta,\eta}\}$, and $\{y_{\delta,\eta}\}$ in conjunction with (3.25) and the weak lower semicontinuity of $u \mapsto \int_0^T (h(u(t)) + \frac{1}{2}|u(t) - u^*(t)|^2)\,dt$, $v \mapsto \int_0^T (k_\delta(t,-v(t)) + I_{K-\psi}(v(t)) + \frac{1}{2}|v(t)|^2)\,dt$, and $y^0 \mapsto \frac{1}{2}|y^0 - y_0|^2_{\mathcal{V}}$, we obtain the following chain of inequalities:

$$\begin{aligned}
&\liminf_{\eta\to 0} J_{\delta,\eta}(u_{\delta,\eta}, v_{\delta,\eta}, y^0_{\delta,\eta}) \geq \liminf_{\eta\to 0} \tilde{J}_{\delta,\eta}(u_{\delta,\eta}, v_{\delta,\eta}, y^0_{\delta,\eta}) \\
&\geq J_\delta(u_\delta, v_\delta, y^0_\delta) \geq J_\delta(u^*, 0, y_0) = \lim_{\eta\to 0} J_{\delta,\eta}(u^*, 0, y_0) \\
&\geq \limsup_{\eta\to 0} J_{\delta,\eta}(u_{\delta,\eta}, v_{\delta,\eta}, y^0_{\delta,\eta}) \geq \limsup_{\eta\to 0} \tilde{J}_{\delta,\eta}(u_{\delta,\eta}, v_{\delta,\eta}, y^0_{\delta,\eta}).
\end{aligned}$$

Consequently,

$$\lim_{\eta\to 0} J_{\delta,\eta}(u_{\delta,\eta}, v_{\delta,\eta}, y^0_{\delta,\eta}) = \lim_{\eta\to 0} \tilde{J}_{\delta,\eta}(u_{\delta,\eta}, v_{\delta,\eta}, y^0_{\delta,\eta}) = J_\delta(u^*, 0, y_0),$$

and hence the conclusion of the lemma follows.

Remark 3.2. As a by-product of the preceding proof we have obtained the following statement: If

$$(u_{\delta,\eta}, v_{\delta,\eta}, y^0_{\delta,\eta}) \to (u_\delta, v_\delta, y^0_\delta) \text{ weakly in } L^2(0,T;\mathcal{U}) \times L^2(0,T;\mathcal{H}) \times \mathcal{V} \text{ as } \eta \to 0,$$

then, on a subsequence of $\{\eta\}$,

$$y_{\delta,\eta} \to y_\delta \text{ strongly in } C([0,T];\mathcal{H})$$

and

$$\beta_\eta(y_{\delta,\eta} - \psi) \to \xi \text{ weakly in } L^2(0,T;\mathcal{H}),$$

where y_δ and ξ satisfy

$$\begin{cases} y'_\delta + Ay_\delta + \xi = Bu_\delta + f + v_\delta \text{ a.e. in } (0,T), \\ y_\delta(0) = y^0_\delta, \end{cases}$$

and

$$\xi \in \beta(y_\delta - \psi) \text{ a.e. in } (0,T).$$

Remark 3.3. One can obtain in addition

$$y_{\delta,\eta} \to y_\delta \text{ strongly in } L^2(0,T;\mathcal{V}) \text{ as } \eta \to 0$$

(see [2, Thm. 4.5]).

In order to obtain the needed optimality conditions for $(\mathrm{P}_{\delta,\eta})$, we shall compare the optimal value with the values of $J_{\delta,\eta}$ corresponding to $u = u_{\delta,\eta} + \lambda\overline{u}, v = v_{\delta,\eta} + \lambda\overline{v}$,

and $y^0 = y^0_{\delta,\eta} + \lambda\overline{y}^0$, where $0 < \lambda < 1$ and $\overline{u} \in L^2(0,T;\mathcal{U}), \overline{v} \in L^2(0,T;\mathcal{H}), \overline{y}^0 \in (\mathcal{V} \cap D(\phi)) - y^0_{\delta,\eta}$ are arbitrary. We have

(3.26)
$$J_{\delta,\eta}(u_{\delta,\eta} + \lambda\overline{u}, v_{\delta,\eta} + \lambda\overline{v}, y^0_{\delta,\eta} + \lambda\overline{y}^0) - J_{\delta,\eta}(u_{\delta,\eta}, v_{\delta,\eta}, y^0_{\delta,\eta}) \geq 0$$
$$\text{for all } \lambda \in (0,1) \text{ and } (\overline{u},\overline{v},\overline{y}^0) \in L^2(0,T;\mathcal{U}) \times L^2(0,T;\mathcal{H}) \times ((\mathcal{V} \cap D(\phi)) - y^0_{\delta,\eta}).$$

(Note that $y^0_{\delta,\eta} + \lambda\overline{y}^0 \in \mathcal{V} \cap D(\phi)$ since $\mathcal{V} \cap D(\phi)$ is convex.)

Let $y_{\delta,\eta,\lambda}$ be the solution of (3.19) corresponding to $u = u_{\delta,\eta} + \lambda\overline{u}, v = v_{\delta,\eta} + \lambda\overline{v}$, and $y^0 = y^0_{\delta,\eta} + \lambda\overline{y}^0$. It is easy to show that

$$\frac{1}{\lambda}(y_{\delta,\eta,\lambda} - y_{\delta,\eta}) \to z_{\delta,\eta} \text{ strongly in } C([0,T];\mathcal{H}) \text{ as } \eta \to 0,$$

where $z_{\delta,\eta}$ is the solution of the following equation (with initial condition):

$$\begin{cases} z' + Az + (\beta^\eta)'(y_{\delta,\eta} - \psi)z = B\overline{u} + \overline{v} \text{ a.e. in } (0,T), \\ z(0) = \overline{y}^0. \end{cases} \tag{3.27}$$

Divide (3.26) by λ and then let $\lambda \to 0$. Since $I_{\delta,\eta}(t, y_{\delta,\eta,\lambda}(t)) = I_{\delta,\eta}(t, y_{\delta,\eta}(t)) = 0$ for all $t \in [0,T]$ if $\eta > 0$ and $\lambda > 0$ are sufficiently small, we obtain for η small enough

$$\begin{aligned} &\int_0^T h'(u_{\delta,\eta}(t); \overline{u}(t))\,dt + \int_0^T (\nabla g_\eta(y_{\delta,\eta}(t)), z_{\delta,\eta}(t))\,dt \\ &+ (\nabla \ell_\eta(y_{\delta,\eta}(T)), z_{\delta,\eta}(T)) + \int_0^T k'_\delta(t, -v_{\delta,\eta}(t); -\overline{v}(t))\,dt \\ &- (\nabla_y V_\eta(0, y^0_{\delta,\eta}), \overline{y}^0) + \int_0^T I'_{K-\psi}(v_{\delta,\eta}(t); \overline{v}(t))\,dt \\ &+ \int_0^T (u_{\delta,\eta}(t) - u^*(t), \overline{u}(t))\,dt + \int_0^T (v_{\delta,\eta}(t), \overline{v}(t))\,dt \\ &+ ((V_\eta(0, y^0_{\delta,\eta}) - V_\eta(0, y_0)) \cdot \nabla_y V_\eta(0, y^0_{\delta,\eta}), \overline{y}^0) + (y^0_{\delta,\eta} - y_0, \overline{y}^0)_{\mathcal{V}} \geq 0 \\ &\text{for all } (\overline{u},\overline{v},\overline{y}^0) \in L^2(0,T;\mathcal{U}) \times L^2(0,T;\mathcal{H}) \times ((\mathcal{V} \cap D(\phi)) - y^0_{\delta,\eta}), \end{aligned} \tag{3.28}$$

where $h'(u_{\delta,\eta}(t); \overline{u}(t))$ and $k'_\delta(t, -v_{\delta,\eta}(t); -\overline{v}(t)), I'_{K-\psi}(v_{\delta,\eta}(t); \overline{v}(t))$ denote the directional derivatives of h and $v \mapsto k_\delta(t,v), I_{K-\psi}$ at $u_{\delta,\eta}(t)$ and $-v_{\delta,\eta}(t)$ in the directions $\overline{u}(t)$ and $-\overline{v}(t)$, respectively.

Next, define $p_{\delta,\eta}$ as the unique solution in $C([0,T];\mathcal{H}) \cap L^2(0,T;\mathcal{V})$ (with $p'_{\delta,\eta} \in L^2(0,T;\mathcal{V}')$) of the following equation (with initial condition):

$$\begin{cases} p' - Ap - (\beta^\eta)'(y_{\delta,\eta} - \psi)p = \nabla g_\eta(y_{\delta,\eta}) \text{ a.e. in } (0,T), \\ p(T) = -\nabla \ell_\eta(y_{\delta,\eta}(T)). \end{cases} \tag{3.29}$$

Some calculation in (3.28) involving (3.27) and (3.29) together with an integration by

parts gives for sufficiently small $\eta > 0$

$$\begin{aligned}
&\int_0^T (h'(u_{\delta,\eta};\overline{u}) + (-B^*p_{\delta,\eta} + u_{\delta,\eta} - u^*, \overline{u}))\,dt \\
&+ \int_0^T (k'_\delta(t, -v_{\delta,\eta}; -\overline{v}) + I'_{K-\psi}(v_{\delta,\eta};\overline{v}) + (-p_{\delta,\eta} + v_{\delta,\eta}, \overline{v}))\,dt \\
&+ (-p_{\delta,\eta}(0) - \nabla_y V_\eta(0, y^0_{\delta,\eta}) + (V_\eta(0, y^0_{\delta,\eta}) - V_\eta(0, y_0)) \cdot \nabla_y V_\eta(0, y^0_{\delta,\eta}), \overline{y}^0) \\
&+ (y^0_{\delta,\eta} - y_0, \overline{y}^0)_{\mathcal{V}} \geq 0 \\
&\qquad \text{for all } (\overline{u}, \overline{v}, \overline{y}^0) \in L^2(0,T;\mathcal{U}) \times L^2(0,T;\mathcal{H}) \times ((\mathcal{V} \cap D(\phi)) - y^0_{\delta,\eta}).
\end{aligned}$$

So, using [3, Prop. 2.3, p. 106], we infer for $\eta > 0$ small enough

$$B^*p_{\delta,\eta}(t) + u^*(t) - u_{\delta,\eta}(t) \in \partial h(u_{\delta,\eta}(t)) \text{ a.e. } t \in (0,T), \tag{3.30}$$

$$-p_{\delta,\eta}(t) + v_{\delta,\eta}(t) \in \partial_v k_\delta(t, -v_{\delta,\eta}(t)) + \partial I_{K-\psi}(v_{\delta,\eta}(t)) \text{ a.e. } t \in (0,T), \tag{3.31}$$

$$\begin{aligned}
&(-p_{\delta,\eta}(0) - \nabla_y V_\eta(0, y^0_{\delta,\eta}) + (V_\eta(0, y^0_{\delta,\eta}) - V_\eta(0, y_0)) \cdot \nabla_y V_\eta(0, y^0_{\delta,\eta}), \overline{z}^0 - y^0_{\delta,\eta}) \\
&\qquad + (y^0_{\delta,\eta} - y_0, \overline{z}^0 - y^0_{\delta,\eta})_{\mathcal{V}} \geq 0 \text{ for all } \overline{z}^0 \in \mathcal{V} \cap D(\phi).
\end{aligned} \tag{3.32}$$

Now, we shall take limits in (3.30)–(3.32) as $\delta \to 0$, $\eta \to 0$. To this end, we use Lemma 3.3. Thus, for each $\delta > 0$, we choose $\eta_\delta > 0$ such that

$$\int_0^T |u_{\delta,\eta_\delta}(t) - u^*(t)|^2\,dt < \delta, \ \int_0^T |v_{\delta,\eta_\delta}(t)|^2\,dt < \delta, \ |y^0_{\delta,\eta_\delta} - y_0|_{\mathcal{V}} < \delta. \tag{3.33}$$

Obviously, we can choose η_δ such that, in addition, $\lim_{\delta\to 0} \eta_\delta = 0$. For simplicity, set $u_{\delta,\eta_\delta} = u_\delta$, $v_{\delta,\eta_\delta} = v_\delta$, and $y^0_{\delta,\eta_\delta} = y^0_\delta$. Let $y_\delta = y_{\delta,\eta_\delta}$ be the solution of (3.19) corresponding to $\eta = \eta_\delta$, $u = u_\delta$, $v = v_\delta$, and $y^0 = y^0_\delta$. Since, by (3.33), we have as $\delta \to 0$

$$u_\delta \to u^* \text{ strongly in } L^2(0,T;\mathcal{U}), \tag{3.34}$$

$$v_\delta \to 0 \text{ strongly in } L^2(0,T;\mathcal{H}), \tag{3.35}$$

$$y^0_\delta \to y_0 \text{ strongly in } \mathcal{V}, \tag{3.36}$$

likewise as in the proof of Lemma 3.3 (see also Remark 3.2), one obtains on a subsequence of $\{\delta\}$:

$$y_\delta \to y^* \text{ strongly in } C([0,T];\mathcal{H}), \tag{3.37}$$

$$\begin{aligned}
&\beta_{\eta_\delta}(y_\delta - \psi) \to \xi \text{ weakly in } L^2(0,T;\mathcal{H}), \\
&\text{where } \xi = Bu^* + f - y^{*'} - Ay^* \in \beta(y^* - \psi) \text{ a.e. in } (0,T).
\end{aligned} \tag{3.38}$$

Also we have

$$V_{\eta_\delta}(0, y^0_\delta) - V_{\eta_\delta}(0, y_0) \to 0 \text{ as } \delta \to 0. \tag{3.39}$$

Let p_δ be the solution of (3.29) corresponding to $\eta = \eta_\delta$. The proof of the following lemma is given in [2, Lem. 5.3]. (Note that only here one needs the assumption (2.5).)

LEMMA 3.4. *For all $\delta > 0$, we have*

$$|p_\delta(t)| \le \text{const. } for \text{ all } t \in [0,T],$$
$$\int_0^T |p_\delta(t)|_{\mathcal{V}}^2 \, dt \le \text{const.},$$
$$\int_Q |(\beta^{\eta_\delta})'(y_\delta - \psi)p_\delta| \, dx \, dt \le \text{const.},$$

where the constants are independent of δ.

By Lemma 3.4, arguing as in [2, p. 181], we find $p \in BV([0,T];\mathcal{Y}') \cap L^\infty(0,T;\mathcal{H}) \cap L^2(0,T;\mathcal{V})$ and $\mu \in (L^\infty(Q))'$ such that, on a subsequence of $\{\delta\}$, $p_\delta(t) \to p(t)$ strongly in $\mathcal{Y}'$ for every $t \in [0,T]$ and

$$p_\delta \to p \text{ weak star in } L^\infty(0,T;\mathcal{H}) \text{ and weakly in } L^2(0,T;\mathcal{V}), \tag{3.40}$$

$$p_\delta \to p \text{ strongly in } L^2(0,T;\mathcal{H}), \tag{3.41}$$

$$(\beta^{\eta_\delta})'(y_\delta - \psi)p_\delta \to \mu \text{ weak star in } (L^\infty(Q))'. \tag{3.42}$$

Using (3.37) in conjunction with (H8), we obtain on a subsequence of $\{\delta\}$:

$$\nabla g_{\eta_\delta}(y_\delta) \to \gamma \text{ weak star in } L^\infty(0,T;\mathcal{H}). \tag{3.43}$$

Next, an application of [2, Lem. 5.4] (do not forget (3.37)) gives

$$\gamma(t) \in \partial g(y^*(t)) \text{ a.e. } t \in (0,T). \tag{3.44}$$

Similarly,

$$\nabla \ell_{\eta_\delta}(y_\delta(T)) = -p_\delta(T) \to -p(T) \text{ weakly in } \mathcal{H} \text{ (and strongly in } \mathcal{Y}').$$

Applying this time [2, Prop. 1.12], we obtain (2.13) (or (2.21)).

Likewise as above (take (3.36) into account),

$$\nabla_y V_{\eta_\delta}(0, y_\delta^0) \to q \text{ weakly in } \mathcal{H}, \text{ where } q \in \partial_y V(0, y_0).$$

By Lemma 3.4, we have on a subsequence of $\{\delta\}$

$$p_\delta(0) \to p(0) \text{ weakly in } \mathcal{H} \text{ (and strongly in } \mathcal{Y}').$$

Hence, as $\delta \to 0$, (3.32) (with $\eta = \eta_\delta$) becomes (do not forget (3.36), (3.39))

$$(p(0) + q, \overline{z}^0 - y_0) \le 0 \text{ for all } \overline{z}^0 \in \mathcal{V} \cap D(\phi).$$

Since $\mathcal{V} \cap D(\phi)$ is dense in $K = \overline{D(\phi)}$, the above inequality holds for all $\overline{z}^0 \in K$. But this means (2.16) or (2.24). (In the case of Theorem 2.1, $K = \mathcal{H}$; therefore $I_K \equiv 0$ and $\partial I_K(y_0) = \{0\}$.)

Now letting $\delta \to 0$ in (3.29) with $\eta = \eta_\delta$, by (3.40)–(3.44), it follows that

$$p' - Ap - \mu \in \partial g(y^*) \text{ a.e. in } (0,T),$$

where $p' - Ap - \mu = \gamma \in L^\infty(0,T;\mathcal{H})$; but this is just (2.11). (Recall that p' is considered in the sense of $\mathcal{V}'$-valued distributions.)

Passing to the limit also in (3.30) with $\eta = \eta_\delta$ as $\delta \to 0$, since $\partial h : \mathcal{U} \to \mathcal{U}$ is demiclosed, by (3.34), (3.41) we obtain (2.14) (or (2.22)).

It remains to take limits only in (3.31) with $\eta = \eta_\delta$. To this end, let us write (3.31) as

$$p_\delta(t) - v_\delta(t) = p_\delta^{(1)}(t) + p_\delta^{(2)}(t) \text{ a.e. } t \in (0,T), \tag{3.45}$$

where

$$-p_\delta^{(1)}(t) \in \partial_v k_\delta(t, -v_\delta(t)) \text{ a.e. } t \in (0,T) \tag{3.46}$$

and

$$-p_\delta^{(2)}(t) \in \partial I_{K-\psi}(v_\delta(t)) \text{ a.e. } t \in (0,T). \tag{3.47}$$

Since $v \mapsto k_\delta(t,v)$ is Lipschitz continuous on $\mathcal{H}$ uniformly with respect to $t \in [0,T]$, $p_\delta^{(1)}$ is bounded in $L^\infty(0,T;\mathcal{H})$; consequently, it is bounded in $L^2(0,T;\mathcal{H})$. But, by (3.40) and (3.35), the same is true also for $p_\delta^{(2)}$. Therefore, on a subsequence of $\{\delta\}$, we have

$$p_\delta^{(1)} \to p^{(1)} \text{ weakly in } L^2(0,T;\mathcal{H}) \text{ (even weak star in } L^\infty(0,T;\mathcal{H})), \tag{3.48}$$

$$p_\delta^{(2)} \to p^{(2)} \text{ weakly in } L^2(0,T;\mathcal{H}). \tag{3.49}$$

Taking (3.40), (3.35), (3.48), and (3.49) into account, let $\delta \to 0$ in (3.45). We obtain

$$p(t) = p^{(1)}(t) + p^{(2)}(t) \text{ a.e. } t \in (0,T). \tag{3.50}$$

Next, (3.35), (3.49), and (3.47) give

$$-p^{(2)}(t) \in \partial I_{K-\psi}(0) = \partial I_K(\psi) \text{ a.e. } t \in (0,T). \tag{3.51}$$

Finally, we shall briefly repeat (with small differences) considerations from [7], but in our infinite-dimensional context. Let us clarify at first the meaning of (3.46). We assert that (3.46) implies

$$-p_\delta^{(1)}(t) \in \overline{\text{co}} \bigcup_{|y-y^*(t)|\le\delta} \partial_y V(t,y) \text{ a.e. } t \in (0,T). \tag{3.52}$$

Indeed, otherwise, there exists a closed hyperplane strictly separating $-p_\delta^{(1)}(t)$ and the closed convex set from the right-hand side of (3.52) (see [3, Cor. 1.9, p. 22]), i.e., there is $v \in \mathcal{H}$ such that

$$\begin{aligned}
&(-p_\delta^{(1)}(t), -v + v_\delta(t)) \\
>\ &\sup\left\{(p, -v + v_\delta(t)) : p \in \overline{\text{co}} \bigcup_{|y-y^*(t)|\le\delta} \partial_y V(t,y)\right\} \\
=\ &\sup\left\{(p, -v + v_\delta(t)) : p \in \bigcup_{|y-y^*(t)|\le\delta} \partial_y V(t,y)\right\} \\
\ge\ &\sup\{(p,-v) - k_\delta(t, -v_\delta(t)) : p \in \partial_y V(t,y), |y - y^*(t)| \le \delta\} \\
=\ &k_\delta(t,-v) - k_\delta(t, -v_\delta(t)),
\end{aligned}$$

which contradicts (3.46). (Here we have used also the definition (3.2).) Now, we take limits in (3.52). Let t be fixed but arbitrary in the subset of full measure of $[0,T]$ in which (3.52) holds on a subsequence of $\{\delta\}$. As before, $p_\delta^{(1)}(t)$ is bounded in $\mathcal{H}$; therefore, on a subsequence of $\{\delta\}$,

$$p_\delta^{(1)}(t) \to p^{(1)}(t) \text{ weakly in } \mathcal{H}.$$

Letting $\delta \to 0$ in (3.52), we obtain

$$-p^{(1)}(t) \in \bigcap_{\delta>0} \overline{\text{co}} \bigcup_{|y-y^*(t)|\le\delta} \partial_y V(t,y)$$

for all t in a subset of full measure of $[0,T]$. But this implies

$$-p^{(1)}(t) \in \partial_y V(t,y^*(t)) \text{ a.e. } t \in (0,T) \tag{3.53}$$

by the same argument as at the end of the proof of Theorem 3.1 from [7]. (Among others things, one uses Propositions 2.1.2 and 2.1.1 from [5].) We can write (3.50), (3.51), and (3.53) as (2.15) or (2.23).

As regards the proofs of (2.12) and (2.19), (2.20) as well as the proof of the last assertion of Theorem 2.1, we refer to [2, Thms. 5.1, 5.2]. (These will require also (3.38).) Thus, the proofs of Theorems 2.1 and 2.2 are complete.

Remark 3.4. A discrete variant of (1.2) also holds when the equation (2.2) is replaced by a certain Trotter product formula approximation (see [10, Thm. 6.1]).

Remark 3.5. While the arising of the normal cone $N_K(y_0)$ in (2.24) is (to all appearances) intrinsic for the case when $K \neq \mathcal{H}$, that of $N_K(\psi)$ in (2.23) seems to be caused by our approach. Indeed, in the proof of Lemma 3.1, we need that $e^{-\mathcal{B}_k t}$ maps K into K (to make sense $e^{-\mathcal{A}_k t}e^{-\mathcal{B}_k t}y$ for all $y \in K$), so we must impose on the auxiliary control v the constraint $v(t) \in K - \psi$ a.e. $t \in (0,T)$. But there are signs (see also the formula (6.24) from [10], to which Remark 3.4 refers) that the presence of $N_K(\psi)$ in (2.23) can be removed. Can a modification of our approximation procedure for the perturbed state equation (3.1) lead to such an effect (without making major changes in our approach)?

4. The convex case. Inspired by [7], we also raise the question: Can the inclusion (1.2) be satisfied *everywhere* on $[0,T]$ (instead of *almost everywhere*) in our infinite-dimensional context? We shall show that the answer is positive at least in the case when the control system is governed by semilinear parabolic equations with β nondecreasing and concave, and the functions g and ℓ are convex. More specifically, we impose on g and ℓ the following hypothesis:

(H8)′ $g, \ell : \mathcal{H} \to \mathbf{R}$ are convex and bounded on bounded subsets.

It is clear (by virtue of some well-known results) that (H8) follows from (H8)′. In particular, (H8)′ implies that g and ℓ are Lipschitz continuous on bounded subsets of $\mathcal{H}$ and, at every point, the generalized gradients of g and ℓ coincide with their convex subdifferentials. An additional monotonicity condition is assumed to be verified by the functions g and ℓ:

(H10) If $y, z \in \mathcal{H}$ satisfy $y \le z$ a.e. on Ω, then $g(y) \le g(z)$ and $\ell(y) \le \ell(z)$.

We are now prepared to state the following result.

THEOREM 4.1. *Besides the hypotheses of Theorem* 2.1 *but with* (H8)′ *replacing* (H8)*, assume that* A *is given by* (2.27) (*consequently* (H1) *and* (H2) *are automatically verified*)*,* β *is concave,* $\psi \equiv 0$*, and* $f \equiv 0$*. Finally, suppose* (H10) *holds. Then, the dual extremal arc* p *given by Theorem* 2.1 *satisfies*

$$-p(t) \in \partial_y V(t,y^*(t)) \text{ for } t \in [0,T] \text{ except possibly a countable subset.} \tag{4.1}$$

If in addition β satisfies (2.17)*, then the above inclusion holds for all $t \in [0,T]$.*

Remark 4.1. The above result may be viewed as a generalization of Proposition 5.5 from [7] to the case of infinite-dimensional and nonlinear systems.

The proof of Theorem 4.1 is, in essence, a succession of three lemmas.

LEMMA 4.1. *Under the assumptions of Theorem* 4.1*, $y \mapsto V(t,y)$ is convex on $\mathcal{H}$ for all $t \in [0,T]$.*

Proof. It is easy to verify that $V(t,y) = W(T-t,y)$, where

$$W(t,y) = \inf\left\{ \int_0^t (h(u(s)) + g(z(s)))\, ds + \ell(z(T)) : \right.$$
$$\left. z' + Az + \beta(z-\psi) \ni Bu + f \text{ a.e. } s \in (0,t), z(0) = y, u \in L^2(0,t;\mathcal{U}) \right\}.$$

So we have to prove that $y \mapsto W(t,y)$ is convex on $\mathcal{H}$ for every $t \in [0,T]$. To this end, we shall use a Trotter-type product formula from [10] for the dynamic programming equation associated with (P).

Let $\varepsilon = T/n$ (n being a positive integer). Define

$$W^\varepsilon(t,y) = \begin{cases} \inf\{\varepsilon h(u) + \varepsilon g((I+\varepsilon A)^{-1}(I+\varepsilon\partial\phi)^{-1}(y+\varepsilon Bu)) \\ \quad + W^\varepsilon(t-\varepsilon, (I+\varepsilon A)^{-1}(I+\varepsilon\partial\phi)^{-1}(y+\varepsilon Bu)) : u \in \mathcal{U}\} \\ \qquad \text{for } (t,y) \in (\varepsilon, T] \times \mathcal{H}, \\ \inf\{th(u) + \varepsilon g((I+\varepsilon A)^{-1}(I+\varepsilon\partial\phi)^{-1}(y+tBu)) \\ \quad + \ell((I+\varepsilon A)^{-1}(I+\varepsilon\partial\phi)^{-1}(y+tBu)) : u \in \mathcal{U}\} \\ \qquad \text{for } (t,y) \in (0, \varepsilon] \times \mathcal{H}, \end{cases}$$

$$W^\varepsilon(0,y) = \ell(y) \text{ for } y \in \mathcal{H}.$$

Obviously,

$$((I+\varepsilon\partial\phi)^{-1}y)(x) = (I+\varepsilon\beta)^{-1}(y(x)) \text{ a.e. } x \in \Omega.$$

By convexity of g, ℓ, and $r \mapsto (I+\varepsilon\beta)^{-1}(r)$ (do not forget that β is concave), using also (H10), we deduce that the functions $y \mapsto g((I+\varepsilon A)^{-1}(I+\varepsilon\partial\phi)^{-1}y)$ and $y \mapsto \ell((I+\varepsilon A)^{-1}(I+\varepsilon\partial\phi)^{-1}y)$ are convex on $\mathcal{H}$. Also, these functions are nondecreasing in the sense of (H10). (Indeed, it suffices to apply the monotonicity of $r \mapsto (I+\varepsilon\beta)^{-1}(r)$ and the maximum principle to the elliptic operator $I + \varepsilon A_0$.) Assume that t/ε is not an integer (the other case is completely similar). Let $y_1, y_2 \in \mathcal{H}$ and $0 \le \lambda \le 1$. We have

$$\lambda W^\varepsilon\left(t - \left[\frac{t}{\varepsilon}\right]\varepsilon, y_1\right) + (1-\lambda) W^\varepsilon\left(t - \left[\frac{t}{\varepsilon}\right]\varepsilon, y_2\right) \ge \left(t - \left[\frac{t}{\varepsilon}\right]\varepsilon\right) h(\lambda u_1 + (1-\lambda)u_2)$$
$$+\varepsilon g\left((I+\varepsilon A)^{-1}(I+\varepsilon\partial\phi)^{-1}\left(\lambda y_1 + (1-\lambda)y_2 + \left(t - \left[\frac{t}{\varepsilon}\right]\varepsilon\right) B(\lambda u_1 + (1-\lambda)u_2)\right)\right)$$
$$+\ell\left((I+\varepsilon A)^{-1}(I+\varepsilon\partial\phi)^{-1}\left(\lambda y_1 + (1-\lambda)y_2 + \left(t - \left[\frac{t}{\varepsilon}\right]\varepsilon\right) B(\lambda u_1 + (1-\lambda)u_2)\right)\right)$$
$$\ge W^\varepsilon\left(t - \left[\frac{t}{\varepsilon}\right]\varepsilon, \lambda y_1 + (1-\lambda)y_2\right).$$

Here $u_1, u_2 \in \mathcal{U}$ are minimizing elements in the expressions defining $W^\varepsilon(t-[t/\varepsilon]\varepsilon, y_1)$ and $W^\varepsilon(t-[t/\varepsilon]\varepsilon, y_2)$, respectively. Also, one readily shows that $y \mapsto W^\varepsilon(t-[t/\varepsilon]\varepsilon, y)$ is nondecreasing. Now we successively argue $[t/\varepsilon]$ times as above to obtain that $y \mapsto W^\varepsilon(t, y)$ is convex on $\mathcal{H}$ (and nondecreasing).

But the following Trotter-type product formula holds (see [10, Thm. 3.2]):

$$\lim_{\varepsilon \to 0} W^\varepsilon(t, y) = W(t, y) \text{ for every } (t, y) \in [0, T] \times \mathcal{H}.$$

So letting ε tend to zero in the inequality proving the convexity of $W^\varepsilon(t, y)$, we obtain the conclusion of Lemma 4.1.

Remark 4.2. As a by-product of the above proof, we have also obtained that $y \mapsto V(t, y)$ is nondecreasing in the sense of (H10).

Remark 4.3. Compare the proof of Lemma 4.1 with that of Proposition 5.5 from [7].

Remark 4.4. Lemma 4.1 remains valid (with the same proof) but with $\overline{D(\phi)}$ instead of $\mathcal{H}$ if we replace β from Theorem 2.1 with β from Theorem 2.2 (given by (2.18)). Also, in the above proof, we can change the places of the resolvents $(I+\varepsilon A)^{-1}, (I+\varepsilon\partial\phi)^{-1}$ between them, but in this case we must apply [10, Thm. 3.1].

Remark 4.5. The hypothesis (H10) is needed only in the proof of Lemma 4.1.

Remark 4.6. We point out that the Trotter-type product formulas for the dynamic programming equation are proved to be adequate tools to establish some qualitative properties for the optimal value function such as convexity or monotonicity.

LEMMA 4.2. *Under the hypotheses of Theorem* 2.1*, the optimal value function* $(t, y) \mapsto V(t, y)$ *is continuous on* $[0, T] \times \mathcal{H}$.

Proof. Similar arguments to those from the proof of Proposition 2.1 lead to the continuity of $t \mapsto V(t, y)$ for all $y \in \mathcal{H}$. But this in conjunction with Proposition 2.1 gives the statement of the lemma.

The following result is an infinite-dimensional version of Corollary 5.2 from [7].

LEMMA 4.3. *Under the assumptions of Theorem* 4.1*, the multifunction* $(t, y) \mapsto \partial_y V(t, y)$ *is strongly-weakly closed.*

Proof. The proof is the same with that of Corollary 5.2 from [7]; however, we give it for the reader's convenience. First let us remark that, by Lemmas 4.1 and 4.2, the gradient $\partial_y V(t, y)$ coincides with the convex subdifferential of $y \mapsto V(t, y)$. Now let $t_m \to t$, $y_m \to y$ strongly in $\mathcal{H}$, and $p_m \to p$ weakly in $\mathcal{H}$, where $p_m \in \partial_y V(t_m, y_m)$, that is,

$$V(t_m, z) - V(t_m, y_m) \geq (z - y_m, p_m) \text{ for all } z \in \mathcal{H}.$$

Letting $m \to \infty$ in the above inequality, by Lemma 4.2, we obtain

$$V(t, z) - V(t, y) \geq (z - y, p) \text{ for all } z \in \mathcal{H},$$

that is, $p \in \partial_y V(t, y)$, and the proof of Lemma 4.3 is complete.

Proof of Theorem 4.1. The conclusion of the theorem will follow by combining Lemma 4.3 with the continuity properties of the function p. Indeed, since $p \in BV([0, T]; \mathcal{Y}')$, $p : [0, T] \to \mathcal{Y}'$ is continuous on $[0, T]$ except a countable subset. Let $t \in [0, T]$ be an arbitrary point of continuity for $p : [0, T] \to \mathcal{Y}'$. By virtue of (2.15) and the fact that p belongs also to $L^\infty(0, T; \mathcal{H})$, there exists a sequence $\{t_m\} \subset [0, T]$ converging to t such that, for every m,

$$-p(t_m) \in \partial_y V(t_m, y^*(t_m)),$$

$$|p(t_m)| \leq ||p||_{L^\infty(0,T;\mathcal{H})}.$$

Hence, extracting a subsequence of $\{m\}$, $p(t_m) \to p \in \mathcal{H}$ weakly in $\mathcal{H}$. Since $p(t_m) \to p(t)$ strongly in $\mathcal{Y}'$, on a subsequence of $\{m\}$, we have

$$p(t_m) \to p(t) \text{ weakly in } \mathcal{H}.$$

But we also have

$$y^*(t_m) \to y^*(t) \text{ strongly in } \mathcal{H},$$

so that, by applying Lemma 4.3, we get

$$-p(t) \in \partial_y V(t, y^*(t)).$$

If also the condition (2.17) holds, then, by Theorem 2.1, $p \in C_w([0,T];\mathcal{H})$, whence, arguing as above, one obtains (4.1) for all $t \in [0,T]$. This completes the proof.

Remark 4.7. When β is given by (2.18), one cannot expect to obtain a similar result to Theorem 4.1 by using the same proof as above and the same definition for $\partial_y V(t,y)$. Indeed, although $y \mapsto V(t,y)$ is convex on $\overline{D(\phi)}$ also in this case (see Remark 4.4), $y \mapsto \tilde{V}(t,y)$ is not in general convex on $\mathcal{H}$, so it is possible that, at some points $y \in \overline{D(\phi)}$, $\partial_y V(t,y)$ does not coincide with the subdifferential of any convex extension of $y \mapsto V(t,y)$; consequently, the argument in the proof of Lemma 4.3 does not work.

Acknowledgment. The author is greatly indebted to Professor V. Barbu for suggesting the problem and for the useful discussions.

REFERENCES

[1] V. Barbu, *Necessary conditions for distributed control problems governed by parabolic variational inequalities*, SIAM J. Control Optim., 19 (1981), pp. 64–86.
[2] V. Barbu, *Optimal Control of Variational Inequalities*, Research Notes in Mathematics 100, Pitman, London, 1984.
[3] V. Barbu and Th. Precupanu, *Convexity and Optimization in Banach Spaces*, 2nd ed., Editura Academiei, Bucureşti, and D. Reidel, Dordrecht, Boston, Lancaster, 1986.
[4] H. Brézis, *Opérateurs maximaux monotones et semigroupes de contractions dans les espaces de Hilbert*, Math. Studies 5, North-Holland, Amsterdam, 1973.
[5] F. H. Clarke, *Optimization and Nonsmooth Analysis*, John Wiley, New York, 1983.
[6] F. H. Clarke and R. B. Vinter, *The Maximum Principle and the Dynamic Programming Technique: How Are They Related*? Technical Report CRM–1300, Univ. of Montréal, Montréal, Quebec, Canada, 1985.
[7] F. H. Clarke and R. B. Vinter, *The relationship between the maximum principle and dynamic programming*, SIAM J. Control Optim., 25 (1987), pp. 1291–1311.
[8] Y. Kobayashi, *Product formula for nonlinear semigroups in Hilbert spaces*, Proc. Japan Acad. Ser. A Math Sci., 58 (1982), pp. 425–428.
[9] C. Popa, *Trotter product formulae for Hamilton-Jacobi equations in infinite dimensions*, Differential Integral Equations, 4 (1991), pp. 1251–1268.
[10] C. Popa, *Feedback laws for nonlinear distributed control problems via Trotter-type product formulae*, SIAM J. Control Optim., 33 (1995), pp. 971–999.
[11] R. T. Rockafellar, *Generalized directional derivatives and subgradients of nonconvex functions*, Canad. J. Math., 32 (1980), pp. 257–280.

SIAM J. CONTROL OPTIM.
Vol. 35, No. 5, pp. 1739–1754, September 1997

016

CONSTRAINED LQR PROBLEMS IN ELLIPTIC DISTRIBUTED CONTROL SYSTEMS WITH POINT OBSERVATIONS—ON CONVERGENCE RATES*

PUHONG YOU† AND JIANXIN ZHOU†

Abstract. In this paper, we continue to study the (bound constrained) linear-quadratic regulator problem in distributed boundary control systems governed by the elliptic equation with point observations. Due to the appearance of singularities in the problem, the traditional Galerkin variational method that leads to an adjoint system is not desirable, and the classical Lagrangian multiplier method is not reliable for providing numerical results. A characterization formula of the optimal control and its singularity decomposition formula are derived in [Z. Ding, L. Ji, and J. Zhou, *SIAM J. Control Optim.*, 34 (1996), pp. 264–294; Z. Ding and J. Zhou, *Appl. Math. Optim.*, to appear] by using the boundary integral equation and potential theory coupled with a variational inequality in a Banach space setting. Based on the characterization formula, a conditioned gradient projection method (CGPM) has been proposed in [Z. Ding and J. Zhou, *Appl. Math. Optim.*, to appear]. Numerical experiment has shown that CGPM is efficient and also insensitive to the partition number of the boundary. In this paper, we estimate the rate of convergence for CGPM. First it is proved that for $\mathcal{N}=2$, CGPM converges exponentially in the L^2 norm, and for $\mathcal{N}=3$, CGPM converges subexponentially in the L^p norm. Then, under a reasonable condition, it is proved that for $\mathcal{N}=2$, CGPM converges uniformly exponentially, and for $\mathcal{N}=3$, CGPM converges uniformly subexponentially.

Key words. LQR, distributed boundary control, point observation, singularity decomposition, (sub)exponential convergence, uniformly (sub)exponential convergence

AMS subject classifications. 49N10, 49J20, 49J22, 49M07, 49M20, 65N38

PII. S0363012994278183

1. Introduction.

Let Ω be an (interior or exterior) open domain in $\Re^{\mathcal{N}}$ ($\mathcal{N}=2,3$) with a bounded, C^∞ boundary Γ. In this paper, we study numerical methods to solve the following constrained linear-quadratic regulator (LQR) problem in a distributed boundary control system governed by the elliptic equation with point observations:

$$\text{(LQR)}\begin{cases}\min\limits_u J(u)=\sum\limits_{k=1}^{M}\mu_k|w(P_k)-Z_k|^2+\gamma\displaystyle\int_\Gamma u^2(x)d\sigma_x\\ \text{subject to (1.1)}\begin{cases}\Delta w(x) = 0 & \text{in } \Omega\\ \dfrac{\partial w(x)}{\partial n} = u(x) & \text{on } \Gamma\end{cases}\\ u\in\mathcal{U},\end{cases}\tag{1.2}$$

where $\mathcal{U}$ is the admissible control set defined by

$$\mathcal{U}=\{v\in L_0^p(\Gamma)\mid Bl(x)\le v(x)\le Bu(x) \text{ on } \Gamma\},\tag{1.3}$$

$$L_0^p(\Gamma)=\left\{f\in L^p(\Gamma)\Big|\int_\Gamma f(\xi)d\sigma_\xi=0\right\},\quad 1\le p<\infty,$$

*Received by the editors December 5, 1994; accepted for publication (in revised form) July 15, 1996. This research was supported in part by NSF grant DMS-9404380 and an IRI Award from Texas A&M University.

http://www.siam.org/journals/sicon/35-5/27818.html

†Department of Mathematics, Texas A&M University, College Station, TX 77843 (pyou@math.tamu.edu, jzhou@math.tamu.edu).

and the lower and upper bounds $Bl \le Bu \in L^p(\Gamma)$ are given s.t.

$$\int_\Gamma Bl(\xi)d\sigma_\xi < 0, \quad \int_\Gamma Bu(\xi)d\sigma_\xi > 0.$$

Throughout this paper, we assume that $p = 2$ for $\mathcal{N} = 2$ and $p > 2$ for $\mathcal{N} = 3$, and $q \le 2$ is given s.t. $\frac{1}{p} + \frac{1}{q} = 1$. In the above setting,

$\frac{\partial}{\partial n}$ is the outward normal derivative,
$u \in \mathcal{U}$ is a Neumann-type boundary control on Γ,
$\gamma, \mu_k > 0, 1 \le k \le M$, are given weighting factors,
$P_k \in \partial\Omega, 1 \le k \le M$, are prescribed "sensor locations,"
$Z_k \in \Re, 1 \le k \le M$, are prescribed "target values" at P_k.

The objective of the above problem is to find the distribution of $u(x)$ on Γ s.t. at sensor locations $P_k, 1 \le k \le M$, the observation values $w(P_k)$ are as close as possible to the target values Z_k with least possible control cost $\int_\Gamma u^2(x)d\sigma_x$. The well-posedness of the problem has been justified in [2, 3, 4].

The study of the above system is motivated by problems in cathodic protection systems in corrosion engineering [20, 21] and contemporary "smart sensors." Since in most applications, point sensors are much cheaper and easier to design than distributed sensors [19, 15, 18], point sensors (observations) are used in our problem setting. Once point sensors are placed on the boundary, singularities will appear in the problem. Mathematically and computationally, it becomes very tough to handle. The usual Galerkin variational method [13] leading to an adjoint system cannot effectively handle this problem.

Recently, Ji and Chen [11] studied a similar problem by using the potential theory and boundary element method (BEM). Their approach has certain important advantages over others. It can provide rather explicit information about the control and state, and it is amenable to direct numerical computation through BEM. In [2, 3], Ding, Ji, and Zhou and Ding and Zhou applied this approach to study the above-constrained LQR problems. They derived a feedback characterization of the optimal control and established a singularity decomposition formula for the characterization of the optimal control to overcome the singularity problem. They also pointed out that the classical Lagrangian multiplier method (LMM) is not reliable for providing a numerical solution for unconstrained LQR; the finite-dimensional gradient projection method is sensitive to the partition number of the boundary. In [4], Ding and Zhou proposed a conditioned gradient projection method (CGPM) to solve the above-constrained LQR problem. Strong convergence and uniform convergence have been verified numerically and mathematically in [4].

In this paper, we estimate the rate of convergence for CGPM. Due to the different space settings, for $\mathcal{N} = 2$, our problem is in a Hilbert space L^2 setting, and for $\mathcal{N} = 3$, it is in a Banach space L^p setting. Thus we have to treat these two cases differently. We prove that for $\mathcal{N} = 2$, CGPM converges exponentially in the L^2 norm and that for $\mathcal{N} = 3$, CGPM converges subexponentially in the L^p norm; i.e., it converges faster than any power of $\frac{1}{n}$, where n is the number of iterations. Finally, under a quite reasonable assumption, we prove that for $\mathcal{N} = 2$, CGPM converges uniformly exponentially, and for $\mathcal{N} = 3$, CGPM converges uniformly subexponentially. The motivation to investigate uniform convergence and uniform convergence rate is due to the concern about the discretization of the above-constrained LQR problem, as partially explained in [2, 4]. First, for a discretization of a constrained optimal control problem, if a numerical algorithm is sensitive to the partition of the boundary

(dimension) (such as the finite-dimensional gradient projection method used in [4]), a refinement of the partition on the boundary will add more constraints to the problem. It then slows down the computation and convergence and ultimately fails to provide reliable numerical solutions if the partition number is very large. While our governing differential equation is a PDE, the partition number on the boundary can be quite large. On the other hand, if the optimal control u^* is in $L^p(\Gamma)$, the strong convergence cannot guarantee the convergence of $u^*(x)$ on a set of zero measure. While in real computation, we can only approximate the optimal control u^* at a finite number of boundary points. Therefore, it is important to study uniform convergence and uniform convergence rate.

Let

$$E(x,\xi) = \begin{cases} -\dfrac{1}{2\pi}\ln|x-\xi|, & x,\xi \in \Re^2, \\ \dfrac{1}{4\pi}\dfrac{1}{|x-\xi|}, & x,\xi \in \Re^3 \end{cases} \tag{1.4}$$

be the fundamental solution of the Laplacian operator. For $f \in L^2(\Gamma)$, define operators

$$\mathcal{S}(f)(x) = \int_\Gamma E(x,\xi) f(\xi) d\sigma_\xi, \quad x \in \Re^{\mathcal{N}}, \tag{1.5}$$

$$\mathcal{K}(f)(x) = \lim_{\varepsilon \to 0+} \int_{\Gamma \bigcap \{|x-\xi|>\varepsilon\}} \frac{\partial}{\partial n_\xi} E(x,\xi) f(\xi) d\sigma_\xi, \quad x \in \Gamma, \tag{1.6}$$

$$\mathcal{K}^*(f)(x) = \lim_{\varepsilon \to 0+} \int_{\Gamma \bigcap \{|x-\xi|>\varepsilon\}} \frac{\partial}{\partial n_x} E(x,\xi) f(\xi) d\sigma_\xi, \quad x \in \Gamma. \tag{1.7}$$

Their basic properties can be found in [2, 3]. According to [1, Chap. 6], a solution w of (1.2) can be written as

$$w(x) = \mathcal{S}(\eta)(x), \quad x \in \overline{\Omega}, \tag{1.8}$$

where η is the layer density to be determined from the boundary condition

$$\frac{\partial w(x)}{\partial n} = \left(\frac{1}{2} I + \mathcal{K}^*\right)(\eta)(x) = u(x), \quad x \in \Gamma. \tag{1.9}$$

Once the layer density η is found, the solution $w(x)$ of (1.1) can be computed from (1.8).

By Theorem 2.4 in [3], for each given $u \in \mathcal{U}$, (1.2) has a unique solution $w \in C(\overline{\Omega})$ s.t.

$$\sum_{k=1}^{M} \mu_k (w(P_k) - Z_k) = 0. \tag{1.10}$$

Once a control u is given, the corresponding optimal state can be obtained by solving (1.2) and (1.10). From now on, we use $w(x,u)$ to stand for this solution.

Remark 1. The difference between the cases $\mathcal{N} = 2$ and $\mathcal{N} = 3$ is that for each fixed x, $E(x,\cdot) \in L^2(\Gamma)$ for $\mathcal{N} = 2$ and $E(x,\cdot) \in L^{2-\varepsilon}(\Gamma)$ $(0 < \varepsilon < 2)$ for $\mathcal{N} = 3$.

LEMMA 1.1. *For any $u \in L_0^p(\Gamma)$, (1.2) and (1.10) have a unique solution $w(x,u)$, and for any $\alpha, \beta \in \Re$, $v_1, v_2 \in L_0^p(\Gamma)$, we have*

$$w(x, \alpha v_1 + \beta v_2) = \alpha w(x, v_1) + \beta w(x, v_2) + (1 - \alpha - \beta)\mu, \tag{1.11}$$

where $\mu = \sum_{k=1}^{M} \mu_k Z_k / \sum_{k=1}^{M} \mu_k$. *In addition, for any* $u \in L_0^p(\Gamma)$, *we have*

$$|w(x,u) - \mu| \le C(x,\Omega)\|u\|_2 \quad \text{for } \mathcal{N} = 2, \tag{1.12}$$

$$|w(x,u) - \mu| \le C(x,\Omega,p)\|u\|_p \quad \text{for } \mathcal{N} = 3, \tag{1.13}$$

where the constant $C(x,\Omega)$ *depends only on* x *and* Ω, *and the constant* $C(x,\Omega,p)$ *depends only on* x, Ω, *and* p.

Proof. (1.11) can be verified directly. (1.12) and (1.13) follow from Theorem 2.4 in [3] and (1.10). ∎

Notice that $\mathcal{U}$ is a bounded, convex, and closed subset in $L^p(\Gamma)$. Let

$$D = \max\{2\|Bl\|_p, 2\|Bu\|_p, 1\};$$

then

$$\|u\|_p \le D, \qquad \|u - v\|_p \le D \qquad \forall\, u, v \in \mathcal{U}.$$

Define the truncation $[u]_{Bl}^{Bu}$ of any function $u(x)$ on Γ by

$$[u]_{Bl}^{Bu}(x) = [u(x)]_{Bl(x)}^{Bu(x)} = \begin{cases} Bl(x) & \text{if} \quad u(x) < Bl(x), \\ u(x) & \text{if} \quad Bl(x) \le u(x) \le Bu(x), \\ Bu(x) & \text{if} \quad u(x) > Bu(x). \end{cases} \tag{1.14}$$

Introduce the projection operator $P_{\mathcal{U}} : L^p(\Gamma) \to \mathcal{U}$:

$$P_{\mathcal{U}}(u) = [u + c_u]_{Bl}^{Bu} \qquad \forall\, u \in L^p(\Gamma), \tag{1.15}$$

where c_u is a constant defined in Lemma 6 in [4] s.t. $P_{\mathcal{U}}(u) \in L_0^p(\Gamma)$. Applying the characterization of truncation from Lemma 8 in [4], we obtain the following lemma.

LEMMA 1.2 (see (2.8) and Theorem 1 in [4]). u^* *is an optimal control to the constrained LQR problem if and only if*

$$\begin{aligned} u^* &= \left[u^* - \frac{1}{2\gamma} \nabla J(u^*) + c_* \right]_{Bl}^{Bu} \\ &= \left[-\frac{1}{\gamma} \left(\frac{1}{2} I + \mathcal{K} \right)^{-1} \sum_{k=1}^{M} \mu_k (w^*(P_k, u^*) - Z_k) E(P_k, \cdot)(\xi) + C^* \right]_{Bl}^{Bu}, \end{aligned} \tag{1.16}$$

where c_* *is defined by Lemma* 6 *in* [4] *s.t.*

$$\int_\Gamma \left[u^*(x) - \frac{1}{2\gamma} \nabla J(u^*)(x) + c_* \right]_{Bl}^{Bu} d\sigma_x = 0. \tag{1.17}$$

Let $\langle \cdot, \cdot \rangle$ be the pairing on $L^q(\Gamma)$ and $L^p(\Gamma)$. The following facts on the objective function J are needed.

LEMMA 1.3. *The functional* $J(u)$ *is convex (quadratic) and differentiable, so for any* $t \in \Re$ *and* $u, v \in L_0^p(\Gamma)$, *we have*

$$J(u + tv) = J(u) + t\langle \nabla J(u), v \rangle + t^2 \left\{ \sum_{k=1}^{M} \mu_k |w(P_k, v) - \mu|^2 + \gamma \int_\Gamma v^2(x) d\sigma_x \right\}, \tag{1.18}$$

$$J(u) - J(v) \le \langle \nabla J(u), u - v \rangle, \tag{1.19}$$

$$J\left(\frac{u+v}{2} \right) \le \max\{J(u), J(v)\} - \frac{1}{4}\gamma \|u - v\|_2^2. \tag{1.20}$$

Proof. (1.18) and (1.19) can be verified directly. By virtue of (1.11) and (1.18), we get

$$J(u)+J(v)-2J\left(\frac{u+v}{2}\right)=\frac{1}{2}\sum_{k=1}^{M}|w(P_k,u)-w(P_k,v)|^2+\frac{1}{2}\gamma\|u-v\|_2^2.$$

(1.20) then follows from the fact that $(J(u)+J(v))\le 2\max\{J(u),J(v)\}$. ∎

2. Numerical algorithm. Based on the characterization of the optimal control, Lemma 1.2, a CGPM has been proposed in [4] to solve the constrained LQR problem. Its strong convergence and uniform convergence are proved in [4]. To estimate its convergence rate, we modify CGPM as follows.

Given $\frac{1}{2}<\alpha<1$. Choose any initial trial $u^1\in\mathcal{U}$.

$$u^{n+1}=u^n+\lambda_n^* d_n,\quad n=1,2,3,\dots, \tag{2.1}$$

where

$$d_n=\bar{u}^n-u^n, \tag{2.2}$$

$$\bar{u}^n=P_{\mathcal{U}}\left(u^n-\frac{1}{2\gamma}\nabla J(u^n)\right)=\left[u^n(x)-\frac{1}{2\gamma}\nabla J(u^n)+C_n\right]_{Bl(x)}^{Bu(x)}, \tag{2.3}$$

$$\lambda_n^*=\min\{\alpha,\lambda_n\}, \tag{2.4}$$

$$\lambda_n=\arg\min_{\lambda}J(u^n+\lambda d_n), \tag{2.5}$$

and C_n has the smallest magnitude s.t.

$$\int_\Gamma\left[u^n(x)-\frac{1}{2\gamma}\nabla J(u^n)+C_n\right]_{Bl}^{Bu}d\sigma_x=0. \tag{2.6}$$

Remark 2.

(i) In order to estimate the convergence rate, the CGPM in [4] has been modified s.t. the constant C_n in (2.6) has the smallest magnitude. Due to Lemma 6 in [4], the function

$$\phi_n(\lambda)=\int_\Gamma\left[u_n(x)-\frac{1}{2\gamma}\nabla J(u_n)+\lambda\right]_{Bl}^{Bu}d\sigma_x$$

is monotonically increasing in λ. So the constant C_n with the smallest magnitude s.t. $\phi_n(C_n)\in L_0^p(\Gamma)$ can be easily found. Once the convergence rates are obtained, we may use Lemma 7 in [4]; i.e., if $\phi_n(\lambda_1),\phi_n(\lambda_2)\in L_0^p(\Gamma)$, then $\phi_n(\lambda_1)=\phi_n(\lambda_2)$, a.e. in Γ, to relax this condition.

(ii) Due to the appearance of singularities in the control variable, an adaptive local refinement scheme has been used in numerical computation to enhance the convergence stability; see [2, 4].

3. Results on the convergence rate. In this section, we estimate the convergence rate of the above CGPM. Due to different natures of the space settings, we have to treat $\mathcal{N}=2$ and $\mathcal{N}=3$ separately.

LEMMA 3.1. *For the above CGPM, we obtain the following:*

(i) $J(u^n)$ *is decreasing, so* $\lim_{n\to\infty}J(u^n)\ge J(u^*)$;

(ii) $J(u^n)-J(u^*)\ge\frac{1}{4}\gamma\|u^n-u^*\|_2^2$;

(iii) $\langle\nabla J(u^n),\bar{u}^n-z\rangle\le 2\gamma\langle u^n-\bar{u}^n,\bar{u}^n-z\rangle\quad\forall\, z\in\mathcal{U}$.

In particular, if $z = u^n$, we have

$$\langle \nabla J(u^n), d_n \rangle \le -2\gamma \|d_n\|_2^2.$$

Proof. (i) follows from a direct computation that leads to

$$J(u^{n+1}) - J(u^n) \begin{cases} = \dfrac{-\frac{1}{4}|\langle \nabla J(u^n), d^n \rangle|^2}{\left\{ \sum_{k=1}^{M} \mu_k |w(P_k, d_n) - \mu|^2 + \gamma \|d_n\|_2^2 \right\}} < 0 & \text{if } \lambda_n^* = \lambda_n, \\ \le \dfrac{\alpha}{2} \langle \nabla J(u^n), d^n \rangle < 0 & \text{if } \lambda_n^* = \alpha, \end{cases} \tag{3.1}$$

where we have used the fact that $\lambda_n^* = \alpha < \lambda_n$ means

$$2\alpha \left\{ \sum_{k=1}^{M} \mu_k |w(P_k, d_n) - \mu|^2 + \gamma \|d_n\|_2^2 \right\} < -\langle \nabla J(u^n), d^n \rangle.$$

Point (ii) can be derived from (1.20), and (iii) follows from Lemma 8 in [4], by taking $u = \bar{u}^n$. ∎

The following lemma is a special case of Lemma 1.4 in [5], which will be used in estimating the convergence rate.

LEMMA 3.2. *Suppose that $\{r_n\}_{n=1}^{\infty} \subset [0, \infty)$ and $q > 0$ satisfy*

$$r_n - q r_n^k \ge r_{n+1}$$

for $n \ge 0$, with k a fixed exponent in the range $(1, \infty)$; then

$$r_n \le r_0 [1 + (k-1) r_0^{k-1} q n]^{-\frac{1}{k-1}}$$

for all n, i.e.,

$$\lim_{n \to \infty} \sup r_n n^{\frac{1}{k-1}} \le [(k-1)q]^{-\frac{1}{k-1}}.$$

Case 1. $\mathcal{N} = 2$.

THEOREM 3.3. *For $\mathcal{N} = 2$, let $\{u_n\}_{n=1}^{\infty}$ and $\{\bar{u}^n\}_{n=1}^{\infty}$ be the sequences generated by the algorithm CGPM, (2.1)–(2.5); then $r_n = J(u^n) - J(u^*)$, u^n, and $\bar{u}^n$ all converge exponentially; i.e.,*

$$r_n \le r_1 \delta^n, \tag{3.2}$$

where $\delta = 1 - \frac{15 - 6\sqrt{6}}{8} \frac{\gamma}{L} \in (0, 1)$, with $L = \sum_{k=1}^{M} \mu_k C^2(P_k, \Omega) + \gamma$. Hence

$$\|u^n - u^*\|_2 \le \sqrt{\frac{4r_1}{\gamma}} \delta^{\frac{n}{2}} \tag{3.3}$$

and

$$\|\bar{u}^n - u^*\|_2 \le \frac{2\sqrt{L} + 2\sqrt{3\gamma}}{\sqrt{3}\gamma} r_1^{\frac{1}{2}} \delta^{\frac{n}{2}}. \tag{3.4}$$

LEMMA 3.4. *For λ_n^* defined in the algorithm* (2.1)–(2.5)*, we have $\lambda_n^* \geq \lambda_0^* \,\forall\, n$, where $\lambda_0^* = \min\{\alpha, \frac{\gamma}{L}\} \in (0,1)$.*

Proof. From (1.12), (1.18), and Lemma 3.1 (iii), we get

$$\begin{aligned}\lambda_n &= \frac{-\langle \nabla J(u_n), d_n\rangle}{2\left\{\sum_{k=1}^M \mu_k |w(P_k,d_n)-\mu|^2 + \gamma \int_\Gamma d_n^2(x)d\sigma_x\right\}} \\ &\geq \frac{\gamma}{\sum_{k=1}^M \mu_k C^2(P_k,\Omega)+\gamma}.\end{aligned}$$

This completes our proof. ∎

Proof of Theorem 3.3. Let $n > 0$; by (1.19) and Lemma 3.1 (iii), we have

$$\begin{aligned} r_n :&= J(u^n) - J(u^*) \\ &\leq \langle \nabla J(u^n), u^n - u^*\rangle \\ &\leq \langle \nabla J(u^n), u^n - \bar{u}^n\rangle + \langle \nabla J(u^n), \bar{u}^n - u^*\rangle \\ &\leq \langle \nabla J(u^n), u^n - \bar{u}^n\rangle + 2\gamma\langle u^n - \bar{u}^n, \bar{u}^n - u^*\rangle \qquad (3.5)\\ &= \langle \nabla J(u^n), u^n - \bar{u}^n\rangle + 2\gamma\langle u^n - \bar{u}^n, u^n - u^*\rangle - 2\gamma\|u^n - \bar{u}^n\|_2^2 \\ &\leq -\langle \nabla J(u^n), d_n\rangle + 2\gamma\|u^n - \bar{u}^n\|_2\|u^n - u^*\|_2. \end{aligned}$$

On the other hand, for any $\beta \in [0, \lambda_0^*]$, from (1.18), Lemma 3.4, and Lemma 3.1 (iii), we obtain

$$\begin{aligned} r_{n+1} - r_n &= J(u^{n+1}) - J(u^n) \\ &\leq \beta\langle \nabla J(u^n), d_n\rangle + L\beta^2\|u^n - \bar{u}^n\|_2^2 \qquad (3.6)\\ &\leq (-2\gamma\beta + L\beta^2)\|u^n - \bar{u}^n\|_2^2. \end{aligned}$$

Let $\beta = \frac{\gamma}{2L} < \lambda_0^* = \min\{\alpha, \frac{\gamma}{L}\}$; then

$$\|u^n - \bar{u}^n\|_2 \leq \frac{2\sqrt{L}}{\sqrt{3}\gamma}(r_n - r_{n+1})^{\frac{1}{2}}. \qquad (3.7)$$

From (3.6), we get

$$-\langle \nabla J(u^n), d_n\rangle \leq \frac{2L}{\gamma}(r_n - r_{n+1}) + \frac{1}{\gamma}\|u^n - \bar{u}^n\|_2^2. \qquad (3.8)$$

Also from Lemma 3.1 (ii), we have

$$\|u^n - u^*\|_2 \leq \frac{2}{\sqrt{\gamma}} r_n^{\frac{1}{2}}. \qquad (3.9)$$

Substituting (3.7), (3.8), and (3.9) into (3.5), we have

$$\begin{aligned} r_n &\leq \frac{2L}{\gamma}(r_n - r_{n+1}) + \frac{\gamma}{2}\|u^n - \bar{u}^n\|_2^2 + 2\gamma\|u^n - \bar{u}^n\|_2\|u^n - u^*\|_2 \\ &\leq \frac{8L}{3\gamma}(r_n - r_{n+1}) + 8\sqrt{\frac{L}{3r}}(r_n - r_{n+1})^{\frac{1}{2}} r_n^{\frac{1}{2}} \qquad (3.10)\\ &= \frac{8L}{3\gamma}\left((r_n - r_{n+1})^{\frac{1}{2}} + \frac{1}{2}\sqrt{\frac{3\gamma}{L}} r_n^{\frac{1}{2}}\right)^2 - 2r_n. \end{aligned}$$

Thus

$$\frac{9\gamma}{8L}r_n \le \left((r_n - r_{n+1})^{\frac{1}{2}} + \frac{1}{2}\sqrt{\frac{3\gamma}{L}}r_n^{\frac{1}{2}}\right)^2.$$

Taking the square root on both sides and rearranging the terms, we obtain

$$\frac{3-\sqrt{6}}{2\sqrt{2}}\sqrt{\frac{\gamma}{L}}r_n^{\frac{1}{2}} \le (r_n - r_{n+1})^{\frac{1}{2}}.$$

From here we get

$$r_{n+1} \le \left(1 - \frac{15-6\sqrt{6}}{8}\frac{\gamma}{L}\right) r_n. \tag{3.11}$$

So, letting $\delta = (1 - \frac{15-6\sqrt{6}}{8}\frac{\gamma}{L})$, we get (3.2). (3.3) follows directly from Lemma 3.1 (ii). For (3.4), notice (3.7) and (3.9). We then have

$$\begin{aligned}\|\bar{u}^n - u^*\|_2 &\le \|\bar{u}^n - u^n\|_2 + \|u^n - u^*\|_2 \\ &\le \frac{2\sqrt{L}}{\sqrt{3}\gamma}(r_n - r_{n+1})^{\frac{1}{2}} + \frac{2}{\sqrt{\gamma}}r_n^{\frac{1}{2}} \\ &\le \left(\frac{2\sqrt{L}}{\sqrt{3}\gamma} + \frac{2}{\sqrt{\gamma}}\right) r_n^{\frac{1}{2}}.\end{aligned}$$

Thus (3.4) follows. ∎

Case 2. $\mathcal{N} = 3$.

For any $2 < p' \le p$, notice that Γ is a compact set. We thus have $L^p(\Gamma) \subset L^{p'}(\Gamma) \subset L^2(\Gamma)$, and the interpolation inequality gives

$$\|u\|_{p'} \le \|u\|_2^{\frac{2(p-p')}{p'(p-2)}} \|u\|_p^{\frac{p(p'-2)}{p'(p-2)}}.$$

Thus

$$\|u\|_{p'} \le D\|u\|_2^{1-\nu} \quad \forall\, u \in \mathcal{U}, \tag{3.12}$$

where $\nu = \frac{p(p'-2)}{p'(p-2)}$.

THEOREM 3.5. *For $\mathcal{N} = 3$, let $\{u^n\}_{n=1}^{\infty}$ and $\{\bar{u}^n\}_{n=1}^{\infty}$ be the sequences generated by the algorithm CGPM,* (2.1)–(2.5)*; then $r_n = J(u^n) - J(u^*)$, u^n, and $\bar{u}^n$ all converge subexponentially, i.e.,*

$$r_n \le E_1 n^{-s}, \tag{3.13}$$

where

$$s = \frac{1-\nu}{2\nu} = \frac{p-p'}{p(p'-2)} \quad \textit{for any} \quad 2 < p' \le p. \tag{3.14}$$

Moreover,

$$\|u^n - u^*\|_2 \le E_2 n^{-\frac{s}{2}} \tag{3.15}$$

and

$$\|\bar{u}^n - u^*\|_2 \le E_3 n^{-\frac{s}{2(1+\nu)}}, \tag{3.16}$$

where E_1, E_2, and E_3 are constants depending only on Ω, γ, p', and P_k, Z_k, μ_k, $k = 1, \dots, M$.

Proof. Letting $n > 0$, by the last inequality in (3.5), we get

$$\begin{aligned} r_n :&= J(u^n) - J(u^*) \\ &\le -\langle \nabla J(u^n), d_n\rangle + 2\gamma \|u^n - \bar{u}^n\|_2 \|u^n - u^*\|_2. \end{aligned} \tag{3.17}$$

On the other hand, we have

$$J(u^{n+1}) - J(u^n) = \lambda_n^* \langle \nabla J(u^n), d_n\rangle + {\lambda_n^*}^2 \left\{ \sum_{k=1}^M \mu_k |w(P_k, d_n) - \mu|^2 + \gamma \|d_n\|_2^2 \right\}, \tag{3.18}$$

where

$$\lambda_n^* = \min\left\{ \alpha, \frac{-\langle \nabla J(u^n), d_n\rangle}{2\{\sum_{k=1}^M \mu_k |w(P_k, d_n) - \mu|^2 + \gamma \|d_n\|_2^2\}} \right\}. \tag{3.19}$$

If $\lambda_n^* = \alpha$, i.e.,

$$-\langle \nabla J(u^n), d_n\rangle \ge 2\alpha \left\{ \sum_{k=1}^M \mu_k |w(P_k, d_n) - \mu|^2 + \gamma \|d_n\|_2^2 \right\},$$

then

$$\begin{aligned} r_{n+1} - r_n &= J(u^{n+1}) - J(u^n) \\ &= \alpha \langle \nabla J(u^n), d_n\rangle + \alpha^2 \left\{ \sum_{k=1}^M \mu_k |w(P_k, d_n) - \mu|^2 + \gamma \|d_n\|_2^2 \right\} \\ &\le \frac{\alpha}{2} \langle \nabla J(u^n), d_n\rangle. \end{aligned}$$

Taking $\alpha \ge \frac{1}{2}$ into account, we have

$$-\langle \nabla J(u^n), d_n\rangle \le \frac{2}{\alpha}(r_n - r_{n+1}) \le 4(r_n - r_{n+1}) \tag{3.20}$$

and

$$\|d_n\|_2 \le \sqrt{\frac{2}{\gamma}} (r_n - r_{n+1})^{\frac{1}{2}}. \tag{3.21}$$

If $\lambda_n^* < \alpha$, i.e., if

$$\begin{aligned} -\langle \nabla J(u^n), d_n\rangle &< 2\alpha \left\{ \sum_{k=1}^M \mu_k |w(P_k, d_n) - \mu|^2 + \gamma \|d_n\|_2^2 \right\} \\ &< 2 \left\{ \sum_{k=1}^M \mu_k |w(P_k, d_n) - \mu|^2 + \gamma \|d_n\|_2^2 \right\}, \end{aligned}$$

then by (1.13), (3.12), and Lemma 3.1 (iii), we get

$$
\begin{aligned}
r_n - r_{n+1} &= J(u^n) - J(u^{n+1}) \\
&= \frac{1}{4}\frac{|\langle \nabla J(u^n), d_n\rangle|^2}{\sum_{k=1}^M \mu_k |w(P_k, d_n) - \mu|^2 + \gamma \|d_n\|_2^2} \\
&\geq \frac{1}{4}\frac{(2\gamma\|d_n\|_2^2)^2}{\sum_{k=1}^M \mu_k C^2(P_k, \Omega, p')\|d_n\|_{p'}^2 + \gamma\|d_n\|_2^2} \\
&\geq \frac{\gamma^2}{\sum_{k=1}^M \mu_k C^2(P_k, \Omega, p')D^2 + \gamma D^{2\nu}} \|d_n\|_2^{2(1+\nu)}.
\end{aligned} \tag{3.22}
$$

Therefore

$$
\|d_n\|_2 \leq \left(\frac{L_1}{\gamma^2}\right)^{\frac{1}{2(1+\nu)}} (r_n - r_{n+1})^{\frac{1}{2(1+\nu)}}, \tag{3.23}
$$

where $L_1 = (\sum_{k=1}^M \mu_k C^2(P_k, \Omega, p') + \gamma)D^2$. And

$$
\begin{aligned}
-\langle \nabla J(u^n), d_n\rangle &< 2\left\{\sum_{k=1}^M \mu_k |w(P_k, d_n) - \mu|^2 + \gamma\|d_n\|_2^2\right\} \\
&< 2\left\{\sum_{k=1}^M \mu_k C^2(P_k, \Omega, p')D^2 + \gamma D^{2\nu}\right\} \|d_n\|_2^{2(1-\nu)} \\
&\leq 2\frac{L_1^{\frac{2}{1+\nu}}}{\gamma^{2\frac{1-\nu}{1+\nu}}} (r_n - r_{n+1})^{\frac{1-\nu}{1+\nu}}.
\end{aligned}
$$

Combining (3.20), (3.21), (3.22), and (3.23), and using the fact that $\nu < 1$ and $\{|r_n - r_{n+1}|\}_{n=1}^\infty$ is bounded by $B = 2r_1$, we obtain that in all cases

$$
\|d_n\|_2 \leq C_1 (r_n - r_{n+1})^{\frac{1}{2(1+\nu)}} \tag{3.24}
$$

and

$$
-\langle \nabla J(u^n), d_n\rangle \leq C_2 (r_n - r_{n+1})^{\frac{1-\nu}{1+\nu}}, \tag{3.25}
$$

where

$$
C_1 = \sqrt{2}\max\left\{\frac{B^{\frac{\nu}{2(1+\nu)}}}{\sqrt{\gamma}}, \left(\frac{L_1}{\gamma^2}\right)^{\frac{1}{2(1+\nu)}}\right\},
$$

$$
C_2 = 2\max\left\{2B^{\frac{2\nu}{1+\nu}}, \frac{L_1^{\frac{2}{1+\nu}}}{\gamma^{2\frac{1-\nu}{1+\nu}}}\right\}.
$$

Substituting (3.24), (3.25), and (3.9) into (3.17), we obtain

$$
\begin{aligned}
r_n &\leq C_2 (r_n - r_{n+1})^{\frac{1-\nu}{1+\nu}} + 4\sqrt{\gamma} C_1 (r_n - r_{n+1})^{\frac{1}{2(1+\nu)}} r_n^{\frac{1}{2}} \\
&\leq C_2 (r_n - r_{n+1})^{\frac{1-\nu}{1+\nu}} + 4\sqrt{\gamma} C_1 (r_n - r_{n+1})^{\frac{1-\nu}{2(1+\nu)}} r_n^{\frac{1}{2}} \\
&\leq \left(\sqrt{C_2}(r_n - r_{n+1})^{\frac{1-\nu}{2(1+\nu)}} + \frac{2\sqrt{\gamma} C_1}{\sqrt{C_2}} r_n^{\frac{1}{2}}\right)^2 - \frac{4\gamma C_1^2}{C_2} r_n.
\end{aligned}
$$

Thus

$$\sqrt{C_2}(r_n - r_{n+1})^{\frac{1-\nu}{2(1+\nu)}} + \frac{2\sqrt{\gamma}C_1}{\sqrt{C_2}} r_n^{\frac{1}{2}} \geq \left(1 + \frac{4\gamma C_1^2}{C_2}\right)^{\frac{1}{2}} r_n^{\frac{1}{2}}.$$

Rearranging the terms and simplifying, we get

$$r_n - r_{n+1} \geq C_3 r_n^{\frac{1+\nu}{1-\nu}}, \tag{3.26}$$

where

$$C_3 = ((C_2 + 4\gamma C_1^2)^{\frac{1}{2}} + 2\sqrt{\gamma}C_1)^{-2\frac{1+\nu}{1-\nu}}.$$

By Lemma 3.2, we have

$$r_n \leq r_0 \left[1 + \left(\frac{2\nu}{1-\nu}\right) C_3 r_0^{\frac{2\nu}{1-\nu}} n\right]^{-\frac{1-\nu}{2\nu}}. \tag{3.27}$$

Now (3.9) gives

$$\|u^n - u^*\|_2 \leq \frac{2}{\sqrt{\gamma}} r_0^{\frac{1}{2}} \left[1 + \left(\frac{2\nu}{1-\nu}\right) C_3 r_0^{\frac{2\nu}{1-\nu}} n\right]^{-\frac{1-\nu}{4\nu}}. \tag{3.28}$$

Combining (3.9) with (3.24) leads to

$$\begin{aligned}
\|\bar{u}^n - u^*\|_2 &\leq \|d_n\|_2 + \|u^n - u^*\|_2 \\
&\leq C_1(r_n - r_{n+1})^{\frac{1}{2(1+\nu)}} + \frac{2}{\sqrt{\gamma}} r_n^{\frac{1}{2}} \\
&\leq \left(C_1 + \frac{2}{\sqrt{\gamma}}\right) r_n^{\frac{1}{2(1+\nu)}} \\
&\leq \left(C_1 + \frac{2}{\sqrt{\gamma}}\right) r_0^{\frac{1}{2(1+\nu)}} \left[1 + \left(\frac{2\nu}{1-\nu}\right) C_3 r_0^{\frac{2\nu}{1-\nu}} n\right]^{-\frac{1-\nu}{4\nu(1+\nu)}}.
\end{aligned}$$

So the proof is complete. ∎

Next we provide the rate estimate of the uniform convergence.

THEOREM 3.6. *Let $\{\bar{u}^n\}_{n=1}^{\infty}$ be the sequence generated by the algorithm CGPM, (2.1)–(2.5). If we assume that $w(P_k, u^*) \neq Z_k$ and that Bu and Bl are locally bounded at P_k for all $k = 1, 2, \ldots, M$, then for $\mathcal{N} = 2$, $\bar{u}^n$ converges uniformly exponentially, and for $\mathcal{N} = 3$, $\bar{u}^n$ converges uniformly subexponentially; i.e., for all $x \in \Gamma$,*

$$|\bar{u}^n(x) - \bar{u}^*(x)| \leq \begin{cases} B_2 \delta^{\frac{n}{2}} & \text{for } \mathcal{N} = 2, \\ B_3 n^{-\frac{s}{2}} & \text{for } \mathcal{N} = 3, \end{cases} \tag{3.29}$$

where the constants B_2 and B_3 are independent of n and x, $\delta = 1 - \frac{15-6\sqrt{6}}{8}\frac{\gamma}{L} \in (0,1)$ with $L = \sum_{k=1}^{M} \mu_k C^2(P_k, \Omega) + \gamma$, and $s = \frac{p-p'}{p(p'-2)}$ for all $2 < p' < p$.

Proof. By taking account of (1.8), (1.9), (1.17), and Lemmas 1 and 2 in [4], we may denote

$$\begin{aligned}
\bar{w}(P_k, u) &= \int_\Gamma E(P_k, \xi) \left[\left(\frac{1}{2}I + \mathcal{K}^*\right)^{-1} u\right](\xi)\, d\sigma_\xi \\
&= \int_\Gamma \left(\frac{1}{2}I + \mathcal{K}\right)^{-1} [E(P_k, \cdot) - C_E](\xi) \cdot u(\xi)\, d\sigma_\xi,
\end{aligned}$$

where the constant C_E is defined s.t. $E(P_k,\cdot)-C_E\in L^q_{\perp f_0}(\Gamma)$; see [4]. Due to the fact that $\mathcal{S}(f_0)=$ constant, C_E is independent of k. We have

$$w(P_k,u^n)=\bar{w}(P_k,u^n)+\rho^n,$$

where by (1.10),

$$\rho^n=-\frac{\sum_{k=1}^M\mu_k(\bar{w}(P_k,u^n)-Z_k)}{\sum_{k=1}^M\mu_k}.$$

By Lemma 2 in [4], we obtain

$$\begin{aligned}&|w(P_k,u^n)-w(P_k,u^*)|\\&=\left|\int_\Gamma\left(\frac{1}{2}I+\mathcal{K}\right)^{-1}(E(P_k,\cdot)-C_E)(\xi)\cdot(u^n(\xi)-u^*(\xi))\,d\sigma_\xi+(\rho^n-\rho^*)\right|\\&\le B_0\|u^n-u^*\|_p,\end{aligned}\tag{3.30}$$

where

$$B_0=2\max_{1\le k\le M}\left\{\left\|\left(\frac{1}{2}I+\mathcal{K}\right)^{-1}(E(P_k,\cdot)-C_E)\right\|_q\right\}.$$

Applying (1.16) and the singularity decomposition formula, Lemma 3 in [2], we write

$$\begin{aligned}\bar{u}^n(x)&=\left[u^n(x)-\frac{1}{2\gamma}\nabla J(u^n)(x)+c_n\right]_{Bl}^{Bu}\\&=\left[-\frac{2}{\gamma}\sum_{k=1}^M\mu_k(w(P_k,u^n)-Z_k)E(P_k,x)+\sum_{k=1}^M\mu_k(w(P_k,u^n)-Z_k)f_0(P_k,x)\right.\\&\quad\left.+\sum_{k=1}^M\mu_k(w(P_k,u^n)-Z_k)f_1(P_k,x)+c_n\right]_{Bl}^{Bu},\end{aligned}$$

where for $\mathcal{N}=2$, $f_0\equiv 0$, and for $\mathcal{N}=3$, $f_0(P_k,x)$ has the only singularity at $x=P_k$, which is dominated by $E(P_k,x)$ and while $f_1(P_k,x)$ is continuous and bounded. Since we assume that $w(P_k,u^*)\neq Z_k$ and both Bu and Bl are locally bounded at P_k for all $k=1,2,\dots,M$ and also notice the fact that $\lim_{x\to P_k}E(P_k,x)=+\infty$ and all c_n and c_* are bounded, there exist $N>0$, $\delta_0>0$, $e_n>0$ and $e_*>0$ s.t. for all $n>N$, $-e_n\le e^n\le e_n$, $-e_*\le e^*\le e_*$ and for all $x\in\Gamma$ with $|x-P_k|<\delta_0$ for some $k=1,2,\dots,M$,

$$\left[u^n(x)-\frac{1}{2\gamma}\nabla J(u^n)(x)+c_n+e^n\right]_{Bl}^{Bu}=\text{ either } Bu(x)\text{ or }Bl(x)$$

and

$$\left[u^*(x)-\frac{1}{2\gamma}\nabla J(u^*)(x)+c_*+e^*\right]_{Bl}^{Bu}=\text{ either } Bu(x)\text{ or }Bl(x),$$

but only one of them is reached for each k. Let

$$\begin{aligned}\Gamma_1&=\{x\in\Gamma:|x-P_k|<\delta_0\ \text{ for some } k=1,2,\dots,M\},\\\Gamma_2&=\{x\in\Gamma:|x-P_k|\ge\delta_0\ \text{ for all } k=1,2,\dots,M\}.\end{aligned}$$

By the uniform convergence theorem, Theorem 3 in [4], we have for all $n > N$ and $x \in \Gamma_1$,

$$\left[u^n(x) - \frac{1}{2\gamma}\nabla J(u^n)(x) + c_n + e^n\right]_{Bl}^{Bu} = \left[u^*(x) - \frac{1}{2\gamma}\nabla J(u^*)(x) + c_* + e^*\right]_{Bl}^{Bu}. \tag{3.31}$$

As for $x \in \Gamma_2$, all functions $E(P_k, x)$, $f_0(P_k, x)$, and $f_1(P_k, x)$ are continuous and bounded, and there exists a constant $B_1 = B_1(\delta_0) > 0$, independent of n, s.t.

$$\begin{aligned}
&\left|\left\{u^n(x) - \frac{1}{2\gamma}\nabla J(u^n)(x)\right\} - \left\{u^*(x) - \frac{1}{2\gamma}\nabla J(u^*)(x)\right\}\right| \\
&= \left|\sum_{k=1}^{M} \mu_k(w(P_k, u^n) - w(P_k, u^*))\left\{\frac{2}{\gamma}E(P_k, x) + f_0(P_k, x) + f_1(P_k, x)\right\}\right| \\
&\le B_1 B_0 \|u^n - u^*\|_p. \qquad (3.32)
\end{aligned}$$

Next we prove that $|c_n - c_*| \le B_1 B_0 \|u^n - u^*\|_p$. Write

$$\delta_n = B_1 B_0 \|u^n - u^*\|_p$$

and

$$\begin{aligned}
F_l(x) &= u^l(x) - \frac{1}{2\gamma}\nabla J(u^l)(x) \\
&= \sum_{k=1}^{M} \mu_k(w(P_k, u^l) - Z_k)\left\{\frac{2}{\gamma}E(P_k, x) + f_0(P_k, x) + f_1(P_k, x)\right\}
\end{aligned}$$

for $l = n$ or $l = *$. Then (3.31) becomes

$$[F_n(x) + c_n]_{Bl}^{Bu} = [F_n(x) + c_n + e^n]_{Bl}^{Bu} = [F_*(x) + c_* + e^*]_{Bl}^{Bu} = [F_*(x) + c_*]_{Bl}^{Bu} \tag{3.33}$$

for all $n > N$, $x \in \Gamma_1$, $-e_n \le e^n \le e_n$, and $-e_* \le e^* \le e_*$. (3.32) now means

$$|F_n(x) - F_*(x)| \le \delta_n \quad \forall\, n > N,\ \forall\, x \in \Gamma_2. \tag{3.34}$$

Notice that c_n and c_* are of the smallest magnitudes s.t.

$$\int_\Gamma [F_n(x) + c_n]_{Bl}^{Bu}\, d\sigma_x = 0 \quad \text{and} \quad \int_\Gamma [F_*(x) + c_*]_{Bl}^{Bu}\, d\sigma_x = 0. \tag{3.35}$$

So, for all $-e_n \le e^n \le e_n$ and $-e_* \le e^* \le e_*$, we have

$$\begin{aligned}
\int_{\Gamma_2} [F_*(x) + c_*]_{Bl}^{Bu}\, d\sigma_x &= -\int_{\Gamma_1} [F_*(x) + c_* + e^*]_{Bl}^{Bu}\, d\sigma_x \\
&= -\int_{\Gamma_1} [F_n(x) + c_n + e^n]_{Bl}^{Bu}\, d\sigma_x \qquad (3.36) \\
&= \int_{\Gamma_2} [F_n(x) + c_n]_{Bl}^{Bu}\, d\sigma_x.
\end{aligned}$$

If $|c_n - c_*| = \delta_n + e_n'$ with $e_n' > 0$ (e.g., $c_* = c_n + \delta_n + e_n'$), then by (3.36) we have

$$\begin{aligned}
\int_{\Gamma_2} [F_n(x) + c_n]_{Bl}^{Bu}\, d\sigma_x &= \int_{\Gamma_2} [F_*(x) + c_*]_{Bl}^{Bu}\, d\sigma_x \\
&= \int_{\Gamma_2} [F_*(x) + c_n + \delta_n + e_n']_{Bl}^{Bu}\, d\sigma_x \\
&\ge \int_{\Gamma_2} [F_n(x) + c_n + e_n']_{Bl}^{Bu}\, d\sigma_x.
\end{aligned}$$

By Lemma 6 in [4], we have

$$\int_{\Gamma_2} [F_n(x) + c_n]_{Bl}^{Bu} \, d\sigma_x = \int_{\Gamma_2} [F_n(x) + c_n + e'_n]_{Bl}^{Bu} \, d\sigma_x$$

and then

$$\int_{\Gamma_2} [F_n(x) + c_n]_{Bl}^{Bu} \, d\sigma_x = \int_{\Gamma_2} [F_n(x) + c_n + e^n]_{Bl}^{Bu} \, d\sigma_x \quad \forall\, 0 \le e^n \le e'_n.$$

It follows that

$$\int_{\Gamma} [F_n(x) + c_n + e^n]_{Bl}^{Bu} \, d\sigma_x = 0, \quad \forall\, 0 \le e^n \le \min\{e_n, e'_n\}. \tag{3.37}$$

Since c_n has the smallest magnitude s.t.

$$\int_{\Gamma} [F_n(x) + c_n]_{Bl}^{Bu} \, d\sigma_x = 0,$$

(3.37) implies $c_n \ge 0$. On the other hand, we have

$$\begin{aligned} \int_{\Gamma_2} [F_*(x) + c_*]_{Bl}^{Bu} \, d\sigma_x &= \int_{\Gamma_2} [F_n(x) + c_n]_{Bl}^{Bu} \, d\sigma_x \\ &= \int_{\Gamma_2} [F_n(x) + c_* - \delta_n - e'_n]_{Bl}^{Bu} \, d\sigma_x \\ &\le \int_{\Gamma_2} [F_*(x) + c_* - e'_n]_{Bl}^{Bu} \, d\sigma_x. \end{aligned}$$

Again by Lemma 6 in [4], we obtain

$$\int_{\Gamma_2} [F_*(x) + c_*]_{Bl}^{Bu} \, d\sigma_x = \int_{\Gamma_2} [F_*(x) + c_* - e'_n]_{Bl}^{Bu} \, d\sigma_x,$$

and then

$$\int_{\Gamma_2} [F_*(x) + c_*]_{Bl}^{Bu} \, d\sigma_x = \int_{\Gamma_2} [F_*(x) + c_* - e^n]_{Bl}^{Bu} \, d\sigma_x \quad \forall\, 0 \le e^n \le e'_n.$$

It follows that

$$\int_{\Gamma} [F_*(x) + c_* - e^n]_{Bl}^{Bu} \, d\sigma_x = 0, \quad \forall\, 0 \le e^n \le \min\{e_*, e'_n\}. \tag{3.38}$$

Since c_* has the smallest magnitude s.t.

$$\int_{\Gamma} [F_*(x) + c_*]_{Bl}^{Bu} \, d\sigma_x = 0,$$

(3.38) implies $c_* \le 0$. It leads to

$$0 \ge c_* = c_n + \delta_n + e_n > 0,$$

which is a contradiction. Similarly, we can show that $c_n = c_* + \delta_n + e'_n$ with $e'_n > 0$ will also lead to a contradiction. Therefore

$$|c_n - c_*| \le \delta_n = B_1 B_0 \|u^n - u^*\|_p. \tag{3.39}$$

Now for all $n > N$ and for all $x \in \Gamma_1$, by applying (3.33), we get

$$|\bar{u}^n(x) - \bar{u}^*(x)| = \left| [F_n(x) + c_n]_{Bl}^{Bu} - [F_*(x) + c_*]_{Bl}^{Bu} \right| = 0,$$

and for all $x \in \Gamma_2$, taking account of (3.34) and (3.39), we have

$$\begin{aligned} |\bar{u}^n(x) - \bar{u}^*(x)| &= \left| [F_n(x) + c_n]_{Bl}^{Bu} - [F_*(x) + c_*]_{Bl}^{Bu} \right| \\ &\le |F_n(x) + c_n - F_*(x) - c_*| \\ &\le 2B_1 B_0 \|u^n - u^*\|_p. \end{aligned}$$

For $n = 1, 2, \ldots, N$, let

$$B_3 = \max \left\{ \max_{1 \le n \le N} \max_{|x - P_k| \ge \delta_0} |F_n(x) + c_n - F_*(x) - c_*|, \right.$$
$$\left. \max_{1 \le k \le M} \sup_{|x - P_k| < \delta_0} [Bu(x) - Bl(x)] \right\}.$$

Then

$$|\bar{u}^n(x) - \bar{u}^*(x)| \le B_3 \le B_4 \|u^n - u^*\|_p$$

with

$$B_4 = \max \left\{ \frac{B_3}{\|u^n - u^*\|_p} : 1 \le n \le N,\ \|u^n - u^*\|_p \ne 0 \right\}.$$

In the above, since $\|u^n - u^*\|_p = 0$ implies $\bar{u}^n(x) = \bar{u}^*(x)$ for all $x \in \Gamma$, the iterate will stop, and the case in which $\|u^n - u^*\|_p = 0$ may be excluded. If we denote

$$B = \max\{B_4, 2B_1 B_0\},$$

then for all $x \in \Gamma$ we have

$$|\bar{u}^n(x) - \bar{u}^*(x)| \le B\|u^n - u^*\|_p \le \begin{cases} 2B\sqrt{\dfrac{r_1}{\gamma}}\delta^{\frac{n}{2}} & \text{for} \quad \mathcal{N} = 2, \\ BE_2 n^{-\frac{s}{2}} & \text{for} \quad \mathcal{N} = 3, \end{cases}$$

where r_1 and δ are defined in (3.3), and E_2 and s are defined in (3.14) and (3.15). Therefore the proof is complete. □

Remark 3. Notice that in (3.13), (3.14), (3.15), (3.16), and (3.29), $s = \frac{p - p'}{p(p'-2)}$ and $2 < p' < p$. So s can be made to be greater than any number by a proper choice of p'. Therefore, the convergence is faster than any power of $\frac{1}{n}$. That is why we call it subexponential convergence. Of course, the constants E_1, E_2, and E_3 in (3.13), (3.15), and (3.16), respectively, and the constant B_3 in (3.29) all depend on s but are independent of n.

Also, the assumption that $w(P_k, u^*) \ne Z_k$ for all $k = 1, 2, \ldots, M$ is important in computing the value of the optimal control $\bar{u}^*$ at $x = P_k$. When $w(P_k, u^*) = Z_k$ for some k, since the term $E(P_k, x)$ has a singularity at $x = P_k$, any error in computing $w(P_k, u^*)$ will result in a huge error in computing $(w(P_k, u^*) - Z_k)E(P_k, x)$, as x is sufficiently close to P_k and then destroys the reliability of the numerical value of the optimal control at $x = P_k$. So the strong convergence of u^n cannot guarantee the convergence of $u^n(P_k)$. Therefore, the condition that $w(P_k, u^*) \ne Z_k$ for all $k = 1, 2, \ldots, M$ is indeed a necessary and sufficient condition for u^n to converge at $P_k,\ k = 1, 2, \ldots, M$.

REFERENCES

[1] G. Chen and J. Zhou, *Boundary Element Methods*, Academic Press, London, 1992.

[2] Z. Ding, L. Ji, and J. Zhou, *Constrained LQR problems in elliptic distributed control systems with point observations*, SIAM J. Control Optim., 34 (1996), pp. 264–294.

[3] Z. Ding and J. Zhou, *Constrained LQR problems governed by the potential equation on Lipschitz domain with point observations*, J. Math. Pures Appl., 74 (1995), pp. 317–344.

[4] Z. Ding and J. Zhou, *Constrained LQR problems in elliptic distributed control systems with point observations—Convergence results*, Appl. Math. Optim., 1997, to appear.

[5] V.F. Demyanov and A.M. Rubinov, *Approximate Methods in Optimization Problems*, American Elsevier, New York, 1971.

[6] J.C. Dunn, *Global and asymptotic convergence rate estimates for a class of projected gradient process*, SIAM J. Control Optim., 19 (1981), pp. 368–400.

[7] E.B. Fabes, M. Jodeit, Jr., and N.M. Riviere, *Potential techniques for boundary value problems on C^1-domains*, Acta Mathematica, 141 (1978), pp. 165–186.

[8] D. Gilbarg and N.S. Trudinger, *Elliptic Partial Differential Equations of Second Order*, Springer-Verlag, New York, 1983.

[9] P. Grisvard, *Elliptic Problems in Nonsmooth Domains*, Pitman, Boston, 1985.

[10] G. Hsiao and R.C. MacCamy, *Solutions of boundary value problems by integral equations of the first kind*, SIAM Rev. 15 (1973), pp. 687–705.

[11] L. Ji and G. Chen, *Point observation in linear quadratic elliptic distributed control systems*, in Proceedings of American Mathematical Society Summer Conference on Control and Identification of Partial Differential Equations, SIAM, Philadelphia, 1993, pp. 155–170.

[12] C.E. Kenig, *Recent Progress on Boundary-Value Problems on Lipschitz Domains*, in AMS Proc. Symp. Pure Math., Vol. 43, 1985, pp. 175–205.

[13] J.L. Lions, *Contrôle optimal des systèms gouvernés par des équations aux dérivées partielles*, Dunod, Gauthier-Villars, Paris, 1968.

[14] J.L. Lions and E. Magenes, *Nonhomogeneous Boundary Value Problems and Applications*, Vol. 1, Springer-Verlag, New York, 1970.

[15] C.E. Lee and H.E. Taylor, *Fiber-optic Fabry-Perot temperature sensor using a low-coherence light source*, J. Lightwave Technol., 9 (1994), pp. 129–134.

[16] G. Verchota, *Layer potentials and regularity for the Dirichlet problems for Laplace's equation in Lipschitz domains*, J. Functional Anal., 59 (1984), pp. 572–611.

[17] G. Verchota, *Layer Potentials and Boundary Value Problems for Laplace's Equation on Lipschitz Domains*, Ph.D. thesis, University of Minnesota, Minneapolis, 1982.

[18] M.C. Wang, *Fiber-Optic Fabry-Perot Temperature and Dynamic Sensor System Using a Low-Coherence LED Light Source*, Ph.D. thesis, Texas A&M University, College Station, TX, 1995.

[19] M.T. Wlodarczyk and G. He, *A fiber-optic combustion pressure sensor system for automotive engine control*, Sensors, 11 (1994), pp. 35–42.

[20] N.G. Zamani and J.M. Chuang, *Optimal control of current in a cathodic protection system: A numerical investigation*, Optim. Control Appl. Meth., 8 (1987), pp. 339–350.

[21] N.G. Zamani, J.F. Porter, and A.A. Mufti, *A survey of computational efforts in the field of corrosive engineering*, J. Numer. Methods Engrg., 23 (1986), pp. 1295–1311.

SIAM J. CONTROL OPTIM.
Vol. 35, No. 5, pp. 1755–1772, September 1997

017

NONLINEAR UNCERTAIN SYSTEMS AND NECESSARY CONDITIONS OF OPTIMALITY*

N. U. AHMED† AND X. XIANG‡

Abstract. In this paper we consider an optimal control problem for a class of systems governed by nonlinear evolution equations containing a nonlinear monotone operator and a nonmonotone operator with uncertain parameters. We prove existence of solutions and present necessary conditions of optimality. Our result is illustrated by an example from quasi-linear partial differential equations with uncertain coefficients. This result is further illustrated by a more practical example.

Key words. uncertain system, monotone operator, existence, Galerkin method, optimal control, necessary conditions

AMS subject classifications. 93C25, 49K24, 49K27

PII. S0363012995285569

1. Introduction.

Many physical systems arising from thermodynamics, electrodynamics, and population biology are modeled by differential equations, integrodifferential equations, and evolution inequalities with uncertain (or undetermined) parameters. Generally, such models can be described by differential inclusions on Banach spaces as follows:

$$\begin{cases} \dot{x} \in -A(t,x) + F(t,x), \\ x(0) = x^0, \end{cases} \tag{1}$$

where A is a linear or a nonlinear unbounded operator in a suitable Banach space and F is a multivalued map. An associated control system may be described as

$$\begin{cases} \dot{x} \in -A(t,x) + \tilde{F}(t,x,u), \\ x(0) = x^0, \end{cases} \tag{2}$$

where $\tilde{F}$ is a multivalued map and u is a suitable function representing the control actions. In recent years optimal control of systems governed by differential inclusions and, more generally, functional differential inclusions has been studied by Ahmed and Papageorgiou (see [2, 8, 9, 10] and the references therein). These studies were mainly concerned with the question of existence of optimal controls or parameters. Here we are concerned with the necessary conditions of optimality for a min-max problem (see [3, 5]).

For each admissible control u, $\mathcal{X}(u)$ denotes the set of solutions of (2) corresponding to u (usually generalized solutions). A natural problem (P) is to find $u_0 \in \mathcal{U}_{ad}$ (admissible controls) so that

$$J_0(u_0) = \inf_{u \in \mathcal{U}_{ad}} J_0(u), \tag{3}$$

where

$$J_0(u) = \sup_{x \in \mathcal{X}(u)} \left\{ \int_I l(t,x,u)dt \right\}.$$

*Received by the editors May 4, 1995; accepted for publication (in revised form) July 15, 1996.
http://www.siam.org/journals/sicon/35-5/28556.html

†Department of Electrical Engineering and Department of Mathematics, University of Ottawa, Ottawa, ON K1N 6N5, Canada (ahmed@trix.genie.uottawa.ca).

‡Department of Mathematics, Guizhou University, Guiyang, Guizhou, People's Republic of China.

The system designer wishes to find a control policy to minimize the maximum risk or maximize the minimum revenue. Since competing interests are involved, this is precisely a problem of game theory—here, conflicting with nature. Similar and closely related problems have been studied by Papageorgiou in [11, 12] and Ahmed and Xiang in [2, 4, 6, 7]. In [11, 4] existence questions were addressed. Here we are concerned with the necessary conditions of optimality.

In the case when $\tilde{F}$ admits a parameterization in the following form:

$$\tilde{F}(t,x,u) \equiv \left\{ \int_{\Sigma} \tilde{g}(t,x,u,\sigma)\ \mu(d\sigma), \quad \mu \in \mathcal{M}(\Sigma) \right\}, \tag{4}$$

where $\mathcal{M}(\Sigma)$ is the space of probability measures, we obtained a series of results. We studied the question of existence of optimal controls for nonlinear uncertain systems (see [4]) and presented necessary conditions of optimality for semilinear uncertain systems whose principal operator is the infinitesimal generator of a strongly continuous semigroup or a linear monotone operator (see [6], [7]).

In this paper we will present necessary conditions of optimality for problem (P) subject to the more general class of systems given by (2). Here A is assumed to be a nonlinear monotone operator and $\tilde{g}(t,x,u,\sigma)$ is a nonlinear but not monotone operator. For this purpose we prove, in section 2, the existence and uniqueness of solutions of nonlinear first-order evolution equations. In section 3, we study some necessary regularity properties of solutions of the associated nonlinear control system. In section 4 we give the necessary conditions of optimality, and in section 5 we present an example of a system, governed by a quasi-linear partial differential equation with uncertain parameters to which our results apply. This is further illustrated by a reaction-diffusion-transport system. The differences between the work of Papageorgiou [11, 12] on this topic and that considered in this paper are as follows. In [11], the system model considered consists of a nonlinear monotone operator perturbed by a nonmonotone but regular operator (mapping within the same Hilbert space, thereby excluding differential expressions) with controls appearing linearly. In [12] the principal operator is also a monotone operator arising from the subdifferential of a proper, convex, lower semicontinuous functional. The lower-order terms are Lipschitz and contain the uncertain parameters. In contrast, in this paper the perturbing operator admits differential expressions (though more regular than the principal operator) with controls appearing nonlinearly. This is a significant difference. Our assumptions, like those of Papageorgiou on the nonlinear monotone operator and the cost integrand, are standard and match those of Papageorgiou. Further, Papageorgiou considers existence questions, and we consider necessary conditions of optimality. These approaches are complementary.

2. Existence and uniqueness of solutions. Let H be a Hilbert space and V be a subspace of H having the structure of a reflexive Banach space, with the embedding $V \hookrightarrow H$ being dense and continuous. Identifying H with its dual, we have $V \hookrightarrow H \hookrightarrow V^*$, where V^* is the topological dual of V. The system model considered here is based on this evolution triple.

Let $\langle x, y\rangle$ denote the pairing of an element $x \in V$ and an element $y \in V^*$. If $x, y \in H$, then $\langle x, y\rangle = (x, y)$, where $(\ ,\ )$ is the scalar product on H. The norm in any Banach space X will be denoted by $\|.\|_X$.

Let $\{e_1, e_2, \ldots\}$ be a basis of V and set

$$H_n = \text{lin.span}\{e_1, \ldots, e_n\}.$$

We introduce in the n-dimensional space H_n the scalar product of Hilbert space H. Note $H_n \subseteq V \subseteq H$.

Let $0 < t \le T < \infty,\ I_t \equiv [0,t],\ I \equiv [0,T]$, and let $p, q \ge 1$ such that

$$1/p + 1/q = 1 \quad \text{and} \quad 2 \le p < \infty.$$

For economy of notation, we write $L_p^t(V) \equiv L_p(I_t, V),\ L_p(V) \equiv L_p(I, V),\ L_q^t(V^*) \equiv L_q(I_t, V^*),\ L_q(V^*) \equiv L_q(I, V^*)$. For p, q satisfying the preceding conditions, it follows from the reflexivity of V that both $L_p^t(V)$ and $L_q^t(V^*)$ are reflexive Banach spaces (see Theorem 1.1.17 of [1]). The pairing of $L_p^t(V)$ and $L_q^t(V^*)$ is denoted by $\langle\langle , \rangle\rangle_t$. In particular, for $t = T$, we use $\langle\langle , \rangle\rangle \equiv \langle\langle , \rangle\rangle_T$. Clearly, for $u, v \in L_2(H), \langle\langle u, v\rangle\rangle = ((u, v))$ is the scalar product in Hilbert space $L_2(H)$.

Define

$$W_{p,q} = \{x : x \in L_p(V), \dot{x} \in L_q(V^*)\},$$

$$\|x\|^2_{W_{p,q}} = \|x\|^2_{L_p(V)} + \|\dot{x}\|^2_{L_q(V^*)}.$$

$\{W_{p,q}, \| \ \|_{W_{p,q}}\}$ is a Banach space, and the embedding $W_{p,q} \hookrightarrow C(I, H)$ is continuous. If the embedding $V \hookrightarrow H$ is compact, then $W_{p,q} \hookrightarrow L_p(H)$ is also compact (see Proposition 23.23 and problem 23.13 of [3]).

Let $\mathcal{L}(X, Z)$ denote the space of bounded linear operators from X to Z and A^* the dual of the operator A.

We introduce the following assumptions.

(A1) $A : I \times V \to V^*$.

(1) $t \to A(t, x)$ is measurable.

(2) $x \to A(t, x)$ is uniformly monotone and hemicontinuous; i.e., there exists a constant $c > 0$ such that

$$\langle A(t, x_1) - A(t, x_2), x_1 - x_2\rangle \ge c\ \|x_1 - x_2\|^p_V \quad \forall\ x_1, x_2 \in V, \quad t \in I;$$

$$A(t, x + sy) \xrightarrow{w} A(t, x) \quad \text{in } V^* \quad \forall\ x, y \in V, \quad \text{as } s \to 0.$$

(3) There exist positive constants c_1, c_2, c_3 and a nonnegative function $c_4(t) \in L_q(I, R)$ such that

$$\langle A(t, x), x\rangle \ge c_1\|x\|^p_V - c_2 \quad \forall\ x \in V,\ t \in I;$$

$$\|A(t, x)\|_{V^*} \le c_4(t) + c_3\|x\|^{p-1}_V \quad \forall\ x \in V,\ t \in I.$$

(G1) $g : I \times H \to\ V^*$.

(1) g is measurable in the first variable and continuous in the second argument.

(2) There exist a constant $\alpha \ge 0$ and $h \in L_q(I, R_+)$ such that

$$\|g(t, x)\|_{V^*} \le h(t) + \alpha\|x\|^{\frac{2}{q}}_H \quad \forall\ x \in V,\ t \in I.$$

(3) g is locally Lipschitz continuous with respect to x, that is, for any $b > 0$, there exists a constant $L(b)$ such that for $x_1, x_2 \in H,\ \|x_1\|_H, \|x_2\|_H \le b$,

$$\|g(t, x_1) - g(t, x_2)\|_{V^*} \le L(b)\|x_1 - x_2\|_H \quad \forall\ t \in I.$$

Remark 2.1. Note that assumption (A1) implies coerciveness (see [3]).

Under the above assumptions we consider the following basic initial value problem:

$$\begin{cases} \dot{x} + A(t,x) = g(t,x), \\ x(0) = x^0. \end{cases} \tag{5}$$

For given $x^0 \in H$, we seek a function $x \in W_{p,q}$ such that (5) is satisfied in a weak sense, as explained later.

For $x \in L_p(V)$, we set

$$A(x)(t) = A(t, x(t)), \quad G(x)(t) = g(t, x(t)), \quad t \in I.$$

It follows from Theorem 30.A of [3] that the operator $A : L_p(V) \to L_q(V^*)$ is bounded, uniformly monotone, hemicontinuous, and coercive. The operator $G : L_p(V) \to L_q(V^*)$ is also bounded (see assumption (G1)) and has the following important properties.

LEMMA 2.2. *Suppose that the embedding $V \hookrightarrow H$ is compact. Then, whenever $x_n \overset{w}{\to} x$ in $W_{p,q}$, $G(x_n) \to G(x)$ in $L_q(V^*)$.*

Proof. Since the embedding $V \hookrightarrow H$ is compact, the embedding $W_{p,q} \hookrightarrow L_p(H)$ is compact. This means that if

$$x_n \overset{w}{\to} x \quad \text{in } W_{p,q},$$

then

$$x_n \to x \quad \text{in } L_p(H).$$

Since $x_n \overset{w}{\to} x$ in $W_{p,q}$, there exists a constant $b > 0$ such that $\|x\|_{C(I,H)} \le b$, $\|x_n\|_{C(I,H)} \le b$. By virtue of assumption (G1) and the embeddings $L_p(H) \hookrightarrow L_q(H) \hookrightarrow L_q(V^*)$, one can easily verify that

$$\begin{aligned} &\|G(x_n) - G(x)\|_{L_q(V^*)} \\ &= \left(\int_I \|g(t, x_n(t)) - g(t, x(t))\|_{V^*}^q dt \right)^{1/q} \\ &\le L(b) \left(\int_I \|x_n(t) - x(t)\|_H^q dt \right)^{1/q} \\ &\le L^* \left(\int_I \|x_n(t) - x(t)\|_H^p dt \right)^{1/p}, \end{aligned}$$

where L^* is a constant depending on p, q, b and the Lebesgue measure of I. Hence the conclusion follows.

Remark 2.3. In fact, following a similar argument as in the proof of Lemma 2.5.1 of [1], we can prove Lemma 2.2 without the local Lipschitz continuity.

It is often convenient to write system (5) as an operator equation in

$$W_{p,q}^0 \equiv \{x \in W_{p,q} : x(0) = x^0\} :$$

$$\begin{cases} \dot{x} + A(x) = G(x), \\ x \in W_{p,q}^0. \end{cases} \tag{6}$$

The purpose of this section is to present an existence result for equation (5) based on Galerkin approximation. At first, we give an a priori bound and prove the uniqueness of the solution.

LEMMA 2.4. *There exists a finite positive number b such that*

$$\|x\|_{C(I,H)} \le b, \quad \|x\|_{L_p(V)} \le b,$$
$$\|\dot{x}\|_{L_q(V^*)} \le b,$$

for any solution x (if one exists) of equation (5).

Proof. If x is any solution of (5), then for each $t \in I$,

$$\langle\langle \dot{x}, x\rangle\rangle_t + \langle\langle A(x), x\rangle\rangle_t = \langle\langle G(x), x\rangle\rangle_t,$$

giving

$$\frac{1}{2}(\|x(t)\|_H^2 - \|x(0)\|_H^2) + \langle\langle A(x), x\rangle\rangle_t = \langle\langle G(x), x\rangle\rangle_t.$$

Using the assumptions and the Cauchy inequality, for any $\varepsilon > 0$, we have

$$\begin{aligned}
&\frac{1}{2}(\|x(t)\|_H^2 - \|x(0)\|_H^2) + c_1 \ \|x\|^p_{L_p^t(V)} \\
&\le c_2 + \int_0^t \|g(t, x(t))\|_{V^*}\|x(t)\|_V dt \\
&\le c_2 + \int_0^t (h(t) + \|x(t)\|_H^{\frac{2}{q}})\|x(t)\|_V dt \\
&\le c_2 + \frac{1}{q\varepsilon^q}\int_0^t (h(t) + \|x(t)\|_H^{\frac{2}{q}})^q dt + \frac{\varepsilon^p}{p}\int_0^t \|x(t)\|_V^p dt.
\end{aligned}$$

Choosing $\varepsilon > 0$ sufficiently small, one can easily verify that there exist positive constants c_5, c_6, c_7 such that

$$\|x(t)\|_H^2 + c_5\|x\|^p_{L_p^t(V)} \le c_6 + c_7\int_0^t \|x(t)\|_H^2 dt. \tag{7}$$

From the Gronwall lemma it follows from the above inequality that

$$\|x(t)\|_H \le c_8 \quad \forall\ t \in I,$$

for some constant c_8 depending on c_6 and c_7. Again, by virtue of assumptions (3) of (A1) and (2) of (G1) and inequality (7), it is easy to verify that there exist positive constants c_9, c_{10} such that

$$\|x\|_{L_p(V)} \le c_9, \quad \|\dot{x}\|_{L_q(V^*)} \le c_{10}.$$

Choosing $b = \max\{c_8, c_9, c_{10}\}$, the assertion follows.

LEMMA 2.5. *The solution of* (5)*, if one exists, is unique.*

Proof. Let $x_1, x_2 \in W^0_{p,q}$ be two solutions of (5). Using integration by parts and the monotonicity of the operator A, we obtain that

$$\begin{aligned}
&\frac{1}{2}\|x_1(t) - x_2(t)\|_H^2 + c\|x_1 - x_2\|^p_{L_p^t(V)} \\
&\le \int_0^t \langle g(t, x_1(t)) - g(t, x_2(t)), x_1(t) - x_2(t)\rangle dt.
\end{aligned}$$

By virtue of assumption (G1), Lemma 2.4, and the Cauchy inequality, for any $\varepsilon > 0$, we have

$$\begin{aligned}
&\frac{1}{2}\|x_1(t) - x_2(t)\|_H^2 + c\|x_1 - x_2\|_{L_p^t(V)}^p \\
&\le \int_0^t \|g(t, x_1(t)) - g(t, x_2(t))\|_{V^*}\|x_1(t) - x_2(t)\|_V dt \\
&\le L(b)\int_0^t \|x_1(t) - x_2(t)\|_H\|x_1(t) - x_2(t)\|_V dt \\
&\le \frac{L(b)\varepsilon}{2}\int_0^t \|x_1(t) - x_2(t)\|_V^2 dt + \frac{L(b)}{2\varepsilon}\int_0^t \|x_1(t) - x_2(t)\|_H^2 dt.
\end{aligned}$$

Using Lemma 2.3 and the continuous embedding $L_p^t(V) \hookrightarrow L_2^t(V)$, we obtain

$$\begin{aligned}
&\|x_1(t) - x_2(t)\|_H^2 + 2c\|x_1 - x_2\|_{L_p^t(V)}^p \\
&\le L_1\varepsilon\int_0^t \|x_1(t) - x_2(t)\|_V^p dt + \frac{L(b)}{\varepsilon}\int_0^t \|x_1(t) - x_2(t)\|_H^2 dt,
\end{aligned}$$

where L_1 is a constant depending on b and the embedding constant. Consequently, for sufficiently small $\varepsilon > 0$, there exists a constant $c' > 0$ such that

$$\begin{aligned}
&\|x_1(t) - x_2(t)\|_H^2 + c'\|x_1 - x_2\|_{L_p^t(V)}^p \\
&\le \frac{L(b)}{\varepsilon}\int_0^t \|x_1(t) - x_2(t)\|_H^2 dt.
\end{aligned}$$

Using the Gronwall lemma, uniqueness follows from the above inequality.

THEOREM 2.A. *Under assumptions* (A1) *and* (G1)*, the evolution equation* (5) *has a unique solution.*

Proof. Let the sequence $\{x_n^0\}$ be an approximation of the given initial state $x^0 \in H$, i.e., $x_n^0 \in H_n$, $x_n^0 \to x^0$ in H, as $n \to \infty$.

Consider the sequence

$$x_n(t) = \sum_{k=1}^n C_{k,n}(t)e_k,$$

and seek a function x_n such that

$$\text{(8)} \qquad \begin{cases} \langle \dot{x}_n(t), e_j\rangle + \langle A(t, x_n(t)), e_j\rangle = \langle g(t, x_n(t)), e_j\rangle, & j = 1, 2, \ldots, n; \\ x_n(0) = x_n^0; \\ x_n \in L_p(I, H_n),\ \dot{x}_n \in L_q(I, H_n). \end{cases}$$

It follows from the existence theorem of Carathéodory for ordinary differential equations in R^n (see problem 30.3 of [3]) and Lemma 2.4 that, for each $n \in N$, the finite-dimensional system (8) has a unique solution x_n. It can be seen from Lemma 2.4 that $\{x_n\}$ is contained in a bounded subset of $W_{p,q}$. Hence, by assumption (A1), $\{A(x_n)\}$ is bounded in $L_q(V^*)$. Since $L_p(V)$ and $L_q(V^*)$ are reflexive Banach spaces, there exists a subsequence, again denoted by $\{x_n\}$; an element $x \in L_p(V)$ with its distributional derivative $\dot{x} \in L_q(V^*)$; and $W \in L_q(V^*)$ such that

$$\begin{aligned}
x_n &\xrightarrow{w} x \qquad \text{in } L_p(V), \\
\dot{x}_n &\xrightarrow{w} \dot{x} \qquad \text{in } L_q(V^*), \\
A(x_n) &\xrightarrow{w} W \quad \text{in } L_q(V^*)
\end{aligned}$$

as $n \to \infty$. Combining the assumptions with Lemma 2.2, we have

$$\begin{aligned} G(x_n) &\to G(x) \quad \text{in } L_q(V^*), \\ x_n(0) &\to x^0 \quad \text{in } H, \\ x_n(T) &\xrightarrow{w} z \quad \text{in } H \end{aligned}$$

as $n \to \infty$.

Let $\psi \in C^\infty(I, R)$ and $v \in H_n$. Using equation (8) and integration by parts, one can obtain

$$\begin{aligned} &(x_n(T), \psi(T)v) - (x_n(0), \psi(0)v) \\ &= \int_I \langle g(t, x_n(t)) - A(t, x_n(t)), \psi(t)v\rangle + \langle \dot{\psi}(t)v, x_n(t)\rangle dt. \end{aligned}$$

Letting $n \to \infty$, we have

$$(z, \psi(T)v) - (x^0, \psi(0)v) = \langle\langle G(x) - W, \psi v\rangle\rangle + \langle\langle \dot{\psi} v, x\rangle\rangle. \tag{9}$$

Using this, one can easily verify that the limit elements x, W, z satisfy

$$\begin{cases} \dot{x} + W = G(x), \quad x \in W_{p,q}, \\ x(0) = x^0, \ x(T) = z. \end{cases}$$

Again using equation (8) and integration by parts, we have

$$\frac{1}{2}(\|x_n(T)\|_H^2 - \|x_n(0)\|_H^2) = \langle\langle G(x_n) - A(x_n), x_n\rangle\rangle.$$

By virtue of the fact that

$$\underline{\lim}_{n\to\infty} \|x_n(T)\|_H \geq \|x(T)\|_H,$$

we obtain

$$\begin{aligned} &\overline{\lim_{n\to\infty}} \langle\langle A(x_n), x_n\rangle\rangle \\ &\leq \langle\langle G(x), x\rangle\rangle + \frac{1}{2}(\|x(0)\|_H^2 - \|x(T)\|_H^2) \\ &= \langle\langle W, x\rangle\rangle. \end{aligned}$$

Since A is monotone and hemicontinuous, A satisfies property (M) (see section 27.1 of [3]), and hence

$$W = A(x).$$

Thus the limit element x satisfies equation (6) and hence is a solution of (5). The uniqueness follows from Lemma 2.5. This finishes the proof of the theorem.

Remark 2.6. It follows from Proposition 21.23 of [3] and Lemma 2.5 that the Galerkin sequence $\{x_n\}$ weakly converges to x in $L_p(V)$. Furthermore, following an argument similar to that in Lemma 30.7 of [3], one can show that $x_n \to x$ in $C(I, H)$.

3. Controlled uncertain system. Let Σ be a compact Polish space and $\mathcal{M}(\Sigma)$ be the space of probability measures on Σ. A sequence $\mu^n \in \mathcal{M}(\Sigma)$ is said to converge weakly to $\mu \in \mathcal{M}(\Sigma)$ if

$$\int_\Sigma g(\sigma)\mu^n(d\sigma) \ \to \int_\Sigma g(\sigma)\mu(d\sigma)$$

for every $g \in C(\Sigma)$.

In addition to assumption (A1) we shall introduce the following assumptions.

(U) (1) Y is a reflexive Banach space.

(2) $U : I \to CC(Y) = \{$class of nonempty, closed, convex subsets of $Y\}$ is a measurable multifunction satisfying $U(t) \subseteq \mathcal{U}$ for almost all $t \in I$, where $\mathcal{U}$ is a fixed weakly compact convex subset of Y. For the admissible controls, we choose the set $\mathcal{U}_{ad} \equiv \{u \in L_r(Y) : u(t) \in U(t)$ a.e.$\}$ $(r \geq 2)$.

(G) $\tilde{g} : I \times H \times Y \times \Sigma \to V^*$, for $\mu \in \mathcal{M}(\Sigma)$, $g(t,x,u,\mu) \equiv \int_\Sigma \tilde{g}(t,x,u,\sigma)\mu(d\sigma)$.

(1) g is measurable in the first variable and continuous in the last three arguments.

(2) $g(.,.,u,\mu)$ satisfies assumption (G1) uniformly with respect to $u \in \mathcal{U}$, $\mu \in \mathcal{M}(\Sigma)$.

(3) g is Fréchet-differentiable with respect to $x \in H$ and $u \in Y$, and the mappings

$$G_1(x,u,\mu) : L_p(H) \times L_r(Y) \times \mathcal{M}(\Sigma) \to \mathcal{L}(L_p(H), L_q(V^*)),$$
$$G_2(x,u,\mu) : L_p(H) \times L_r(Y) \times \mathcal{M}(\Sigma) \to \mathcal{L}(L_r(Y), L_q(V^*))$$

are bounded and continuous, where

$$G_1(x,u,\mu)(t) \equiv g_x(t,x(t),u(t),\mu),$$
$$G_2(x,u,\mu)(t) \equiv g_u(t,x(t),u(t),\mu).$$

(L) $l : I \times H \times Y \to R \bigcup \{\infty\}$ is continuous and Fréchet-differentiable in both x and u on H and Y, respectively, so that $l_x \in L_q(I,H)$ and $l_u \in L_{r'}(I,Y^*)$ $(r' = r/r-1)$ in the neighborhood of the optimal control state pair (u_0, x_0), whenever such a pair exists.

Under the above assumptions, we consider the following controlled uncertain system:

$$\text{(10)} \qquad \begin{cases} \dot{x}(t) + A(t,x(t)) = g(t,x(t),u(t),\mu) \quad \text{for almost all } t \in I, \\ x(0) = x^0 \in H. \end{cases}$$

DEFINITION 3.1. *A function x is said to be a solution of the problem* (10) *corresponding to $u \in \mathcal{U}_{ad}$ if*

(1) $x \in W_{p,q}$,

(2) *for a given $\mu \in \mathcal{M}(\Sigma)$, x satisfies* (10).

Define the solution set

$$\mathcal{X}(u) \equiv \{x \mid x \text{ is a solution of (10) corresponding to } u\}.$$

Our problem, called (P), is to find $u_0 \in \mathcal{U}_{ad}$ such that

$$J_0(u_0) = \inf_{u \in \mathcal{U}_{ad}} J_0(u)$$

where

$$J_0(u) = \sup_{x \in \mathcal{X}(u)} \left\{ \int_I l(t,x,u)dt \right\}.$$

Occasionally, we use the notation $x(u,\mu)$ to denote the solution of (10) corresponding to $u \in \mathcal{U}_{ad}$ and $\mu \in \mathcal{M}(\Sigma)$. Then we can write

$$\mathcal{X}(u) = \bigcup_{\mu \in \mathcal{M}(\Sigma)} x(u,\mu).$$

Set

$$\mathcal{X} = \bigcup_{u \in \mathcal{U}_{ad}} \mathcal{X}(u).$$

Using the results of section 2, one can easily obtain the following theorem.

THEOREM 3.A. *Suppose that assumptions* (A1), (G), *and* (U) *hold. Then for every* $u \in \mathcal{U}_{ad}$ *and* $\mu \in \mathcal{M}(\Sigma)$, *equation* (10) *has a unique solution. Further, the set* $\mathcal{X}$ *is a bounded subset of* $W_{p,q}$.

For necessary conditions, we need some results on continuous dependence of solutions on controls u and parameters μ.

PROPOSITION 3.2. *Suppose that assumptions* (A1), (G), *and* (U) *hold. For any fixed* $\mu_0 \in \mathcal{M}(\Sigma)$, *let* $u, u_0 \in \mathcal{U}_{ad}$ *and* $u_\varepsilon = u_0 + \varepsilon(u - u_0)$ $(0 \le \varepsilon \le 1)$. *Then we have*

$$x_\varepsilon \equiv x(u_\varepsilon, \mu_0) \to x_0 \equiv x(u_0, \mu_0) \quad \text{in } C(I,H) \bigcap L_p(V),$$

as $\varepsilon \to 0$.

Proof. Define

$$G_1^\varepsilon = \int_0^1 G_1(x_0 + s(x_\varepsilon - x_0), u_\varepsilon, \mu_0) ds,$$
$$G_2^\varepsilon = \int_0^1 G_2(x_0, u_0 + s(u_\varepsilon - u_0), \mu_0) ds.$$

By assumptions (G) and (U), there exists a constant K such that

$$\|G_1^\varepsilon\|_{\mathcal{L}(L_p(H), L_q(V^*))} \le K, \quad \|G_2^\varepsilon\|_{\mathcal{L}(L_r(Y), L_q(V^*))} \le K.$$

Following similar steps as in the proof of Lemma 2.5, we have

$$\begin{aligned}
&\frac{1}{2}\|x_\varepsilon(t) - x_0(t)\|_H^2 + \langle\langle A(x_\varepsilon) - A(x_0), x_\varepsilon - x_0\rangle\rangle_t \\
&= \int_0^t \langle g(t, x_\varepsilon, u_\varepsilon, \mu_0) - g(t, x_0, u_0, \mu_0), x_\varepsilon - x_0\rangle dt \\
&= \int_0^t \langle g(t, x_\varepsilon, u_\varepsilon, \mu_0) - g(t, x_0, u_\varepsilon, \mu_0), x_\varepsilon - x_0\rangle dt \\
&\quad + \int_0^t \langle g(t, x_0, u_\varepsilon, \mu_0) - g(t, x_0, u_0, \mu_0), x_\varepsilon - x_0\rangle dt \\
&= \langle\langle G_1^\varepsilon(x_\varepsilon - x_0), x_\varepsilon - x_0\rangle\rangle_t + \varepsilon\langle\langle G_2^\varepsilon(u_\varepsilon - u_0), x_\varepsilon - x_0\rangle\rangle_t.
\end{aligned}$$

Using assumptions (3) of (A1) and (3) of (G1), Lemma 2.4, and the Cauchy inequality, one can verify that there exist positive constants c'', K_1, and K_2 such that

$$\begin{aligned}
&\|x_\varepsilon(t) - x_0(t)\|_H^2 + c''\|x_\varepsilon - x_0\|_{L_p^t(V)}^p \\
&\le K_1\varepsilon + K_2 \int_0^t \|x_\varepsilon - x_0\|_H^2 dt.
\end{aligned}$$

The conclusion now follows from the Gronwall lemma.

Remark 3.3. Note that Proposition 3.2 remains valid if g is merely locally Lipschitz continuous with respect to x and u.

PROPOSITION 3.4. *Suppose that the assumptions* (A1), (G), *and* (U) *hold. For any fixed* $u_0 \in \mathcal{U}_{ad}$, *let* $\mu, \mu_0 \in \mathcal{M}(\Sigma)$ *and* $\mu^\varepsilon = \mu_0 + \varepsilon(\mu - \mu_0)$ $(0 \le \varepsilon \le 1)$. *Then*

$$x^\varepsilon \equiv x(u_0, \mu^\varepsilon) \to x_0 \equiv x(u_0, \mu_0) \quad \text{in } C(I, H) \bigcap L_p(V),$$

as $\varepsilon \to 0$.

Proof. The proof is similar to that of Proposition 3.2.

4. Necessary conditions of optimality. In this section, we present our main results, the necessary conditions of optimality for the problem (P) as stated in section 3. In what follows, we shall assume that an optimal control exists (see [4, 11, 12]).

DEFINITION 4.1. *The pair* $(u_0, \mu_0) \in \mathcal{U}_{ad} \times \mathcal{M}(\Sigma)$ *is said to be the optimal strategy pair for the problem* (P) *if*

$$\inf_{u \in \mathcal{U}_{ad}} \sup_{\mu \in \mathcal{M}(\Sigma)} J(u, \mu) = \sup_{\mu \in \mathcal{M}(\Sigma)} \inf_{u \in \mathcal{U}_{ad}} J(u, \mu) = J(u_0, \mu_0),$$

where $J(u, \mu) = \int_I l(t, x(u, \mu), u)dt$, *with* $x(u, \mu)$ *being the solution of equation* (10). *In other words, the optimal strategy pair is the saddle point of the loss (cost) functional* J.

Clearly, if (u_0, μ_0) is the saddle point, the following system of inequalities must hold:

$$J(u_0, \mu) \le J(u_0, \mu_0) \le J(u, \mu_0) \quad \forall\ u \in \mathcal{U}_{ad},\ \mu \in \mathcal{M}(\Sigma). \tag{11}$$

In the game-theoretic language, $J(u_0, \mu_0)$ is called the value of the game.

Letting x_0 denote the solution corresponding to the saddle point $\{u_0, \mu_0\}$, we have the optimal triple $\{u_0, x_0, \mu_0\}$. In this section we assume that $p = q = 2$.

In order to derive the necessary optimality conditions, we need some additional assumptions for the operator A.

(A) $A : I \times V \to V^*$.

(1) A satisfies conditions (A1).

(2) A is Fréchet-differentiable with respect to $x \in V$ and for each $\xi \in W_{2,2}$ the Nemytski operator A_x defined by $A_x(\xi)(t) \equiv A_x(t, \xi(t))$ belongs to $\mathcal{L}(L_2(V), L_2(V^*))$. Further, the map $\xi \longrightarrow A_x(\xi)$ is continuous and bounded on bounded subsets of $W_{2,2}$.

Remark 4.2. Under assumption (A), for any $x \in V$, $A_x(t, x)$ is strictly positive, i.e.,

$$\langle A_x(t, x)y, y\rangle \ge c\|y\|_V^2 \quad \forall\ t \in I, \quad \forall\ y \in V.$$

In fact, by the monotonicity of A, we have, for $s > 0$,

$$\langle A(t, x + sy) - A(t, x), sy\rangle \ge c\|sy\|_V^2;$$

hence

$$\left\langle \frac{A(t, x + sy) - A(t, x)}{s}, y \right\rangle \ge c\|y\|_V^2.$$

The assertion follows upon letting $s \to 0$.

For the optimal triple $\{u_0, x_0, \mu_0\}$, define

$$A^0 \equiv A_x(x_0), \quad G_1^0 \equiv G_x(x_0, u_0, \mu_0),$$
$$G_2^0 \equiv G_u(x_0, u_0, \mu_0).$$

LEMMA 4.3. *For each $f \in L_2(V^*)$, the linear evolution equation*

$$\begin{cases} \dot{y} + A^0 y = G_1^0 y + f, \\ y(0) = 0 \end{cases} \tag{12}$$

has a unique solution $y \in W_{2.2}$.

Proof. By virtue of assumptions (A) and (G), it is easy to verify that A^0 and G_1^0 satisfy (A1) and (G1), respectively. Then, by Theorem 2.A, equation (12) has a unique solution $y \in W_{2.2}$.

As usual, in the study of optimal control problems, we need an associated adjoint problem. Here we present an existence result for the following adjoint Cauchy problem.

THEOREM 4.A. *Define $l_x^0(t) \equiv l_x(t, x_0(t), u_0(t))$. The adjoint problem*

$$\begin{cases} -\dot{\psi} + (A^0)^* \psi = (G_1^0)^* \psi + l_x^0, \\ \psi(T) = 0 \end{cases} \tag{13}$$

has a unique solution $\psi \in W_{2.2}$.

Proof. By assumption (L), $l_x^0 \in L_2(H)$. Since $L_2(H) \hookrightarrow L_2(V^*)$, $l_x^0 \in L_2(V^*)$. Reversing the flow of time $t \to T - t$, the conclusion follows from Lemma 4.3.

THEOREM 4.B (necessary conditions). *Suppose that assumptions* (A), (G), (U), *and* (L) *hold. In order that $\{u_0, x_0, \mu_0\}$ be the optimal triple for problem* (P), *it is necessary that there exist a $\psi \in W_{2,2}$ such that the following equations and inequalities hold:*

(1) $\dot{x_0} + A(t, x_0) = g(t, x_0, u_0, \mu_0), \quad x_0(0) = x^0;$
(2) $-\dot{\psi} + (A^0)^* \psi = (G_1^0)^* \psi + l_x^0(t), \quad \psi(T) = 0;$
(3) $\int_I \langle (G_2^0)^* \psi + l_u^0, u - u_0 \rangle_{Y^*, Y} dt \geq 0 \quad \forall\ u \in \mathcal{U}_{ad},$
where $l_u^0(t) = l_u(t, x_0(t), u_0(t))$;
(4) $\int_I \langle \psi, G(x_0, u_0, \mu) \rangle dt \leq \int_I \langle \psi, G(x_0, u_0, \mu_0) \rangle dt \ \forall\ \mu \in \mathcal{M}(\Sigma)$.

Proof. Let (u_0, μ_0) be a saddle point for the problem (P) and $x_0 = x(u_0, \mu_0)$ be the corresponding optimal trajectory. For any $u \in \mathcal{U}_{ad}$, by the convexity of $\mathcal{U}_{ad}$, $u_\varepsilon \equiv u_0 + \varepsilon(u - u_0) \in \mathcal{U}_{ad}$ for $0 \leq \varepsilon \leq 1$. By Theorem 3.A, the state equation (10) has a unique strong solution $x_\varepsilon = x(u_\varepsilon, \mu_0)$ corresponding to the control u_ε and parameter μ_0.

Using the second part of the inequality (11), we have

$$\int_I l(t, x_\varepsilon(t), u_\varepsilon(t)) dt - \int_I l(t, x_0(t), u_0(t)) dt \geq 0 \quad \forall\ u \in \mathcal{U}_{ad}. \tag{14}$$

Define $y_\varepsilon = (x_\varepsilon - x_0)/\varepsilon$, $A_x^\varepsilon = \int_0^1 A_x(x_0 + s(x_\varepsilon - x_0)) ds$, $G_1^\varepsilon = \int_0^1 G_1(x_0 + s(x_\varepsilon - x_0), u_\varepsilon, \mu_0) ds$, $G_2^\varepsilon = \int_0^1 G_2(x_0, u_0 + s(u_\varepsilon - u_0), \mu_0) ds$. Note that y_ε satisfies the following equation:

$$\begin{aligned} \dot{y}_\varepsilon + \frac{A(x_\varepsilon) - A(x_0)}{\varepsilon} &= [G(x_\varepsilon, u_\varepsilon, \mu_0) - G(x_0, u_0, \mu_0)]/\varepsilon \\ &= [G(x_\varepsilon, u_\varepsilon, \mu_0) - G(x_0, u_\varepsilon, \mu_0)]/\varepsilon \\ &\quad + [G(x_0, u_\varepsilon, \mu_0) - G(x_0, u_0, \mu_0)]/\varepsilon, \end{aligned}$$

which can be written as

$$\begin{cases} \dot{y}_\varepsilon + A_x^\varepsilon y_\varepsilon = G_1^\varepsilon y_\varepsilon + G_2^\varepsilon (u - u_0), \\ y_\varepsilon(0) = 0. \end{cases} \tag{15}$$

Following similar steps as in the proof of Lemma 2.4, using assumptions (A), (G), and (U), we can obtain the following a priori estimate: there exists a constant d such that for all $0 \le \varepsilon \le 1$,

$$\|y_\varepsilon\|_{C(I,H)} \le d, \quad \|y_\varepsilon\|_{L_2(V)} \le d.$$

Consider the following equation:

$$\begin{cases} \dot{y} + A^0 y = G_1^0 y + G_2^0 (u - u_0), \\ y(0) = 0. \end{cases} \tag{16}$$

Since $G_2^0(u - u_0) \in L_2(V^*)$, by Lemma 4.3, equation (16) has a unique solution $y_0 \in W_{2.2}$.

To show that $y_\varepsilon \longrightarrow y_0$, use integration by parts to obtain

$$\begin{aligned} &\frac{1}{2}\|y_\varepsilon(t) - y_0(t)\|_H^2 + \langle\langle A_x^\varepsilon y_\varepsilon - A^0 y_0, y_\varepsilon - y_0\rangle\rangle_t \\ &= \frac{1}{2}\|y_\varepsilon(t) - y_0(t)\|_H^2 + \langle\langle A_x^\varepsilon (y_\varepsilon - y_0), y_\varepsilon - y_0\rangle\rangle_t \\ &\quad + \langle\langle (A_x^\varepsilon - A^0) y_0, y_\varepsilon - y_0\rangle\rangle_t \\ &\le \langle\langle G_1^\varepsilon (y_\varepsilon - y_0), y_\varepsilon - y_0\rangle\rangle_t + \langle\langle (G_1^\varepsilon - G_1^0) y_0, y_\varepsilon - y_0\rangle\rangle_t \\ &\quad + \langle\langle (G_2^\varepsilon - G_2^0)(u - u_0), y_\varepsilon - y_0\rangle\rangle_t. \end{aligned}$$

From assumptions (A) and (G), Remark 4.2, and the Cauchy inequality, it follows that there exist constants $c^*, \eta_0, \eta_1, \eta_2$, and η_3 such that

$$\begin{aligned} &\|y_\varepsilon(t) - y_0(t)\|_H^2 + c^* \|y_\varepsilon - y_0\|_{L_2^t(V)}^2 \\ &\le \eta_0 \int_0^t \|y_\varepsilon(t) - y_0(t)\|_H^2 dt + \eta_1 \|(A_x^\varepsilon - A^0) y_0\|_{L_2^t(v^*)}^2 \\ &\quad + \eta_2 \|(G_1^\varepsilon - G_1^0) y_0\|_{L_2^t(v^*)}^2 + \eta_3 \|(G_2^\varepsilon - G_2^0)(u - u_0)\|_{L_2^t(v^*).}^2 \end{aligned} \tag{17}$$

Since $u_\varepsilon \to u_0$ in $L_r(Y)$ and $x_\varepsilon \to x_0$ in $C(I,H) \bigcap L_2(V)$ (see Proposition 3.2), we have

$$\begin{aligned} &\|(A_x^\varepsilon - A^0) y_0\|_{L_2^t(v^*)}^2 \to 0, \\ &\|(G_1^\varepsilon - G_1^0) y_0\|_{L_2^t(v^*)}^2 \to 0, \\ &\|(G_2^\varepsilon - G_2^0)(u - u_0)\|_{L_2^t(v^*)}^2 \to 0. \end{aligned}$$

By virtue of the Gronwall lemma, it follows from (17) that

$$y_\varepsilon \to y_0 \quad \text{in } C(I,H) \bigcap L_2(V),$$

as $\varepsilon \to 0$, where y_0 is the Gâteaux differential of x with respect to u in the direction $u - u_0$.

By the use of hypothesis (L), after some elementary computations, one obtains from inequality (14) that

$$\int_I [(l_x^0, y_0) + \langle l_u^0, u - u_0\rangle_{Y^*,Y}]dt \geq 0. \tag{18}$$

By Theorem 4.A, the adjoint equation (13) has a unique solution $\psi \in W_{2.2}$, and $y_0 \in W_{2,2}$ is the solution of (16). Hence

$$\begin{aligned}\langle\langle l_x^0, y_0\rangle\rangle &= \langle\langle -\dot{\psi} + (A^0)^*\psi - (G_1^0)^*\psi, y_0\rangle\rangle \\ &= \langle\langle \psi, \dot{y_0} + A^0 y_0 - G_1^0 y_0\rangle\rangle \\ &= \langle\langle \psi, G_2^0(u - u_0)\rangle\rangle \\ &= \langle\langle (G_2^0)^*\psi, u - u_0\rangle\rangle.\end{aligned}$$

It follows from (14) and the preceding arguments that

$$\int_I \langle (G_2^0)^*\psi + l_u^0, u - u_0\rangle dt \geq 0 \quad \forall\ u \in \mathcal{U}_{ad}.$$

This proves the inequality (3) as stated in Theorem 4.B.

Since (u_0, μ_0) is an optimal solution of problem (P), it follows from the first part of inequality (11) that

$$\int_I l(t, x(u_0, \mu), u_0)\ dt - \int_I l(t, x_0, u_0)\ dt \leq 0 \quad \forall\ \mu \in \mathcal{M}(\Sigma). \tag{19}$$

For any $\mu \in \mathcal{M}$, $\mu^\varepsilon = \mu_0 + \varepsilon(\mu - \mu_0) \in \mathcal{M}(\Sigma)$ $(0 \leq \varepsilon \leq 1)$, $x^\varepsilon \equiv x(u_0, \mu^\varepsilon)$ is the unique solution of the system (10) corresponding to the control u_0 and parameter μ^ε. Clearly, we have

$$\int_I l(t, x^\varepsilon, u_0)\ dt - \int_I l(t, x_0, u_0)\ dt \leq 0. \tag{20}$$

Define $\omega^\varepsilon = (x^\varepsilon - x_0)/\varepsilon$, $\tilde{A}_x^\varepsilon = \int_0^1 A_x(x_0 + s(x^\varepsilon - x_0))ds$, $\tilde{G}_1^\varepsilon = \int_0^1 G_1(x_0 + s(x^\varepsilon - x_0), u_0, \mu^\varepsilon)ds$. Clearly, ω^ε satisfies the following equation:

$$\begin{aligned}\dot{\omega}^\varepsilon + \tilde{A}_x^\varepsilon\omega^\varepsilon &= [G(x^\varepsilon, u_0, \mu^\varepsilon) - G(x_0, u_0, \mu_0)]/\varepsilon \\ &= \tilde{G}_1^\varepsilon\omega^\varepsilon + [G(x_0, u_0, \mu) - G(x_0, u_0, \mu_0)].\end{aligned}$$

Following similar steps, as in the case of controls, one can verify that the Gâteaux differential of x with respect to μ in the direction $\mu - \mu_0$ is the solution of the following equation:

$$\begin{cases} \dot{\omega} + A^0\omega = G_1^0\omega + [G(x_0, u_0, \mu) - G(x_0, u_0, \mu_0)], \\ \omega(0) = 0. \end{cases} \tag{21}$$

Again, following similar arguments as in the case of control, proving inequality (3), one can verify that

$$\int_I \langle \psi, G(x_0, u_0, \mu)\rangle dt \leq \int_I \langle \psi, G(x_0, u_0, \mu_0)\rangle dt \quad \forall\ \mu \in \mathcal{M}(\Sigma),$$

where ψ satisfies the adjoint equation (13). This completes the proof of Theorem 4.B.

5. An example. Let Ω be an open connected bounded region in R^n with smooth boundary $\partial\Omega$, $Q_T = I \times \Omega$, $(I = (0,T),\ 0 < T < \infty)$. Let $\alpha = (\alpha_1, \ldots, \alpha_n)$ be a multi-index with $\{\alpha_i\}$ nonnegative integers and $\mid \alpha \mid = \sum_1^n \alpha_i$. Suppose $p \geq 2$ and $q = p/(p-1)$, $W^{m,p}(\Omega)$, $mp > n$, denotes the standard Sobolev space with the usual norm:

$$\|\varphi\|_{W^{m,p}} = \left(\sum_{|\alpha| \leq m} \|D^\alpha \varphi\|^p_{L^p(\Omega)} \right)^{1/p}, \qquad m = 0, 1, \ldots.$$

Let $W_o^{m,p} \equiv \{\varphi \in W^{m,p} \mid D^\beta \varphi \mid_{\partial\Omega} = 0, \mid \beta \mid \leq m-1\}$. It is well known that $C_0^\infty \subset W_0^{m,p} \subset L^2 \subset W^{-m,p}$ and the embedding $W_0^{m,p} \hookrightarrow L^2$ is compact.

We consider a minimax problem of a typical uncertain system governed by the following controlled quasi-linear parabolic partial differential equation of order $2m$ with uncertain coefficients as described below:

$$\text{(22)} \qquad \begin{cases} \partial/\partial t \varphi(\xi,t) + \sum_{|\alpha| \leq m} (-1)^{|\alpha|} D^\alpha A_\alpha(\xi, t, \eta(\varphi)(\xi,t)) \\ \qquad = \sum_{|\alpha| \leq m} D^\alpha [\sigma_\alpha(\xi) f_\alpha(\xi, t, \varphi, u)], \\ D^\beta \varphi(\xi,t) = 0 \quad \text{on} \quad \partial\Omega \times [0,T] \qquad \forall \beta : \mid \beta \mid \leq m-1, \\ \varphi(\xi, 0) = \varphi^0(\xi) \in L^2(\Omega), \end{cases}$$

where $\eta(\varphi) \equiv \{(D^\alpha \varphi),\ \mid \alpha \mid \leq m\}$. Let $M \equiv card\{\mid \alpha \mid \leq m\}$.

The coefficients $\{\sigma_\alpha, \mid \alpha \mid \leq m\}$ are the uncertain parameters. Suppose $\{\lambda_{i,\alpha}\}(i = 1, 2; \mid \alpha \mid \leq m)$ are given constants and $\sigma_\alpha : \Omega \to R$ are bounded measurable functions satisfying

$$\lambda_{1,\alpha} \leq \sigma_\alpha(\xi) \leq \lambda_{2,\alpha} \qquad \forall\ \xi \in \Omega, \mid \alpha \mid \leq m.$$

We denote this class of functions by Σ. It is assumed that these functions are not known to the control system designer, except their range Σ. In other words, the system has unknown (uncertain) coefficients. For example, in the mathematical models for pollutant transport equations described by reaction-diffusion-transport equations (see the comments following this example), the three-dimensional velocity distribution of the water (or air) is impossible to determine at every point of the medium. However, it is possible to set lower and upper limits to these quantities.

Let $\beta_1(\xi,t), \beta_2(\xi,t) : Q_T \to R$ be bounded continuous functions so that

$$\beta_1(\xi,t) \leq \beta_2(\xi,t) \quad \forall\ (\xi,t) \in Q_T.$$

Set

$$U(t) \equiv \{w \in L_\infty(\Omega, R^{M_1}) \equiv Z : \beta_1(\xi,t) \leq w_i(\xi) \leq \beta_2(\xi,t), \quad \xi \in \Omega, 1 \leq i \leq M_1\}.$$

The set of admissible controls $\mathcal{U}_{ad}$ is chosen as

$$\mathcal{U}_{ad} \equiv \{u \in L_\infty(I, Z) : u(t) \in U(t) \text{ a.e.}\}.$$

The cost function for the control problem is given by

$$J_0(u) \equiv \sup_{\sigma \in \Sigma} \int_I \int_\Omega l_0(\xi, t, \eta(\varphi(\sigma)(\xi,t)), u(\xi,t)) d\xi dt.$$

Our problem (P*) is to find u_0 such that

$$J_0(u_0) = \inf\{J_0(u), u \in \mathcal{U}_{ad}\}.$$

For $\varphi, \psi \in W_0^{m,p},\ t \in I$, we set

$$a(t,\varphi,\psi) \equiv \int_\Omega \sum_{|\alpha|\le m} A_\alpha(\xi,t,\eta(\varphi(\xi)))D^\alpha\psi d\xi$$

and assume that the function $A_\alpha(|\ \alpha\ | \le m) : Q_T \times R^M \to R$ satisfies the following properties.

(A1) $(\xi,t) \to A_\alpha(\xi,t,\eta)$ is measurable on Q_T for $\eta \in R^M, \eta \to A_\alpha(\xi,t,\eta)$ is continuously differentiable on R^M for almost all $(\xi,t) \in Q_T$.

(A2) There exist positive constants c, c_1, c_2, c_3, c_4 for $\eta = (\eta^\alpha) \in R^M$ such that

$$\sum_{|\alpha|\le m} (A_\alpha(\xi,t,\eta) - A_\alpha(\xi,t,\tilde\eta))(\eta_\alpha - \tilde\eta_\alpha) \ge c \sum_{|\alpha|\le m} |\ \eta_\alpha - \tilde\eta_\alpha\ |^p;$$

$$\sum_{|\alpha|\le m} A_\alpha(\xi,t,\eta)\eta_\alpha \ge c_1 \sum_{|\alpha|\le m} |\ \eta_\alpha\ |^p - c_2;$$

$$|\ A_\alpha(\xi,t,\eta)\ | \le c_4 + c_3 \sum_{|\alpha|\le m} |\ \eta_\alpha\ |^{p-1}\ .$$

Under the above assumptions one can verify that for each $\varphi \in W_0^{m,p}$ and $t \in I$, $\psi \to a(t,\varphi,\psi)$ is a continuous linear form on $W_0^{m,p}$ and hence there exists exactly one operator $A : I \times W_0^{m,p} \to W^{-m,q}$ such that

$$\langle A(t,\varphi),\psi\rangle_{W^{-m,q},W_0^{m,p}} = a(t,\varphi,\psi).$$

Identifying $V \equiv W_0^{m,p}, H \equiv L_2(\Omega), V^* \equiv W^{-m,q}$, it follows from the above assumptions that the operator A as defined above satisfies the assumption (A) of section 4.

Clearly, the set Σ is a closed bounded convex subset of $L_\infty(\Omega, R^M)$, and hence w^* is compact. Thus, with respect to the w^* topology, it is a compact Hausdorff space. Since $L_1(\Omega, R^M)$ is separable, the set Σ is metrizable with respect to which it is a complete separable metric space and hence a compact Polish space. Therefore $\mathcal{M}(\Sigma)$ is a compact Polish space. For our control problem we choose this as our parameter set.

Assume the function $f_\alpha : Q_T \times R \times R^{M_1} \to R \quad (|\ \alpha\ | \le m)$ satisfies the following properties:

(F1) $(\xi,t) \to f_\alpha(\xi,t,\gamma,v)$ is measurable on Q_T for all $(\gamma,v) \in R \times R^{M_1}$; $(\gamma,v) \to f_\alpha(\xi,t,\gamma,v)$ is continuously differentiable on $R \times R^{M_1}$ for almost all $(\xi,t) \in Q_T$.

(F2) There exist $b_1 \in L_q(Q_T)$ and $b_2 \ge 0$ such that

$$|\ f_\alpha(\xi,t,\gamma,v)\ | \le b_1(\xi,t) + b_2\ |\ \gamma\ |^{2/q}$$

for all $(\xi,t) \in Q_T,\ v \in U(t)$.

For $\varphi \in L^2(\Omega),\ v \in U(t),\ t \in I$, set

$$b^{v,\sigma}(t,\varphi,\psi) \equiv \int_\Omega \sum_{|\alpha|\le m} \sigma_\alpha(\xi) f_\alpha(\xi,t,\varphi,v) D^\alpha\psi d\xi.$$

Under the above assumptions it is not difficult to verify that $\psi \to b^{v,\sigma}(t,\varphi,\psi)$ is a continuous linear form on $W_0^{m,p}$, and hence, for each $\sigma \in \Sigma$, there exists an operator $F_\sigma : I \times L^2(\Omega) \times U(t) \to W^{-m,q}$ such that

$$b^{v,\sigma}(t,\varphi,\psi) = \langle F_\sigma(t,\varphi,v),\psi\rangle_{V^*,V},$$

and F_σ satisfies assumption (G) of section 3.

Using the operators A and F_σ as defined above, equation (22) can be written as the abstract evolution equation

$$\begin{cases} \dot{\varphi} + A(t,\varphi) = F_\sigma(t,\varphi,u), \\ \varphi(0) = \varphi^0. \end{cases} \tag{23}$$

For each $u \in \mathcal{U}_{ad}$, define the set

$$\mathcal{X}(u) \equiv \{\varphi \in C(I,H) \mid \varphi \text{ is a generalized solution of (23)}\}$$

(see section 30.4 of [3]). Theorem 3.A shows that $\mathcal{X}(u) \neq \emptyset$.

For $\phi \in W_0^{m,p},\ v \in U(t)$, define

$$l(t,\phi,v) = \int_\Omega l_0(\xi,t,\eta(\phi(\xi)),v)d\xi.$$

Suppose that the function $l_0 : Q_T \times R^M, \times R^{M_1} \to R \cup \{\infty\}$ is continuous and continuously differentiable with respect to the last $M + M^1$ variables and, further, the inequality

$$\mid l_0(\xi,t,\eta,v) \mid \leq k(1+ \mid \eta \mid^2_{R^M}), \quad v \in U(t), \quad (\xi,t) \in Q_T,$$

holds. Under the above hypothesis, it is easy to see that the function $l(t,\varphi,v) : I \times W_0^{m,p} \times U(t) \to R$ satisfies assumption (L) of section 4. Thus the problem (P*) can be restated as follows.

Minimize $J_0(u), u \in \mathcal{U}_{ad}$, subject to the differential equation (23), where

$$J_0(u) = \sup\left\{\int_I l(t,\varphi,u)dt, \varphi \in \mathcal{X}(u)\right\}.$$

Hence our result of Theorem 4.B can be used to solve this problem.

Further remarks. For further illustration we present here a more practical example dealing with a reaction-diffusion-transport system, as remarked earlier. The system is governed by

$$\begin{cases} (\partial/\partial t)C(\xi,t) - D\triangle C + \nabla C\sigma = f(\sigma,C,u) \quad \text{on} \quad \Omega \times (0,T], \\ C(\xi,t) = 0 \quad \text{on} \quad \partial\Omega \times [0,T], \\ C(\xi,0) = C^0(\xi), \quad \xi \in \Omega. \end{cases} \tag{24}$$

Here, Ω is an open bounded subset of R^3 with smooth boundary representing, for example, the aquatic body (lakes, rivers); C represents the concentration of n different biochemical pollutants in the aquatic system. $D \equiv \text{diag}\{d_1, d_2, \ldots, d_n\}$ is the diffusion matrix, $\sigma = \sigma(\xi)$ represents the three-dimensional velocity field of the water body, and $u = u(t,\xi)$ represents the control vector. The second term in the equation accounts for pollutant movement by diffusion, the third represents transport (of pollutants) by the flow field, and the last term on the right-hand side of the equation represents interaction between the pollutants and the control agents. Note that ∇C is an $n \times 3$ matrix. The controls may consist of (a) biological agents such as micro-organisms capable of producing biodegradation of pollutants, (b) chemicals, or (c) simple physical means of extraction. For the mathematical setting, we take $p = q = 2$, $H = L_2(\Omega, R^n)$, $V = H_0^1(\Omega, R^n)$, with V^* being the dual of V. Assuming that D is positive definite, we take, for the operator A, the $L_2(\Omega, R^n)$ realization of

the Laplacian $-D\triangle$ with Dirichlet boundary condition. It is obvious in this case that the operator $A \in \mathcal{L}(V, V^*)$, and it is coercive and hence monotone. For the nonlinear operator F_σ, we take

$$F_\sigma(C, u) \equiv -\nabla C\sigma + f(\sigma, C, u).$$

Here we assume that the velocity field σ is unknown and that

$$\sigma \in \Sigma \equiv \{\sigma \in L_\infty(\Omega, R^3) : \ \mathrm{div}\sigma = 0, \lambda_1 \le \sigma_i(\xi) \le \lambda_2, \xi \in \Omega, i = 1, 2, 3\},$$

where the upper and lower bounds are known. For the admissible controls, define

$$U \equiv \{v \in L_\infty(\Omega, R^m) \equiv Z : \beta_i(\xi) \le v_i(\xi) \le \gamma_i(\xi), i = 1, 2, \ldots, m\},$$

where $\beta_i, \gamma_i \in L_\infty(\Omega)$ are a given set of functions satisfying $\beta_i(\xi) \le \gamma_i(\xi), \xi \in \Omega$. The admissible controls are given by

$$\mathcal{U}_{ad} \equiv \{u \in L_\infty([0, T], Z) : u(t) \in U \text{ a.e.}\}.$$

We assume that f is locally Lipschitz in $y \in R^n$ and continuous in all the variables satisfying

$$|f(\sigma, y, v)|_{R^n} \le b_1(\xi) + b_2|y|_{R^n}, \quad \text{for all } \sigma \in \Sigma, v \in U,$$

where $b_1 \in L_2^+(\Omega), b_2 > 0$. For Lotka–Volterra-type logistical interactions, the local Lipschitz property holds. Under the above assumptions it is easy to verify that for each $\sigma \in \Sigma$,

$$F_\sigma : H \times U \mapsto V^* \ \text{ and also } F_\sigma : V \times U \mapsto H.$$

Hence equation (24) can be written in abstract form as

$$\left\{ \begin{array}{l} \dot{C} + AC = F_\sigma(C, u), \\ C(0) = C^0. \end{array} \right. \tag{25}$$

For the cost integrand l, one may choose the quadratic function

$$l(C, u) \equiv \int_\Omega \{(Q(\xi, t)(C - C^d), C - C^d)_{R^n} + (R(\xi, t)u, u)_{R^m}\} d\xi,$$

where Q and R are positive semidefinite matrix-valued functions on $\Omega \times [0, T]$, possibly with essentially bounded measurable entries and C^d denotes the acceptable pollution level. Note that variables Q and R signify regional and temporal variation in the cost of pollutant extraction. Since all the assumptions of our abstract result and those of the example are satisfied in this particular case, our result applies.

Remark. The optimal strategy pair can be computed by use of the computational algorithm for the min-max problem given in [6]. This algorithm is based on similar optimality conditions as given in our Theorem 4.B. For details, the reader may see [6].

REFERENCES

[1] N. U. Ahmed and K. L. Teo, *Optimal Control of Distributed Parameter Systems*, North-Holland, Amsterdam, 1981.

[2] N. U. Ahmed, *Optimization and Identification of Systems Governed by Evolution Equations on Banach Space*, Pitman Res. Notes Math. Ser. 184, Longman, Harlow, UK, 1988.
[3] E. Zeidler, *Nonlinear Functional Analysis and Its Applications*, II, Springer-Verlag, New York, 1990.
[4] N. U. Ahmed and X. Xiang, *Optimal control of infinite dimensional uncertain systems*, J. Optim. Theory Appl., 80 (1994), pp. 261–272.
[5] J. Warga, *Optimal Control of Differential and Functional Equations*, Academic Press, New York, 1972.
[6] N. U. Ahmed and X. Xiang, *Necessary conditions of optimality for infinite dimensional uncertain systems*, Math. Problems Engrg., 1–2 (1995), pp. 179–191.
[7] N. U. Ahmed and X. Xiang, *Necessary conditions of optimality for a class of semilinear uncertain systems with unbounded monotone operator*, WSSIAA, 4 (1995), pp. 27–43.
[8] N. U. Ahmed, *Optimal control of infinite-dimensional systems governed by functional differential inclusions*, Disc. Math. Diff. Inclusions, 15 (1995), pp. 75–94.
[9] N. S. Papageorgiou, *Nonlinear Volterra integrodifferential evolution inclusions and optimal control*, Kodai Math. J., 14 (1991), pp. 254–280.
[10] N. S. Papageorgiou, *Identification of parameters in systems governed by nonlinear evolution equations*, Publ. Math. Debrecen, 46 (1995), pp. 215–237.
[11] N. S. Papageorgiou, *Optimization of parametric controlled nonlinear evolution equations*, Yokohoma Math. J., 42 (1994), pp. 107–120.
[12] N. S. Papageorgiou, *Minimax control of nonlinear evolution equations*, Comm. Math. Univ. Carolinae, 36 (1995), pp. 39–56.

SIAM J. CONTROL OPTIM.
Vol. 35, No. 5, pp. 1773–1789, September 1997

018

EXACT CONTROLLABILITY OF THE DAMPED WAVE EQUATION*

MARIANNA A. SHUBOV†, CLYDE F. MARTIN†, JERALD P. DAUER‡, AND BORIS P. BELINSKIY‡

Abstract. We study the controllability problem for a distributed parameter system governed by the damped wave equation

$$u_{tt} - \frac{1}{\rho(x)}\frac{d}{dx}\left(p(x)\frac{du}{dx}\right) + 2d(x)u_t + q(x)u = g(x)f(t),$$

where $x \in (0, a)$, with the boundary conditions

$$u(0) = 0, \quad (u_x + hu_t)(a) = 0, \quad h \in \mathbb{C} \cup \{\infty\}.$$

This equation describes the forced motion of a nonhomogeneous string subject to a viscous damping with the damping coefficient $d(x)$ and with damping (if Re $h > 0$) or energy production (if Re $h < 0$) at one end. (All results extend to the case when a similar condition is imposed at the other end as well.) The function $f(t)$ is considered as a control. Generalizing well-known results by D. Russell concerning the string with $d(x) = 0$, we give necessary and sufficient conditions for exact unique controllability and approximate controllability of the system. Our proofs are based on recent results by M. Shubov concerning the spectral analysis of a class of nonselfadjoint operators and operator pencils generated by the above equation.

Key words. nonselfadjoint operators, exact and approximate controllability, eigenvectors and root vectors, Riesz basis, nonharmonic exponential basis, hyperbolic equations, damping, distributed parameter control, stability

AMS subject classifications. 35C20, 35J10, 35L20, 93B05, 93B60, 93C20

PII. S0363012996291616

1. Introduction.

We consider the one-dimensional wave equation which governs the vibrations of a string with spatially nonhomogeneous positive damping, modulus of elasticity, and density coefficients. The equation is defined on a finite interval with linear first-order nonselfadjoint boundary conditions containing damping terms. We use the spectral decomposition method to construct a control law that brings the system to rest in a specified finite time. This control law is implemented through the forcing term of the form $g(x)f(t)$, where g is the force distribution function and f is the control. We give necessary and sufficient conditions on the initial data and the force distribution function under which the construction of the desired control is possible, and then we give an explicit construction of the control. If these conditions are not satisfied but are replaced by weaker conditions, the system is, nevertheless, approximately controllable.

Our results can be considered as generalization of classic results by D. Russell [1]–[4], who solved the above controllability problem for the one-dimensional wave equation without damping term. The main difference between the damped wave equation and the undamped one [1]–[6] is the fact that our equation generates a nonselfadjoint

*Received by the editors March 4, 1996; accepted for publication (in revised form) July 15, 1996. http://www.siam.org/journals/sicon/35-5/29161.html

†Department of Mathematics, Texas Tech University, Lubbock, TX 79409-1042 (mshubov@math.ttu.edu, gqcfm@ttacs1.ttu.edu). The research of the first author was supported in part by ARP-95 of Texas grant 0206-44-3817. The research of the second author was supported in part by NASA grants NAG2–902 and NAG2–899.

‡Department of Mathematics, University of Tennessee at Chattanooga, 615 McCallie Avenue, Chattanooga, TN 37403-2598 (jdauer@utcvm.utc.edu, bbelinsk@utcvm.utc.edu).

operator, for which the spectral theory has only recently been developed. Our solution is based on recent results obtained by M. Shubov [7]–[9], who showed that the system of eigenvectors and associated vectors of the aforementioned nonselfadjoint operator forms a Riesz basis in the energy space. The case of a string with constant density, nonconstant damping, and the Dirichlet boundary conditions was recently studied in [23], [24]. For the case of a nonconstant density, zero damping coefficient and damping in the boundary conditions see [25]–[28], [6]. However, the combination of nonconstant damping and density with the damping in the boundary conditions makes the problem significantly more complicated even if the coefficients are smooth.

This paper is organized as follows. In section 2 we provide all necessary definitions, collect the necessary information from [7]–[9], and formulate our main results. In section 3 we reformulate the original problem in operator format and reduce it to a nonclassical moment problem. The solution of the moment problem is based on the fact that the system of nonharmonic exponentials $\{e^{i\lambda_n t}\}$, where λ_n are the complex eigenfrequencies of the string, forms a Riesz basis in the space $L^2(0,T)$, where T is the control time. This fact follows from the asymptotic formulas for λ_n obtained in [7] in the general case (even if the density has zeros or singularities) and in [6], [23]–[27] in the aforementioned particular cases.

In section 3 we, also, define the strong solution of the original initial-boundary problem following the spirit of works by J.-L. Lions [10], [11] and mention that our solution obtained by the method of spectral decomposition satisfies all the requirements of this definition if the initial data are sufficiently regular. In section 4 we use the variational approach to define the weak solution. However, our definition is nonstandard. It is adapted to our operator formulation of the control problem, and we suggest that it might be of interest in itself. We prove the uniqueness of our weak solution using the spectral properties of the dynamics generator and show that the solution of the control problem satisfies this definition for a wider class of initial data.

Using the spectral decomposition method, we give an explicit construction of the control function in terms of the eigenvalues and the eigenvectors of the nonselfadjoint operator, which is the dynamic generator of the damped string. We also give the formula for the minimal time which is required to damp the motion. This time turns out to be equal to twice the time it takes for a wave to travel from one end of the string to the other. This means that the control time for the damped string is the same as in the case of undamped string [2].

We mention in conclusion that the controllability problem for hyperbolic equations is treated in an extensive literature (see, e.g., the classic book [29] and references therein). The general controllability results for linear hyperbolic equations are given in [30]. However, as far as we know, the controllability problem for the damped wave equation has never been treated by the spectral decomposition method. This method allows one to consider the control of the form $g(x)f(t)$ and to provide an explicit construction of the control function.

2. Auxiliary propositions and statement of main results. Our main object of interest is the following wave equation:

$$u_{tt} - \frac{1}{\rho(x)}\frac{d}{dx}\left(p(x)\frac{du}{dx}\right) + 2d(x)u_t + q(x)u = g(x)f(t), \quad x \in [0,a,], \quad t \geq 0. \tag{2.1}$$

Equation (2.1) describes forced motion of a string with density $\rho(x)$ and modulus of elasticity $p(x)$, subject to an external harmonic force with the rigidity coefficient $q(x)$ and in the presence of viscous damping $2d(x)$. Precise conditions for these functions

will be formulated later. The force distribution function is denoted by $g \in L^2(0,a)$, and $f(t)$ is called an admissible control function on the interval $[0,T]$ if it is an element of $L^2(0,T)$.

We will assume that $u(x,t)$ satisfies the boundary conditions

$$u(0,t) = 0, \quad (u_x + hu_t)(a) = 0, \quad h \in \mathbb{C} \cup \{\infty\}, \tag{2.2}$$

and the initial conditions

$$u(x,0) = u_0^0(x), \quad u_t(x,0) = u_1^0(x). \tag{2.3}$$

Let us mention some particular cases of problem (2.1)–(2.3). If $h = 0$ (the Neumann boundary condition at $x = a$), the problem describes the vibrations of a finite string with the left end fixed and the right end free. If $h = \infty$ (the Dirichlet boundary condition $u(a) = 0$), we deal with the vibrations of a finite string with both ends fixed. If $h = 1$ (the so-called Sommerfeld radiation boundary condition), the problem describes the resonance phenomena in the scattering of acoustical waves on the semi-infinite string. (For more details on the resonance phenomenon see [6], [8]). We consider the Dirichlet condition at $x = 0$ only for the simplification of the exposition. All our proofs can be extended to the case $(u_x + ku_t)(0) = 0$, $k \in \mathbb{C} \cup \{\infty\}$. We consider the following problem.

Let initial conditions (2.3) *and* $T > 0$ *be given. Does there exist an admissible control* f *on* $[0,T]$ *such that the solution of problem* (2.1)–(2.3) *also satisfies an additional condition at* $t = T$*:*

$$u(x,T) = 0, \quad u_t(x,T) = 0, \ x \in [0,a]? \tag{2.4}$$

Due to the uniqueness theorem, (2.4) *means that* $u(x,t) = 0$ *for* $t \geqslant T$.

For a satisfactory solution of the control problem it is necessary that the set of aforementioned initial conditions be sufficiently large. Conditions (2.13), (2.14) of our main Theorem 2.4 below define a dense linear subspace of the energy space $\mathcal{H}$ which consists of initial data that can be steered to zero in time $T = 2\mathcal{M}$ (see (2.10) below).

To describe the energy space we first formulate the precise conditions on the coefficients of (2.1):

$$\rho(x), p(x), d(x) > 0, \quad q(x) \geq 0, \quad \rho, p \in H^2(0,a), \quad d \in H^1(0,a), \quad q \in L^\infty(0,a). \tag{2.5}$$

Remark 2.1. (a) As was already mentioned in the introduction, in this paper we consider the case of a nonconstant positive viscous damping. It is certainly possible that $d(x) = \text{const} > 0$, but it is important that $d(x)$ does not vanish identically. Assumptions (2.5) about the coefficients of our problem are sufficient for Theorems 2.1–2.3 below to be correct. These theorems are also correct in the case $d(x) = 0$ for all $x \in [0,a]$ (see the aforementioned works [25]–[28], [6]) or if $d(x)$ vanishes on a set of a positive Lebesgue measure. However, in these cases certain additional assumptions about the parameters of the problem should be satisfied. See Remark 2.2 below for a more detailed commentary.

(b) In fact, the assumption $d(x) > 0$ is also unnecessary. It is introduced in order to treat $d(x)u_t$ as a dissipative term. $d(x)$ may take negative values on $[0,a]$. In this case we should require that $d(x) = 0$ only on a set of Lebesgue measure zero and a certain integral involving d, p, and ρ does not vanish (see [7], [8]). (If $d(x) = 0$ on a set of a positive measure, then the aforementioned conditions from Remark 2.2 must be satisfied.)

(c) Note that due to (2.5) and the embedding $H^1(0,a) \subset C[0,a]$, the function $d(x)$ is continuous. So, the condition $d(x) > 0$ for $x \in [0,a]$ means that $d(x) \geq d_0 > 0$.

As the space $\mathcal{H}$ we take the closure of all smooth, two-component functions $U(x) = \binom{u_0(x)}{u_1(x)}$, such that $u_0(x) = 0$ in a vicinity of $x = 0$ with respect to the standard energy metric:

$$\|U\|^2_{\mathcal{H}} = \frac{1}{2}\int_0^a \left[p(x)|u_0'|^2 + q(x)\rho(x)|u_0|^2 + \rho(x)|u_1|^2\right] dx. \tag{2.6}$$

Under our assumption ($q \geqslant 0$) the above quantity is always positive and can be treated as the energy of the system. The case $q \leq 0$ is also treatable as a problem with indefinite metric [12].

Now we reproduce some important corollaries of the results from [7]–[9]. Let us look for a solution of (2.1), (2.2) in the form $e^{i\lambda t}u(x)$, and for $u(x)$ we obtain the following one-dimensional boundary value problem:

$$\frac{d}{dx}\left(p(x)\frac{du}{dx}\right) + \lambda^2\rho(x)u - 2i\lambda d(x)\rho(x)u - q(x)\rho(x)u = 0, \quad x \in [0,a], \tag{2.7}$$

$$u(0) = 0, \quad (u_x + i\lambda h u)(a) = 0, \quad h \in \mathbb{C} \cup \{\infty\}. \tag{2.8}$$

Problem (2.7), (2.8) is nonselfadjoint for two reasons: the damping term in the equation ($d(x) \neq 0$) and the complex parameter h in the boundary condition.

DEFINITION 2.1. *We say that $\lambda \in \mathbb{C}$ is an eigenvalue of the nonselfadjoint quadratic operator pencil generated by* (2.1), (2.2) *if problem* (2.7), (2.8) *has a nontrivial solution, which is called an* eigenfunction.

The necessary results from [7]–[9] are given by Theorems 2.1–2.3 below.

THEOREM 2.1. (1) *The problem defined by* (2.7), (2.8) *has an infinite discrete spectrum $\{\lambda_n\}_{n\in\mathbb{Z}}$. All eigenvalues, except for at most a finite number of finite multiplicity, are simple. There are no real eigenvalues.*

(2) *All eigenvalues are located in a strip parallel to the real axis, and* $\mathrm{Re}\lambda_n \to \pm\infty$ *as $n \to \pm\infty$. The simple eigenvalues are strongly separated:*

$$\inf_{n,m,n\neq m} |\lambda_n - \lambda_m| > 0. \tag{2.9}$$

The following asymptotics are valid for $\{\lambda_n\}_{n\in\mathbb{Z}}$:

$$\lim_{n\to\pm\infty} \lambda_n/n = \mathcal{M}, \quad \mathcal{M} = \int_0^a \sqrt{\rho(x)/p(x)}dx < \infty. \tag{2.10}$$

The next statement (Theorem 2.2 below) describes the geometry of the set of eigenfunctions. We notice that each eigenvalue of the quadratic operator pencil (2.7), (2.8) may have only a finite number of associated functions. (For the definition of associated functions of an operator pencil see [13, p. 56].) In the future, we will denote by $\{\varphi_n\}_{n\in\mathbb{Z}}$ the set of all linearly independent eigenfunctions and associated functions, making no difference between them. (For more information about the structure of this set see [8]).

THEOREM 2.2. (1) *The eigenfunctions $\{\varphi_n\}_{n\in\mathbb{Z}}$, specified by the condition $\varphi_n'(0) = d_0\lambda_n$, where d_0 is an absolute constant that can be given in terms of $\rho, p, d,$ and q, and corresponding associated functions are almost normalized (i.e., their norms are bounded from above and below by positive constants) in the space $L^2_\eta(0,a)$, where $\eta = \rho/p$.*

(2) *The set of two-component vectors*

$$\Phi_n(x) = \begin{pmatrix} \dfrac{1}{i\lambda_n} & \varphi_n(x) \\ & \varphi_n(x) \end{pmatrix}, \quad -\infty < n < \infty, \tag{2.11}$$

forms a Riesz basis in the energy space (*see* (2.6)).

THEOREM 2.3. *The set of nonharmonic exponentials* $\{e^{i\overline{\lambda}_n t}\}_{n\in\mathbb{Z}}$ *forms a Riesz basis in the space* $L^2(0, 2\mathcal{M})$, *where* $\mathcal{M}$ *is given in* (2.10).

This theorem follows immediately from the results established in [8]. It is shown in [8] that the points $\{\lambda_n\}_{n\in\mathbb{Z}}$ coincide with the roots of the so-called generalized Jost function. The latter function (see [8, Theorem 2.1]) is an entire function of the first order and of the exponential type $\mathcal{M}$, which is also a sine-type function with the width of the indicator diagram equal to $2\mathcal{M}$. All these properties imply the statement of Theorem 2.3 due to the Levin–Golovin theorem (see, e.g., [28]).

An alternative proof of Theorem 2.3 can be obtained based on the combination of Theorem 2.1 with the general approach to bases of exponentials developed in [28] (see also references in [28]). The latter approach gives additional information about the above set of exponentials. Namely, this set forms a Riesz basis in its closed linear span in $L^2(0, \ell)$ for any $\ell : 2\mathcal{M} \le \ell > \infty$. For $\ell < 2\mathcal{M}$, the set of exponentials is overloaded in $L^2(0, \ell)$. The only value of ℓ for which the above exponentials form a Riesz basis in the whole space $L^2(0, \ell)$ is $\ell = 2\mathcal{M}$. We briefly outline the argument which leads to all of the above conclusions. It follows immediately from statement (2) of Theorem 2.1 that the set of points $\{\lambda_n\}_{n\in\mathbb{Z}}$ satisfies the well-known Carleson condition [14], [15]:

$$\inf_i \prod_{\substack{j=-\infty \\ j\ne i}}^{\infty} \left| \frac{\lambda_j - \lambda_i}{\lambda_j - \overline{\lambda}_i} \right| = \delta > 0. \tag{C}$$

Condition (C) is necessary and sufficient for a system of exponentials to form a Riesz basis in its closed linear span E in the space $L^2(0, \infty)$. Consider the natural projection $P_\ell : L^2(0, \infty) \to L^2(0, \ell)$, $0 < \ell < \infty$. If the restriction $P_\ell|_E$ of this projection to the subspace E is a linear isomorphism, then the set $\{e^{i\lambda_n t}\}_{n\in\mathbb{Z}}$ is a Riesz basis in its closed span in the space $L^2(0, \ell)$ as well. The information about the spectrum, contained in Theorem 2.1, combined with the results in [28] allows us to conclude that the latter statement is true precisely for $2\mathcal{M} \le \ell < \infty$. If $\ell = 2\mathcal{M}$, then the range of $P_\ell|_E$ coincides with $L^2(0, 2\mathcal{M})$, and therefore, the exponentials form a Riesz basis in $L^2(0, 2\mathcal{M})$. We refer to [28] for more details.

Remark 2.2. Theorems 2.1–2.3 are also correct without any additional assumptions in the case when $d(x) \equiv 0$, $x \in [0, a]$, and $\operatorname{Im} h \ne 0$. However, if $d(x) \equiv 0$ or vanishes on a set of positive Lebesgue measure and h is real, one must assume, in addition to (2.5), that at least one of the following three conditions is satisfied:

(a) $\sqrt{p(a)/\rho(a)} \ne |h|$;

(b) $\lim\limits_{x\to a^-} \dfrac{d}{dx}\sqrt{\dfrac{p(x)}{\rho(x)}} \ne 0$;

(c) $\lim\limits_{x\to a^-} \dfrac{d^2}{dx^2}\sqrt{\dfrac{p(x)}{\rho(x)}} \ne 0$.

If all of the above relations (a)–(c) are not satisfied, i.e., we have equalities, and $p, \rho \in C^2(0,a) \cap C[0,a]$, then Theorems 2.1–2.3 do not take place. In this case the spectrum (it is not empty if $p(x)/\rho(x)$ is not identically equal to $|h|$) does not belong to a strip parallel to the real axis: Im $\lambda_n \to \infty$. This behavior of the spectrum destroys the basis property of both the eigenfunctions and the nonharmonic exponentials. In particular, the above conditions are not satisfied if $p(x) = \rho(x) = \text{const}$, $h = 1$. In this case, since it is easy to check by an elementary computation, the spectrum of the problem is empty. However, we stress once again that if $d(x) > 0$ for $x \in [0,a]$, then the behavior of the spectrum described in Theorem 2.1 is immediately restored and all three theorems take place. We refer to [6]–[8] for more detailed information.

The above-described situation has a natural physical interpretation. Assume for simplification that $p(x) = 1$, $d(x) = 0$, $h = 1$, but $\rho(x)$ may be nonconstant. Let us extend $\rho(x)$ to the whole semiaxis $[0,\infty)$ by the rule $\rho(x) = 1$ for $x \in (a,\infty)$. As was mentioned above, in this case our problem describes the scattering of elastic waves on an obstacle concentrated on the interval $[0,a]$. Problem (2.7), (2.8) describes the so-called resonances and resonance states, i.e., quasi-stationary oscillations of the string with finite lifetime. The lifetime of a particular resonance is $\tau_n = |\mathrm{Im}\lambda_n|^{-1}$. If the eigenvalues λ_n behave as described in Theorem 2.1, then $\tau_n \geq \tau^\circ > 0$. So we have an infinite series of "long-lived" resonance states (eigenfunctions of pencil (2.7), (2.78)) ($\tau_n \nrightarrow 0$). The above conditions (a)–(c) adapted to this case say that for an existence for such a series of resonances it is necessary that at least one of the functions $\rho(x)$, $\rho'(x)$, or $\rho''(x)$ have a jump at $x = a$; i.e., the obstacle is not "too smooth." If the conditions (a)–(c) are not satisfied and the density (extended to $[0,\infty)$ by the above rule) is smooth ($\rho \in C^2[0,\infty)$) but nonconstant ($\rho(x) \neq$ const for $x \in [0,a]$), then the resonance spectrum $\{\lambda_n\}$ still exists. However, in this case $\mathrm{Im}\,\lambda_n \to \infty$ and the lifetime of resonance states $\tau_n \to 0$. In other words, for a smooth obstacle we have an infinite series of "short-lived" resonance states. Such a behavior of the spectrum destroys the Riesz basis property of the resonance states and of the corresponding nonharmonic exponentials. Finally, if $\rho(x) = 1$ for all $x \in [0,\infty)$, then there is no obstacle at all and, therefore, there are no resonances, i.e., the spectrum is empty.

We remark in conclusion that the case when Im $h \neq 0$ also has a natural physical interpretation. In this case we have a jump of the "refraction coefficient" at $x = a$. In other words, there is a refraction of elastic waves at the boundary of the obstacle. This explains why the nonsmoothness conditions (a)–(c) are not necessary for the existence of long-lived resonance states if Im $h \neq 0$.

Now we make the following important remark. From this moment on, we assume that for each eigenvalue we have only one normalized eigenvector and no associated vectors. The above assumption makes the exposition below more transparent. If we took into account a possible finite number of nonsimple eigenvalues, then the formulation of our main result would become somewhat more complicated. However, its proof would not become significantly more difficult.

Now we formulate our main results.

THEOREM 2.4 (unique controllability). *Let*

$$U_0(x) = \begin{pmatrix} u_0^0(x) \\ u_1^0(x) \end{pmatrix}$$

be initial data (2.3) *and* $G(x) = \begin{pmatrix} 0 \\ g(x) \end{pmatrix}$, *where* $g(x)$ *is the force distribution function from* (2.1). *Assume that* U_0, $G \in \mathcal{H}$ *have the following expansions with respect to*

basis (2.11):

$$U_0(x) = \sum_{n\in\mathbb{Z}} u_n^0 \Phi_n(x), \quad G(x) = \sum_{n\in\mathbb{Z}} g_n \Phi_n(x). \tag{2.12}$$

(1) *Assume that*

$$g_n \neq 0 \text{ for all } n \in \mathbb{Z}. \tag{2.13}$$

The following statements hold.

(a) *System* (2.1)–(2.3) *is controllable in the time interval* $[0, 2\mathcal{M}]$ *if and only if*

$$\{\gamma_n = u_n^0/g_n\}_{n\in\mathbb{Z}} \in \ell^2(\mathbb{Z}), \text{ i.e., } \sum_{n\in\mathbb{Z}} |\gamma_n|^2 < \infty. \tag{2.14}$$

(b) *Let* $\{\omega_n(t)\}_{n\in\mathbb{Z}}$ *be the Riesz basis in* $L^2(0, 2\mathcal{M})$, *biorthogonal* [13], [15] *to the basis* $\{e^{i\overline{\lambda}_n t}\}_{n\in\mathbb{Z}}$, *i.e.,* $\int_0^{2\mathcal{M}} e^{-i\lambda_m t}\omega_n(t)dt = \delta_{m,n}$. *The desired control function which brings the system to zero state on the time interval* $[0, 2\mathcal{M}]$ *is uniquely defined by the formula*

$$f(t) = -\sum_{n\in\mathbb{Z}} \gamma_n \omega_n(t). \tag{2.15}$$

(c) *If* $T < 2\mathcal{M}$, *then the system is not controllable in time* T.

(d) *If* $T > 2\mathcal{M}$, *then the system is controllable in time* T *and our control problem has infinitely many solutions* $f \in L^2(0, T)$.

(2) *Assume now that* (2.13) *is not satisfied, and let* $R = \{n \in \mathbb{Z} : g_n = 0\}$. *Let* γ_n *be defined by* (2.14) *only for* $n \in \mathbb{Z}\backslash R$. *Let* $S = \{n \in \mathbb{Z} : u_n^\circ = 0\}$.

The following statements hold.

(a) *The system is controllable in time* $T = 2\mathcal{M}$ *if and only if*

$$R \subseteq S \text{ and } \sum_{n\in\mathbb{Z}\backslash S} |\gamma_n|^2 < \infty. \tag{2.16}$$

(b) *The desired control function is not unique and can be given by*

$$f(t) = -\sum_{n\in\mathbb{Z}\backslash S} \gamma_n \omega_n(t) + \sum_{m\in R} \alpha_m \omega_m(t), \tag{2.17}$$

where $\alpha_m \in \mathbb{C}$ *are arbitrary coefficients such that* $\sum_{m\in R} |\alpha_m|^2 < \infty$.

(c) *If the set* $R\backslash(R \cap S)$ *is not empty, then the problem is not controllable in any time.*

Remark 2.1. (a) An explicit formula for $\omega_n(t)$ (in terms of the truncated Blaschke product) is known but is rather complicated [17], [18]. (b) Statement (2) of Theorem 2.4 is a generalization of statement (1). We formulate statement (1) separately because of its importance.

THEOREM 2.5 (approximate controllability). *Assume that condition* (2.14) *is not satisfied but*

$$\{\gamma_n\}_{n\in\mathbb{Z}} \in \ell^p(\mathbb{Z}) \text{ for some } p \in (2, \infty], \tag{2.18}$$

$$\text{i.e., } \sum_{n\in\mathbb{Z}} |\gamma_n|^p < \infty \quad (\text{if } 2 < p < \infty) \text{ or } \sup_{n\in\mathbb{Z}} |\gamma_n| < \infty \quad (\text{if } p = \infty).$$

Then for any $\epsilon > 0$, there exists N such that for the control function

$$f_N(t) = -\sum_{|n|\leqslant N} \gamma_n \omega_n(t) \tag{2.19}$$

we have

$$\|U(\cdot,T)\|_{\mathcal{H}} \leqslant \epsilon \quad \text{for } T = 2\mathcal{M}. \tag{2.20}$$

However, $\|f_N\|_{L^2(0,2\mathcal{M})} \to \infty$ as $N \to \infty$.

3. Reduction to the moment problem and strong solution of the control problem. In this section we first give an operator reformulation of the problem and then using the generalized Fourier method obtain a formal solution. This allows us to prove our main results from section 2. If the initial data are sufficiently regular (Theorem 3.2 below), then our formal solution is, in fact, the strong solution of the control problem.

Let us represent our initial-boundary problem in the form of the following nonselfadjoint operator equation for the function $U = \left(\begin{smallmatrix} u \\ u_t \end{smallmatrix}\right) \equiv \left(\begin{smallmatrix} u_0 \\ u_1 \end{smallmatrix}\right)$:

$$U_t = i\mathcal{L}U + \widehat{G}, \quad U|_{t=0} = U_0 \in \mathcal{H}, \tag{3.1}$$

where

$$U_0(x) = \begin{pmatrix} u_0^0(x) \\ u_1^0(x) \end{pmatrix}, \quad \widehat{G}(x,t) = f(t)G(x), \quad G(x) = \begin{pmatrix} 0 \\ g(x) \end{pmatrix},$$

and $\mathcal{L}$ is the following matrix differential operator in $\mathcal{H}$:

$$\mathcal{L} = -i \begin{pmatrix} 0 & 1 \\ \dfrac{1}{\rho(x)} \dfrac{d}{dx}\left(p(x)\dfrac{d}{dx}\right) - q(x) & -2d(x) \end{pmatrix}. \tag{3.2}$$

$\mathcal{L}$ is a closed maximal (dissipative if $\operatorname{Re} h \geq 0$) nonselfadjoint operator in $\mathcal{H}$ with the domain

$$D(\mathcal{L}) = \{U \in \mathcal{H} : u_0 \in H^2(0,a), u_1 \in H^1(0,a), u_1(0) = 0, (u_0' + hu_1)(a) = 0.\} \tag{3.3}$$

LEMMA 3.1. *The set of all root vectors (eigenvectors and associated vectors together) of $\mathcal{L}$ forms a Riesz basis (linear isomorphic image of an orthonormal basis) in $\mathcal{H}$. It coincides with the set of two-component functions* (2.11).

Proof. If $\Phi \in \mathcal{H}$ is an eigenvector of $\mathcal{L}$, then we see that the second component of $\Phi = \left(\begin{smallmatrix} \varphi_1 \\ \varphi_2 \end{smallmatrix}\right)$ satisfies (2.1) and conditions (2.2), while the first component $\varphi_1 = (i\lambda)^{-1}\varphi_2$. If we normalize the function φ_1 by the condition $\varphi_1'(0) = 1$, then Theorems 2.1 and 2.2 provide the Riesz-basis property of the set of root-vectors of $\mathcal{L}$. ∎

Utilizing the Fourier method and the Riesz basis property of vectors (2.11), we look for a solution of (3.1) in the form

$$U(x,t) = \sum_{n\in\mathbb{Z}} a_n(t)\Phi_n(x). \tag{3.4}$$

We also decompose both functions $\widehat{G}$ and U_0 with respect to $\{\Phi_n\}_{n\in\mathbb{Z}}$:

$$\widehat{G}(x,t) = f(t)\begin{pmatrix} 0 \\ g(x) \end{pmatrix} = f(t)\sum_{n\in\mathbb{Z}} g_n\Phi_n(x), \quad \sum |g_n|^2 < \infty, \tag{3.5}$$

$$U_0(x) = \begin{pmatrix} u_0^0(x) \\ u_1^0(x) \end{pmatrix} = \sum_{n\in\mathbb{Z}} u_n^0\Phi_n, \quad \sum_{n\in\mathbb{Z}} |u_n^0|^2 < \infty. \tag{3.6}$$

Substituting (3.4)–(3.6) into (3.1) gives the following formal series for $U(x,t)$:

$$U(x,t) = \sum_{n\in\mathbb{Z}} \left[u_n^0 e^{i\lambda_n t} + g_n \int_0^t e^{i\lambda_n(t-\tau)} f(\tau)d\tau \right] \Phi_n(x). \tag{3.7}$$

Assume for now that series (3.7) does represent the solution of (3.1). (All necessary justifications are given below.) *Can we find the moment T such that $u(x,T) = u_t(x,T) = 0$?* From (3.7) and the Riesz basis property of $\{\Phi_n\}$ it follows that $U(x,T) = 0$ if and only if the following infinite system has a solution $f \in L^2(0,T)$:

$$u_n^0 + g_n \int_0^T e^{-i\lambda_n\tau} f(\tau)d\tau = 0, \quad n \in \mathbb{Z}. \tag{3.8}$$

Theorem 3.1 below provides the proof of Theorem 2.4 (up to necessary justifications given at the end of this section and in section 4).

THEOREM 3.1.

(1) *Assume that* (2.13) *is satisfied and $\gamma_n = u_n^0/g_n$ are defined by* (2.12). *Then moment problem* (3.8) *has a unique solution if and only if condition* (2.14) *is satisfied. This solution is given by* (2.15).

(2) *Assume that* (2.13) *is not satisfied. Then the moment problem* (3.8) *has a solution if and only if conditions* (2.16) *are satisfied. This solution is not unique and is given by* (2.19), (2.20).

Proof. If (2.13) is satisfied, (3.8) can be written as

$$(f, \chi_n)_{L^2(0,T)} = -\gamma_n, \quad n \in \mathbb{Z}. \tag{3.9}$$

Thus $\{-\gamma_n\}_{n\in\mathbb{Z}}$ is just the sequence of generalized Fourier coefficients of f with respect to the Riesz basis $\{\chi_n = e^{i\overline{\lambda}_n t}\}_{n\in\mathbb{Z}}$. It follows from the Riesz-basis property of the system $\{\chi_n\}_{n\in\mathbb{Z}}$ that (3.9) has a unique solution $f \in L^2(0,T)$ if and only if (2.13) is satisfied. This solution (see [16, 21]) has the form

$$f(t) = \sum_{n\in\mathbb{Z}} (f, \chi_n)_{L^2(0,T)}\omega_n(t), \tag{3.10}$$

where $\{\omega_n\}_{n\in\mathbb{Z}}$ is the basis in $L^2(0,T)$, which is biorthogonal to $\{\chi_n\}_{n\in\mathbb{Z}}$. Substituting (3.9) into (3.10) we obtain precisely (2.15).

Assume now that (2.13) is not satisfied, i.e., $g_n = 0$ for $n \in R \subset \mathbb{Z}$, but (2.16) takes place. Then (3.18) is equivalent to the system

$$(f, \chi_n)_{L^2(0,T)} = -\gamma_n, \quad n \in \mathbb{Z}\backslash R \quad \text{and} \quad (f, \chi_n)_{L^2(0,T)} = \alpha_n, \quad n \in R, \tag{3.11}$$

where α_n, $n \in R$, are arbitrary complex numbers. Since $\{\chi_n\}_{n\in\mathbb{Z}}$ is a Riesz basis, the solution to (3.11) exists if and only if (2.16) and is satisfied. This solution again has

the form (3.10). Substituting (3.11) into (3.10) and taking into account that $\gamma_n \neq 0$ only if $n \in \mathbb{Z}\backslash S$, we obtain (2.17). □

In the conclusion we prove the proposition, which implies Theorem 2.5.

PROPOSITION 3.1. *Assume that the sequence* $\{\gamma_n\}_{n\in\mathbb{Z}}$ *from* (2.9) *is bounded:*

$$|\gamma_n| \leq \Gamma < \infty. \tag{3.12}$$

Assume also that the control function $f(t) = f_N(t)$ *where* $f_N(t)$ *is defined in* (2.19). *Then the following statement takes place for the function* $U(x,t)$ *defined by* (3.7)*: for every* $\varepsilon > 0$ *there exists* N_0 *such that*

$$\|U(\cdot,T)\|_{\mathcal{H}} \leqslant \varepsilon \quad \textit{for} \quad N \geq N_0. \tag{3.13}$$

Proof. Substituting control function (2.19) into (3.7) and taking $t = T$, we obtain

$$U(x,T) = \sum_{|n|=N+1}^{\infty} e^{i\lambda_n t} u_n^0 \Phi_n(x). \tag{3.14}$$

Now we estimate the $\mathcal{H}$-norm of (3.14) using (3.12) and the fact that the system $\{\Phi_n\}$ is almost normalized (we write $\varphi \asymp \psi$ if $C_1\varphi \leq \psi \leq C_2\varphi$, C_1, $C_2 > 0$):

$$\begin{aligned} \|U(\cdot,T)\|_{\mathcal{H}}^2 \asymp \sum_{|n|=N+1}^{\infty} |u_n^0|^2 &= \sum_{|n|=N+1}^{\infty} |g_n|^2 |\gamma_n|^2 \\ &\leqslant \Gamma^2 \sum_{|n|=N+1}^{\infty} |g_N|^2 \to 0 \text{ as } N \to \infty. \end{aligned} \tag{3.15}$$

□

Remark 3.1. (a) If $\{\gamma_n\}_{n\in\mathbb{Z}} \in \ell^q(\mathbb{Z})$ with $q > 2$, then we can make the estimate (3.15) more precise. Indeed, due to the R-basis property of $\{\gamma_n\}_{n\in\mathbb{Z}}$ we can change the numeration of the sequence $\{\lambda_n\}$ such that each sequence $\{|\gamma_n|\}_{n=1}^{\infty}$ and $\{|\gamma_n|\}_{n=-\infty}^{-1}$ becomes nonincreasing: $|\gamma_{|n|}| \geqslant |\gamma_{|n+1|}|$. Then estimate (3.15) can be continued:

$$\|U(\cdot,T)\|_{\mathcal{H}}^2 \leqslant \max\{|\gamma_{N+1}|^2, |\gamma_{-N-1}|^2\} \sum_{|n|=N+1}^{\infty} |g_n^0|^2. \tag{3.16}$$

As $|\gamma|_{|N+1|} \to 0$, we see that $\|U\|_{\mathcal{H}}$ from (3.16) decreases much faster than in (3.15).

(b) As follows from (2.19), for the control function f_N we have $\|f_N\|_{L^2(0,T)}^2 \asymp \sum_{|n|=1}^{N} |\gamma_n|^2$. If $|\gamma_n| \leqslant \Gamma < \infty$, then $\|f_N\| \asymp N \to \infty$ as $N \to \infty$. This means that we can "kill" as many low-frequency harmonics as we need but at the expense of the increasing of the norm of f_N.

(c) If $\{\gamma_n\}_{n\in\mathbb{Z}} \in \ell^q(\mathbb{Z})$, then we can give an approximation for the number of harmonics, which should "be killed" to keep estimate (3.16). Namely, if $\alpha = \|G\|_{\mathcal{H}}$, then we can take $N = [(\varepsilon\alpha)^{-q}] + 1$, where ε is the same as in (3.13) and $[x]$ denotes the greatest integer that does not exceed x.

Now we come back to formula (3.7) and show that under appropriate conditions on U_0 and $\widehat{G}$ it defines the strong solution of problem (3.1). The existence and uniqueness theorems for a linear evolutionary problem (3.1) are well known both for abstract Hilbert space setting [21] and for hyperbolic equations [22]. Using the semigroup approach we can construct a mild or a strong solution of (3.1). However, in our case a complete spectral information about the nonselfadjoint operator $\mathcal{L}$ allows

us to point out some additional properties of the solution which cannot be obtained by purely semigroup methods.

The operator $\mathcal{L}$ generates a continuous scale of Hilbert spaces $\mathcal{H}_\alpha (\alpha \geq 0)$ in the following way. Based on Theorem 2.2 and the Bari's theorem [15], any vector $\Psi \in \mathcal{H}$ can be expanded as

$$\Psi(x) = \sum_{n\in\mathbb{Z}} \psi_n \Phi_n(x), \quad \sum_{n\in\mathbb{Z}} |\psi_n|^2 < \infty. \tag{3.17}$$

We define $\mathcal{H}_\alpha (\alpha \geq 0)$ as the set of all vectors $\Psi \in \mathcal{H}$ such that

$$\|\Psi\|^2_{\mathcal{H}_\alpha} \equiv \sum_{n\in\mathbb{Z}} |\lambda_n|^\alpha |\psi_n|^2 < \infty. \tag{3.18}$$

Since $\lambda = 0$ does not belong to the spectrum of $\mathcal{L}$, $\mathcal{H}_\alpha$ is a Hilbert space with the norm given by (3.18). It is clear that $\mathcal{H}_0 = \mathcal{H}$ and the norm $\|\cdot\|_{\mathcal{H}_0}$ is equivalent to the energy norm (2.6). It is possible to show that $\mathcal{H}_2 = D(\mathcal{L})$ (see (3.3)). Moreover, $\mathcal{H}_\alpha = D(\mathcal{L}^{\alpha/2})$, where powers of $\mathcal{L}$ can be defined based on its spectral decomposition.

DEFINITION 3.1. *Let $\widehat{G} \in L^2(0,T;\mathcal{H}_2)$ ($\widehat{G}$ is not necessarily of form* (3.5)*). A function $U : [0,T] \to \mathcal{H}_2$ is a strong solution of problem* (3.1) *on the interval $[0,T]$ if it satisfies the following conditions:* (i) *$U \in C(0,T;\mathcal{H}_2)$;* (ii) *$U$ is differentiable with respect to t in the sense of distributions with values in $\mathcal{H}$ and the derivative $U_t \in L^2(0,T;\mathcal{H})$;* (iii) *$U$ is strongly differentiable with respect to t in the sense of $\mathcal{H}$-norm a.e. on $[0,T]$ and the strong derivative coincides a.e. with the distributional derivative U_t;* (iv) *U satisfies* (3.1) *for almost all $t \in [0,T]$;* (v) *$U = U_0 \in \mathcal{H}_2$ for $t = 0$.*

There are two classic approaches to the notion of a strong solution of an evolution equation $U_t + AU = \widehat{G}(t)$ in a Banach space X [11, p. 405]. If $\widehat{G} \in C(0,T;X)$, it is required that U is strongly differentiable for all $t \in [0,T]$ and $U(t) \in D(A)$ for all t. If, however, $\widehat{G} \in L^p(0,T;X)$ $(p \geq 1)$, then the first definition is not applicable. It is required that U_t exists in the sense of distributions with values in X, $U(t) \in D(A)$ a.e., and the equation is satisfied a.e. on $[0,T]$. In our problem the control $f \in L^2(0,T)$. So we have to use the second approach. However, our Definition 3.1 differs from the classical one in two aspects. First, $U \in C(0,T;D(A))$, and second, we point out (property (iii)) that the strong $\mathcal{H}$-derivative of U exists a.e. and coincides with the distributional derivative. This property cannot be obtained based on purely semigroup argument, and it takes place due to our spectral representation for A.

THEOREM 3.2. *Assume that $\widehat{G} \in L^2(0,T;\mathcal{H}_2)$ and $U_0 \in \mathcal{H}_2$. Then the series in* (3.7) *is convergent in the norm of $\mathcal{H}_2$ for any $t \in [0,T]$ and defines the unique strong solution of problem* (3.1)*. (If $\widehat{G}$ is not of form* (3.5)*, then formula* (3.7) *can be easily modified and the statement of the theorem remains valid.)*

We omit the proof of this theorem since it is based on standard estimates. The only step which might require some commentary is the proof of the property (iii). It is obvious that the distributional derivative of U has the form

$$U_t(x,t) = \sum_{n\in\mathbb{Z}} \dot{a}_n(t)\Phi_n(x) \ \text{ for almost all } \ t \in [0,T],$$

since distributional derivatives commute with limits. So to prove (iii) one estimates the $\mathcal{H}$ norm of the difference between this expression and $\delta^{-1}[U(x,t+\delta) - U(x,t)]$.

We also recall that, according to Theorem 2.1 (statement (2)), the spectrum belongs to a strip parallel to the real axis and, therefore, the imaginary parts of

the eigenvalues are bounded. The latter fact is necessary for the series (3.7) to be convergent.

4. Weak solution of the control problem. In this section we show that for a wider class of U_0 and $\widehat{G}$ than in Theorem 3.2 formula (3.7) gives a weak solution of problem (3.1). The definition of the weak solution, as well as its existence and uniqueness, for problem (2.1), (2.2) are well known [10, 11, 22]. However, the standard definition is inconvenient for us. We need a modified definition of the weak solution which is adjusted to the operator formulation (3.1) of the problem. To simplify the exposition we give proofs only for the case of the Dirichlet boundary conditions at $x = a (h = \infty)$.

Since $\{\Phi_n\}_{n\in\mathbb{Z}}$ forms a Riesz basis in $\mathcal{H}$, its biorthogonal system $\{\mathcal{X}_n\}_{n\in\mathbb{Z}}$ (defined by $(\Phi_n, \mathcal{X}_m)_{\mathcal{H}} = \delta_{nm}$) also forms a Riesz basis [19]. From the Riesz-basis property it follows that the operator $\mathcal{L}$ admits the spectral representation which allows us to define powers of $\mathcal{L}$:

$$\mathcal{L}^\alpha \Psi = \sum_{n\in\mathbb{Z}} \lambda_n^\alpha (\Psi, \chi_n)_{\mathcal{H}} \Phi_n, \quad \Psi \in D(\mathcal{L}^\alpha), \quad \alpha \geq 0. \tag{4.1}$$

To define λ_n^α uniquely one should use the principal branch of the logarithm. $\mathcal{L}$ generates the scale of Hilbert spaces

$$\mathcal{H}_\alpha = D(\mathcal{L}^{\alpha/2}) = \left\{ \Psi \in \mathcal{H} : \|\Psi\|^2_{\mathcal{H}_\alpha} = \sum_{n\in\mathbb{Z}} |\lambda_n|^\alpha |(\Psi, \mathcal{X}_n)_{\mathcal{H}}|^2 < \infty \right\}. \tag{4.2}$$

The operator $\mathcal{L}^*$ adjoint to $\mathcal{L}$ in $\mathcal{H}$ can be described as the differential operator

$$\mathcal{L}^* = -i \begin{pmatrix} 0 & 1 \\ \dfrac{1}{\rho(x)} \dfrac{d}{dx} \left(p(x) \dfrac{d}{dx} \right) - q(x) & 2d(x) \end{pmatrix} \tag{4.3}$$

on the domain

$$D(\mathcal{L}^*) = D(\mathcal{L}) = \{U \in \mathcal{H} : u_0 \in H^2(0,a), \quad u_1 \in H^1(0,a), \quad u_1(0) = u_0(a) = 0\}.$$

The operator $\mathcal{L}$ has the discrete spectrum $\{\bar{\lambda}_n\}_{n\in\mathbb{Z}}$ of the same multiplicity as the spectrum of $\mathcal{L}$. The system of the root vectors of $\mathcal{L}^*$ coincides with the biorthogonal basis $\{\mathcal{X}_n\}_{n\in\mathbb{Z}}$. The adjoint operator $\mathcal{L}^*$ and its powers have the spectral representation

$$(\mathcal{L}^*)^\alpha \Psi = \sum_{n\in\mathbb{Z}} \bar{\lambda}_n^\alpha (\Psi, \Phi_n)_{\mathcal{H}} \mathcal{X}_n, \quad \Psi \in D((\mathcal{L}^*)^\alpha), \quad \alpha \geq 0, \tag{4.4}$$

and $\mathcal{L}^*$ also defines a scale of Hilbert spaces:

$$\widehat{\mathcal{H}}_\alpha = D((\mathcal{L}^*)^{\alpha/2}) = \left\{ \Psi \in \mathcal{H} : \|\Psi\|^2_{\widehat{\mathcal{H}}_\alpha} = \sum_{n\in\mathbb{Z}} |\lambda_n|^\alpha |(\Psi, \Phi_n)_{\mathcal{H}}|^2 < \infty \right\}. \tag{4.5}$$

It follows from the general properties of the biorthogonal Riesz bases that $\mathcal{H}_\alpha$ and $\widehat{\mathcal{H}}_\alpha$ coincide as vector spaces. We use different notation for these spaces to emphasize that norms (4.2) and (4.5) are different. These norms are, however, equivalent.

In the following we will use the space $W(0,T;\widehat{\mathcal{H}}_\alpha)$. It consists of all functions $V \in L^2(0,T;\widehat{\mathcal{H}}_\alpha)$ which have derivative, V_t, in the sense of distributions with values in $\mathcal{H}_\alpha$ (see, e.g., [11, Chap. 3]) and $V_t \in L^2(0,T;\mathcal{H}_\alpha)$. The norm in $W(0,T;\widehat{\mathcal{H}}_\alpha)$ is

$$\|V\|_W^2 = \int_0^T (\|V(\cdot,t)\|_{\widehat{\mathcal{H}}_\alpha}^2 + \|V_t(\cdot,t)\|_{\widehat{\mathcal{H}}_\alpha}^2)dt. \tag{4.6}$$

It is equivalent to the norm

$$|V|_W^2 = \sum_{n\in\mathbb{Z}} |\lambda_n|^\alpha \|\widehat{v}_n\|_{H^1(0,T)}^2, \quad \widehat{v}_n(t) = (V(\cdot,t),\Phi_n)_{\mathcal{H}}. \tag{4.7}$$

The following bounded embedding takes place: $W(0,T;\widehat{\mathcal{H}}_\alpha) \subset C(0,T;\widehat{\mathcal{H}}_\alpha)$ ([26] or [11, Chap. 3]). We use the notation $(\cdot,\cdot)_{\mathbb{L}}$ for the inner product in $L^2(0,T;\mathcal{H})$.

We give a formal "derivation" of the definition of the weak solution. Let us take the $L^2(0,T;\mathcal{H})$-inner product of (3.1) with arbitrary two-component function $V(x,t) = \left(\begin{smallmatrix} v_0(x,t) \\ v_1(x,t) \end{smallmatrix}\right)$, $v_0,\ v_1 \in C^\infty(\Omega\times[0,T])$. Using the initial condition (3.1) we obtain for $(U_t,V)_{\mathbb{L}}$

$$(U_t,V)_{\mathbb{L}} = (U(\cdot,T),V(\cdot,T))_{\mathcal{H}} - (U_0,V(\cdot,0))_{\mathcal{H}} - (U,V_t)_{\mathbb{L}} = i(\mathcal{L}U,V)_{\mathbb{L}} + (\widehat{G},V)_{\mathbb{L}}. \tag{4.8}$$

From (4.8) we get the first necessary condition $V(x,T)=0$ for all $x\in\overline{\Omega}$. We rewrite $i(\mathcal{L}U,V)_{\mathbb{L}}$ from (4.8) as follows:

$$i(\mathcal{L}U,V)_{\mathbb{L}} = \frac{1}{2}\int_0^T dt \int_0^a [p(x)(u_1'\overline{v}_0' - u_0'\overline{v}_1') + q(x)\rho(x)(u_1\overline{v}_0 - u_0\overline{v}_1)$$

$$-2d(x)\rho(x)u_1\overline{v}_1]dx + \frac{1}{2}\int_0^T [p(x)u_0'(x,t)v_1(x,t)]_{x=a}^{x=0}dt. \tag{4.9}$$

To get rid of the boundary term we impose the second restriction on V: $v_1(0,t) = v_1(a,t) = 0$. Our next step is to rewrite (4.9) in the operator form. Setting $\tau(x) = i\sqrt{p(x)/\rho(x)}$, we introduce the following matrix operators on $\mathcal{H}$:

$$\mathcal{A} = \begin{pmatrix} 0 & 1 \\ -\tau(x)\frac{d}{dx} & 0 \end{pmatrix}, \quad \mathcal{B} = \begin{pmatrix} 1 & 0 \\ 0 & \tau(x)\frac{d}{dx} \end{pmatrix}, \quad \mathcal{C} = \begin{pmatrix} 0 & 0 \\ -q(x) & -2d(x) \end{pmatrix}. \tag{4.10}$$

It can be verified by a straightforward calculation that by taking into account the boundary conditions on V and the fact $V(x,T)=0$, we can rewrite (4.9) as follows:

$$(U_0,V(\cdot,0))_{\mathcal{H}} + (U,V_t)_{\mathbb{L}} = (\mathcal{A}U,\mathcal{B}V)_{\mathbb{L}} + (\mathcal{C}U,V)_{\mathbb{L}} + (\widehat{G},V)_{\mathbb{L}}. \tag{4.11}$$

DEFINITION 4.1. *A function $U\in L^2(0,T;\mathcal{H}_1)$ is a weak solution of problem* (3.1) *if for any $V\in W(0,T;\widehat{\mathcal{H}}_1)$, $V(\cdot,T)=0$, $v_1(0,t)=v_1(a,t)=0$ integral identity* (4.11) *holds.*

To show that the definition makes sense, we describe the domains of $\mathcal{A}$ and $\mathcal{B}$ and give an alternative description of $\mathcal{H}_1 = \mathcal{D}(\mathcal{L}^{1/2})$ in terms of Sobolev spaces.

LEMMA 4.1. (a) *$\mathcal{A}$ and $\mathcal{B}$ are closed operators in $\mathcal{H}$ with the domains*

$$D(\mathcal{A}) = H_0^1(0,a)\times H_0^1(0,a); \quad D(\mathcal{B}) = H_0^1(0,a)\times H^1(0,a). \tag{4.12}$$

(*With our assumptions about* p, ρ, *and* d *we have* $\mathcal{H} = H_0^1(0,a) \times L^2(0,a)$).

(b) *The space* $\mathcal{H}_1$ *can be described in terms of Sobolev spaces as follows:*

$$\mathcal{H}_1 = D(\mathcal{L}^{1/2}) = D((\mathcal{L}^*)^{1/2}) = H^{3/2}(0,a) \times H_0^1(0,a). \tag{4.13}$$

Here and below $H_0^\alpha(0,a) = \{u \in H^\alpha(0,a), u(0) = u(a) = 0\}, \alpha > 1/2$. For $\alpha > 1/2$ the restriction $u(0) = u(a) = 0$ makes sense due to the embedding theorems.

Remark 4.1. First, from (4.12) and (4.13) we have $\mathcal{H}_1 \subset \mathcal{D}(\mathcal{A}) \subset \mathcal{D}(\mathcal{B})$ and, therefore, definition (4.11) makes sense. Second, it is clear from (4.13) that all functions $U \in L^2(0,T;\mathcal{H}_1)$ satisfy the necessary boundary condition a.e. on $[0,T]$.

Proof of Lemma 4.1. Formula (4.12) follows immediately from the definitions of the operators $\mathcal{A}$ and $\mathcal{B}$. To prove (4.13) we note that it can be verified by a straightforward calculation that the operator $\mathcal{L}^{1/2}$ can be represented in the form $\mathcal{L}^{1/2} = (I + K)\mathcal{L}_0^{1/2}$, with K given by the formula

$$K = I - (I + \mathcal{P}\mathcal{L}_0^{-1})^{1/2}, \qquad \mathcal{P} = -i \begin{pmatrix} 0 & 0 \\ -q(x) & -2d(x) \end{pmatrix}, \tag{4.14}$$

where $\mathcal{L}_0$ is the positive selfadjoint operator

$$\mathcal{L}_0 = -i \begin{pmatrix} 0 & I \\ \dfrac{1}{\rho(x)} \dfrac{d}{dx} \left(p(x) \dfrac{d}{dx} \right) & 0 \end{pmatrix} \tag{4.15}$$

with the domain $D(\mathcal{L}_0) = D(\mathcal{L}) = \mathcal{H}_2 = H_0^2(\Omega) \times H_0^1(\Omega)$. We see from (4.14) that K is a compact Hilbert–Schmidt operator [20]. Therefore, $D(\mathcal{L}^{1/2}) = D(\mathcal{L}_0^{1/2})$. It is not difficult to check by a straightforward computation that

$$\mathcal{L}_0^{1/2} = \frac{i}{2} L^{1/4} \begin{pmatrix} I & -L^{1/2} \\ L^{1/2} & I \end{pmatrix}, \tag{4.16}$$

where L is the selfadjoint operator on $L^2(0,a)$ given by the expression

$$L = \frac{1}{\rho(x)} \frac{d}{dx} \left(p(x) \frac{d}{dx} \right), \qquad D(L) = H_0^2(0,a). \tag{4.17}$$

Using (4.16), (4.17), for any $U \in \mathcal{H}$ we have

$$\mathcal{L}_0^{1/2} U = \frac{i}{2} \begin{pmatrix} L^{1/4} u_0 - L^{-1/4} u_1 \\ L^{3/4} u_0 + L^{1/4} u_1 \end{pmatrix} \in \mathcal{H} = H_0^1(0,a) \times L^2(0,a).$$

From the latter relation, (4.13) follows immediately. □

THEOREM 4.1. *The weak solution of the problem* (3.1) *in the sense of Definition* 4.1 *is unique.*

Proof. Using a contradiction argument, assume that we have two different weak solutions. Their difference U satisfies the homogeneous integral identity

$$(U, V_t)_{\mathbb{L}} + (\mathcal{A}U, \mathcal{B}V)_{\mathbb{L}} + (\mathcal{C}U, V)_{\mathbb{L}} = 0. \tag{4.18}$$

Due to the completeness of $\{\mathcal{X}_n\}_{n\in\mathbb{Z}}$ in $\mathcal{H}$, we claim that there exists a number n_0 such that $u(t) = (U, \mathcal{X}_{n_0})_{\mathcal{H}} \neq 0$, i.e., $\|u\|_{L^2(0,T)}0$. Let us choose a particular test function $V(x,t)$:

$$V(x,t) = \widehat{g}(t)\mathcal{X}_{n_0}, \quad \widehat{g}(t) = \int_t^T e^{-i\overline{\lambda}_{n_0}(t-\tau)}\overline{u}(t)dt. \tag{4.19}$$

Since $\mathcal{X}_{n_0} \in D(\mathcal{L}^*)$, we have $V(0,t) = V(a,t) = V(x,T) = 0$. So V can be used as a test function. For this test function relation (4.18) has the form

$$(U, (\widehat{g}_t + i\overline{\lambda}_{n_0}\widehat{g})\mathcal{X}_{n_0})_{\mathbb{L}} = \int_0^T (\widehat{g}_t + i\overline{\lambda}_{n_0}\widehat{g})\overline{(U, \mathcal{X}_{n_0})}_{\mathcal{H}}dt = 0. \tag{4.20}$$

A straightforward computation shows that for $\widehat{g}(t)$ from (4.20) we have

$$\widehat{g}_t + i\overline{\lambda}_{n_0}\widehat{g} = -u(t). \tag{4.21}$$

Substituting (4.21) into (4.20) we obtain $\|u\|_{L^2(0,T)} = 0$, which contradicts our assumption. □

THEOREM 4.2. *Let $\widehat{G} \in L^2(0,T;\mathcal{H}_1)$ and $U_0 \in \mathcal{H}_1$. Then the series in* (3.7) *is convergent in the norm of $\mathcal{H}_1$ for any $t \in [0,T]$ and defines the unique weak solution U of problem* (3.1). *Moreover, $U \in C(0,T;\mathcal{H}_1)$.*

Proof. First, the fact that series (3.17) defines a function from $C(0,T;\mathcal{H}_1)$ can be shown by means of standard estimates. The only nontrivial step is to show that the function given by (3.17) satisfies integral identity (4.11).

Consider the following sequences of functions:

$$U_m = \sum_{|n|\leqslant m} a_n(t)\Phi_n(x), \quad U_{0m} = \sum_{|n|\leqslant m} u_n^0\Phi_n(x), \quad \widehat{G}_m = f(t)\sum_{|n|\leqslant m} g_n\Phi_n(x), \tag{4.22}$$

where u_n^0 and g_n are defined by (3.5), (3.6) and $a_n(t)$ are the coefficients from (3.7). The following integral identity holds for every $m \geqslant 1$ and any $V \in W(0,T;\widehat{\mathcal{H}}_1)$, $V(\cdot,T) = 0$:

$$(U_{0m}, V(\cdot,0))_{\mathcal{H}} + (U_m, V_t)_{\mathbb{L}} = (\mathcal{A}U_m, \mathcal{B}V)_{\mathbb{L}} + (\mathcal{C}U_m, V)_{\mathbb{L}} + (\widehat{G}_m, V)_{\mathbb{L}}. \tag{4.23}$$

Now we show that every sequence in (4.23) is a Cauchy sequence. In order to conserve space, we will prove the convergence for the two of the five sequences which are the most complicated: $\{\mathbf{S}_1(m)\}_{m-1}^{\infty} = \{(U_m, V_t)\}_{m=1}^{\infty}$ and $\{\mathbf{S}_2(m)\}_{m=1}^{\infty} = \{(\mathcal{A}U_m, \mathcal{B}V)_{\mathbb{L}}\}_{m=1}^{\infty}$. For any positive integers p and m we have

$$|\mathbf{S}_1(p) - \mathbf{S}_1(m)| \leqslant |(U_p - U_m)_t, V)_{\mathbb{L}}| + |(U_{0p} - U_{0m}, V(\cdot,0))_{\mathcal{H}}|$$

$$\leqslant \left|\left(f, \sum_{|n|=m}^{p} \overline{g}_n v_n(t)\right)_{L^2(0,T)}\right| + \left|\sum_{|n|=m}^{p} i\lambda_n(a_n(t), v_n(t))_{L^2(0,T)}\right| + \left|\sum_{|n|=m}^{p} u_n^0\overline{v}_n(0)\right|.$$

The following estimates are valid:

(a)

$$\left|\left(f(t), \sum_{|n|=m}^{p} \overline{g}_n v_n(t)\right)_{L^2(0,T)}\right| \leqslant \|f\|_{L^2(0,T)}\sqrt{\sum_{|n|=m}^{p} |g_n|^2}\sqrt{\sum_{n\in\mathbb{Z}} \|v_n\|^2_{H^1(0,T)}}.$$

(b)

$$\left|\sum_{|n|=m}^{p} i\lambda_n(a_n(t),v_n(t))_{L^2(0,T)}\right| \leqslant \sqrt{\sum_{|n|=m}^{p}|\lambda_n|\|a_n\|^2_{L^2(0,T)}}\sqrt{\sum_{n\in\mathbb{Z}}|\lambda_n|\|v_n\|^2_{H^1(0,T)}}.$$

Since $V\in W(0,T;\widehat{\mathcal{H}}_1)$, the latter series in (b) converges.

(c)

$$\left|\sum_{|n|=m}^{p} u_n^0\overline{v}_n(0)\right| \leqslant \sqrt{\sum_{|n|=m}^{p}|u_n^0|^2}\sqrt{\sum_{n\in\mathbb{Z}}|v_n(0)|^2}.$$

Collecting together (a)–(c) we obtain with some absolute constant d:

$$|((U_p-U_m),V_t)_{\mathbb{L}}| \leqslant d\left[\|V(\cdot,0)\|_{\mathbb{L}}\sqrt{\sum_{|n|=m}^{p}|u_n^0|^2}\right.$$

$$\left.+\|V\|_{W(0,T;\widehat{\mathcal{H}}_1)}\left(\sqrt{\sum_{|n|=m}^{p}|g_n|^2}+\sqrt{\sum_{|n|=m}^{p}|\lambda_n|\ \|a_n\|^2_{L^2(0,T)}}\right)\right]. \tag{4.24}$$

Due to (4.22) we can pass to the limit: $\lim_{m\to\infty}\mathbf{S}_1(m)=(\lim U_m,V_t)_{\mathbb{L}}=(U,V_t)_{\mathbb{L}}$. At last we justify the passage to the limit for $\mathbf{S}_2$. We have

$$|(\mathbf{S}_2(p)-\mathbf{S}_2(m))+(\mathcal{C}(U_p-U_m),V)_{\mathbb{L}}| \leqslant \sum_{|n|=m}^{p}|\lambda_n|\left|(a_n(t),v_n(t))_{L^2(0,T)}\right|$$

$$\leqslant\ d_1\|U\|_{L^2(0,T;\mathcal{H}_1)}\|V\|_{W(0,T;\widehat{\mathcal{H}}_1)}\sqrt{\sum_{|n|=m}^{p}|\lambda_n|\ \|a_n\|^2_{L^2(0,T)}}. \tag{4.25}$$

Based on (4.22) and $\lim_{m\to\infty}(\mathcal{C}U_m,V)_{\mathbb{L}}=(\mathcal{C}U,V)_{\mathbb{L}}$ we have $\lim_{m\to\infty}\mathbf{S}_2(m)=(\mathcal{A}U,\mathcal{B}V)_{\mathbb{L}}$, and thus, the integral identity is shown. □

REFERENCES

[1] D.L. Russell, *On boundary-value controllability of linear symmetric hyperbolic system*, in Mathematical Theory of Control, A.V. Balakrishnan and L.W. Neustadt, eds., Academic Press, New York, (1967), pp. 312–321.

[2] D.L. Russell, *Nonharmonic Fourier series in the control theory of distributed parameter systems*, J. Math. Anal. Appl., 18 (1967), pp. 542–559.

[3] D.L. Russell, *Boundary value control of the higher dimensional wave equation*, SIAM J. Control, 9 (1971), pp. 29–42, pp. 401–419.

[4] D.L. Russell, *A unified boundary controllability theory for hyperbolic and parabolic partial differential equations*, Stud. Appl. Math., LII (1973), pp. 189–211.

[5] K. Graham and D.L. Russell, *Boundary value control of the wave equation in a spherical region*, SIAM J. Control, 13 (1975), pp. 174–196.

[6] M.A. Shubov, *Asymptotics of resonances and geometry of resonance states in the problem of scattering of acoustical waves by a spherically symmetric inhomogeneity of the density*, Differential Integral Equations, 8 (1995), pp. 1073–1115.

[7] M.A. Shubov, *Asymptotics of resonances and eigenvalues for nonhomogeneous damped string*, Asymptotic Anal., 13 (1996), pp. 31–78.

[8] M.A. Shubov, *Basis property of eigenfunctions of nonselfadjoint operator pencils generated by the equation of nonhomogeneous damped string*, Integral Equations Operator Theory, 25 (1996), pp. 289–328.
[9] M.A. Shubov, *Transformation Operators and Basis Property of Root Vectors for Damped Wave Equation*, Texas Tech University, Lubbock, TX, 1996, preprint.
[10] J.-L. Lions, *Optimal Control of Systems Governed by Partial Differential Equations*, Springer-Verlag, New York, 1971.
[11] R. Dautray and J.-L. Lions, *Mathematical Analysis and Numerical Methods for Science and Technology*, Vol. 5, Springer-Verlag, New York, 1992.
[12] P.D. Lax and R.S. Phillips, *Scattering Theory*, rev. ed., Academic Press, New York, 1989.
[13] A.S. Marcus, *Introduction to the Spectral Theory of Polynomial Pencils*, Transl. Math. Monographs, 71, AMS, Providence, RI, 1988.
[14] N.K. Nikol'skii and B.S. Pavlov, *Eigenvector bases of completely nonunitary contractions and the characteristic function*, Math. USSR-Izv., 4 (1970), pp. 90–133.
[15] N.K. Nikol'skii, *Treatise on the Shift Operator*, Springer-Verlag, Berlin, 1986.
[16] S.A. Ivanov and B.S. Pavlov, *Carleson series of resonances and the Regge problem*, Math. USSR-Izv., 12(1), (1978), pp. 26–55.
[17] B. Ya. Levin, *Distribution of Zeroes on Entire Functions*, Transl. Math. Monographs 5, AMS, Providence, RI, 1964.
[18] A.M. Sedletskii, *On convergence of nonharmonic Fourier series in systems of exponentials, cosines and sines*, Soviet Math. Dokl., 38 (1989), pp. 179–183.
[19] I. Ts. Gohberg and M.G. Krein, *Introduction to the Theory of Linear Nonselfadjoint Operators*, Transl. Math. Monographs 18, AMS, Providence, RI, 1969.
[20] R.M. Young, *An Introduction to Nonharmonic Fourier Series*, Academic Press, New York, 1980.
[21] A. Pazy, *Semigroup of Linear Operators and Applications to Partial Differential Equations*, Appl. Math. Sciences 44, Springer-Verlag, New York, 1983.
[22] J.-L. Lions and E. Magenes, *Problèmes aux Limites non Homogènes et Applications*, Vol. I, II, Dunod, Paris, 1968, Vol. III, 1970.
[23] S. Cox and E. Zuazua, *Estimations sur le taux de decroissance exponentielle de l'energie dans des équations des ondes dissipatives linéaires*, C. R. Acad. Sci. Paris Sér. I Math., 317 (1993), pp. 249–254.
[24] S. Cox and E. Zuazua, *The rate at which the energy decays in a damped string*, Comm. Partial Differential Equations, 19 (1994), pp. 213–243.
[25] S. Cox and E. Zuazua, *The rate at which the energy decays in the string damped at one end*, Indiana Univ. Math. J., 44 (1995), pp. 545–573.
[26] M.A. Pekker (M.A. Shubov), *Resonances in the scattering of acoustic waves by a spherical inhomogeneity of the density*, Soviet Math. Dokl., 15 (1974), pp. 1131–1134.
[27] M.A. Pekker (M.A. Shubov), *Resonances in the scattering of acoustic waves by a spherical inhomogeneity of the density*, Amer. Math. Soc. Transl. (2), 115 (1980), pp. 143–163.
[28] S.V. Hruščev, N.K. Nikol'skii, and B.S. Pavlov, *Unconditional Bases of Exponentials and of Reproducing Kernels*, Lecture Notes in Math. 864, Springer-Verlag, New York, 1981, pp. 214–335.
[29] J.-L. Lions, *Contrôlabilité exacte, perturbations et stabilisation de systemes distribues*, T.1: *Contrôlabilité exacte*, Rech. Math. Appl. 8, Masson, Paris, 1988; T.2, *Perturbations*, Rech. Math. Appl. 9, Masson, Paris, 1988.
[30] C. Bardos, G. Lebeau, and J. Rauch, *Contrôle et stabilisation dans les problèmes hyperboliques*, Appendix II, in J.-L. Lions, Contrôlabilité exacte et stabilisation de systèmes distribués, Vols. 1,2, Masson, 1990.

SIAM J. CONTROL OPTIM.
Vol. 35, No. 5, pp. 1790–1810, September 1997

019

RISK-SENSITIVE CONTROL OF FINITE STATE MACHINES ON AN INFINITE HORIZON I*

W. H. FLEMING† AND D. HERNÁNDEZ-HERNÁNDEZ‡

Abstract. In this paper we consider robust and risk-sensitive control of discrete time finite state systems on an infinite horizon. The solution of the state feedback robust control problem is characterized in terms of the value of an average cost dynamic game. The risk-sensitive stochastic optimal control problem is solved using the policy iteration algorithm, and the optimal rate is expressed in terms of the value of a stochastic dynamic game with average cost per unit time criterion. By taking a small noise limit, a deterministic dynamic game which is closely related to the robust control problem is obtained.

Key words. risk-sensitive control, robust control, finite state machines, large deviations

AMS subject classifications. 93E20, 93B36, 93C10, 90D25, 60F10, 49L25

PII. S0363012995291622

1. Introduction.

There are various approaches to treating disturbances in control systems. In stochastic control, disturbances are modelled as stochastic processes (random noise). On the other hand, in H_∞/robust control theory, disturbances are modelled deterministically. The theory of risk-sensitive optimal control provides a link between stochastic and deterministic approaches. The link is made by considering small noise limits for stochastic control problems with exponential cost criteria. For continuous variable, finite time horizon problems this idea was introduced by Whittle [18], [19]. For the state feedback (complete state observation) case, Whittle's idea was put on a mathematically rigorous basis in [14], [9] using viscosity solution methods. Discrete time, output feedback (partial state information) problems on a finite time horizon were treated in [15]. See also [7], where a solution approach for risk-sensitive control problems for hidden Markov models is given.

In [1] robust and risk-sensitive control of discrete time finite state systems on a finite horizon is considered. The purpose of the present paper is to study such systems on an infinite time horizon. We consider here only the state feedback case. In a sequel, output feedback control on an infinite horizon will be considered. Our approach is similar in spirit to [10], [11], where continuous variable systems modelled by differential equations are considered. However, the technical details are quite different for the discrete (finite state machine) case.

To illustrate the ideas in a simple setting, we begin in section 2 with uncontrolled finite state machines described by the difference equation (2.1). In the robust/H_∞ formulation, deterministic perturbations are described by (2.2). The H_∞-norm is characterized in terms of a deterministic optimal control problem, in which the

*Received by the editors September 13, 1995; accepted for publication (in revised form) July 22, 1996.

http://www.siam.org/journals/sicon/35-5/29162.html

†Division of Applied Mathematics, Brown University, Providence, RI 02912 (whf@cfm.brown.edu). The research of this author was supported in part by AFOSR grant F49620-92-J-0081, ARO grant DAAL03-92-G-0115, and NSF grant DMS-9301048.

‡Departamento de Matemáticas, CINVESTAV-IPN, Apartado Postal 14-740, México, D.F. 07000, México (dher@math.cinvestav.mx). The research of this author was supported by the Consejo Nacional de Ciencia y Tecnología (CONACYT) and the Centro de Investigación y de Estudios Avanzados (CINVESTAV), México. This research was performed while the author visited the Division of Applied Mathematics, Brown University.

perturbations (or disturbances) are chosen to maximize some long run average cost per unit time criterion. The maximum average cost λ_0 is nonnegative, and is zero if and only if H_∞-control is achieved. The corresponding cost potential function $W_0(x)$ has the role of a "storage function" in H_∞-control terminology. It is unique, provided that the H_∞-norm parameter γ exceeds the critical value γ^*.

Following [1], stochastic perturbations of the finite state machine model are introduced in section 2.2. The strength of the perturbations are described through a parameter ε. Stochastic analogues $\lambda_\varepsilon, W_\varepsilon$ of λ_0, W_0 are introduced, with λ_ε the maximal expected average cost per unit time for a corresponding ergodic stochastic control problem. In section 3, convergence of $\lambda_\varepsilon, W_\varepsilon$ to λ_0, W_0 as $\varepsilon \to 0$ is proved.

In section 4, we consider controlled finite state machines. With the deterministic robust $/H_\infty$ formulation, the optimal H_∞-norm is characterized in terms of a dynamic game with average cost per unit time. In the stochastic risk-sensitive formulation, a corresponding stochastic difference game is introduced, with payoff involving a relative entropy function. Results like those in section 3 are obtained as the noise intensity parameter $\varepsilon \to 0$.

2. Risk-sensitive analysis. In this section we consider both deterministic and stochastic perturbations of a discrete time finite state system. In order to measure the size of the deterministic perturbations, an analogous definition of the H_∞-norm for nonlinear continuous variable systems is given. This norm is characterized in terms of the value of a long run average cost deterministic control problem. The risk-sensitive index is used to measure the effect of the stochastic perturbations, and it is expressed as the value of an average cost stochastic optimal control problem.

2.1. The H_∞-norm. Consider the deterministic finite state machine

$$x_{t+1} = f(x_t), \quad t = 0, 1, \ldots; \quad x_0 = x, \tag{2.1}$$

where x_t takes values in the finite set X and f is a function from X into itself that defines the dynamics of the system. The set X has N elements, $X = \{x_1, \ldots, x_N\}$.

Let us now define a perturbed system Σ,

$$x_{t+1} = b(x_t, w_t), \quad t = 0, 1, \cdots; \quad x_0 = x. \tag{2.2}$$

Here the exogenous inputs (disturbances) w_t take values in a finite set W, the state variable x_t evolves in X, and $b : X \times W \to X$ is a given function.

Remark 2.1 (notation). Throughout this paper we denote by $[0, T]$ the time interval $\{0, 1, \ldots, T\}$. If Z is a generic set, then $Z[0, T]$ denotes the set of functions $\underline{z} : [0, T] \to Z$. Moreover, given any function $v : Z \to \mathbb{R}$, $\|v\|$ stands for the supremum norm of v, i.e., $\|v\| = \sup_{z \in Z} |v(z)|$.

We assume the following.

(A1) There exists a null state x_ϕ and a null disturbance $w_\phi \in W$ such that

$$\text{(i)} \quad f(x_\phi) = x_\phi \quad \text{and} \qquad \text{(ii)} \quad b(x, w_\phi) = f(x) \text{ for all } x \in X.$$

(A2) Given $x, x'' \in X$, there exist T_1, $0 < T_1 \leq N$, and $\underline{w} \in W[0, T_1]$ such that, for the initial condition $x_0 = x$ and input $\underline{w}$, the system Σ reaches x'' after T_1 steps.

(A3) There exists a positive integer N_0 such that for any initial condition $x_0 = x$, the unperturbed system (2.1) reaches the null state x_ϕ after N_0 steps.

Let us introduce a pair of functions $\theta : W \to \mathbb{R}$ and $\ell : X \to \mathbb{R}$ such that

$$\begin{cases} \theta(w_\phi) = 0, \\ \theta(w) > 0 \quad \text{if} \quad w \neq w_\phi \in W \end{cases} \tag{2.3}$$

and

$$\begin{cases} \ell(x_\phi) = 0, \\ \ell(x) > 0 \qquad \text{if} \quad x \neq x_\phi \in X. \end{cases}$$

The values $\theta(w)$ and $\ell(x)$ represent the magnitude of disturbance w and the cost per stage generated by the system (2.2), respectively.

Now, let us give the definition of the H_∞-norm of the discrete system Σ; this definition is analogous to the one given for nonlinear continuous variable systems; see, e.g., [17, 2].

DEFINITION 2.2. *We say that the H_∞-norm $\|\Sigma\|_{H_\infty}$ is less than or equal to a positive number γ if and only if for every $x_0 = x$, there exists a nonnegative constant $K(x)$, with $K(x_\phi) = 0$, such that*

$$K(x) + \sum_{t=0}^{T} \left[\gamma^2 \theta(w_t) - \ell(x_t)\right] \geq 0 \text{ for all } \underline{w} \in W[0,T], T \geq 0. \tag{2.4}$$

Then $\|\Sigma\|_{H_\infty}$ is the smallest γ such that $\|\Sigma\|_{H_\infty} \leq \gamma$.

Straightforward calculations show that $\|\Sigma\|_{H_\infty} \leq \gamma$ if and only if there exists a nonnegative function W_0 defined on X, called a *storage function*, such that

$$\begin{cases} W_0(x) \geq \sup\limits_{T>0} \sup\limits_{\underline{w} \in W[0,T]} \left\{ W_0(x_{T+1}) - \sum\limits_{t=0}^{T} \left[\gamma^2 \theta(w_t) - \ell(x_t)\right] \right\}, \\ W_0(x_\phi) = 0. \end{cases} \tag{2.5}$$

If there exists a storage function, then the system Σ is called *dissipative* with respect to the supply rate $(x, w) \to \gamma^2 \theta(w) - \ell(x)$. Actually, the inequality (2.5) can be rewritten as

$$\begin{cases} W_0(x) \geq \max\limits_{w \in W} \{W_0(b(x,w)) + \ell(x) - \gamma^2 \theta(w)\}, \\ W_0(x_\phi) = 0. \end{cases} \tag{2.5$'$}$$

The H_∞-norm shall be characterized in terms of the value of an average cost optimal control problem, but first we introduce some preliminary results.

PROPOSITION 2.3. *Assume* (A2). *Then there exist a nonnegative number λ_0 and a function $W_0 : X \to \mathbb{R}$ such that*

$$\lambda_0 + W_0(x) = \max_{w \in W} [W_0(b(x,w)) + \ell(x) - \gamma^2 \theta(w)]. \tag{2.6}$$

Proof. The proof is based on the standard vanishing discount approach.

Define the value function

$$V_\beta(x) := \sup_{\underline{w} \in W[0,\infty)} \sum_{t=0}^{\infty} \beta^t [\ell(x_t) - \gamma^2 \theta(w_t)], \tag{2.7}$$

where $\beta \in (0,1)$ and x_t obeys the dynamic described by (2.2). This function satisfies the dynamic programming equation

$$V_\beta(x) = \sup_{w \in W} [\beta V_\beta(b(x,w)) + \ell(x) - \gamma^2 \theta(w)] \text{ for all } x \in X. \tag{2.8}$$

On the other hand, from (2.7), we have

$$0 \le (1-\beta)V_\beta(x) \le \|\ell\|. \tag{2.9}$$

Then, for each $x \in X$, (2.8)–(2.9) yield

$$\begin{aligned} V_\beta(x) &\ge \beta V_\beta(b(x,w)) - \gamma^2\|\theta\| \\ &= V_\beta(b(x,w)) - [(1-\beta)V_\beta(b(x,w)) + \gamma^2\|\theta\|] \\ &\ge V_\beta(b(x,w)) - C_1 \text{ for all } w \in W, \end{aligned} \tag{2.10}$$

with $C_1 = \|\ell\| + \gamma^2\|\theta\|$. Thus, given $x, x'' \in X$, from (A2) and iterating (2.10) T_1 times, we have

$$V_\beta(x) \ge V_\beta(x'') - T_1 C_1.$$

Thus, interchanging the roles of x and x'', for some constant T' depending on x and x'',

$$|V_\beta(x) - V_\beta(x'')| \le T' C_1. \tag{2.11}$$

Let $\beta_n \to 1$ be a sequence in $(0,1)$. Then, (2.9)–(2.11) imply that, by a suitable diagonalization, we may pick a subsequence $\{\beta_n\}$ (denoting it again by $\{\beta_n\}$) along which $V_{\beta_n}(x) - V_{\beta_n}(x_\phi), x \in X$, and $(1-\beta_n)V_{\beta_n}(x_\phi)$ converge to some limits $W_0(x)$ and λ_0, respectively.

Set $\bar{V}_\beta(x) := V_\beta(x) - V_\beta(x_\phi), x \in X$. Then, $\bar{V}_\beta(x_\phi) = 0$ and

$$(1-\beta)V_\beta(x_\phi) + \bar{V}_\beta(x) = \max_{w\in W}\left[\beta\bar{V}_\beta(b(x,w)) + \ell(x) - \gamma^2\theta(w)\right].$$

Passing to the limit $\beta_n \to 1$, we get

$$\lambda_0 + W_0(x) = \max_{w\in W}[W_0(b(x,w)) + \ell(x) - \gamma^2\theta(w)]. \qquad \square$$

Remark 2.4. The number λ_0 in the above theorem is unique, as follows from the next theorem. Regarding the uniqueness of the function W_0 (up to an additive constant), at the end of this subsection we prove that if $\gamma > \gamma^*$, for γ^* being the H_∞-norm, then this holds. For $\gamma \le \gamma^*$ we still do not have any uniqueness result.

The equation (2.6) corresponds to the dynamic programming equation of the following average cost deterministic optimal control problem. The dynamic is given by

$$x_{t+1} = b(x_t, w_t), \quad t = 0, 1, \ldots; \quad x_0 = x, \tag{2.12}$$

where the disturbances $\underline{w} = \{w_t\} \in W[0,\infty)$ play the role of a maximizing control. The cost per stage is $(x,w) \to \ell(x) - \gamma^2\theta(w)$, and the cost functional we try to maximize is given by

$$J^{\underline{w}}(x) = \limsup_{T\to\infty} \frac{1}{T}\sum_{t=0}^{T-1}[\ell(x_t) - \gamma^2\theta(w_t)].$$

The next theorem is a straightforward application of standard dynamic programming arguments.

THEOREM 2.5. *For any* $x \in X$,

$$\lambda_0 = \sup_{\underline{w} \in W[0,\infty)} J^{\underline{w}}(x),$$

and an optimal control is $w_t^* = w^*(x_t)$, *where* $w^*(x)$ *achieves the maximum in* (2.6).

The link between the above optimal control problem and the H_∞-norm of the system Σ is given in the next theorem.

THEOREM 2.6. *Assume that* (A.1)–(A.3) *hold. Then* $\lambda_0 = 0$ *if and only if* $\|\Sigma\|_{H_\infty} \leq \gamma$.

Proof. Assume $\lambda_0 = 0$. Let W_0 be the function defined in Proposition 2.3. Since this function satisfies (2.6), it solves in particular the first part of (2.5′), and the second part follows from the construction of W_0. To prove that W_0 is nonnegative, let $T > 0$ and $\underline{w} \in W[0,T]$ with $w_t = w_\phi$ for all $t = 0, 1, \ldots, T$. Then, in view of (2.6), it follows that

$$W_0(x) \geq \sum_{t=0}^{T} \ell(x_t) + W_0(x_{T+1}),$$

where x_t evolves according to the dynamic (2.12) (or equivalently (2.1)) with initial condition $x_0 = x$. However, from (A3), $x_T = x_\phi$ for $T \geq N_0$. Hence

$$W_0(x) \geq W_0(x_\phi) = 0.$$

Conversely, assume (2.5′). Then, for any $T > 0$ and $\underline{w} \in W[0,T]$,

$$\sum_{t=0}^{T} [\ell(x_t) - \gamma^2 \theta(w_t)] \leq W_0(x_\phi) - W_0(x_{T+1}) \leq 0,$$

where the state dynamics (2.12) start at the initial condition $x_0 = x_\phi$. This implies that

$$\limsup_{T\to\infty} \frac{1}{T} \sum_{t=0}^{T-1} [\ell(x_t) - \gamma^2 \theta(w_t)] \leq 0.$$

Then, from Theorem 2.5, $\lambda_0 \leq 0$. However, in view of Proposition 2.3, $\lambda_0 \geq 0$. Hence, $\lambda_0 = 0$. □

For the rest of this subsection let us make explicit the dependence on γ of λ_0 and W_0 in Proposition 2.3, denoting them by λ_0^γ and W_0^γ, respectively.

Define the set

$$\Gamma = \{\gamma : \lambda_0^\gamma = 0\}.$$

Note that this set is not empty, since for γ great enough, the optimal control w^* in Theorem 2.5 is such that $w^*(x_\phi) = w_\phi$, and in view of (A1), $\lambda_0^\gamma = 0$. Note also that, from Theorem 2.5, λ_0^γ is a nonincreasing function of γ. Let $\gamma^* = \inf \Gamma$. Then, in view of the above facts, $\Gamma = [\gamma^*, \infty)$, and by Theorem 2.6,

$$\gamma^* = \|\Sigma\|_{H_\infty}. \tag{2.13}$$

PROPOSITION 2.7. *For any* $\gamma > \gamma^*$ *the function* W_0^γ, *with* $W_0^\gamma(x_\phi) = 0$, *is the unique solution of* (2.6).

Proof. Let $\gamma > \gamma^*$ be arbitrary but fixed, and define the value function

$$\tilde{W}_0(x) = \sup_{\underline{w} \in \tilde{W}[0,\infty)} \left\{ \sum_{t=0}^{\infty} \left[\ell(x_t) - \gamma^2 \theta(w_t) \right] \right\}, \ x \in X.$$

Here x_t obeys the dynamic (2.12), and $\tilde{W}[0,\infty) = \{\underline{w} \in W[0,\infty) : \ w_t = w_\phi$ except for finitely many $t\}$. Then, by well-known dynamic programming methods (see, e.g., [3]), it follows that $\tilde{W}_0$ is a solution of (2.6) with $\lambda_0 = 0$.

We claim that $\tilde{W}_0 = W_0^\gamma$. Actually, by (2.6) with $\lambda_0 = 0$ and the definition of $\tilde{W}_0$, it follows immediately that $W_0^\gamma \geq \tilde{W}_0$. Thus, we shall just prove the reverse inequality, i.e., that $W_0^\gamma(x) \leq \tilde{W}_0(x) \ \forall \ x \in X$. Let $x \in X$ and $\gamma_1 \in (\gamma^*, \gamma)$. Then, (2.6) yields

$$W_0^{\gamma_1}(x) \geq \sum_{t=0}^{T-1} \left[\ell(x_t) - \gamma_1^2 \theta(w_t) \right] + W_0^{\gamma_1}(x_T) \text{ for all } \underline{w} \in W[0,\infty), T > 0.$$

On the other hand, if $\underline{w}^*$ is the optimal control defined in Theorem 2.5, then

$$\begin{aligned} W_0^\gamma(x) &= \sum_{t=0}^{T-1} \left[\ell(x_t^*) - \gamma^2 \theta(w_t^*) \right] + W_0^\gamma(x_T^*) \\ &= \sum_{t=0}^{T-1} \left[\ell(x_t^*) - \gamma_1^2 \theta(w_t^*) \right] - (\gamma^2 - \gamma_1^2) \sum_{t=0}^{T-1} \theta(w_t^*) + W_0^\gamma(x_T^*) \end{aligned} \tag{2.14}$$

for all $T > 0$. Therefore,

$$\begin{aligned} \sum_{t=0}^{T-1} \theta(w_t^*) &\leq \frac{W_0^{\gamma_1}(x) - W_0^\gamma(x) + W_0^\gamma(x_T^*)}{\gamma^2 - \gamma_1^2} \\ &\leq \frac{C}{\gamma^2 - \gamma_1^2} \end{aligned} \tag{2.15}$$

for some suitable constant C.

Thus, in view of (2.3), (2.15) implies that $\underline{w}^* \in \tilde{W}[0,\infty)$, and by (A3), if T is large enough, $x_t^* = x_\phi$ for $t \geq T$. Finally, the above facts and (2.14) imply that $W_0^\gamma(x) \leq \tilde{W}_0(x)$. This completes the proof. ∎

2.2. Stochastic perturbation. In this subsection we define a finite state Markov chain, which represents a stochastic perturbation of the system (2.1). This model has been described in [1].

Throughout this subsection we assume (A2). However, we will need (A1) and (A3) for the small noise limit analysis in section 3.

Let $V : X \times X \to \mathbb{R} \cup \{+\infty\}$ be the function defined by

$$V(x, x'') = \min\{\theta(w) : x'' = b(x,w)\},$$

with the standard convention that the minimum over an empty set equals $+\infty$. The value $V(x, x'')$ represents the minimum "magnitude" associated with the disturbances (see (2.3)) to go from x to x'' in one time step.

Let us define the following stochastic matrix Π_ε: given $x, x'' \in X$,

$$\Pi_\varepsilon(x, x'') = \frac{1}{Z_\varepsilon(x)} e^{-\frac{V(x,x'')}{\varepsilon}},$$

where $\varepsilon > 0$ is a small noise parameter and $Z_\varepsilon(x)$ is a normalizing constant satisfying $\sum_{x'' \in X} \Pi_\varepsilon(x, x'') = 1$.

This stochastic matrix satisfies the consistency condition

$$\lim_{\varepsilon \to 0} \Pi_\varepsilon(x, x'') = \begin{cases} 1 \text{ if } x'' = f(x), \\ 0 \text{ otherwise.} \end{cases}$$

Remark 2.8. Note that (A2) implies that Π_ε is irreducible. Remember that an $N \times N$ nonnegative matrix M is irreducible if, for every $x, x'' \in X$, there exist $T > 0$ such that $M^T(x, x'') > 0$, where M^T denotes the T-power of M.

DEFINITION 2.9. *The risk-sensitive index for* Π_ε *is defined by*

$$\lambda_\varepsilon = \lim_{T \to \infty} \frac{\varepsilon}{\mu} \cdot \frac{1}{T} \log E_x \exp\left\{ \frac{\mu}{\varepsilon} \sum_{t=0}^{T-1} \ell(x_t) \right\}, \tag{2.16}$$

where $\mu > 0$ *has the role of a risk-averse factor.*

The existence of the limit in (2.16) is implied by Sanov's theorem (see, e.g., [4]), and it coincides with the optimal value of an average cost infinite horizon optimal control problem, which we will define later in this section. Keeping this in mind, we shall prove that $e^{\frac{\mu}{\varepsilon}\lambda_\varepsilon}$ is the dominant eigenvalue of the nonnegative matrix defined by

$$L(x, x'') = e^{\frac{\mu}{\varepsilon}\ell(x)} \Pi_\varepsilon(x, x'') \quad \text{for} \quad x, x'' \in X.$$

Note that, since Π_ε is irreducible, so is L.

Remark 2.10 (notation). If M is any matrix on X and $h : X \to \mathbb{R}$ is any function, we denote by Mh their product, that is,

$$Mh(x) = \sum_{x'' \in X} M(x, x'') h(x'') \quad \text{for each} \quad x \in X.$$

On the other hand, given $x \in X$, the x-row vector of M is denoted by $M(x)$, i.e., $M(x) = (M(x, x_1), \ldots, M(x, x_N))$.

THEOREM 2.11. *There exist* $\alpha_\varepsilon > 0$ *and a unique strictly positive function* $\Psi_\varepsilon : X \to \mathbb{R}$, *with* $\Psi_\varepsilon(x_\phi) = 1$, *such that*

$$\alpha_\varepsilon \Psi_\varepsilon = L \Psi_\varepsilon. \tag{2.17}$$

Further,

$$\lambda_\varepsilon = \frac{\varepsilon}{\mu} \log \alpha_\varepsilon. \tag{2.18}$$

Proof. The first part follows from the Perron–Frobenios theorem; see, e.g., [16]. To prove the rest, let x_t be the Markov chain governed by Π_ε with initial condition $x_0 = x$.

Thus, (2.17) yields

$$\begin{aligned} E_x \exp\{\frac{\mu}{\varepsilon} \sum_{t=0}^{T-1} \ell(x_t)\} &= E_x \prod_{t=0}^{T-1} \alpha_\varepsilon \frac{\Psi_\varepsilon(x_t)}{\Pi_\varepsilon \Psi_\varepsilon(x_t)} \\ &= \alpha_\varepsilon^T E_x \prod_{t=0}^{T-1} \left[\frac{\Psi_\varepsilon(x_t)}{\Pi_\varepsilon \Psi_\varepsilon(x_t)} \right]. \end{aligned} \tag{2.19}$$

Using the Markovian property of x_t, we have

$$E_x\left[\prod_{t=0}^{T-1}\frac{\Psi_\varepsilon(x_t)}{\Pi_\varepsilon\Psi_\varepsilon(x_t)}\cdot\Pi_\varepsilon\Psi_\varepsilon(x_{T-1})\right]=\Psi_\varepsilon(x). \tag{2.20}$$

Since Ψ_ε is strictly positive, so is $\Pi_\varepsilon\Psi_\varepsilon$, and in view of (2.20), there follows the existence of suitable positive constants K_1 and K_2 such that

$$K_1\le\frac{\Psi_\varepsilon(x)}{\max_{x\in X}\Pi_\varepsilon\Psi_\varepsilon(x)}\le E_x\left[\prod_{t=0}^{T-1}\frac{\Psi_\varepsilon(x_t)}{\Pi_\varepsilon\Psi_\varepsilon(x_t)}\right]\le\frac{\Psi_\varepsilon(x)}{\min_{x\in X}\Pi_\varepsilon\Psi_\varepsilon(x)}\le K_2. \tag{2.21}$$

Therefore, from (2.19)–(2.21) we have

$$\frac{\varepsilon}{\mu}\log\alpha_\varepsilon=\lim_{T\to\infty}\frac{\varepsilon}{\mu}\cdot\frac{1}{T}\log E_x\exp\left\{\frac{\mu}{\varepsilon}\sum_{t=0}^{T-1}\ell(x_t)\right\}.$$

This completes the proof. □

Let us define $W_\varepsilon(x)=\frac{\varepsilon}{\mu}\log\Psi_\varepsilon(x), x\in X$, and rewrite (2.17) as

$$\lambda_\varepsilon+W_\varepsilon(x)=\frac{\varepsilon}{\mu}\log\Pi_\varepsilon e^{\frac{\mu}{\varepsilon}W_\varepsilon(x)}+\ell(x)\text{ for all }x\in X. \tag{2.22}$$

In order to transform (2.22) into the dynamic programming equation of some ergodic cost optimal control problem, we now introduce the relative entropy function. Let $P(X)$ be the set of probability vectors on X, i.e.,

$$P(X)=\left\{\pi=(\pi_1,\dots,\pi_N):\pi_i\ge0,\sum_{i=1}^{N}\pi_i=1\right\}.$$

Let us fix $\nu\in P(X)$. We define the relative entropy function $I(\cdot\|\nu):P(X)\to\mathbb{R}\cup\{+\infty\}$ by

$$I(\pi\|\nu)=\begin{cases}\displaystyle\sum_{x''\in X}\log[r(x'')]\pi(x'') & \text{if }\pi<<\nu,\\ +\infty & \text{otherwise,}\end{cases}$$

where

$$r(x'')=\begin{cases}\frac{\pi(x'')}{\nu(x'')} & \text{if }\nu(x'')>0,\\ 1 & \text{otherwise.}\end{cases}$$

The next lemma is proved in an appendix at the end of the paper. Actually, it is a particular case of Proposition II.4.2 in [5].

LEMMA 2.12. *The pair $\lambda_\varepsilon, W_\varepsilon$ satisfies the equation*

$$\lambda_\varepsilon+W_\varepsilon(x)=\sup_{\pi\in P(X)}\left\{\sum_{x''\in X}W_\varepsilon(x'')\pi(x'')+\ell(x)-\frac{\varepsilon}{\mu}I(\pi\|\Pi_\varepsilon(x))\right\}. \tag{2.23}$$

Moreover, the supremum in the right-hand side (r.h.s.) is attained at the unique probability vector $\eta^*(x)$ defined by

$$\eta^*(x,x'')=\frac{e^{\frac{\mu}{\varepsilon}W_\varepsilon(x'')}}{\Pi_\varepsilon e^{\frac{\mu}{\varepsilon}W_\varepsilon(x)}}\Pi_\varepsilon(x,x''),\ x''\in X. \tag{2.24}$$

Now we describe the stochastic optimal control problem associated with the dynamic programming equation (2.23). The set $P(X)$ plays the role of a control set, and we denote by Q the set of stationary controls, i.e., the set of functions $\eta : X \to P(X)$. Actually, the set Q can be identified with the set of stochastic matrices on X, which we denote again by Q.

Given a stationary control $\eta \in Q$ and an initial condition $x_0 = x$, the controlled process is the Markov chain x_t with stochastic matrix η and initial condition $x_0 = x$. On the other hand, $(x, \eta(x)) \to \ell(x) - \frac{\varepsilon}{\mu} I(\eta(x) \| \Pi_\varepsilon(x))$ is the reward per stage. Note that we are allowing rewards equal to $-\infty$. Finally, the associated reward functional is

$$J^\eta(x) = \limsup_{T\to\infty} \frac{1}{T} \sum_{t=0}^{T-1} E_x^\eta \left[\ell(x_t) - \frac{\varepsilon}{\mu} I(\eta(x_t) \| \Pi_\varepsilon(x_t)) \right].$$

Then, by standard dynamic programming methods, it follows that, for any $x \in X$,

$$\lambda_\varepsilon = \sup_{\eta \in Q} J^\eta(x),$$

and the optimal control is given by (2.24).

3. Small noise limit. In order to relate the H_∞-norm of the system Σ and the risk-sensitive index λ_ε, we will take the limit when the noise intensity ε goes to zero in (2.23). Therefore, we are first concerned about the existence of a limit point of the family $\{\lambda_\varepsilon, W_\varepsilon\}$; that is, we want to prove the existence of a sequence $\{\varepsilon_n\}$, with $\varepsilon_n \to 0$, and a pair λ_0, W_0 such that

$$\begin{cases} \lim\limits_{\varepsilon_n \to 0} \lambda_{\varepsilon_n} = \lambda_0, \\ \lim\limits_{\varepsilon_n \to 0} W_{\varepsilon_n}(x) = W_0(x) \text{ for all } x \in X. \end{cases} \tag{3.1}$$

Since X is finite, in order to get (3.1) we just need to prove that the family $\{\lambda_\varepsilon, W_\varepsilon\}$ is uniformly bounded.

By (2.16), we have

$$0 \le \lambda_\varepsilon \le \|\ell\|. \tag{3.2}$$

So it just remains to prove that $\{W_\varepsilon\}$ is uniformly bounded.

Throughout this section we assume the following condition.

(A4) There exists a positive integer T_2 such that $\Pi_\varepsilon^{T_2}(x, x'') > 0$ for all $x, x'' \in X$.

Remark 3.1. Note that (A1)–(A2) imply that the stochastic matrix Π_ε is aperiodic irreducible. In particular, (A1)–(A2) imply (A4).

THEOREM 3.2. *Let W_ε be the solution of* (2.23) *satisfying the normalizing condition $W_\varepsilon(x_\phi) = 0$. Then, for any $\varepsilon_0 > 0$, there exists a constant K_1 such that for $\varepsilon < \varepsilon_0$*

$$\|W_\varepsilon\| \le K_1.$$

Proof. Let x_t be the Markov chain with stochastic matrix Π_ε with initial condition $x_0 = x$, and let $T > 0$ be arbitrary.

Let us define

$$e^{\frac{\mu}{\varepsilon} V_\varepsilon(x,T)} := E_x e^{\frac{\mu}{\varepsilon} \sum_{t=0}^{T-1} \ell(x_t) + \frac{\mu}{\varepsilon} W_\varepsilon(x_T)}.$$

Then, one can prove by induction that

$$e^{\frac{\mu}{\varepsilon}V_\varepsilon(x,T)} = L^T e^{\frac{\mu}{\varepsilon}W_\varepsilon(x)}, \tag{3.3}$$

where L^T is the T-power of the matrix L; see Theorem 2.11. Indeed, by (2.17)

$$L^T e^{\frac{\mu}{\varepsilon}W_\varepsilon(x)} = e^{\frac{\mu}{\varepsilon}[T\lambda_\varepsilon + W_\varepsilon(x)]}.$$

Substituting this in (3.3), and making the logarithmic transformation on both sides, we get

$$V_\varepsilon(x,T) - T\lambda_\varepsilon = W_\varepsilon(x).$$

Since x was chosen arbitrarily, in particular we have

$$V_\varepsilon(x,T) - V_\varepsilon(x_\phi,T) = W_\varepsilon(x). \tag{3.4}$$

Now we shall get a uniform bound for $x \to V_\varepsilon(x,T) - V_\varepsilon(x_\phi,T)$. Let $x, x'' \in X$, and let ν_0 ($\tilde{\nu}_0$) be the distribution of x_{T_2} with initial condition $x_0 = x$ ($x_0 = x''$, respectively), with T_2 as in (A4). Then, in view of the definition of the stochastic matrix Π_ε,

$$\frac{1}{N^{T_2}} e^{\frac{-T_2\|\theta\|}{\varepsilon}} \le \tilde{\nu}_0(y) \le 1 \text{ for all } y \in X,$$

and therefore,

$$\frac{\nu_0(y)}{\tilde{\nu}_0(y)} \le e^{\frac{T_2\|\theta\|}{\varepsilon}} N^{T_2} \text{ for all } y \in X.$$

Thus,

$$\begin{aligned} e^{\frac{\mu}{\varepsilon}V_\varepsilon(x,T)} &\le e^{\frac{\mu}{\varepsilon}T_2\|\ell\|} E_{\tilde{\nu}_0}\left[e^{\frac{\mu}{\varepsilon}\sum_{t=T_2}^{T-1}\ell(x_t) + \frac{\mu}{\varepsilon}W_\varepsilon(x_T)} \cdot \frac{\nu_0(x_{T_2})}{\tilde{\nu}_0(x_{T_2})}\right] \\ &\le N^{T_2} e^{\frac{1}{\varepsilon}T_2(\mu\|\ell\| + \|\theta\|)} \cdot e^{\frac{\mu}{\varepsilon}V_\varepsilon(x'',T)} \text{ for all } T > T_2. \end{aligned}$$

Then,

$$V_\varepsilon(x,T) \le V_\varepsilon(x'',T) + T_2\left(\|\ell\| + \frac{1}{\mu}\|\theta\|\right) + \frac{\varepsilon}{\mu}T_2 \log N,$$

and therefore, given any $\varepsilon_0 > 0$ arbitrary but fixed,

$$V_\varepsilon(x,T) - V_\varepsilon(x'',T) \le T_2\left(\|\ell\| + \frac{1}{\mu}\|\theta\| + \frac{\varepsilon_0}{\mu}\log N\right)$$

for $\varepsilon < \varepsilon_0$. Hence, interchanging the roles of x and x'', we get

$$|V_\varepsilon(x,T) - V_\varepsilon(x'',T)| \le T_2\left(\|\ell\| + \frac{1}{\mu}\|\theta\| + \frac{\varepsilon_0}{\mu}\log N\right). \tag{3.5}$$

Therefore, in view of (3.4)–(3.5), taking $K_1 = T_2(\|\ell\| + \frac{1}{\mu}\|\theta\| + \frac{\varepsilon_0}{\mu}\log N)$ and $x'' = x_\phi$, the theorem follows. □

THEOREM 3.3. *There exist a number $\lambda_0 \geq 0$ and a function $W_0 : X \to \mathbb{R}$ limit point of the family $\{(\lambda_\varepsilon, W_\varepsilon)\}$, such that*

$$\lambda_0 + W_0(x) = \max_{w \in W} \left\{ W_0(b(x,w)) + \ell(x) - \frac{1}{\mu}\theta(w) \right\}. \tag{3.6}$$

Proof. Estimate (3.2) and Theorem 3.2 imply the existence of a limit point (λ_0, W_0) of the family $(\lambda_\varepsilon, W_\varepsilon)$ through a sequence $\varepsilon_n \to 0$. Now we rewrite (2.22) as follows:

$$\lambda_\varepsilon + W_\varepsilon(x) = \frac{\varepsilon}{\mu} \log \sum_{x'' \in X} e^{\frac{\mu}{\varepsilon} W_\varepsilon(x'')} \cdot \frac{e^{-\frac{1}{\varepsilon}V(x,x'')}}{Z_\varepsilon(x)} + \ell(x). \tag{3.7}$$

Using a version of the Laplace–Varadhan lemma (see the appendix), it follows that the r.h.s. of (3.7) converges to

$$\sup_{w \in W} \left\{ W_0(b(x,w)) + \ell(x) - \frac{1}{\mu}\theta(w) \right\} \text{ as } \varepsilon_n \to 0.$$

Thus, letting $\varepsilon = \varepsilon_n$ and $\varepsilon_n \to 0$ in (3.7), we get that the pair (λ_0, W_0) solves (3.6). □

Remark 3.4. Note that (3.6) is the equation (2.6) we had introduced in Proposition 2.3 with $\gamma^2 = \frac{1}{\mu}$. Actually, assuming (A1)–(A3), W_0 is the same as in section 2.1 for μ small enough (by uniqueness); see Proposition 2.7. Note also that uniqueness of $\lambda_0 (= 0)$ and W_0 implies convergence of λ_ε to 0 and $W_\varepsilon(x)$ to $W_0(x)$ as $\varepsilon \to 0$—not just convergence for sequences $\varepsilon_n \to 0$.

4. Risk-sensitive control problem. In this section we set up the state feedback robust control problem. Paralleling the approach of the previous sections, we introduce an infinite horizon risk-sensitive control problem that is solved using the policy iteration algorithm. The optimal rate is interpreted as the upper value of a stochastic dynamic game with average cost per unit time criterion.

4.1. State feedback control problem. Consider the finite state controlled machine defined by

$$x_{t+1} = f(x_t, u_t), \quad t = 0, 1, \ldots; \quad x_0 = x, \tag{4.1}$$

where the state x_t takes values in the finite set X, the control u_t evolves in the finite set U, and $f : X \times U \to X$ is a given function. We recall that N is the number of states in X.

We now define a deterministic perturbation of the system (4.1). Let $b : X \times U \times W \to X$ be the function that defines the dynamics of the system Σ^u given by

$$x_{t+1} = b(x_t, u_t, w_t), \tag{4.2}$$

where x_t and u_t take values in X and U, respectively, and the disturbance w_t takes values in a finite set W.

We assume the following.

(H1) There exist a null control $u_\phi \in U$, an equilibrium state $x_\phi \in X$, and a null disturbance $w_\phi \in W$ such that

(i) $f(x_\phi, u_\phi) = x_\phi$ and (ii) $b(x, u, w_\phi) = f(x, u)$ for all $x \in X, u \in U$.

(H2) Let $\mathcal{U}$ be the *finite* set of all stationary control policies $\tilde{u} : X \to U$. Given $\tilde{u} \in \mathcal{U}$ and $x, x'' \in X$, there exist $T_1,\ 0 < T_1 \leq N$, and $\underline{w} \in W[0, T_1]$ such that for the initial condition $x_0 = x$, the system $\sum^{\tilde{u}}$ reaches x'' after T_1 steps.

Let $\theta : W \to \mathbb{R}$ and $\ell : X \times U \to \mathbb{R}$ be functions such that

$$\begin{cases} \theta(w_\phi) = 0, \\ \theta(w) > 0 \ \text{ for } \ w \neq w_\phi \in W \end{cases}$$

and

$$\begin{cases} \ell(x_\phi, u_\phi) = 0, \\ \ell(x, u) > 0 \ \text{ for } \ (x, u) \neq (x_\phi, u_\phi) \in X \times U. \end{cases} \tag{4.3}$$

The functions θ and ℓ play the same role as in previous sections.

Let $\mathcal{U}_1 \subset \mathcal{U}$ be the subset of stationary policies $\tilde{u}$ such that the following condition is satisfied. For each initial condition $x_0 = x$, there exists a positive integer N_0 such that the system (4.1) reaches the equilibrium state x_ϕ after N_0 steps, and $\tilde{u}(x_\phi) = u_\phi$. Note that, given $\tilde{u} \in \mathcal{U}$ ($\tilde{u} \in \mathcal{U}_1$), letting $f^{\tilde{u}}(x) = f(x, \tilde{u}(x))$, with $b^{\tilde{u}}(x, w)$ and $\ell^{\tilde{u}}(x)$ defined similarly, then (A2) ((A1) and (A3), respectively) of section 2 holds, with f, b replaced by $f^{\tilde{u}}, b^{\tilde{u}}$. Indeed, for $\tilde{u} \in \mathcal{U}_1$, the H_∞-norm $\| \sum^{\tilde{u}} \|_{H_\infty}$ is defined (see (2.13)).

The state feedback robust control problem is the following; see, e.g., [2]. Given $\gamma > 0$, find a control $\tilde{u} \in \mathcal{U}_1$ such that for each initial condition $x_0 = x$, there exists a nonnegative constant $K(x)$, with $K(x_\phi) = 0$, satisfying

$$K(x) + \sum_{t=0}^{T} \left[\gamma^2 \theta(w_t) - \ell(x_t)\right] \geq 0 \ \text{ for all } \ \underline{w} \in W[0, T], T > 0.$$

In other words, given $\gamma > 0$ we want to find $\tilde{u} \in \mathcal{U}_1$ such that $\|\Sigma^{\tilde{u}}\|_{H_\infty} \leq \gamma$.

Following the same arguments as in section 2, we deduce that the existence of a nonnegative function $W_0 : X \to \mathbb{R}$ such that

$$\begin{cases} W_0(x) \geq \min\limits_{u \in U} \ \max\limits_{w \in W} \{W_0(b(x, u, w)) + \ell(x, u) - \gamma^2 \theta(w)\}, \\ W_0(x_\phi) = 0 \end{cases}$$

is a necessary and sufficient condition for the existence of solution to the state feedback robust control problem. Actually, the solution can be characterized in terms of the value of an average cost dynamic game, as we shall see later in this section.

The proof of the next proposition is structurally similar to the one given for Proposition 2.3, and we sketch it in the appendix at the end of the paper.

PROPOSITION 4.1. *If* (H2) *holds, then there exist a nonnegative number* λ_0 *and a function* $W_0 : X \to \mathbb{R}$ *such that*

$$\lambda_0 + W_0(x) = \min_{u \in U} \max_{w \in W} \left[W_0(b(x, u, w)) + \ell(x, u) - \gamma^2 \theta(w)\right]. \tag{4.4}$$

COROLLARY 4.2. *Let* $\tilde{u}^* \in \mathcal{U}$ *be a control achieving the minimum in the r.h.s. of* (4.4). *If* (H1)–(H2) *hold, then the following conditions are equivalent:*

(i) $\lambda_0 = 0$,

(ii) $\|\Sigma^{\tilde{u}^*}\| \leq \gamma$.

Proof. The proof of this corollary is the same as that of Theorem 2.6, noting that $\lambda_0 = 0$ implies that $\tilde{u}^* \in \mathcal{U}_1$. To see this, note that, for any $x \in X$ and $T > 0$,

$$\sum_{t=0}^{T} \ell(x_t, \tilde{u}^*(x_t)) \leq W_0(x) - W_0(x_{T+1}),$$

where x_t obeys the dynamic (4.1). Thus, for T great enough, $\ell(x_T, \tilde{u}^*(x_T)) = 0$, and in view of (4.3), $x_T = x_\phi$ and $\tilde{u}^*(x_T) = u_\phi$. □

The equation (4.4) is the Isaacs equation of the following average cost zero-sum dynamic game.

Dynamic game. Consider the difference equation

$$x_{t+1} = b(x_t, u_t, w_t), \quad t = 0, 1, \ldots; \quad x_0 = x.$$

Here $\underline{u} = \{u_t\} \in U[0, \infty)$ and $\underline{w} = \{w_t\} \in W[0, \infty)$, and they play the role of controls for Player 1 (minimizer) and Player 2 (maximizer), respectively. The associated cost functional is

$$J(x; \underline{u}, \underline{w}) = \limsup_{T \to \infty} \frac{1}{T} \sum_{t=0}^{T-1} \left[\ell(x_t, u_t) - \gamma^2 \theta(w_t) \right]. \tag{4.5}$$

We use the following definition of value of a dynamic game [8], which is given in terms of strategies. A strategy $\vec{u}$ for Player 1 consists of a sequence of functions $\bar{u}_0, \bar{u}_1, \ldots$, with values in U such that $\bar{u}_t$ is a function of $x_s, x_t, w_s, 0 \leq s < t$. A strategy $\vec{w}$ for Player 2 is a sequence of functions $\bar{w}_0, \bar{w}_1, \ldots$, with values in W such that $\bar{w}_t$ is a function of x_s, u_s, $0 \leq s \leq t$. We say that a strategy $\vec{u}$ is *stationary feedback* if $\bar{u}_t$ depends only on the current state x_t, i.e., $\bar{u}_t : X \to U$, and $\bar{u}_t = \tilde{u}$ is independent of t. Analogously, $\vec{w}$ is *stationary feedback* if $\bar{w}_t$ depends only on the current state x_t and the current control u_t of Player 1, and $\bar{w}_t = \tilde{w} : X \times U \to W$ is independent of t.

Given a pair of strategies $(\vec{u}, \vec{w})$ and the initial condition $x_0 = x$, the controls for Player 1 and Player 2 are generated recursively as

$$u_0 = \bar{u}_0(x), \ w_0 = \bar{w}_0(x, u_0), \ u_1 = \bar{u}_1(x, x_1, w_0), \ w_1 = \bar{w}_1(x, x_1, u_0, u_1), \ldots.$$

DEFINITION 4.3. *When there exists a pair of strategies* $(\vec{u}^*, \vec{w}^*)$ *such that*

$$J(x; \vec{u}^*, \vec{w}) \leq J(x, \vec{u}^*, \vec{w}^*) \leq J(x, \vec{u}, \vec{w}^*) \text{ for all } \vec{u}, \vec{w},$$

the value $V(x) = J(x, \vec{u}^*, \vec{w}^*)$ *is called the* value of the game*, and* $(\vec{u}^*, \vec{w}^*)$ *are referred to as* optimal strategies*.*

Remark 4.4. $V(x)$ is often called the upper value of this infinite horizon dynamic game, since the maximizing Player 2 has the advantage of knowing Player 1's choice u_t before choosing w_t. This is also reflected in the order min max (rather than max min) in the Isaacs equation (4.4). For continuous variable differential games, an alternative definition due to Elliott and Kalton is often used. See, e.g., [6]. The discrete time version of the Elliott–Kalton definition is as follows. The minimizing Player 1 chooses any control sequence $\underline{u} \in U[0, \infty)$, while the maximizing Player 2 chooses a map $\xi : U[0, \infty) \to W[0, \infty)$ such that $\xi[\underline{u}]_t$ depends only on $u_0, u_1, \ldots, u_t$. The Elliott–Kalton upper value of our discrete time dynamic game can be easily shown to be the same as the one defined above.

PROPOSITION 4.5. *For every $x \in X$,*

$$\lambda_0 = V(x).$$

Furthermore, the stationary strategies

$$\tilde{u}^*(x) \in \arg\min_{u\in U}\left\{\max_{w\in W}\left[W_0(b(x,u,w)) + \ell(x,u) - \gamma^2\theta(w)\right]\right\}$$

and

$$\tilde{w}^*(x,u) \in \arg\max_{w\in W}\left\{W_0(b(x,u,w)) + \ell(x,u) - \gamma^2\theta(w)\right\}$$

are optimal.

The proof of this proposition is an immediate application of (4.4), and we omit it.

4.2. Risk-sensitive optimal control problem. We regard system (4.1) as a deterministic controlled Markov chain and define a random perturbation as follows. Given $x, x'' \in X$ and $u \in U$ we define

$$V(x,u;x'') = \min\left\{\theta(w) : x'' = b(x,u,w)\right\}.$$

Here the minimum over an empty set is defined as $+\infty$.

For each $u \in U$, we define the stochastic matrix

$$\Pi_\varepsilon^u(x,x'') = \frac{1}{Z_\varepsilon(x,u)}\exp\left\{-\frac{1}{\varepsilon}V(x,u;x'')\right\}, \tag{4.6}$$

where $\varepsilon > 0$ is a noise parameter and $Z_\varepsilon(x,u)$ is a normalizing constant satisfying the condition $\Sigma_{x''\in X}\Pi_\varepsilon^u(x,x'') = 1$.

Throughout this subsection we assume (H2).

For each $\tilde{u} \in \mathcal{U}$ the cost functional (to be minimized) is the infinite horizon exponential growth criterion

$$\lambda_\varepsilon(\tilde{u}) = \lim_{T\to\infty}\frac{\varepsilon}{\mu}\cdot\frac{1}{T}\log E_x\exp\left\{\frac{\mu}{\varepsilon}\sum_{t=0}^{T-1}\ell(x_t,\tilde{u}(x_t))\right\}, \tag{4.7}$$

where $\mu > 0$ is given.

The risk-sensitive optimal control problem is to find a control $\tilde{u}^* \in \mathcal{U}$ that minimizes $\lambda_\varepsilon(\tilde{u})$. Let

$$\Lambda_\varepsilon := \inf_{\tilde{u}\in\mathcal{U}}\lambda_\varepsilon(\tilde{u}).$$

Next we have a verification theorem.

THEOREM 4.6. *Suppose that there exist a number $\alpha > 0$ and a strictly positive function $\Psi : X \to \mathbb{R}$ such that*

$$\alpha\Psi(x) = \min_{u\in U}\left\{e^{\frac{\mu}{\varepsilon}\ell(x,u)}\Pi_\varepsilon^u\Psi(x)\right\} \text{ for all } x \in X.$$

Then $\Lambda_\varepsilon = \frac{\varepsilon}{\mu}\log\alpha$, and the control $\tilde{u}^* \in \mathcal{U}$, with $\tilde{u}^*(x)$ achieving the minimum on the r.h.s., is optimal.

Proof. Let $\tilde{u} \in \mathcal{U}$. Following the same arguments as in the proof of Theorem 2.11, we have

$$\frac{\varepsilon}{\mu} \log \alpha \leq \lambda_\varepsilon(\tilde{u}),$$

with equality for the control $\tilde{u}^*$. □

Now, in order to get an optimal policy, we use the policy iteration algorithm, which is described as follows. Given an arbitrary policy $\tilde{u}_0 \in \mathcal{U}$, we have proved already the existence of a number $\alpha_0 > 0$ and a function $\Psi_0 : X \to \mathbb{R}$ strictly positive (see Theorem 2.11), such that for all $x \in X$

$$\begin{aligned} \alpha_0 \Psi_0(x) &= e^{\frac{\mu}{\varepsilon}\ell(x,\tilde{u}_0(x))} \Pi_\varepsilon^{\tilde{u}_0} \Psi_0(x) \\ &\equiv T_0 \Psi_0(x). \end{aligned}$$

Let $\tilde{u}_1 \in \mathcal{U}$ be defined by

$$\tilde{u}_1(x) \in \arg\min_{u \in U} \left\{ e^{\frac{\mu}{\varepsilon}\ell(x,u)} \Pi_\varepsilon^u \Psi_0(x) \right\}.$$

Calculate α_1 and Ψ_1, and repeat the process. If we reach a point where

$$T_k \Psi_k(x) = \min_{u \in U} [e^{\frac{\mu}{\varepsilon}\ell(x,u)} \Pi_\varepsilon^u \Psi_k(x)] \text{ for all } x \in X,$$

then, according to Theorem 4.6, $\tilde{u}_k$ is optimal, and we stop.

THEOREM 4.7. *The policy iteration generates a finite sequence of controls* $\{\tilde{u}_0, \tilde{u}_1, \ldots, \tilde{u}_m = \tilde{u}^*\}$ *with strictly monotonically decreasing* $\lambda_\varepsilon(\tilde{u}_k)$ *until the iteration reaches a stopping point.*

Proof. Let $\tilde{u}_k$ and $\tilde{u}_{k+1}$ be control policies generated by the policy iteration algorithm, and α_k, Ψ_k (α_{k+1}, Ψ_{k+1}) be the dominant eigenvalue and eigenfunction of T_k (T_{k+1}, respectively). Thus,

$$T_{k+1} \Psi_k \leq \alpha_k \Psi_k (= T_k \Psi_k). \tag{4.8}$$

If $T_{k+1}\Psi_k = \alpha_k \Psi_k$, then

$$T_k \Psi_k(x) = \min_{u \in U} \left\{ e^{\frac{\mu}{\varepsilon}\ell(x,u)} \Pi_\varepsilon^u \Psi_k(x) \right\} \text{ for all } x \in X,$$

and the iteration terminates. In this case, $\tilde{u}_k$ is optimal by Theorem 4.6.

So, assume that there exists some component $x_0 \in X$ such that the inequality (4.8) is strict, i.e.,

$$T_{k+1} \Psi_k(x_0) < \alpha_k \Psi_k(x_0).$$

Then, Theorem 1.6 in [16] implies that $\alpha_{k+1} < \alpha_k$, and therefore,

$$\lambda_k(\tilde{u}_{k+1}) < \lambda_\varepsilon(\tilde{u}_k).$$

Thus, since there exist just a finite number of policies, the iteration will stop after a finite number of steps. □

COROLLARY 4.8. *There exist* $\Lambda_\varepsilon > 0$ *and a function* $W_\varepsilon : X \to \mathbb{R}$ *such that*

$$\exp\left\{ \frac{\mu}{\varepsilon} (\Lambda_\varepsilon + W_\varepsilon(x)) \right\} = \min_{u \in U} \left[e^{\frac{\mu}{\varepsilon}\ell(x,u)} \Pi_\varepsilon^u e^{\frac{\mu}{\varepsilon} W_\varepsilon(x)} \right]. \tag{4.9}$$

Proof. The proof follows immediately taking $e^{\frac{\mu}{\varepsilon}\Lambda_\varepsilon} = \alpha_m$ and $e^{\frac{\mu}{\varepsilon}W_\varepsilon} = \Psi_m$, where m is the step where the iteration finishes. □

Using the variational equality (A.1) in the appendix, we rewrite (4.9) as

$$(4.9') \quad \Lambda_\varepsilon + W_\varepsilon(x) = \min_{u\in U}\max_{\pi\in P(X)}\left\{\sum_{x''\in X} W_\varepsilon(x'')\pi(x'') + \ell(x,u) - \frac{\varepsilon}{\mu}I(\pi\|\Pi_\varepsilon^u(x))\right\}.$$

This equation is the Isaacs equation associated with the following stochastic game with average cost per unit time criterion.

Stochastic dynamic game. Let X be the state space, U be the control set for Player 1 (minimizer), and $P(X)$ (the set of probability vectors on X) be the control set for Player 2 (maximizer). On the other hand, $(x,u,\pi) \to \ell(x,u) - I(\pi\|\Pi_\varepsilon^u(x))$, with $(x,u,\pi) \in X\times U\times P(X)$, is the reward function. Note that the reward function is allowed to take the value $-\infty$.

The system evolves as follows: at each time $t \in \{0,1,\ldots\}$ the state of the system is observed, say $x_t = x \in X$. Then, a control $u_t \in U$ is chosen for Player 1, and $\pi_t \in P(X)$ is chosen for Player 2. (Actually, π_t will turn out to be a conditional probability based on the past up to time t.) Then, a reward $\ell(x_t,u_t) - I(\pi_t\|\Pi_\varepsilon^{u_t}(x_t))$ is earned, and the state of the system moves to the state x_{t+1} according to the probability distribution π_t.

Strategies. For each $t \geq 0$, let H_t and K_t be the set of feasible histories up to time t for Player 1 and Player 2, respectively. That is, $H_0 = X$ and $H_t = (X\times P(X))^t \times X$, while $K_0 = X\times U$ and $K_t = (X\times U)^t \times (X\times U)$. A strategy for Player 1 is a sequence $\vec{u} = \{\bar{u}_t\}$ of functions $\bar{u}_t$ from H_t to U. We say that $\vec{u}$ is *stationary feedback* (or stationary Markov policy) if, for all $t \geq 0$, $\bar{u}_t$ depends only on the current state x_t, and $\bar{u}_t = \tilde{u}$ is independent of t; i.e., $\bar{u}_t \equiv \tilde{u} : X \to U$ for all $t \geq 0$. A strategy for Player 2 is a sequence $\vec{\pi} = \{\bar{\pi}_t\}$ of functions $\bar{\pi}_t$ from K_t to $P(X)$. Analogously, $\vec{\pi}$ is *stationary feedback* (or stationary Markov policy) if for all $t \geq 0$, $\bar{\pi}_t \equiv \tilde{\pi} : X\times U \to P(X)$.

Let $\Omega := X[0,\infty)$ and $\mathcal{B}(\Omega) = \sigma$-field generated by the subsets of the form $A_1\times A_2\times\cdots\times A_T\times X\cdots$, with $A_t \subset X, t = 1,\ldots,T$. A generic element of Ω is an infinite sequence $\omega = (\xi_0,\xi_1,\ldots)$, and the state process is defined as the projection from Ω to X, i.e., $x_t(\omega) = \xi_t$. Given the initial condition $x_0 = x$ and the strategies $\vec{u},\vec{\pi}$ being used, there exists a unique probability measure $P^{\vec{u},\vec{\pi}}$ on $(\Omega,\mathcal{B}(\Omega))$ such that

$$\text{(i)} \quad P^{\vec{u},\vec{\pi}}(x_0 = x) = 1,$$

$$\text{(ii)} \quad P^{\vec{u},\vec{\pi}}(x_{t+1} = \xi_{t+1}|x_t = \xi_t,\ldots,x_0 = \xi_0) = \bar{\pi}_t[\xi_0,u_0,\xi_1,u_1,\ldots,\xi_t,u_t](\xi_{t+1}).$$

We denote by $E_x^{\vec{u},\vec{\pi}}$ the corresponding expectation operator. Finally, we observe that, under the action of stationary strategies, the state process $\{x_t\}_{t=0}^\infty$ is a Markov chain with stationary transition matrix. Now let us define the associated cost functional

$$J(x,\vec{u},\vec{\pi}) = \limsup_{T\to\infty}\frac{1}{T}\sum_{t=0}^{T-1} E^{\vec{u},\vec{\pi}}\left[\ell(x_t,u_t) - \frac{\varepsilon}{\mu}I(\pi_t\|\Pi_\varepsilon^{u_t}(x_t))\right].$$

The same definition of value given for deterministic dynamic games is applied for stochastic dynamic games; see Definition 4.3. We denote by $V_\varepsilon(x)$ the value of the game.

THEOREM 4.9. *Let Λ_ε and W_ε be as in (4.9′). Then, for every $x \in X$,*

$$\Lambda_\varepsilon = V_\varepsilon(x).$$

Furthermore, the stationary strategies $\tilde{u}^*$ and $\tilde{\pi}^*$, with

$$\tilde{u}^*(x) \in \arg\min_{u\in U}\left[e^{\frac{\mu}{\varepsilon}\ell(x,u)}\Pi^u_\varepsilon e^{\frac{\mu}{\varepsilon}W_\varepsilon(x)}\right]$$

and

$$\tilde{\pi}^*[x,u](x'') = \frac{e^{\frac{\mu}{\varepsilon}W_\varepsilon(x'')}}{\Pi^u_\varepsilon e^{\frac{\mu}{\varepsilon}W_\varepsilon(x)}}\Pi^u_\varepsilon(x,x'')$$

are optimal.

The proof of this theorem is based on standard dynamic programming arguments and the variational equation (A.1) in the Appendix.

Remark 4.10. A stochastic analogue of the Elliott–Kalton definition of strategy for differential games has been given in [12]. In order to make an analogous definition here, it would be necessary to model the disturbances w_t as exogenous random inputs (in analogy with the Brownian motions, which drive the stochastic differential game dynamics). This could be done. However, following [1], we have instead modelled the stochastic effects via conditional state probabilities.

4.3. Small noise limit and deterministic dynamic game. Paralleling section 3, in this section we relate the state feedback robust control problem and the risk-sensitive optimal control problem taking small noise limit. First, we introduce the following condition, which is analogous to (A4) in section 3.

(H4) For each $\tilde{u} \in \mathcal{U}$, there exists a positive integer T_2 such that all the entries of the T_2-power of $\Pi^{\tilde{u}}_\varepsilon$ are strictly positive.

THEOREM 4.11. *If* (H4) *holds, then there exist a number $\lambda_0 \geq 0$ and a function $W_0 : X \to \mathbb{R}$ limit point of the family $\{\lambda_\varepsilon, W_\varepsilon\}$, such that*

$$\lambda_0 + W_0(x) = \min_{u\in U}\max_{w\in W}\left\{W_0(b(x,u,w)) + \ell(x,u) - \frac{1}{u}\theta(w)\right\}.$$

Sketch of proof. The proof of this theorem will be reduced to the one given for Theorem 3.3. Let $\{\varepsilon_n\}$ be a sequence converging to zero as $n \to \infty$, and denote by $\tilde{u}^*_{\varepsilon_n}$ the optimal policy defined in Theorem 4.9. Since there exist just a finite number of policies, there exist a subsequence of $\{\varepsilon_n\}$, which we denote again as $\{\varepsilon_n\}$, and a policy $\tilde{u}^* = \tilde{u}^*_{\varepsilon_n}$ independent of n. Then, in view of (H4), the same arguments used in the proof of Theorem 3.3 can be applied again. □

THEOREM 4.12. *Assume* (H2) *and that $\mathcal{U}_1$ is not empty. Then there exists $\tilde{\mu}^*$ with the following property. For $\mu < \tilde{\mu}^*$ there exist a sequence $\{\varepsilon_n\}$, with $\varepsilon_n \to 0$ as $n \to \infty$, and a policy $\tilde{u}^* \in \mathcal{U}_1$ not depending on n, such that*

$$\Lambda_{\varepsilon_n} = \lambda_{\varepsilon_n}(\tilde{u}^*) \text{ for } n = 1, 2, \ldots.$$

Furthermore,

$$\lim_{n\to\infty}\Lambda_{\varepsilon_n} = 0.$$

Proof. Let $\tilde{u} \in \mathcal{U}_1$. Then (2.13) implies the existence of $\tilde{\mu}^*$ such that $\lim_{\varepsilon\to 0}\lambda_\varepsilon(\tilde{u}) = 0$ for $\mu < \tilde{\mu}^*$. Therefore, since $\lambda_\varepsilon(\tilde{u}) \geq \Lambda_\varepsilon \geq 0$, letting $\varepsilon \to 0$, we get

$$\lim_{\varepsilon\to 0}\Lambda_\varepsilon = 0. \tag{4.10}$$

Let $\{\varepsilon_n\}$ be a sequence converging to zero as $n \to \infty$, and $\tilde{u}^*_{\varepsilon_n}$ be the optimal policy defined in Theorem 4.9. Since there are just a finite number of policies, there exist a subsequence of $\{\varepsilon_n\}$, which we denote again as $\{\varepsilon_n\}$, and a policy $\tilde{u}^* = \tilde{u}^*_{\varepsilon_n}$ independent of n. Now, to prove that $\tilde{u}^*$ belongs to $\mathcal{U}_1$, we need to verify that (i) given any initial condition $x_0 = x \in X$, there exists a positive integer N_0 such that the system (4.1), with $u_t = \tilde{u}^*(x_t)$, reaches the null state x_ϕ after N_0 steps, and (ii) $\tilde{u}^*(x_\phi) = u_\phi$. We will just prove the first part, since the argument to prove the second one is the same.

So, suppose that $\tilde{u}^*$ does not satisfy (i). Since $\tilde{u}^*$ is optimal for each ε_n,

$$\Lambda_{\varepsilon_n} = \lim_{T\to\infty} \frac{\varepsilon_n}{\mu} \cdot \frac{1}{T} \log E_x \left\{ \frac{\mu}{\varepsilon_n} \sum_{t=0}^{T-1} \ell(x_t, \tilde{u}^*(x_t)) \right\} \text{ for all } n = 1, 2, \ldots.$$

Then, in view of our assumption, and noting that the functions Z_ε and V defined in (4.6) satisfy

$$1 \le Z_\varepsilon \le N \quad \text{and} \quad V(x, u, x'') = 0 \text{ if } x'' = f(x, u),$$

we have

$$E_x \exp \left\{ \frac{\mu}{\varepsilon_n} \sum_{t=0}^{T-1} \ell(x_t, \tilde{u}^*(x_t)) \right\} \ge \left[\frac{e^{\frac{\mu}{\varepsilon_n}\ell^*}}{N} \right]^T,$$

where $\ell^* := \min_{\substack{x \neq x_\phi \in X \\ u \in U}} \ell(x, u)$ and N is the number of states.

Therefore,

$$\frac{1}{T} \frac{\varepsilon_n}{\mu} \log E_x \exp \left\{ \frac{\mu}{\varepsilon_n} \sum_{t=0}^{T-1} \ell(x_t, \bar{u}^*(x_t)) \right\} \ge \ell^* - \frac{\varepsilon_n}{\mu} \log N.$$

Letting $T \to \infty$ and $n \to \infty$, we have

$$\lim_{n\to\infty} \Lambda_{\varepsilon_n} \ge \ell^* > 0,$$

which contradicts (4.10). Thus, $\tilde{u}^*$ satisfies (i). □

COROLLARY 4.13. *If $\mathcal{U}_1$ is not empty and μ is small enough, then there exists a unique nonnegative function $W_0 : X \to \mathbb{R}$, with $W_0(x_\phi) = 0$, such that*

$$W_0 = \min_{u\in U} \max_{w\in W} \left\{ W_0(b(x, u, w)) + \ell(x, u) - \frac{1}{\mu}\theta(w) \right\}.$$

In particular, the control $\tilde{u}^*$, where $\tilde{u}^*$ achieves the minimum in the r.h.s., solves the state feedback robust control problem.

This corollary is a straightforward consequence of Theorems 4.12 and 3.3 and Proposition 2.7.

Appendix.

Proof of Lemma 2.12. From (2.22), it is sufficient to prove that

$$\frac{\varepsilon}{\mu} \log \Pi^\varepsilon e^{\frac{\mu}{\varepsilon} W_\varepsilon(x)} = \sup_{\pi \in P(X)} \left\{ \sum_{x'' \in X} W_\varepsilon(x'')\pi(x'') - \frac{\varepsilon}{\mu} I(\pi \| \Pi_\varepsilon(x)) \right\}, \tag{A.1}$$

and that the supremum on the r.h.s. is achieved at the unique probability vector $\eta^*(x)$ defined by

$$\eta^*(x,x'') = \frac{e^{\frac{\mu}{\varepsilon}W_\varepsilon(x'')}}{\Pi_\varepsilon e^{\frac{\mu}{\varepsilon}W_\varepsilon(x)}}\Pi_\varepsilon(x,x'').$$

Let $\pi \in P(X)$, and fix $x \in X$. Let us assume that π is absolutely continuous with respect to $\Pi_\varepsilon(x)$. Then, since $\Pi_\varepsilon(x)$ is absolutely continuous with respect to $\eta^*(x)$, we have

$$\begin{aligned}
-I(\pi\|\Pi_\varepsilon(x)) \quad + \quad & \frac{\mu}{\varepsilon}\sum_{x''\in X} W_\varepsilon(x'')\pi(x'') = -\sum_{x''\in X}\log\left[\frac{\pi(x'')}{\Pi_\varepsilon(x,x'')}\right]\pi(x'') \\
& +\frac{\mu}{\varepsilon}\sum_{x''\in X} W_\varepsilon(x'')\pi(x'') \\
= \quad & -\sum_{x''\in X}\log\left[\frac{\pi(x'')}{\eta^*(x,x'')}\right]\pi(x'') - \sum_{x''\in X}\log\left[\frac{\eta^*(x,x'')}{\Pi_\varepsilon(x,x'')}\right]\pi(x'') \\
& +\frac{\mu}{\varepsilon}\sum_{x''\in X} W_\varepsilon(x'')\pi(x'') \\
= \quad & -I(\pi\|\eta^*(x)) + \log\Pi_\varepsilon e^{\frac{\mu}{\varepsilon}W_\varepsilon(x)}.
\end{aligned}$$

Noting that $I(\pi\|\eta^*(x)) \geq 0$, and that $I(\pi\|\eta^*(x)) = 0$ if and only if $\pi = \eta^*(x)$, (A.1) follows. □

The next lemma is a version of the Varadhan–Laplace lemma. See [13].

LEMMA A.1. *Let A be a finite set, and F_a^ε, F_a be real valued functions defined on a finite set X such that*

$$\lim_{\varepsilon\to 0}\max_{a\in A}\max_{x\in X}|F_a^\varepsilon(x) - F_a(x)| = 0.$$

Then

$$\lim_{\varepsilon\to 0}\max_{a\in A}\left|\varepsilon\log\sum_{x\in X} e^{\frac{F_a^\varepsilon(x)}{\varepsilon}} - \max_{x\in X}F_a(x)\right| = 0 \tag{A.2}$$

Proof. Define

$$\bar{F}_a^\varepsilon := \max_{x\in X}F_a^\varepsilon(x),\ \ \bar{F}_a := \max_{x\in X}F_a(x).$$

Then

$$\lim_{\varepsilon\to 0}\max_{a\in A}\bar{F}_a^\varepsilon = \bar{F}_a.$$

Therefore,

$$e^{\frac{\bar{F}_a}{\varepsilon}} \leq \sum_{x\in X} e^{\frac{F_a^\varepsilon(x)}{\varepsilon}} \leq Ne^{\frac{\bar{F}_a}{\varepsilon}},$$

where N is the number of elements of X, and (A.2) follows. □

Sketch of proof of Proposition 4.1. Consider the dynamic game described at the end of section 4.2, but with the cost functional given by

$$J(x, \underline{u}, \underline{w}) = \sum_{t=0}^{\infty} \beta^t [\ell(x_t, u_t) - \gamma^2 \theta(w_t)],$$

where β is the discount factor, $0 < \beta < 1$.

Let $V_\beta(x)$ be the value of the game; see Definition 4.3. Then V_β is the unique solution of the Isaacs equation

$$V_\beta(x) = \min_{u \in U} \max_{w \in W} [\beta V_\alpha(b(x, u, w)) + \ell(x, u) - \gamma^2 \theta(w)]. \tag{A.3}$$

Following the same arguments as in the proof of Proposition 2.3, we get

$$\text{(i)} \quad |V_\beta(x) - V_\beta(x_\phi)| < C^1,$$
$$\text{(ii)} \quad 0 \le (1 - \beta) V_\beta(x) \le \|\ell\|$$

for some suitable constant C^1. Then, the above estimates imply the existence of a sequence $\{\beta_n\}$, with $\beta_n \to 1$ as $n \to \infty$, such that $(1 - \beta_n) V_{\beta_n}(x_\phi)$ and $V_{\beta_n}(x) - V_{\beta_n}(x_\phi)$ converge to some limit λ_0 and $W_0(x)$ as $n \to \infty$. Then, rewriting (A.3) as

$$(1 - \beta) V_\beta(x_\phi) + \bar{V}_\beta(x) = \min_{u \in U} \max_{w \in W} [\beta \bar{V}_\beta(b(x, u, w)) + \ell(x, u) - \gamma^2 \theta(u)],$$

with $\bar{V}_\beta(x) := V_\beta(x) - V_\beta(x_\phi)$, and passing the limit $\beta_n \to 1$,

$$\lambda_0 + W_0(x) = \min_{u \in U} \max_{w \in W} [\alpha W_0(b(x, u, w)) + \ell(x, u) - \gamma^2 \theta(w)]. \qquad \square$$

REFERENCES

[1] J. S. Baras and M. R. James, *Robust and risk-sensitive output feedback control for finite state machines and hidden Markov models*, J. Math. Systems Estimation Control, to appear.

[2] T. Basar and P. Bernhard, *H^∞-Optimal Control and Related Minimax Design Problems*, Birkhauser, Boston, 1991.

[3] D. P. Bertsekas, *Dynamic Programming and Stochastic Control*, Academic Press, New York, 1976.

[4] A. Dembo and O. Zeitouni, *Large Deviations Techniques and Applications*, Jones and Bartlett, London, 1993.

[5] P. Dupuis and R. S. Ellis, *A Weak Convergence Approach to the Theory of Large Deviations*, John Wiley, New York, 1997.

[6] L. C. Evans and P. E. Souganidis, *Differential games and representation formulas for solutions of Hamilton-Jacobi-Isaacs equations*, Indiana Univ. Math. J., 33 (1984), pp. 773–797.

[7] E. Fernández-Gaucherand and S. I. Marcus, *Risk sensitive optimal control of hidden Markov models: A case study*, in Proc. 33rd IEEE CDC, Lake Buena Vista, FL, 1994, pp. 1657–1662.

[8] W. H. Fleming, *The Cauchy problem for degenerate parabolic equations*, J. Math. Mech., 13 (1964), pp. 987–1008.

[9] W. H. Fleming and W. M. McEneaney, *Risk Sensitive Control and Differential Games*, Lecture Notes in Control and Info. Sci. 184, Springer-Verlag, New York, 1992, pp. 185–197.

[10] W. H. Fleming and W. M. McEneaney, *Risk sensitive control with ergodic cost criteria*, in Proc. 31st IEEE CDC, Tucson, AZ, 1992.

[11] W. H. Fleming and W. M. McEneaney, *Risk sensitive control on an infinite time horizon*, SIAM J. Control Optim., 33 (1995), pp. 1881–1915.

[12] W. H. Fleming and P. E. Souganidis, *On the existence of value functions of two-player, zero-sum stochastic differential games*, Indiana Univ. Math. J., 38 (1989), pp. 293–314.

[13] M. I. Freidlin and A. D. Wentzell, *Random Perturbations of Dynamical Systems*, Springer-Verlag, New York, 1984.
[14] M. R. James, *Asymptotic analysis of nonlinear stochastic risk-sensitive control and differential games*, Math. Control Signals Systems, 5 (1992), pp. 401–417.
[15] M. R. James, J. S. Baras, and R. J. Elliott, *Risk-sensitive control and dynamic games for partially observed discrete-time nonlinear systems*, IEEE Trans. Automat. Control, AC (1994), pp. 780–792.
[16] E. Seneta, *Non-Negative Matrices and Markov Chains*, Springer-Verlag, New York, 1973.
[17] A. J. van der Schaft, *Nonlinear state space H^∞ control theory*, in Perspectives in Control, H. L. Trentelman and J. C. Willems, eds., Progress in Systems and Control, 2nd ECC, Groningen, the Netherlands, Birkhauser, Basel, 1993.
[18] P. Whittle, *Risk-Sensitive Optimal Control*, John Wiley, New York, 1990.
[19] P. Whittle, *A risk sensitive maximum principle*, Systems Control Lett., 15 (1990), pp. 183–192.

SIAM J. CONTROL OPTIM.
Vol. 35, No. 5, pp. 1811–1827, September 1997

020

WEIGHTED MEANS IN STOCHASTIC APPROXIMATION OF MINIMA*

J. DIPPON† AND J. RENZ‡

Abstract. Weighted averages of Kiefer–Wolfowitz-type procedures, which are driven by larger step lengths than usual, can achieve the optimal rate of convergence. A priori knowledge of a lower bound on the smallest eigenvalue of the Hessian matrix is avoided. The asymptotic mean squared error of the weighted averaging algorithm is the same as would emerge using a Newton-type adaptive algorithm. Several different gradient estimates are considered; one of them leads to a vanishing asymptotic bias. This gradient estimate applied with the weighted averaging algorithm usually yields a better asymptotic mean squared error than applied with the standard algorithm.

Key words. stochastic approximation, acceleration by weighted averaging, weak invariance principle, consistency, Kiefer–Wolfowitz procedure, gradient estimation, optimization

AMS subject classifications. Primary, 60L20; Secondary, 60F05, 60F17, 93E23

PII. S0363012995283789

1. Introduction.

In stochastic approximation the minimizer ϑ of an unknown regression function $f : \mathbb{R}^d \to \mathbb{R}$ can be estimated by running the recursion

$$X_{n+1} = X_n - a_n Y_n, \tag{1.1}$$

where Y_n is a gradient estimate of f at the point X_n and a_n are positive step lengths decreasing to zero. For instance, for $d = 1$ and decreasing span c_n, Kiefer and Wolfowitz [11] used divided differences $Y_n = (Y_{n,1} - Y_{n,2})/(2c_n)$ as approximation of $f'(X_n)$, where $Y_{n,1}$ and $Y_{n,2}$ are error-contaminated observations of $f(X_n + c_n)$ and $f(X_n - c_n)$, respectively. If f is p-times differentiable at ϑ, and if the gradient estimates Y_n are constructed appropriately, one can obtain

$$n^{\frac{\alpha}{2}\left(1-\frac{1}{p}\right)}(X_n - \vartheta) \xrightarrow{\mathcal{D}} N(\mu, K) \quad (n \to \infty)$$

with step lengths $a_n = an^{-\alpha}$ for some $a > 0$ and $\alpha \in (0, 1]$ (see Fabian [8] for $p \geq 3$ odd). Hence, for step lengths $a_n = a/n$, the convergence rate $n^{(1-1/p)/2}$ is obtained. This is the exact minimax order in the problem of estimating the minimizer of f for f belonging to a certain class of p-times differentiable functions (Polyak and Tsybakov [18]).

In this paper we investigate *weighted means*

$$\widetilde{X}_{n,\delta} = \tfrac{1+\delta}{n^{1+\delta}} \sum_{i=1}^{n} i^{\delta} X_i \tag{1.2}$$

of Kiefer–Wolfowitz-type processes (X_n) generated by recursion (1.1) with some gradient estimates Y_n for p-times differentiable regression functions and step lengths converging slower to zero than $1/n$. We obtain

$$n^{\frac{1}{2}\left(1-\frac{1}{p}\right)}\left(\widetilde{X}_{n,\delta} - \vartheta\right) \xrightarrow{\mathcal{D}} N(\widetilde{\mu}, \widetilde{K}) \quad (n \to \infty)$$

*Received by the editors March 27, 1995; accepted for publication (in revised form) July 23, 1996. The research of the second author was supported by a Deutsche Forschungsgemeinschaft grant.
http://www.siam.org/journals/sicon/35-5/28378.html
†Mathematisches Institut A, Universität Stuttgart, 70511 Stuttgart, Germany (dippon@mathematik.uni-stuttgart.de).
‡Landesgirokasse, 70144 Stuttgart, Germany (RiskManagement@t-online.de).

for some weight parameters δ and various types of gradient estimates (Theorems 3.2 and 4.2). The main advantages are the following. First, a priori knowledge of a lower bound on the smallest eigenvalue λ_0 of the Hessian $Hf(\vartheta)$ of f at ϑ is avoided. If, in the standard algorithm with $a_n = a/n$, the constant a is chosen too small, i.e., $a \leq (1 - 1/p)/(2\lambda_0)$, convergence can be very slow. To be safe one might choose a pretty large. But the asymptotic mean squared error (AMSE) produced by the standard algorithm grows approximately linearly in a. These problems do not arise when the averaging algorithm is applied. On the other side, if an asymptotic bias is present, the AMSE of the averaging algorithm cannot be greater than four times the AMSE of the standard algorithm with the optimal, but usually unknown, constant a. In this sense the averaging algorithm can be considered to be more stable than the standard one. Furthermore, the averaging algorithm shows the same limit distribution as the Newton-type adaptive procedure suggested by Fabian [9] (section 5).

The method proposed in this paper is inspired by an idea of Ruppert [21] and Polyak [16], who suggested considering the arithmetic mean of the trajectories of a Robbins–Monro process, which is driven by step lengths slower than $1/n$, too. In this case one obtains the best possible convergence rate and the optimal covariance of the asymptotic distribution in a certain sense [17]. Since then Yin [27], Pechtl [15], Kushner and Yang [13], Györfi and Walk [10], Nazin and Shcherbakov [14], and others have studied this idea.

A further contribution of this paper is a new design to estimate the gradient which leads to a vanishing asymptotic bias $\widetilde{\mu}$ (for $d = 1$ see Renz [19]) regardless of which method (with or without averaging, or with adaptation) is used. Applying the weighted averaging algorithm together with this gradient estimate leads to a second moment of the asymptotic distribution which is minimal within a large class of procedures (relation (5.4)).

Spall [22] introduced another gradient estimate Y_n, the so-called simultaneous gradient perturbation method. It uses only two observations at each step instead of $2d$ observations, as in the standard Kiefer–Wolfowitz method in $\mathbb{R}^d$. This makes it suitable for certain optimization problems in high-dimensional spaces $\mathbb{R}^d$. Taking weighted averages of the process generated with Spall's gradient estimate stabilizes the performance as discussed below (Theorem 4.2 and section 5).

All these central limit theorems require consistency of the stochastic approximation method (Propositions 3.1 and 4.1). To prove the central limit theorems we apply a weak invariance principle stated in Lemma 7.1. Taking weighted averages of the trajectories leads to an accumulation of terms due to the nonlinearity of the regression function. To cope with this effect the assumptions of this lemma are partly stronger than those of a functional central limit theorem for the nonweighted case (see Walk [24]). But fortunately, the additional conditions can be shown to be fulfilled for many stochastic approximation procedures. The assertions of both central limit theorems in this paper can be formulated as invariance principles in the spirit of Lemma 7.1.

As already indicated in Dippon and Renz [4], taking weighted averages of the trajectories works well with the original gradient estimate of Kiefer and Wolfowitz ($p = 3$).

2. Notations. For a d-dimensional Euclidean space the linear space of $d \times d$ matrices is denoted by $\mathcal{L}(\mathbb{R}^d)$. x^* is the transposed vector of $x \in \mathbb{R}^d$, A^* is the adjoint matrix, and $\operatorname{tr} A$ is the trace of $A \in \mathcal{L}(\mathbb{R}^d)$. The tensor product $x \otimes y : \mathbb{R}^d \to \mathbb{R}^d$ is defined by $\langle y, \cdot \rangle x$, where $x, y \in \mathbb{R}^d$ and $\langle \cdot, \cdot \rangle$ is the usual inner product. The space $C([0,1], \mathbb{R}^d)$ of $\mathbb{R}^d$-valued continuous functions on $[0,1]$ is equipped with the maximum

norm. $Hf(\vartheta)$ is the Hessian of a function $f : \mathbb{R}^d \to \mathbb{R}$ at $\vartheta \in \mathbb{R}^d$. For $x \in \mathbb{R}$ we use $\lfloor x \rfloor$ and $\lceil x \rceil$, denoting the integer part of x and the least integer greater than or equal to x, respectively.

Let $(\Omega, \mathcal{A}, P)$ be a probability space. Then a sequence (X_n) of $\mathbb{R}^d$-valued random variables (r.v.'s) is called bounded in probability whenever $\lim_{R\to\infty} \overline{\lim}_n P(\|X_n\| \geq R) = 0$; (X_n) converges to zero almost in L^r or is bounded almost in L^r ($r \in (0,\infty)$) if for each $\varepsilon > 0$ there exists an $\Omega_\varepsilon \in \mathcal{A}$ with $P(\Omega_\varepsilon) \geq 1-\varepsilon$ such that $(\int_{\Omega_\varepsilon} \|X_n\|^r \, dP)^{1/r} = o(1)$ or $= O(1)$, respectively. Convergence almost in L^r implies convergence in probability, but it is weaker than a.s. convergence or convergence in the rth mean.

3. A Kiefer–Wolfowitz procedure with an improved gradient estimate. The Kiefer–Wolfowitz procedure, which finds the minimizer ϑ of a regression function $f : \mathbb{R}^d \to \mathbb{R}$, has been modified by Fabian [6] in such a way that the rate of convergence nearly reaches the rate of a Robbins–Monro procedure if f is assumed to be sufficiently smooth in a neighborhood of ϑ. The method uses multiple observations per step.

We consider here, including the Fabian procedure, a modified Kiefer–Wolfowitz procedure which is given by recursion (1.1). There Y_n is an estimate of the gradient $\nabla f(X_n)$ based on error-contaminated observations of f. It is defined by

$$(3.1)\quad Y_n = c_n^{-1} \sum_{j=1}^{m} v_j \Big(\{ f(X_n + c_n u_j e_i) - V_{n,2j-1}^{(i)} \} - \{ f(X_n - c_n u_j e_i) - V_{n,2j}^{(i)} \} \Big)_{i=1,\dots,d},$$

where the following definitions and relations are used throughout section 3: $m \in \mathbb{N}$, $0 < u_1 < \cdots < u_m \leq 1$, $v_1, \dots, v_m$ are real numbers with $\sum_{j=1}^m v_j u_j^{2i-1} = (1/2)\delta_{1i}$ for all $i = 1, \dots, m$ (as to the existence, compare Fabian [6]), and $c_n = cn^{-\gamma}$ with $c > 0$ and $0 < \gamma < 1/2$. The unit vectors in $\mathbb{R}^d$ are denoted by $e_1, \dots, e_d$.

For future reference, we state the following additional conditions:

(A) ∇f exists on $\mathbb{R}^d$ with $\nabla f(\vartheta) = 0$.

Concerning the local differentiability of f at ϑ we consider two cases. In the first case ($p = 2$) we assume that there exists $\varepsilon > 0$, $\tau \in (0,1]$, K_1 and K_2 such that

(B1a) $Hf(\vartheta)$ exists with $\|\nabla f(x) - Hf(\vartheta)(x-\vartheta)\| \leq K_1 \|x-\vartheta\|^{1+\tau}$ for all $x \in U_\varepsilon(\vartheta)$,
(B1b) $\|\nabla f(x) - \nabla f(y)\| \leq K_2 \|x-y\|$ for all $x, y \in U_\varepsilon(\vartheta)$.

(B1b) holds, for instance, if all second partial derivatives of f exist and are bounded on $U_\varepsilon(\vartheta)$. For the second case ($p \geq 3$), we assume that there exist $\varepsilon > 0$ and L such that

(B2a) derivatives of f up to order $p-1$ exist on $U_\varepsilon(\vartheta)$,
(B2b) the pth derivative of f at ϑ exists,
(B2c) $\|Hf(x) - Hf(y)\| \leq L\|x-y\|$ for all $x, y \in U_\varepsilon(\vartheta)$.

A sufficient condition for (B2c) to hold is that all third partial derivatives of f exist and are bounded on $U_\varepsilon(\vartheta)$.

For brevity, (B1) stands for (B1a) and (B1b), and (B2) for (B2a), (B2b), and (B2c). We use (B) to indicate that either (B1) or (B2) holds.

So far, m has not been specified. The number m must be adapted to the particular value of p given by (B1) or (B2). Fabian [6] considers in this connection the case

(C1) $m := \lfloor p/2 \rfloor = (p-1)/2$ for an odd $p \geq 3$, $\gamma := 1/(2p)$.

We will consider in addition the following case (for $d = 1$ see Renz [19]):

(C2) $m := \lceil p/2 \rceil$ for $p \geq 2$ (p not necessarily odd), $\gamma := 1/(2p)$,

which will result in an unbiased limit distribution, whereas (C1) generally leads to a nonzero bias (Theorem 3.2).

Similarly as above, (C) means that either (C1) or (C2) holds. We note here that the assumptions (B1) and (C1) do not occur together.

The sequence (W_n) of random variables $W_n := \sum_{j=1}^m v_j \left(V_{n,2j-1}^{(i)} - V_{n,2j}^{(i)}\right)_{i=1,\dots,d}$ satisfies

$$\text{(D)} \qquad \underset{n\geq m}{\forall} \quad \|EW_m \otimes W_n\| \leq \varrho_{n-m}(E\|W_m\|^2 E\|W_n\|^2)^{\frac{1}{2}}$$
$$\text{with } \sum_{l=0}^{\infty} \varrho_l < \infty \text{ and } E\|W_n\|^2 = O(1).$$

Regarding assumption (B2b) it is worthwhile to note that this condition is invariant under rotation of coordinates (compare Fabian [8]). As a further comparison with related work (Fabian [8], Spall [23]), we remark that our results, Theorems 3.2 and 4.2, do not assume continuity of the highest-order partial derivatives.

Results about asymptotic normality in stochastic approximation usually rely on local smoothness of the regression function f around ϑ and on the consistency of the procedure. The next proposition shows consistency of the modified procedure. The assumptions imposed on f allow us to decouple the influence of the r.v.'s W_n and to use the weak dependence condition (D).

PROPOSITION 3.1. *Let $a_n = a/n^\alpha$ with $\alpha \in (\max\{1/2 + 1/(2p), 1 - 1/p\}, 1)$ or $a_n = (a \ln n)/n$, $a > 0$. For recursion (1.1) with gradient estimate (3.1), assume that conditions (A) and (D) hold, f is bounded from below and has a Lipschitz continuous derivative with $\nabla f(x) \neq 0$ for all $x \neq \vartheta$, and $\sup\{\|x\| : f(x) \leq \lambda\} < \infty$ for all $\lambda > \inf\{f(x) : x \in \mathbb{R}^d\}$. Then $X_n \to \vartheta$ $(n \to \infty)$ a.s.*

Under condition (C1) a nonweighted analogue of the next theorem can be found in Fabian [8].

THEOREM 3.2. *Let $a_n = (a \ln n)/n$ for $p = 2$ and $a_n = a/n^\alpha$ with $\alpha \in (1/2+1/(2p), 1)$ for $p \geq 3$. For recursion (1.1) with gradient estimate (3.1), assume that conditions (A)–(D) hold, $A := Hf(\vartheta)$ is positive definite, and $X_n \to \vartheta$ a.s. Let $B_n(t) := n^{-1/2}\left\{\sum_{i=1}^{\lfloor nt\rfloor} W_i + (nt - \lfloor nt\rfloor)W_{\lfloor nt\rfloor+1}\right\}$. Suppose the existence of a Brownian motion B with covariance matrix S of $B(1)$ and*

$$B_n \xrightarrow{\mathcal{D}} B \quad \text{in } C([0,1], \mathbb{R}^d) \quad (n \to \infty).$$

Then, for all $\delta > -(p+1)/(2p)$,

$$n^{\frac{1}{2}\left(1-\frac{1}{p}\right)}\left(\widetilde{X}_{n,\delta} - \vartheta\right) \xrightarrow{\mathcal{D}} N\left(\tfrac{2p(1+\delta)}{p+1+2p\delta}\, c^{p-1}A^{-1}b, \tfrac{p(1+\delta)^2}{p+1+2p\delta}\, c^{-2}A^{-1}SA^{-1}\right) \quad (n \to \infty),$$

where $b = -\frac{1}{p!}\left(\sum_{j=1}^m v_j u_j^p (1 + (-1)^{p+1}) \frac{\partial^p}{(\partial x_i)^p} f(\vartheta)\right)_{i=1,\dots,d}$ and $\widetilde{X}_{n,\delta}$ is defined in (1.2). In particular, under condition (C2), $b = 0$.

REMARK 3.3. *The choices $\delta = 0$ and $\delta = -2\gamma = -1/p$ are of special interest. Provided $b \neq 0$, the pair $(\delta, c) = (0, c_0)$ with c_0 as given in (5.1) minimizes the second moment of the limit distribution. However, for fixed $c > 0$, the limit's covariance is minimized by $\delta = -2\gamma = -1/p$. In particular, Theorem 3.2 yields for $n \to \infty$*

$$n^{\frac{1}{2}\left(1-\frac{1}{p}\right)}\left(n^{-1}\sum_{k=1}^n X_k - \vartheta\right) \xrightarrow{\mathcal{D}} N\left(\tfrac{2p}{p+1}\, c^{p-1}A^{-1}b, \tfrac{p}{p+1}\, c^{-2}A^{-1}SA^{-1}\right),$$

$$n^{\frac{1}{2}\left(1-\frac{1}{p}\right)}\left(\tfrac{p-1}{p}\, n^{-\frac{p-1}{p}}\sum_{k=1}^n k^{-\frac{1}{p}} X_k - \vartheta\right) \xrightarrow{\mathcal{D}} N\left(2\, c^{p-1}A^{-1}b,\ \tfrac{p-1}{p}\, c^{-2}A^{-1}SA^{-1}\right).$$

4. A Kiefer–Wolfowitz procedure with simultaneous perturbation gradient approximation. The classical Kiefer–Wolfowitz (finite difference) stochastic approximation method (FDSA) needs $2d$ observations to obtain a finite difference approximation of the gradient belonging to the function $f : \mathbb{R}^d \to \mathbb{R}$ of which the minimizer ϑ is sought. To reduce the number of observations in each step, randomized gradient approximation methods have been considered in the literature. Two examples are random direction stochastic approximation (RDSA), suggested by Kushner and Clark [12], and simultaneous perturbation stochastic approximation (SPSA), suggested by Spall [22]. Both methods are based on only two observations in each iteration. Depending on the dimension d and the third derivatives of the regression function f the AMSE of the SPSA method can be better or worse than that of the FDSA and RDSA methods. At least for second-order polynomials f, the FDSA method needs d times more observations than the SPSA method to achieve the same level of mean squared error asymptotically, when the same span $c_n = cn^{-\gamma}$ is used (Spall [23]).

Before the idea of weighted averages is applied to the SPSA method, we will describe this algorithm in more detail. Again recursion (1.1) is used, but with step lengths $a_n = an^{-\alpha}$ and with the following so-called *simultaneous perturbation gradient estimate* of $\nabla f(X_n)$:

$$Y_n = \frac{1}{2c_n} \begin{pmatrix} (\Delta_n^{(1)})^{-1} \\ \vdots \\ (\Delta_n^{(d)})^{-1} \end{pmatrix} ([f(X_n + c_n\Delta_n) - W_{n,1}] - [f(X_n - c_n\Delta_n) - W_{n,2}]) \tag{4.1}$$

consisting of (artificially generated) random vectors $\Delta_n \in \mathcal{M}(\Omega, \mathbb{R}^d)$, observation errors $W_{n,1}$, $W_{n,2} \in \mathcal{M}(\Omega, \mathbb{R})$, and span c_n.

We consider the following set of conditions.

(E) The components $\Delta_n^{(i)}$ of Δ_n, $i = 1, \dots, d$, for $n \in \mathbb{N}$ fixed, form a set of independent, identically and symmetrically distributed r.v.'s with $|\Delta_n^{(l)}|$ having values between fixed positive numbers $\alpha_0 < \alpha_1$. The r.v. Δ_n is assumed to be independent of $\{X_1, \dots, X_n, \Delta_1, \dots, \Delta_{n-1}\}$. Furthermore, we use $\xi^2 = E|\Delta_n^{(l)}|^2$ and $\rho^2 = E|\Delta_n^{(l)}|^{-2}$. For simplicity, the column vector appearing in (4.1) is denoted by Δ_n^{-1}.

(F) The difference $W_n = W_{n,1} - W_{n,2}$ of the observation errors satisfies $E(W_n \mid \mathcal{F}_n) = 0$ and $\sup_n E\left(W_n^2 \mid \mathcal{G}_n\right) < \infty$ a.s., where $\mathcal{F}_n$ and $\mathcal{G}_n$ denote the σ-fields generated by $\{X_1, \dots, X_n, \Delta_1, \dots, \Delta_n\}$ and $\{X_1, \dots, X_n, \Delta_1, \dots, \Delta_{n-1}\}$, respectively.

(G) $\infty > E(W_n^2 \mid \mathcal{F}_n) \to \sigma^2$ a.s. and $E(W_n^2 1_{[W_n^2 \geq rn]} \mid \mathcal{F}_n) \to 0$ a.s. for every $r > 0$.

(H) (B2) holds for $p = 3$, and $A = Hf(\vartheta)$ is a positive definite matrix.

The proposition below presents conditions for the recursion's consistency. It is related to Blum's result [2] on multivariate Kiefer–Wolfowitz procedures. Under different and less intuitive assumptions and with a different method of proof, Spall [23] asserts consistency as well.

PROPOSITION 4.1. *Let $a_n = a/n^\alpha$ with $\alpha \in (\max\{\gamma + 1/2, 1 - 2\gamma\}, 1]$ and $\gamma > 0$. For recursion* (1.1) *with gradient estimate* (4.1)*, assume that conditions* (A), (E), *and* (F) *hold, and that f is bounded from below and has a Lipschitz continuous gradient.*

(a) *If $\sup\{\|x\| : f(x) \leq \lambda\} < \infty$ for all $\lambda > \inf\{f(x) : x \in \mathbb{R}^d\}$, then $\sup_n \|X_n\| < \infty$ a.s.*

(b) *Assume* $\nabla f(x) \neq 0$ *and* $f(x) > f(\vartheta)$ *for all* $x \neq \vartheta$. *If* $\sup_n \|X_n\| < \infty$ *a.s., then* $X_n \to \vartheta$ $(n \to \infty)$ *a.s.*

A nonweighted analogue of the following theorem is stated in Spall [23].

THEOREM 4.2. *Let* $\alpha \in (2/3, 1)$ *and* $\gamma = 1/6$. *For recursion* (1.1) *with gradient estimate* (4.1), *assume conditions* (A), (E)–(H), *and* $X_n \to \vartheta$ *a.s. Then, for all* $\delta > -2/3$,

$$n^{\frac{1}{3}} \left(\widetilde{X}_{n,\delta} - \vartheta \right) \xrightarrow{\mathcal{D}} N \left(\tfrac{1+\delta}{2/3+\delta} c^2 A^{-1} b, \tfrac{(1+\delta)^2}{4/3+2\delta} c^{-2} A^{-1} S A^{-1} \right) \quad (n \to \infty),$$

where

$$S = \frac{\sigma^2 \rho^2}{4} I, \qquad b = -\frac{1}{6} \xi^2 \left(\frac{\partial^3}{(\partial x_i)^3} f(\vartheta) + 3 \sum_{j=1,\, j \neq i}^{d} \frac{\partial^3}{\partial x_i (\partial x_j)^2} f(\vartheta) \right)_{i=1,\ldots,d},$$

and $\widetilde{X}_{n,\delta}$ *is as defined in* (1.2).

5. Comparison of stochastic approximation procedures with respect to their asymptotic mean squared error and further comments. Based on recursion (1.1) with any of the gradient estimates Y_n discussed in this paper, we consider the following three variants of algorithms:

(i) the basic recursion with $a_n = a/n$,
(ii) an adaptive variant obtained from the basic recursion with $a_n = (a/n) M_n$ and random matrices M_n converging to $M = Hf(\vartheta)^{-1}$ a.s.,
(iii) the basic recursion with a_n converging to zero slower than $1/n$ combined with averaging of the trajectories

and compare the corresponding estimators with regard to their asymptotic behavior. Some of these estimators have been treated in the literature (Fabian [6], [9], Spall [22], [23]). The adaptive procedure has been introduced by Fabian [9] to improve the limit distribution. There the auxiliary sequence M_n is built up from information available up to stage n. For both algorithms (i) and (ii), with any gradient estimate considered in this paper, the limit distribution can be obtained by Theorem 1 in Walk [24] and by the representations derived in the proofs of Theorems 3.2 and 4.2.

Assuming that $A = Hf(\vartheta)$ is a positive definite matrix, the related second moments of the asymptotic distributions turn out to be

$$E(a,c) := \left(2c^{p-1} a \|(2aA - \beta)^{-1} b\| \right)^2 + \frac{a^2}{c^2} \operatorname{tr} \left((2aA - \beta)^{-1} S \right), \quad a > \beta/(2\lambda_0),$$

$$\widehat{E}(a,c) := \left(\frac{2c^{p-1} a}{2a - \beta} \|A^{-1} b\| \right)^2 + \frac{a^2}{c^2 (2a - \beta)} \operatorname{tr} \left(A^{-1} S A^{-1} \right), \quad a > \beta/2,$$

$$\widetilde{E}(\delta,c) := \left(\frac{2c^{p-1}(1+\delta)}{2 - \beta + 2\delta} \|A^{-1} b\| \right)^2 + \frac{(1+\delta)^2}{c^2 (2 - \beta + 2\delta)} \operatorname{tr} \left(A^{-1} S A^{-1} \right), \quad \delta > \beta/2 - 1,$$

respectively, where $\beta = 1 - 1/p$, $\lambda_0 = \min\{\lambda : \lambda \in \operatorname{spec} A\}$, and $c > 0$ (concerning E and $\widehat{E}$, use Theorem 5.8 and Remark 5.9 of [25]). Under appropriate assumptions these quantities are equal to the AMSEs $\lim_n E\|n^{\frac{1}{2}(1-\frac{1}{p})}(X_n - \vartheta)\|^2$ shown by algorithm (i) or (ii), and $\lim_n E\|n^{\frac{1}{2}(1-\frac{1}{p})}(\widetilde{X}_{n,\delta} - \vartheta)\|^2$ shown by algorithm (iii). Apparently it holds that $\widehat{E}(a,c) = \widetilde{E}(a-1,c)$.

If $b \neq 0$, the asymptotic distribution is biased. In this case $\widehat{E}$ and $\widetilde{E}$ are minimized by $(a, c) = (1, c_0)$ and $(\delta, c) = (0, c_0)$, respectively, with

$$c_0 = \left(\frac{(2-\beta)}{4(p-1)} \frac{\operatorname{tr}(A^{-1}SA^{-1})}{\|A^{-1}b\|^2} \right)^{\frac{1}{2p}}, \tag{5.1}$$

which is usually unknown. At the end of section 6 we show that

$$\forall c > 0 \qquad \frac{1}{4} < \left(\frac{p+1}{2p} \right)^2 < \min_{a > \beta/(2\lambda_0)} \frac{E(a,c)}{\widetilde{E}(0,c)} < \sup_{a > \beta/(2\lambda_0)} \frac{E(a,c)}{\widetilde{E}(0,c)} = \infty \tag{5.2}$$

and

$$\frac{1}{4} < \left(\frac{p+1}{2p} \right)^2 < \min_{a > \beta/(2\lambda_0)} \min_{c > 0} \frac{E(a,c)}{\widetilde{E}(0,c_0)} < \sup_{a > \beta/(2\lambda_0)} \min_{c > 0} \frac{E(a,c)}{\widetilde{E}(0,c_0)} = \infty. \tag{5.3}$$

Noticing the last equation of the preceding paragraph, these relations can be rewritten in terms of $\widehat{E}$ instead of $\widetilde{E}$. Thus the AMSE of the adaptive algorithm (ii) and the averaging algorithm (iii) is less than four times the AMSE of the standard algorithm (i) for any admissible a, no matter whether a common c is used or the optimal values of c are chosen. On the opposite side, a bad choice of a (a close to $\beta/(2\lambda_0)$ or a too large) results in an arbitrarily large AMSE of the standard algorithm (i), whereas this difficulty does not arise when the adaptive or averaging method is used. In this sense one may say that the averaging and adaptive algorithms are more stable than the standard algorithm.

In the one-dimensional case the AMSE of the standard algorithm (i) is minimized by $a_0' = 1/A$ and $c_0' = (\frac{2-\beta}{4(p-1)} \frac{S}{b^2})^{1/(2p)}$. Hence, for $d = 1$, the second relation in (5.3) can be sharpened to $E(a_0', c_0')/\widetilde{E}(0, c_0) = 1$.

In section 3 the design $(u_1, \ldots, u_m)$ was fixed. If condition (C1) holds, the gradient estimate (3.1) usually produces an asymptotic bias. In this case Fabian [7] and Erickson, Fabian, and Mařik [5] investigated how the AMSE can be further reduced by the choice of an optimal design.

If the gradient estimate (3.1) is constructed under condition (C2), the bias is vanishing (since $b = 0$). Then, for a fixed positive c, the AMSEs $\widehat{E}$ and $\widetilde{E}$ attain their minimum $c^{-2}(1 - 1/p) \operatorname{tr}(A^{-1}SA^{-1})$ for $a = 1 - 1/p$ and $\delta = -1/p$, respectively. We get

$$\forall c > 0 \qquad 1 \leq \min_{a > \beta/(2\lambda_0)} \frac{E(a,c)}{\widetilde{E}(-1/p, c)} < \sup_{a > \beta/(2\lambda_0)} \frac{E(a,c)}{\widetilde{E}(-1/p, c)} = \infty. \tag{5.4}$$

Assume that a_0 ($> \beta/(2\lambda_0)$) minimizes $E(a, c)$ for a fixed c. Then, only in special cases can $E(a_0, c) = \widetilde{E}(-1/p, c)$ be achieved. In any of the three variants the related AMSE can be made arbitrarily small by choosing c sufficiently large. Hence, with respect to the AMSE criterion, the procedures using the gradient estimate leading to $b = 0$ are superior to those leading to $b \neq 0$, although they need $2d$ more observations per step.

It must be emphasized that in the case $b \neq 0$ the adaptive recursion employing consistent estimators (M_n) of $M = A^{-1}$ instead of some other matrix M is, due to (5.2) and (5.3), a fairly good choice but not the best one with respect to the optimal AMSE. For fixed $c > 0$, a better choice would require consistent estimators M_n of

a matrix M which minimizes $c^{2p-2}\|(MA-\frac{\beta}{2}I)^{-1}Mb\|^2 + c^{-2}\operatorname{tr} Z$ where $\min\{\operatorname{re}\lambda : \lambda \in \operatorname{spec}(MA-\frac{\beta}{2}I)\} > 0$ and Z is the unique solution of $(MA-\frac{\beta}{2}I)Z + Z(MA-\frac{\beta}{2}I)^* = MSM^*$. Hence, averaging applied in the Kiefer–Wolfowitz situation with nonvanishing asymptotic bias does not optimize the AMSE. This is in contrast to the Robbins–Monro situation. Minimizing the expression above in both M and c would lead to an even better AMSE, but we do not pursue this possibility here.

Let $\vartheta_n(f)$ be an estimator of the minimum $\vartheta(f)$ of a p-times differentiable regression function f on $\mathbb{R}$ using n observations. Consider, for fixed $c > 0$,

$$\sup_f P[n^{\frac{p-1}{2p}}|\vartheta_n(f) - \vartheta(f)| > c],$$

where the supremum is taken over all regression functions f satisfying conditions (A) and (B) and some further boundedness conditions. Then, according to results by Chen [3] and by Polyak and Tsybakov [18], this supremum as a function of n has a universal positive lower bound independent of the choice of $\vartheta_n(f)$. This raises the interesting problem of determining which type of algorithm, together with which type of gradient estimate, leads to the smallest supremum above.

The condition $\sup\{\|x\| : f(x) \le \lambda\} < \infty$ for all $\lambda > \inf\{f(x) : x \in \mathbb{R}^d\}$ appearing in Propositions 3.1 and 4.1 is equivalent to $\inf\{f(x) : \|x\| \ge K\} \to \infty$ as $K \to \infty$. In applications this condition can be satisfied by adding the function values of an appropriate increasing and differentiable function to the basic observations taken at x. A possible choice is $x \mapsto \|x - d\frac{x}{\|x\|}\|^2 1_{[\|x\|\ge d]}$ for a fixed d large enough.

Finally, it is worth mentioning that the weighted means $\widetilde{X}_{n,\delta}$ can easily be recursified by

$$\begin{pmatrix} X_{n+1} \\ \widetilde{X}_{n+1,\delta} \end{pmatrix} = \begin{pmatrix} 1 & 0 \\ \frac{1+\delta}{n+1} & \left(\frac{n}{n+1}\right)^{1+\delta} \end{pmatrix} \begin{pmatrix} X_n \\ \widetilde{X}_{n,\delta} \end{pmatrix} - a_n \begin{pmatrix} Y_n \\ \frac{1+\delta}{n+1} Y_n \end{pmatrix}, \quad \widetilde{X}_{1,\delta} = (1+\delta)X_1.$$

6. Proofs.

Proof of Proposition 3.1. For $x \in \mathbb{R}^d$ and $h > 0$ define

$$g(x,h) := \left(g^{(i)}(x,h)\right)_{i=1,\dots,d} := \sum_{j=1}^m v_j (f(x+hu_je_i) - f(x-hu_je_i))_{i=1,\dots,d}.$$

Then, with $V_n := c^{-1}W_n$ and $H_n := \nabla f(X_n) - c_n^{-1}g(X_n, c_n)$, we obtain

$$X_{n+1} = X_n - a_n(\nabla f(X_n) - n^\gamma V_n - H_n). \tag{6.1}$$

By Lipschitz continuity of ∇f we have $f \in C^1(\mathbb{R}^d)$. This leads, with a Lipschitz constant K, to

$$\begin{aligned} \left|h^{-1}g^{(i)}(x,h) - \tfrac{\partial}{\partial x_i} f(x)\right| &= \left|\sum_{j=1}^m v_j u_j \int_{-1}^1 (\tfrac{\partial}{\partial x_i} f(x + shu_je_i) - \tfrac{\partial}{\partial x_i} f(x))\, ds\right| \\ &\le \sum_{j=1}^m |v_j| u_j \int_{-1}^1 \|\nabla f(x + shu_je_i) - \nabla f(x)\|\, ds \le K \sum_{j=1}^m |v_j| u_j^2 h. \end{aligned}$$

Therefore $\|H_n\| \le \sqrt{d}K \sum_{j=1}^m |v_j| u_j^2 c_n$. Our assumptions yield $\sum a_n^2 n^{1/p} (\log n)^2 < \infty$ and $\sum a_n n^{-1/p} < \infty$. Proposition 4.1 in Dippon and Renz [4] implies the assertion. ∎

Proof of Theorem 3.2. *First step*: *Expansions for* $h^{-1}g(x,h)$. In what follows, we assume $x \in U_{\varepsilon/2}(\vartheta)$ and $h \in (0, \varepsilon/2)$. As a consequence we have $x + she_i$, $\vartheta + t(x - \vartheta) \pm hu_j e_i \in U_\varepsilon(\vartheta)$ for all $t \in [0,1]$ and $s \in [-1,1]$.

First we consider the case of an at least three-times differentiable function f ($p \geq 3$). Using (B2c), we obtain $f \in C^2(U_\varepsilon(\vartheta))$, and therefore, according to Taylor's formula,

$$
\begin{aligned}
(6.2)\quad g^{(i)}(x,h) &= \sum_{j=1}^{m} v_j (f(x + hu_j e_i) - f(x - hu_j e_i)) \\
&= \sum_{j=1}^{m} v_j (f(\vartheta + hu_j e_i) - f(\vartheta - hu_j e_i)) \\
&\quad + \sum_{j=1}^{m} v_j (\nabla f(\vartheta + hu_j e_i) - \nabla f(\vartheta - hu_j e_i))^* (x - \vartheta) \\
&\quad + (x - \vartheta)^* \int_0^1 (1-t) \sum_{j=1}^{m} v_j (Hf(\vartheta + t(x - \vartheta) + hu_j e_i) \\
&\qquad\qquad - Hf(\vartheta + t(x - \vartheta) - hu_j e_i))\, dt\, (x - \vartheta)\,.
\end{aligned}
$$

Let us denote the first term of this sum by $t^{(i)}(h)$ $(i = 1, \ldots, d)$. (B2a) implies

$$
\tfrac{d^l}{(dh)^l} t^{(i)}(h) = \sum_{j=1}^{m} v_j u_j^l \left(\tfrac{\partial^l}{(\partial x_i)^l} f(\vartheta + hu_j e_i) + (-1)^{l+1} \tfrac{\partial^l}{(\partial x_i)^l} f(\vartheta - hu_j e_i) \right)
$$

for $l = 0, \ldots, p-1$. Then $\frac{d^l}{(dh)^l} t^{(i)}(0) = 0$ for all $l = 0, \ldots, p-1$. This is obvious for l even. For l odd with $1 \leq l \leq 2m-1$, this follows from (A) and the choice of the v_k. In the case $m := \lceil p/2 \rceil$, we have $p - 1 \leq 2m - 1$, and in the case $m := \lfloor p/2 \rfloor = (p-1)/2$, p odd, we have $2m - 1 = p - 2$ and $p - 1$ is even. (B2b) implies

$$
\tfrac{d^p}{(dh)^p} t^{(i)}(0) = \sum_{j=1}^{m} v_j u_j^p \left(1 + (-1)^{p+1}\right) \tfrac{\partial^p}{(\partial x_i)^p} f(\vartheta).
$$

In the case $m := \lceil p/2 \rceil$ we obtain $\frac{d^p}{(dh)^p} t^{(i)}(0) = 0$. For p even, this is again obvious, and for p odd, it follows from $2m - 1 = p$ and from the choice of the v_k. Taylor's formula yields

$$
(6.3)\qquad t^{(i)}(h) = \tfrac{h^p}{p!} \left(\tfrac{d^p}{(dh)^p} t^{(i)}(0) + o(1) \right) \quad (h \to 0).
$$

For the discussion of the second term of the sum on the right-hand side (r.h.s.) in (6.2) we define

$$
s^{(i,k)}(h) := \sum_{j=1}^{m} v_j \left(\tfrac{\partial}{\partial x_k} f(\vartheta + hu_j e_i) - \tfrac{\partial}{\partial x_k} f(\vartheta - hu_j e_i) \right) \quad (i,k = 1, \ldots, d).
$$

Using (B2a) we obtain by a consideration analogous to that above

$$
\tfrac{d^l}{(dh)^l}\, s^{(i,k)}(h) = \sum_{j=1}^{m} v_j u_j^l \left(\tfrac{\partial^l}{(\partial x_i)^l} \tfrac{\partial}{\partial x_k} f(\vartheta + hu_j e_i) + (-1)^{l+1} \tfrac{\partial^l}{(\partial x_i)^l} \tfrac{\partial}{\partial x_k} f(\vartheta - hu_j e_i) \right)
$$

for $l = 0, \dots, p-2$, where $s^{(i,k)}(0) = 0$, $\frac{d}{dh} s^{(i,k)}(0) = \frac{\partial}{\partial x_i} \frac{\partial}{\partial x_k} f(\vartheta)$ and $\frac{d^l}{(dh)^l} s^{(i,k)}(0) = 0$ for all $l = 2, \dots, p-2$. (B2b) implies, by reasoning similar to that in the case of $\frac{d^p}{(dh)^p} t^{(i)}(0)$,

$$\tfrac{d^{p-1}}{(dh)^{p-1}} s^{(i,k)}(0) \;=\; \sum_{j=1}^{m} v_j u_j^{p-1} \left(1 + (-1)^p\right) \tfrac{\partial^{p-1}}{(\partial x_i)^{p-1}} \tfrac{\partial}{\partial x_k} f(\vartheta) \;=\; 0.$$

Again, by using Taylor's formula, we obtain

$$s^{(i,k)}(h) = h \tfrac{\partial}{\partial x_i} \tfrac{\partial}{\partial x_k} f(\vartheta) + h^{p-1} o(1) \quad (h \to 0). \tag{6.4}$$

Finally, for every $i = 1, \dots, d$, the expression

$$q^{(i)}(x,h) \;:=\; (x-\vartheta)^* \int_0^1 (1-t) \sum_{j=1}^{m} v_j \left(Hf(\vartheta + t(x-\vartheta) + h u_j e_i) - Hf(\vartheta + t(x-\vartheta) - h u_j e_i)\right) dt$$

can be bounded in the following way by using (B2c):

$$\|q^{(i)}(x,h)\| \;\le\; h \sum_{j=1}^{m} |v_j|\, u_j\, L \|x - \vartheta\|. \tag{6.5}$$

Because of (6.2), (6.3), (6.4), and (6.5) we obtain the following representation:

$$h^{-1} g(x,h) = \left(Hf(\vartheta) + h^{p-2} P(h) + Q(x,h)\right)(x-\vartheta) - \tfrac{h^{p-1}}{c^{p-1}} T(h) \tag{6.6}$$

with matrices $P(h)$, $Q(x,h)$ and a vector $T(h)$ satisfying the relations $\|P(h)\| = o(1)$ $(h \to 0)$, $\|Q(x,h)\| \le \sqrt{d} L \sum_{j=1}^{m} |v_j|\, u_j \, \|x - \vartheta\|$, and $T(h) \to T = c^{p-1} b$ $(h \to 0)$. Notice that Q is a measurable function $(x \in U_{\varepsilon/2}(\vartheta),\, h \in (0, \varepsilon/2))$.

Now we are going to consider the case of a twice differentiable function f $(p = 2)$. (B1a) implies the existence of a measurable matrix-valued function R with

$$\nabla f(x) = \left(Hf(\vartheta) + R(x)\right)(x - \vartheta), \;\; \text{where } \|R(x)\| \le K_1 \|x - \vartheta\|^\tau \text{ for } x \in U_\varepsilon(\vartheta).$$

Because of $p = 2$ we have $m = 1$. Without loss of generality, we may assume that $u_1 = 1$. Then we have $v_1 = 1/2$. By (B1b) we obtain $f \in C^1(U_\varepsilon(\vartheta))$, and therefore

$$\begin{aligned} g^{(i)}(x,h) &= \tfrac{1}{2}(f(x + h e_i) - f(x - h e_i)) = \tfrac{1}{2} h \int_{-1}^{1} \tfrac{\partial}{\partial x_i} f(x + s h e_i)\, ds \\ &= \tfrac{1}{2} h \int_{-1}^{1} (\tfrac{\partial}{\partial x_i} f(x + s h e_i) - \tfrac{\partial}{\partial x_i} f(x))\, ds + h \tfrac{\partial}{\partial x_i} f(x). \end{aligned}$$

Putting the last two relations together gives the following representation:

$$h^{-1} g(x,h) = \left(Hf(\vartheta) + R(x)\right)(x - \vartheta) + s(x,h) \tag{6.7}$$

with vector $s(x,h)$ satisfying $\|s(x,h)\| \le 0.5 \sqrt{d} K_2\, h$. Notice that s is also a measurable function $(x \in U_{\varepsilon/2}(\vartheta),\, h \in (0, \varepsilon/2))$.

While the last representation and Lemma 7.1(b) yield a rate of convergence, we will need a second representation to apply Lemma 7.1(a). We obtain

$$g^{(i)}(x,h) = \tfrac{1}{2}\,h\int_{-1}^{1}\Big(\tfrac{\partial}{\partial x_i}f(x+she_i) - (\nabla\tfrac{\partial}{\partial x_i}f(\vartheta))^*(x-\vartheta+she_i)\Big)\,ds \\ +h\,(\nabla\tfrac{\partial}{\partial x_i}f(\vartheta))^*(x-\vartheta).$$

For $r^{(i)}(x,h) := \frac{1}{2}\int_{-1}^{1}(\frac{\partial}{\partial x_i}f(x+she_i)-(\nabla\frac{\partial}{\partial x_i}f(\vartheta))^*(x-\vartheta+she_i))\,ds$, our assumptions imply

$$|r^{(i)}(x,h)| \\ \le \tfrac{1}{2}\int_{-1}^{1}\|\nabla f(x+she_i) - Hf(\vartheta)(x-\vartheta+she_i)\|\,ds \;\le\; \tfrac{1}{2}\int_{-1}^{1}K_1\,\|x-\vartheta+she_i\|^{1+\tau}\,ds \\ \le 2^{\tau-1}K_1\int_{-1}^{1}\big(\|x-\vartheta\|^{1+\tau}+|s|^{1+\tau}h^{1+\tau}\big)\,ds \;\le\; 2^{\tau}K_1\left(\|x-\vartheta\|^{1+\tau}+\tfrac{h^{1+\tau}}{2+\tau}\right).$$

As a consequence of the last two relations we obtain the following representation:

$$h^{-1}g(x,h) = Hf(\vartheta)(x-\vartheta) + r(x,h), \tag{6.8}$$

where the vector $r(x,h) = (r^{(i)}(x,h))$ is bounded by $\|r(x,h)\| \le 2^{\tau}\sqrt{d}K_1(\|x-\vartheta\|^{1+\tau} + h^{1+\tau}/(2+\tau))$. Again, r is a measurable function ($x \in U_{\varepsilon/2}(\vartheta)$, $h \in (0,\varepsilon/2)$).

Second step: *Rate of convergence for $X_n \to \vartheta$ and proof of asymptotic normality.* Let us define $U_n := X_n - \vartheta$, $V_n := c^{-1}W_n$, and $\Omega(n) := [\|U_n\| < \varepsilon/2 \text{ and } c_n < \varepsilon/2]$.

First we consider the case $p \ge 3$. We define $A_n := (Hf(\vartheta) + c_n^{p-2}P(c_n) + Q(X_n,c_n))1_{\Omega(n)}$ and $T_n := T(c_n)1_{\Omega(n)} - n^{1/2}c^{-1}g(X_n,c_n)1_{\Omega(n)^c}$. In the case $p=2$ we define $A_n := (Hf(\vartheta)+R(X_n))1_{\Omega(n)}$ and $T_n := -n^{1/4}s(X_n,c_n)1_{\Omega(n)} - n^{1/2}c^{-1}g(X_n,c_n)\cdot 1_{\Omega(n)^c}$. Regarding properties (6.1), (6.6), and (6.7), the so-defined quantities fulfill recursion (7.2) and satisfy $A_n \to A = Hf(\vartheta)$ a.s. and $T_n = O(1)$ almost in L^2. Condition (7.12) holds by assumption (D). Therefore, Lemma 7.1(b) yields $U_n = O(a_n^{1/2}n^{1/(2p)})$ almost in L^2.

For $p \ge 3$ the above quantities satisfy $T_n \to T$ a.s. and $\|A_n - A\| \le C_1\,n^{-(p-2)/(2p)} + C_2\,\|U_n\| + C_3 1_{\Omega(n)^c}$. In the case $p=2$ we have to alter the definition of A_n and T_n. Let $A_n := Hf(\vartheta)1_{\Omega(n)}$ and $T_n := -n^{1/4}r(X_n,c_n)1_{\Omega(n)} - n^{1/2}c^{-1}g(X_n,c_n)1_{\Omega(n)^c}$. Due to (6.8) recursion (7.2) holds. Note that $\|n^{1/4}r(X_n,c_n)1_{\Omega(n)}\| \le C_4\,n^{1/4}\|U_n\|^{1+\tau} + C_5\,n^{-\tau/4}$.

In both cases we get $T_n - T = o(1)$ almost in L^1 and $A_n - A = o(1/\sqrt{na_n})$ almost in L^2 , where we have used $2/p \le 1/2 + 1/(2p) < \alpha$ for $p \ge 3$. Thus the assertion follows from Lemma 7.1 (a). □

Proof of Proposition 4.1. We may assume $\vartheta = 0$ and $f(\vartheta) = 0$. Lipschitz continuity of ∇f implies

$$|f(x+h) - f(x-h) - \langle 2h, \nabla f(x)\rangle| = \left|\int_{-1}^{1}\langle h, \nabla f(x+sh) - \nabla f(x)\rangle\,ds\right| \\ \le \|h\|\int_{-1}^{1}K|s|\,\|h\|\,ds \;=\; K\|h\|^2, \tag{6.9}$$

where K is, here and in the following inequalities, a constant that may vary from formula to formula. The last inequality, together with

$$E\left(\tfrac{\Delta_n^{(k)}}{\Delta_n^{(l)}}\tfrac{\partial}{\partial x_k}f(X_n) \mid \mathcal{G}_n\right) = \delta_{kl}\tfrac{\partial}{\partial x_k}f(X_n) \text{ a.s.},$$

proves

$$\|E\,(Y_n \mid \mathcal{G}_n) - \nabla f(X_n)\| \le K c_n \quad \text{a.s.} \tag{6.10}$$

Due to (6.9), (6.10), and assumption (F), we obtain

$$E(\|Y_n\|^2 \mid \mathcal{G}_n) \le K c_n^2 + K\,\|\nabla f(X_n)\|^2 + K\,c_n^{-2}\,E(W_n^2 \mid \mathcal{G}_n) \quad \text{a.s.} \tag{6.11}$$

and

$$\langle \nabla f(X_n), E\,(Y_n \mid \mathcal{G}_n)\rangle \ge \|\nabla f(X_n)\|^2 - K c_n\,\|\nabla f(X_n)\| \quad \text{a.s.} \tag{6.12}$$

Lipschitz continuity of ∇f implies, as above,

$$f(X_{n+1}) \le f(X_n) - a_n\langle \nabla f(X_n), Y_n\rangle + K a_n^2\,\|Y_n\|^2.$$

Taking conditional expectations and using inequalities (6.11) and (6.12), we obtain

$$\begin{aligned}
&E\,(f(X_{n+1}) \mid \mathcal{G}_n)\\
&\le f(X_n) - a_n\left(\|\nabla f(X_n)\|^2 - K c_n\,\|\nabla f(X_n)\|\right)\\
&\qquad\qquad + K a_n^2\,\|\nabla f(X_n)\|^2 + K a_n^2/c_n^2\left(E(W_n^2 \mid \mathcal{G}_n) + 1\right)\\
&\le f(X_n) - a_n/2\,\left(\|\nabla f(X_n)\| - K c_n\right)^2 + K^2/2\,a_n c_n^2 + K a_n^2/c_n^2\left(E(W_n^2 \mid \mathcal{G}_n) + 1\right) \quad \text{a.s.}
\end{aligned}$$

for all n with $K a_n < 1/2$. Let $A_n := a_n/2\,(\|\nabla f(X_n)\| - K c_n)^2$ and $B_n := K^2/2\,a_n c_n^2 + K a_n^2/c_n^2\left(E(W_n^2 \mid \mathcal{G}_n) + 1\right)$. For n large enough

$$E\,(f(X_{n+1}) \mid \mathcal{G}_n) \le f(X_n) - A_n + B_n \quad \text{a.s.},$$

where $A_n \ge 0$, $B_n \ge 0$, and $\sum_{n=1}^{\infty} B_n < \infty$ a.s. On a set Ω_0 of measure 1 we have convergence of $f(X_n)$ and $\sum_{n=1}^{\infty} A_n$ according to a theorem of Robbins and Siegmund [20] for nonnegative almost-supermartingales.

Fix $\omega \in \Omega_0$ and denote $x_n := X_n(\omega)$. Then for almost all n the relation $f(x_n) \le \lambda := \lim f(x_n) + 1$ holds. Since $\{x : f(x) \le \lambda\}$ is bounded, (x_n) is bounded as well.

To prove (b) fix $\omega \in \Omega_0$ with $\sup_n \|x_n\| < \infty$. Select a subsequence $(x_{n'})$ with $\nabla f(x_{n'}) \to 0$. Then there exists a convergent subsequence $(x_{n''})$ of $(x_{n'})$. Since $\nabla f(x_{n''}) \to 0$ and ∇f is continuous, $(x_{n''})$ converges to zero. Hence $f(x_{n''}) \to 0$ and $f(x_n) \to 0$. Choose $\varepsilon > 0$ such that $\|x_n\| < 1/\varepsilon$ for all n. For n sufficiently large we have $f(x_n) < \inf\{f(x) : \varepsilon < \|x\| < 1/\varepsilon\}$. This proves $x_n \to 0$. □

Proof of Theorem 4.2. We will verify the assumptions of Lemma 7.1. For this purpose let us define $U_n := X_n - \vartheta$, $D_n := (2c_n)^{-1}\Delta_n^{-1}(f(X_n + c_n\Delta_n) - f(X_n - c_n\Delta_n))$, $V_{n,1} := (2c)^{-1}\Delta_n^{-1}W_n$, $\Omega(n) := [\|U_n\| < \varepsilon/2$ and $c_n < \varepsilon/(2d^{1/2}\alpha_1)]$, $V_{n,2} := -n^{-1/6}(D_n 1_{\Omega(n)} - E(D_n 1_{\Omega(n)} \mid \mathcal{G}_n))$, $V_n := V_{n,1} + V_{n,2}$, and $T := c^2 b$.

For $x, z \in \mathbb{R}^d$ and $h > 0$ with $x, x \pm hz, \vartheta \pm hz \in U_\varepsilon(\vartheta)$, we obtain by condition (H)

$$\begin{aligned}
&f(x + hz) - f(x - hz)\\
&\quad = f(\vartheta + hz) - f(\vartheta - hz) + \langle \nabla f(\vartheta + hz) - \nabla f(\vartheta - hz), x - \vartheta\rangle\\
&\qquad + (x - \vartheta)^* \int_0^1 (1-t)(Hf(\vartheta + t(x - \vartheta) + hz) - Hf(\vartheta + t(x - \vartheta) - hz))dt\,(x - \vartheta)
\end{aligned}$$

where

$$f(\vartheta + hz) - f(\vartheta - hz) = \frac{2h^3}{6}\sum_{i,j,k} \tfrac{\partial^3}{\partial x_i \partial x_j \partial x_k} f(\vartheta)\, z_i z_j z_k \;+\; o(h^3\|z\|^3),$$

$$\nabla f(\vartheta + hz) - \nabla f(\vartheta - hz) = 2h\,Hf(\vartheta)z \;+\; o(h^2\|z\|^2),$$

and

$$\left\| \int_0^1 (1-t)(Hf(\vartheta + t(x-\vartheta) + hz) - Hf(\vartheta + t(x-\vartheta) - hz))dt \right\| \leq Lh \, \|z\|.$$

This expansion, together with condition (E), leads to the following representation:

$$E\left(D_n 1_{\Omega(n)} \mid \mathcal{G}_n\right) = \left(Hf(\vartheta) + o(c_n) + O(\|U_n\|)\right) U_n 1_{\Omega(n)} - n^{-\frac{1}{3}} \left(T + o(1)\right) 1_{\Omega(n)}.$$

With $A_n := (Hf(\vartheta)+o(c_n)+O(\|U_n\|))1_{\Omega(n)}$, $T_n := (T+o(1))1_{\Omega(n)} - n^{1/3} D_n 1_{\Omega(n)^c}$, and the quantities defined at the beginning of the proof, recursion (1.1) can be rewritten in the form of recursion (7.2).

Let $B_{n,j}(t) := 1/\sqrt{n}\,(\sum_{i=1}^{\lfloor nt \rfloor} V_{i,j} + (nt - \lfloor nt \rfloor) V_{\lfloor nt \rfloor + 1, j})$, $t \in [0,1]$, $j \in \{1,2\}$. To show that $B_{n,1}$ converges in distribution to a Brownian motion B, and that $B_{n,2}$ converges to zero in probability, we apply an invariance principle for martingale difference sequences of Berger [1].

We first consider the case $j = 1$. Since

$$E(V_{n,1} \mid \mathcal{F}_n) = \tfrac{1}{2c} \Delta_n^{-1} E(W_n \mid \mathcal{F}_n) = 0 \text{ a.s.}$$

and $V_{n,1}$ is $\mathcal{F}_{n+1}$-measurable, $(V_{n,1})$ is a martingale difference sequence with respect to $(\mathcal{F}_{n+1})$. Similarly, we get from the assumptions,

$$\begin{aligned}
&\tfrac{1}{n} \sum_{i=1}^n E(V_{i,1} \otimes V_{i,1} \mid \mathcal{F}_i) \\
&\quad = \tfrac{1}{4c^2} \tfrac{1}{n} \sum_{i=1}^n \Delta_i^{-1} \otimes \Delta_i^{-1} E(W_i^2 \mid \mathcal{F}_i) \\
&\quad = \tfrac{\sigma^2}{4c^2} \tfrac{1}{n} \sum_{i=1}^n \Delta_i^{-1} \otimes \Delta_i^{-1} + \tfrac{1}{4c^2} \tfrac{1}{n} \sum_{i=1}^n \Delta_i^{-1} \otimes \Delta_i^{-1} \left(E(W_i^2 \mid \mathcal{F}_i) - \sigma^2\right) \\
&\quad \to \frac{\sigma^2 \rho^2}{4c^2} I \quad (n \to \infty) \text{ a.s.}
\end{aligned}$$

according to Kolmogorov's strong law of large numbers. Further, we obtain

$$\begin{aligned}
\tfrac{1}{n} \sum_{i=1}^n E(\|V_{i,1}\|^2 1_{[\|V_{i,1}\|^2 \geq ri]} \mid \mathcal{F}_i) &\leq \tfrac{1}{n} \sum_{i=1}^n E(\tfrac{d}{4c^2\alpha_0^2} W_i^2 1_{\left[W_i^2 \geq \frac{4c^2\alpha_0^2 r}{d} i\right]} \mid \mathcal{F}_i) \\
&\xrightarrow{P} 0 \quad (n \to \infty)
\end{aligned}$$

since $E(W_i^2 1_{[W_i^2 \geq \tilde{r} i]} \mid \mathcal{F}_i)$ is converging to zero a.s.

To get (7.4) for the sequence $(V_{n,1})$, we check that

$$\sup_n E\left(\|V_{n,1}\|^2 \mid \mathcal{F}_n\right) \leq \tfrac{1}{4c^2} \sup_n \|\Delta_n^{-1}\|^2 \sup_n E\left(W_n^2 \mid \mathcal{F}_n\right) < \infty \quad \text{a.s.},$$

which holds in view of the assumptions.

Likewise, we treat the case $j = 2$. Note that $V_{n,2}$ is $\mathcal{G}_{n+1}$-measurable, and $E(V_{n,2} \mid \mathcal{G}_n) = 0$ a.s. Condition (H) implies

$$\begin{aligned}
\|E\left(V_{n,2} \otimes V_{n,2} \mid \mathcal{G}_n\right)\| &\leq \tfrac{d}{4c^2\alpha_0^2} E\left(|f(X_n + c_n\Delta_n) - f(X_n - c_n\Delta_n)|^2 \; 1_{\Omega(n)} \mid \mathcal{G}_n\right) \\
&\leq (dL\alpha_1/\alpha_0)^2 n^{-2\gamma} \to 0 \quad (n \to \infty) \quad \text{a.s.}
\end{aligned}$$

This implies $\sup_n E(\|V_{n,2}\|^2 \mid \mathcal{G}_n) < \infty$ a.s., and thus validity of (7.4) for the sequence $(V_{n,2})$. Additionally

$$E\left(\|V_{n,2}\|^2 1_{[\|V_{n,2}\|^2 \geq rn]} \mid \mathcal{G}_n\right) \to 0 \quad (n \to \infty) \quad \text{a.s.}$$

Once more, the invariance principle in Berger [1] can be applied to prove the desired result.

Since $X_n \to \vartheta$ a.s., we obtain $A_n \to A$ and $T_n \to T = c^2 b$ a.s. The latter is sufficient for (7.5) and (7.9). Furthermore, the sequence (V_n) fulfills (7.10) and (7.11) with respect to $(\mathcal{G}_{n+1})$. Now Lemma 7.1(b) asserts $U_n = O(n^{(1/6)-(\alpha/2)})$ almost in L^2. To obtain $A_n - A = o(n^{(\alpha-1)/2})$ almost in L^2, one has to choose $\alpha > 2/3$. This completes the proof. □

Proof of relations (5.2)–(5.4). Since $\operatorname{spec}(2aA - \beta) \subset (0, \infty)$ we obtain

$$\begin{aligned} a\|(2aA-\beta)^{-1}b\| &= \tfrac{1}{2} \left\| \left(\tfrac{2a}{\beta}A - I\right)^{-1} \left(\tfrac{2a}{\beta}A - I\right) A^{-1}b + \left(\tfrac{2a}{\beta}A - I\right)^{-1} A^{-1}b \right\| \\ &= \tfrac{1}{2} \left\| \left(I + \left(\tfrac{2a}{\beta}A - I\right)^{-1}\right) A^{-1}b \right\| \\ &\geq \tfrac{1}{2} \min\left\{ \lambda \in \operatorname{spec}\left(I + \left(\tfrac{2a}{\beta}A - I\right)^{-1}\right)\right\} \|A^{-1}b\| \\ &\geq \tfrac{1}{2} \|A^{-1}b\| \end{aligned}$$

and, by Theorem 1 in Wei [26],

$$\operatorname{tr}\left(a^2(2aA-\beta)^{-1}S\right) \geq \operatorname{tr}\left(\beta A^{-1}SA^{-1}\right). \tag{6.13}$$

Noticing that $4\beta/(2-\beta) > 1$ and $(2-\beta)/2 > 1/2$, this yields

$$E(a,c) \geq c^{2p-2}\|A^{-1}b\|^2 + \tfrac{\beta}{c^2}\operatorname{tr}(A^{-1}SA^{-1}) \;>\; \left(\tfrac{2-\beta}{2}\right)^2 \widetilde{E}(0,c) \;>\; \tfrac{1}{4}\widetilde{E}(0,c).$$

The last relation of (5.2) is obvious.

The first two relations of (5.3) follow from (5.2). To prove the last one, we find for a given admissible a that

$$c_0(a) = \left(\frac{\operatorname{tr}((2aA-\beta)^{-1}S)}{4(p-1)\|(2aA-\beta)^{-1}b\|^2}\right)^{\frac{1}{2p}}$$

minimizes $E(a,c)$. Now observe that $E(a, c_0(a)) \to \infty$ as $a \to \infty$ or $a \searrow \beta/(2\lambda_0)$.

The first relation of (5.4) follows from (6.13), and the third one is as shown above. □

7. Appendix: A weak invariance principle for weighted means in stochastic approximation. For the following lemma, which is a consequence of Theorems 3.1 and 4.1 in Dippon and Renz [4], let (a_n) be a sequence decreasing to 0 with

$$na_n \nearrow \infty \quad (n \to \infty) \tag{7.1}$$

and satisfying the relation $a_n - a_{n+1} = o_n a_n^2$ with

$$\sum_{n=1}^{\infty} |o_n - o_{n+1}| < \infty.$$

Note that under (7.1) $o_n \le 1/(na_n) \to 0$ as $n \to \infty$. Examples for sequences having all these properties are (a/n^α) and $(a \log n/n)$ with $\alpha \in (0,1)$, $a > 0$.

LEMMA 7.1. *Let $\gamma \in [0, 1/2)$ and $\delta > -1/2 - \gamma$. For $\mathbb{R}^d$-valued random variables U_n, V_n, T_n and $\mathcal{L}(\mathbb{R}^d)$-valued random variables A_n, assume the following recursion:*

$$U_{n+1} = (I - a_n A_n)\, U_n + a_n n^\gamma \left(V_n + n^{-\frac{1}{2}} T_n \right). \tag{7.2}$$

Suppose that $A \in \mathcal{L}(\mathbb{R}^d)$ satisfies

$$\min\{\operatorname{re} \lambda : \lambda \in \operatorname{spec} A\} > 0.$$

(a) *Let $B_n(t) := n^{-1/2} \left\{ \sum_{i=1}^{\lfloor nt \rfloor} V_i + (nt - \lfloor nt \rfloor) V_{\lfloor nt \rfloor + 1} \right\}$, $t \in [0,1]$, $n \in \mathbb{N}$. Assume the existence of a centered Brownian motion B with covariance matrix S of $B(1)$ and with*

$$B_n \xrightarrow{\mathcal{D}} B \quad \text{in } C([0,1], \mathbb{R}^d) \qquad (n \to \infty), \tag{7.3}$$

$$B_n(1) = O(1) \quad \text{almost in } L^1. \tag{7.4}$$

If there exists $T \in \mathbb{R}^d$ such that

$$T_n - T = o(1) \quad \text{almost in } L^1, \tag{7.5}$$

$$U_n = O\left(n^\gamma \sqrt{a_n}\right) \quad \text{almost in } L^2, \tag{7.6}$$

$$A_n - A = o\left(1/\sqrt{na_n}\right) \quad \text{almost in } L^2, \tag{7.7}$$

then

$$n^{1/2-\gamma}\, t^{-\min\{1,\gamma+\delta\}} \tfrac{1+\delta}{n^{1+\delta}} \left(\sum_{k=1}^{\lfloor nt \rfloor} k^\delta U_k + (nt - \lfloor nt \rfloor)(\lfloor nt \rfloor + 1)^\delta U_{\lfloor nt \rfloor + 1} \right)$$

$$\xrightarrow{\mathcal{D}} G(t) := (1+\delta)\, t^{\max\{0,\gamma+\delta-1\}} A^{-1} \left(\int_{(0,1]} u^{\gamma+\delta}\, dB(tu) + \tfrac{t^{1/2}}{1/2+\gamma+\delta} T \right)$$

in $C([0,1], \mathbb{R}^d)$ for $n \to \infty$, where $G(1)$ is a Gaussian distributed random variable in $\mathbb{R}^d$ with expectation $2(1+\delta)/(1+2\gamma+2\delta) A^{-1} T$ and covariance matrix $(1+\delta)^2/(1+2\gamma+2\delta)$ $A^{-1} S A^{-1*}$.

(b) *Assume*

$$A_n \to A \ \text{ a.s. } (n \to \infty), \tag{7.8}$$

$$T_n = O(1) \ \text{ almost in } L^2, \tag{7.9}$$

and

$$E(V_n \mid \mathcal{F}_{n-1}) = 0 \ \text{ a.s.}, \tag{7.10}$$

$$\sup_n E(\|V_n\|^2 \mid \mathcal{F}_{n-1}) < \infty \ \text{ a.s. or } \ E\|V_n\|^2 = O(1), \tag{7.11}$$

where $(\mathcal{F}_n)$ is a filtration and (V_n) is adapted to $(\mathcal{F}_n)$, or, instead of (7.10) *and* (7.11), *alternatively:*

$$\underset{n \ge m}{\forall} \ \|EV_m \otimes V_n\| \le \varrho_{n-m} (E\|V_m\|^2 E\|V_n\|^2)^{\frac{1}{2}} \tag{7.12}$$

$$\text{with } \sum_{l=0}^{\infty} \varrho_l < \infty \ \text{ and } \ E\|V_n\|^2 = O(1).$$

Then condition (7.6) *holds.*

REMARK 7.2. (a) *For conditions implying* (7.3) *in case of a martingale difference sequence* (V_n)*, see Theorem* 5.1 *in Berger* [1].

(b) *Condition* (7.4) *is implied by* (7.10) *and* (7.11)*, or by* (7.12).

(c) *In applications,* (7.6) *can often be used to show* (7.7). *Usually,* (7.8) *follows from the consistency of the stochastic approximation procedure.*

Acknowledgments. Part of this work was done while the second author was visiting the Department of Statistics and Probability, Michigan State University, East Lansing, and the Department of Statistics, University of Illinois at Urbana-Champaign. He thanks the members of both departments for their hospitality.

The authors wish to thank Professor V. Fabian and the referees for helpful comments.

REFERENCES

[1] E. BERGER, *Asymptotic behaviour of a class of stochastic approximation procedures*, Probab. Theory Related Fields, 71 (1986), pp. 517–552.

[2] J.R. BLUM, *Multidimensional stochastic approximation methods*, Ann. Math. Statist., 25 (1954), pp. 737–744.

[3] H. CHEN, *Lower rate of convergence for locating a maximum of a function*, Ann. Statist., 16 (1988), pp. 1330–1334.

[4] J. DIPPON AND J. RENZ, *Weighted means of processes in stochastic approximation*, Math. Meth. Statist., 5 (1996), pp. 32–60.

[5] R.E. ERICKSON, V. FABIAN, AND J. MAŘIK, *An optimum design for estimating the first derivative*, Ann. Statist., 23 (1995), pp. 1234–1247.

[6] V. FABIAN, *Stochastic approximation of minima with improved asymptotic speed*, Ann. Math. Statist., 38 (1967), pp. 191–200.

[7] V. FABIAN, *On the choice of design in stochastic approximation*, Ann. Math. Statist., 39 (1968), pp. 457–465.

[8] V. FABIAN, *On asymptotic normality in stochastic approximation*, Ann. Math. Statist., 39 (1968), pp. 1327–1332.

[9] V. FABIAN, *Stochastic approximation*, in Optimizing Methods in Statistics, J.S. Rustagi, ed., Academic Press, New York, 1971, pp. 439–470.

[10] L. GYÖRFI AND H. WALK, *On the averaged stochastic approximation for linear regression*, SIAM J. Control Optim., 34 (1996), pp. 31–61.

[11] J. KIEFER AND J. WOLFOWITZ, *Stochastic estimation of the maximum of a regression function*, Ann. Math. Statist., 23 (1952), pp. 462–466.

[12] H.J. KUSHNER AND D.S. CLARK, *Stochastic Approximation Methods for Constrained and Unconstrained Systems*, Springer-Verlag, New York, 1978.

[13] H.J. KUSHNER AND J. YANG, *Stochastic approximation with averaging the iterates: Optimal asymptotic rate of convergence for general processes*, SIAM J. Control Optim., 31 (1993), pp. 1045–1062.

[14] A.V. NAZIN AND P.S. SHCHERBAKOV, *Method of averaging along trajectories in passive stochastic approximation*, Prob. Inform. Trans., 29 (1993), pp. 328–338.

[15] A. PECHTL, *Arithmetic means and invariance principles in stochastic approximation*, J. Theoret. Probab., 6 (1993), pp. 153–173.

[16] B.T. POLYAK, *New method of stochastic approximation type*, Automat. Remote Control, 51 (1990), pp. 937–946.

[17] B.T. POLYAK AND A.B. JUDITSKY, *Acceleration of stochastic approximation by averaging*, SIAM J. Control Optim., 30 (1992), pp. 838–855.

[18] B.T. POLYAK AND A.B. TSYBAKOV, *Optimal orders of accuracy for search algorithms of stochastic optimization*, Prob. Inform. Trans., 26 (1990), pp. 126–133.

[19] J. RENZ, *Konvergenzgeschwindigkeit und asymptotische Konfidenzintervalle in der stochastischen Approximation*, Ph.D. thesis, Universität Stuttgart, Germany, 1991.

[20] H. ROBBINS AND D. SIEGMUND, *A convergence theorem for nonnegative almost supermartingales and some applications*, in Optimizing Methods in Statistics, J.S. Rustagi, ed., Academic Press, New York, 1971, pp. 233–257.

[21] D. Ruppert, *Efficient Estimators from a Slowly Converging Robbins-Monro Process*, Tech. Rep. No. 781, School of Oper. Res. and Ind. Engrg., Cornell University, Ithaca, NY, 1988. (See also §2.8 of D. Ruppert, *Stochastic approximation*, in Handbook of Sequential Analysis, B.K. Ghosh and P.K. Sen, eds., Marcel Dekker, New York, 1991, pp. 503–529.)

[22] J.C. Spall, *A stochastic approximation algorithm for large-dimensional systems in the Kiefer-Wolfowitz setting*, in Proc. IEEE Conf. Decision Contr., 1988, pp. 1544–1548.

[23] J.C. Spall, *Multivariate stochastic approximation using a simultaneous perturbation gradient approximation*, IEEE Trans. Automat. Control, 37 (1992), pp. 332–341.

[24] H. Walk, *Limit behaviour of stochastic approximation processes*, Statist. Decisions, 6 (1988), pp. 109–128.

[25] H. Walk, *Foundations of stochastic approximation*, in Stochastic Approximation and Optimization of Random Systems, L. Ljung, G. Pflug, and H. Walk, eds., Birkhäuser, Basel, 1992, pp. 1–51.

[26] C.Z. Wei, *Multivariate adaptive stochastic approximation*, Ann. Statist., 15 (1987), pp. 1115–1130.

[27] G. Yin, *On extensions of Polyak's averaging approach to stochastic approximation*, Stochastics Stochastic Rep., 36 (1991), pp. 245–264.

SIAM J. CONTROL OPTIM.
Vol. 35, No. 6, pp. 1829–1846, November 1997

001

NECESSARY CONDITIONS FOR OPTIMAL IMPULSIVE CONTROL PROBLEMS*

G. N. SILVA† AND R. B. VINTER‡

Abstract. Necessary conditions of optimality, in the form of a maximum principle, are derived for a class of optimal control problems, certain of whose controls are represented by measures and whose state trajectories are functions of bounded variation. State trajectories are interpreted as robust solutions of the dynamic equations, a concept of solutions which takes account of the interaction between the measure control and the state variables during the jumps. The maximum principle which is derived improves on earlier optimality conditions for problems of this nature, by allowing nonsmooth data, measurable time dependence, and a possibly time-varying constraint set for the conventional controls.

Key words. impulsive control, necessary conditions, nonsmooth analysis, maximum principle

AMS subject classifications. 65F10, 65F15

PII. S0363012995281857

1. Introduction.

In this paper, optimal control problems in which certain control variables are represented by measures ("impulsive" control variables) and the state trajectories are functions of bounded variation are studied. Necessary conditions of optimality in the form of a maximum principle are derived for these problems. Specifically, we consider the following problem:

$$\text{(P)} \qquad \begin{cases} \text{Minimize} \quad h\,(x(0), x(1)) \\ \text{subject to} \quad dx(t) = f\,(t, x(t), u(t))\,dt + g(t, x(t))\mu(dt), \quad t \in [\,0, 1], \\ (x(0),\ x(1)) \in C, \\ u(t) \in U_t, \quad \mathcal{L} - \text{a.e. } t \in [\,0, 1], \quad \text{and} \quad \mu \geq 0. \end{cases}$$

Here $h : \Re^n \times \Re^n \longrightarrow \Re$, $f : [0,1] \times \Re^n \times \Re^m \longrightarrow \Re^n$, and $g : [\,0, 1] \times \Re^n \longrightarrow \Re^n$ are given functions. U is a Borel subset of $[\,0, 1] \times \Re^m$ (U_t denotes the "section" $\{x : (t, x) \in U\}$), and C is a closed subset of $\Re^n \times \Re^n$.

A control policy is taken to be a pair of elements $(u : [\,0, 1] \longrightarrow \Re, \mu)$, in which the "conventional control" component u is a Lebesgue measurable function satisfying $u(t) \in U_t$ a.e. with respect to Lebesgue measure and the "impulsive" control μ is a regular, Borel, nonnegative valued measure. A *process* is a triple (x, u, μ), comprising a control policy (u, μ) and a corresponding *state trajectory* x; that is, x is a function of bounded variation which is a solution (appropriately defined) of the dynamical equation of problem (P). The control problem is to minimize the cost function $h(x(0), x(1))$ over processes (x, u, μ) for which $(x(0), x(1)) \in C$.

We particularly stress that, in our formulation, the coefficient $g(t, x)$ associated with the impulsive control is allowed to be x-dependent. This raises at the outset

*Received by the editors February 21, 1995; accepted for publication (in revised form) July 30, 1996. This research was supported by the SERC and the Brazilian National Science Council.
http://www.siam.org/journals/sicon/35-6/28185.html

†Interdisciplinary Research Centre for Process Systems and Department of Electrical and Electronic Engineering, Imperial College, London SW7 2BY, England. Current address: Universidade Estadual Paulista–UNESP, Caixa Postal 136, Campus de Sao Jose do Rio Preto, Sao Jose do Rio Preto, SP, CEP 15054-000, Brazil (gsilva@nimitz.dcce.ibilce.unesp.br).

‡Interdisciplinary Research Centre for Process Systems and Department of Electrical and Electronic Engineering, Imperial College, London SW7 2BY, England (r.vinter@ps.ic.ac.uk).

questions of how we should interpret state trajectories to take account of the interaction between the evolving state trajectory and the impulsive control at times when jumps occur.

It is natural to regard the state equation as shorthand for the integral equation

$$x(t) = x(0) + \int_0^t f(\tau, x(\tau), u(\tau))d\tau + \int_{[0,t]} g(\tau, x(\tau))\mu(d\tau) \quad \forall\ t \in (0,1]. \tag{1.1}$$

The second term on the right is affected by the value of the integral at an atom τ of the impulsive control μ. But here an ambiguity arises, because x can be expected to jump at τ and it is not obvious how we should interpret $g(t, x(t))$: a left limit, $g(t, x(t^-))$ as in [23], some averaged value, or what?

Our approach to defining state trajectories is in the spirit of recent work by Bressan, Dal Maso, and Rampazzo [1], [2], [6] and Miller [13], which in turn has its origins in the reparameterization techniques of Rishel [16] and refinements due to Warga [22]. The control problem studied here arises in the calculation of optimal flight trajectories, "midcourse guidance problems" [8], [10], in which context an impulsive control μ is the idealization of a "conventional" control which, at the "atoms" of μ, takes large values over a small interval of time. To be consistent with this view of an impulsive control, we need to be sure that the state trajectory corresponding to the idealized μ approximates the conventional control with which it is associated. We might require, say,

$$x_i(t) \to x(t) \qquad \forall\ t \in C_\mu \cup \{0,1\},$$

where C_μ denotes the points in $[\,0,1]$ which are not atoms of μ (the "continuity points" of μ). In this relationship, $\{x_i(t)\}$ is a sequence of state trajectories, with initial value $x(0)$, corresponding to a sequence of conventional controls $\{m_i(t)\} \subset L^1$ (think of m_i as defining a Borel measure $m_i(t)dt$), with $m_i(t) \geq 0$ a.e. for each i, for which

$$m_i(t)dt \xrightarrow{*} \mu(dt) \quad \text{weak}^*.$$

As shown by Dal Maso and Rampazzo [6] there is a concept of solutions with this continuous dependence property; we call them *robust solutions* to the state equation. A robust solution x is a solution of the integral equation

$$x(t) = x(0) + \int_0^t f(\tau, x(\tau), u(\tau))d\tau + \int_{[0,t]} \tilde{g}(\tau, x(\tau^-); \mu(\{t\}))\mu(d\tau) \quad \text{for} \quad t \in (\,0,1], \tag{1.2}$$

(in which $x(\tau^-)$ denotes limit from the left). Here one takes account of interaction between the state trajectory and the measure by choosing the integrand $\tilde{g}(\tau, x(\tau^-); \mu(\{t\}))$ of the impulsive term on the right to depend, at time τ, on $\mu(\{\tau\})$. $\tilde{g}(\tau, x(\tau^-); \mu(\{t\}))$ is determined by motion along integral curves of the impulsive dynamics, from the initial state $x(\tau^-)$. The magnitude of $\mu(\{\tau\})$ governs how far we move along an integral curve. If τ is not an atom of μ, then $\mu(\{\tau\}) = 0$ and there is no motion. In this case $\tilde{g}(\tau, x(\tau^-); \mu(\{t\})) = g(\tau, x(\tau))$.

The central result in this paper is a necessary condition, in the form of a maximum principle, governing minimizers for (P) over control policies and corresponding state trajectories, when the latter are interpreted as robust solutions of the dynamic equations. Our derivation of the condition covers problems involving a general, time-dependent control constraint set U_t associated with the conventional control and

having dynamics which are nonsmooth in the state variable and measurably time dependent.

There is a substantial literature on necessary conditions of optimality for problem (P), relating to robust solutions of the dynamical equations. (See [11] and the survey article, with extensive bibliography, provided by Miller [12].) Earlier necessary conditions have been proven under the assumptions that the control constraint set is time independent and that the dynamics are at least Lipschitz continuous in the time variable. Indeed these conditions do not even make sense when the data are merely measurably dependent on the time variable, because they involve time derivatives of the data. Earlier approaches have been to reparameterize the independent variable in such a manner that the measure-driven dynamic equation reduces to a conventional equation, and (P) is transformed into a standard optimal control problem. Necessary conditions for (P) follow directly from the maximum principle (as conventionally understood) applied to the transformed problem. The transformation renders the original time variable a state variable: the extra restrictions arise then, because the conventional maximum principle does not allow for a state dependent control constraint set or data measurable in the state variable.

The proof given here follows a different route, in which we approximate (P) by a conventional problem with the help of Ekeland's theorem and pass to the limit. No reparameterization of the independent variable is involved in this approximation itself (this would appear to be crucial for problems with data measurably dependent on the time variable and with a time-dependent control set), although a delicate consideration of the dynamical equations in both their original and their reparameterized forms is involved in the convergence analysis. A similar, simpler approximation procedure was followed in [21], which does not allow state dependence of g.

The optimality conditions assert the existence of a costate function p which satisfies, among other things, a measure-driven Hamiltonian system involving state derivatives of the data. Because the data are nonsmooth, and the derivatives are set valued, the Hamiltonian system is a measure-driven inclusion. The formulation and derivation of the optimality conditions make use of the concept of "robust solutions" to measure driven differential inclusions, and associated closure properties, provided in a recent paper [18].

The impulse controls considered here are scalar valued. For problems involving vector-valued controls the picture is complicated by the possibility that different sequences of approximating measures (absolutely continuous with respect to Lebesgue measure) of the same measure μ can give different "state trajectories" (for a fixed initial state and conventional control); see, e.g., [2], [3], [4], [19]. The optimal control problem posed over processes (x, u, μ) still makes sense even though a unique "state trajectory" x no longer is associated with a fixed control policy and initial state. In the case that the data are smooth with respect to the time variable and the conventional control constraint set is constant, necessary conditions can once again be derived by applying the conventional maximum principle to a standard optimal control problem obtained by transforming the time variable (see Motta and Rampazzo [15] and Miller [12]). It would appear that necessary conditions for problems involving vector-valued impulse controls, when the data are assumed merely measurable in time, can be derived by reformulating the problem as one with scalar impulsive controls but one whose dynamics involve a measure differential inclusion, and adapting the techniques of this paper to allow for these more general dynamics.

We list notation and conventions adhered to below.

B denotes the open unit ball in Euclidean space. $C([0,1];\Re^n)$ denotes the vector space of continuous $\Re^n$-valued functions on $[\,0,1]$ with supremum norm, and $C^*([0,1];\Re^n)$ its topological dual.

$C^+([0,1];\Re^n) \subset C^*([0,1];\Re^n)$ is the cone of functionals taking nonnegative values on nonnegative functions.

$AC\,([\,0,1]\,;\Re^n)$ is the space of absolutely continuous $\Re^n$-valued functions on $[\,0,1]$.

$BV^+([0,1];\Re^n)$ denotes the vector space of $\Re^n$-valued functions on $[\,0,1]$, of bounded variation, which are continuous from the right on $(0,1)$. The Borel measure associated with some $x \in BV^+\,([0,1];\Re^n)$ is denoted dx.

For brevity we often do not distinguish between elements in $C^*([0,1];\Re^n)$ and the Borel measures which represent them.

The weak* topology on $BV^+\,([0,1];\Re^n)$ refers to the weak* topology on $(\Re^n \times C([0,1];\Re^n))^*$ under the isomorphism

$$x \to (x(0), dx).$$

Thus "$x_i \to x$ (weakly*)" indicates that $x_i(0) \to x(0)$ and $dx_i \to dx$ (weakly* in $C^*([0,1];\Re^n)$). For simplicity we write $C(0,1)$ in place of $C([0,1];\Re^1)$, $C^*(0,1)$ in place of $C^*([0,1];\Re^1)$, and so on.

$\mathcal{L}$ denotes the Lebesgue subsets of $[\,0,1]$, $\mathcal{B}$ the Borel sets in $\Re^k$, and $\mathcal{L} \times \mathcal{B}$ the product σ-field.

The following concepts from nonsmooth analysis are required. Consider a closed set $A \in \Re^k$ and points $x \in A,\ p \in \Re^k$. We say that p is a *limiting normal* to A at x if and only if there exist $p_i \to p$ and $x_i \overset{A}{\to} x$ such that for each i we have

$$p_i \cdot (z - x_i) \le o(|z - x_i|) \quad \forall\ z \in A$$

(i.e., limiting normals are limits of vectors which support A at points near x, to first order). The *limiting normal cone* to A at x, written $N_A(x)$, comprises the limiting normals to A at x.

Given a lower semicontinuous function $f : \Re^k \to \Re \cup \{+\infty\}$ and a point $x \in \Re^k$ such that $f(x) < \infty$, we define the *limiting subdifferential* of f at x, written $\partial f(x)$, to be

$$\partial f(x) := \{\xi : (-1, \xi) \in N_{\mathrm{epi}\{f\}}(f(x), x)\},$$

in which epi$\{f\}$ denotes the epigraph set $\{(\eta, x) : \eta \ge f(x)\}$. In the event f is Lipschitz continuous on a neighborhood of x, co$\partial\{f(x)\}$ coincides with the *(Clarke) generalized gradient* of f at x, which may be defined directly [5].

The properties of limiting normal cones, limiting subdifferentials, and generalized gradients are developed in [9], [14] and [5].

2. Change of variables. We outline a change of variables technique, previously used in Rishel [16], Warga [22], Dal Maso and Rampazzo [6], and elsewhere, whose role will be to reduce measure-driven differential equations and inclusions to ordinary differential equations and inclusions.

Fix a measure $\mu \in C^+(0,1)$. Let F be its distribution function

$$F(t) := \begin{cases} \int_{[0,t]} \mu(ds), & t \in (0,1], \\ 0 & \text{if} \quad t = 0. \end{cases}$$

Define the *reparameterization function* η corresponding to μ to be

$$\eta(t) := \begin{cases} (t + \int_{[0,t]} \mu(d\tau))/(1 + \mu([\,0,1])), & t \in (0,1], \\ 0 & \text{if} \quad t = 0. \end{cases}$$

Evidently, η is a $BV^+(0,1)$ function which is strictly increasing. Define also the continuous, nondecreasing function $\theta : [0,1] \to [0,1]$ to be

$$\theta(s) := \sup\{t \in [0,1] : \eta(t) \le s\} \quad \forall\ s \in [0,1].$$

Let $\{t_i\}$ be an enumeration of the atoms of μ, and let $S_i (= [\sigma_i', \sigma_i''])$ be the subintervals $S_i := \theta^{-1}(\{t_i\})$ for $i = 1, 2, \ldots$. Now define the function $\gamma : [0,1] \to \Re^+$ to be

$$\gamma(s) := \begin{cases} F(\theta(s)) & \text{if} \quad s \in [0,1] \setminus \bigcup_{i=1}^{\infty} S_i, \\ F(t_i^-) + \frac{(s-\sigma_i')}{(\sigma_i''-\sigma_i')}(F(t_i) - F(t_i^-)) & \text{for} \quad s \in S_i, \quad i = 1, 2, \ldots. \end{cases}$$

(In this formula $F(t_i^-)$ and $F(t_i)$ are interpreted as $F(0)$ and $F(0^+)$ if $t_i = 0$.)

The function $(\theta, \gamma) : [0,1] \to (\Re^+)^2$ is called the *graph completion* of the measure μ. It is so called because it corresponds to filling in with straight line segments the graph of F and reparameterizing the resulting curve in $\Re^2$.

Basic properties of the graph completion are as follows.

PROPOSITION 2.1. *Let (θ, γ) be the graph completion of $\mu \in C^+(0,1)$. Then*

(i) *θ and γ are Lipschitz continuous, nonnegative, nondecreasing functions and*

$$\dot\theta(s) + \dot\gamma(s) = 1 + \mu([0,1]) \quad \mathcal{L} - a.e.$$

(ii) *For any Borel measurable function h which is μ integrable and any Borel set $T \subset [0,1]$ we have*

$$\int_{\theta^{-1}(T)} h(\theta(s))\dot\gamma(s)ds = \int_T h(\tau)\mu(d\tau).$$

(iii) *For any $\mathcal{L}$-integrable function g and Borel set $S \subset [0,1]$, $\theta(S)$ is also a Borel set and*

$$\int_S g(\theta(s))\dot\theta(s)ds = \int_{\theta(S)} g(\tau)d\tau.$$

(iv) *Let $\{\mu_i\}$ be a sequence of elements in $C^+(0,1)$, and let $\{(\theta_i, \gamma_i)\}$ be the corresponding graph completions. Suppose that $\mu_i \to \mu$ (weakly*). Then $(\theta_i, \gamma_i) \to (\theta, \gamma)$ uniformly and $(\dot\theta_i, \dot\gamma_i) \to (\dot\theta, \dot\gamma)$ weakly in L^1.*

Parts (i) and (iv) are proven by Dal Maso and Rampazzo [6]. We comment briefly on the other assertions. (ii) will be recognized as an example of the "change of variables" lemma [7, Theorem 6.9], since μ can be interpreted as the measure induced by the measure $\dot\gamma(s)ds$ under the mapping θ, i.e.,

$$\mu(A) = \int_{\theta^{-1}(A)} \dot\gamma(s)ds \qquad \forall\ A \in \mathcal{B}.$$

As for (iii), $\theta(S)$ is a Borel set since θ is monotone. The identity is another consequence of the change-of-variables lemma in view of the fact that

$$\int_{\theta^{-1}\circ\theta(S)} h\dot\theta ds = \int_S h\dot\theta ds.$$

This last relationship comes about because $\dot\theta(s) = 0$ almost everywhere on the set where θ is not one-to-one.

3. Measure-driven differential inclusions. In this section we give precise meaning to robust solutions of measure-driven differential inclusions (MDIs) of the form

$$\left\{ \begin{array}{lcl} dx(t) & \in & F_1(t,x(t))dt + F_2(t,x(t))\mu(dt), \quad t \in [\,0,1\,], \\ x(0) & = & x_0 \end{array} \right. \tag{3.1}$$

for which the data are multifunctions $F_1 : [\,0,1\,] \times \Re^n \longmapsto \Re^n$ and $F_2 : [\,0,1\,] \times \Re^n \longmapsto \Re^n$. In the equation, $\mu \in C^+(0,1)$ is a given measure, and $x_0 \in \Re^n$ a given initial state.

The definition of robust solution involves the multifunction $\tilde{F}_2 : [\,0,1\,] \times \Re^n \times [0,\infty) \longmapsto \Re^n$:

$$\tilde{F}_2(t,v;\alpha) := \big\{\alpha^{-1}[\xi(1)-\xi(0)] : \xi \in AC([0,1];\Re^n),\ \dot{\xi}(\sigma) \in \alpha F_2(t,\xi(\sigma)) \text{ a.e.}, \text{ and } \xi(0) = v\big\}$$

if $\alpha > 0$ and $\tilde{F}_2(t,v;0) := F_2(t,v)$.

DEFINITION 3.1. *We say that a function $x \in BV^+([0,1];\Re^n)$ is a* robust solution *to* (3.1) *if there exist an $\mathcal{L}$-integrable function ϕ_1 and μ-integrable function ϕ_2 such that*

$$\begin{aligned} &\phi_1(t) \in F_1(t,x(t)), \qquad \mathcal{L}\text{-a.e.}, \\ &\phi_2(t) \in \tilde{F}_2(t,x(t^-);\mu(\{t\})), \qquad \mu\text{-a.e.}, \end{aligned}$$

and

$$x(t) = x(0) + \int_0^t \phi_1(\tau)d\tau + \int_{[0,t]} \phi_2(\tau)\mu(d\tau) \qquad \forall \quad t \in (0,1].$$

Reparameterization by means of the graph completion of μ results in a (conventional) differential inclusion as described in the following proposition. Here η is the reparameterization function of μ, and $(\theta(\cdot),\gamma(\cdot))$ is the graph completion of this measure (see section 2).

PROPOSITION 3.1. *Suppose that the data for MDI* (3.1) *satisfies the following:*

- *F_1 has values-closed sets and is $\mathcal{L} \times \mathcal{B}$ measurable*

and

- *F_2 has values-closed sets and is Borel measurable.*

Fix a measure $\mu \in C^+(0,1)$ and an initial state x_0. We have the following:

(i) *Suppose that $x(\cdot) \in BV^+([0,1];\Re^n)$ is a robust solution to MDI* (3.1). *Then there exists a solution $y(\cdot) \in AC([\,0,1\,];\Re^n)$ to*

$$\left\{ \begin{array}{lcl} \dot{y}(s) & = & F_1(\theta(s),y(s))\dot{\theta}(s) + F_2(\theta(s),y(s))\dot{\gamma}(s), \quad s \in [0,1], \\ y(0) & = & x_0 \end{array} \right. \tag{3.2}$$

for which

$$x(t) = y(\eta(t)) \qquad \forall \quad t \in [\,0,1\,]. \tag{3.3}$$

Conversely,

(ii) *suppose that $y(\cdot) \in AC([\,0,1\,];\Re^n)$ is a solution to* (3.2). *Then there exists a robust solution $x(\cdot) \in BV^+([0,1];\Re^n)$ to MDI* (3.1) *for which* (3.3) *is satisfied.*

Proof. See [18, Theorem 4.1]. □

Solutions to the MDI (3.1) (as defined above) are "robust" in the sense that the set of solutions has desirable "closure" properties with respect to perturbations of the driving measure μ and the initial state. A result of this kind, suitable for future applications, is conveniently described in terms of a sequence of MDIs approximating (3.1), namely,

$$\left\{ \begin{array}{lcl} dx_i(t) & \in & F_1^{(i)}(t, x_i(t))dt + F_2(t, x_i(t))\mu_i(dt) \quad \text{on} \quad [\,0,1\,], \\ x_i(0) & = & x_0^i, \end{array} \right. \tag{3.4}$$

$i = 1, 2, \ldots$.

Here $F_1^{(i)} : [0,1] \times \Re^n \rightrightarrows \Re^n, \quad i = 1, 2, \ldots;\ F_1 : [0,1] \times \Re^n \rightrightarrows \Re^n$; and $F_2 : [0,1] \times \Re^n \longmapsto \Re^n$ are given multifunctions. $\mu_i,\ i = 1, 2, \ldots$, and μ are elements in $C^+(0,1)$, and $x_0^i, i = 1, 2, \ldots$, and x_0 are n-vectors.

The associated reparameterized equations are

$$\left\{ \begin{array}{lcl} \dot{y}_i(s) & \in & F_1^{(i)}(\theta_i(s), y_i(s))\dot{\theta}_i(s) + F_2(\theta_i(s), y_i(s))\dot{\gamma}_i(s), \quad s \in [0,1], \\ y_i(0) & = & x_0^i, \end{array} \right. \tag{3.5}$$

in which (θ_i, γ_i) is the graph completion of $\mu_i,\ i = 1, 2, \ldots$, and (θ, γ) is the graph completion of μ. Denote by $\eta : [\,0,1\,] \to [\,0,1\,]$ the reparameterization function for μ. We refer also to the reparameterization function $\eta_i : [0,1] \to [0,1]$ of $\mu_i,\ i = 1, 2, \ldots$. Reparameterization of the nominal MDI (3.1) results in the differential inclusion

$$\left\{ \begin{array}{lcl} \dot{y}(s) & \in & F_1(\theta(s), y(s))\dot{\theta}(s) + F_2(\theta(s), y(s))\dot{\gamma}(s), \quad s \in [0,1], \\ y(0) & = & x_0. \end{array} \right. \tag{3.6}$$

We have the following proposition.

PROPOSITION 3.2. *Consider multifunctions $F_1,\ F_1^{(i)},\ i = 1, 2, \ldots$, and F_2 with domain $[\,0,1\,] \times \Re^n$ and taking values-compact subsets of $\Re^n$. Assume that*

- *$F_1^{(i)}(t, \cdot),\ i = 1, 2, \ldots$, and $F_1(t, \cdot)$ have closed graph and $F_1^{(i)}(\cdot, \cdot),\ i = 1, 2, \ldots$, and $F_1(\cdot, \cdot)$ are $\mathcal{L} \times \mathcal{B}$ measurable.*
- *$F_1(t, x)$ is convex for all (t, x).*
- *$F_2(\cdot, \cdot)$ has closed graph and takes values convex sets.*

Assume further that

- *$\mathcal{L}$-measure$\{t : F_1^{(i)}(t, x) = F_1(t, x) \ \ \forall\ x \in \Re^n\} \to 1$ as $t \to \infty$.*

Take a sequence $\{x_0^i\}$ in $\Re^n$, a sequence $\{\mu_i\}$ in $C^+(0,1)$, and elements $x_0 \in \Re^n$ and $\mu \in C^+(0,1)$. Take also a sequence $\{x_i\} \in BV^+([0,1]; \Re^n)$ such that x_i is a robust solution to (3.4) *for each i and*

$$x_0^{(i)} \to x_0 \quad \textit{and} \quad \mu_i \to \mu \ \ (\textit{weakly}^*) \quad \textit{as} \ \ i \to \infty.$$

Assume that there exists $\beta(t) \in L^1$ and $c > 0$ such that $F_1^{(i)}(t, x_i(t)) \subset \beta(t)B$ a.e. and $F_2(t, x_i(t)) \subset cB$ for all t.

Then there exists a sequence $\{y_i\} \subset AC\,([\,0,1\,]\,; \Re^n)$ such that y_i is a solution to (3.5) *for each i, a solution y to* (3.6)*, and a robust solution x to* (3.1) *such that*

$$x_i(t) = y_i(\eta_i(t)) \quad \forall \quad t \in [0,1]$$

and

$$x(t) = y(\eta(t)) \quad \forall \quad t \in [0,1].$$

Along a subsequence we have

$$x_i \longrightarrow x \quad (\textit{weakly}^*),$$
$$x_i(t) \longrightarrow x(t) \quad \forall \quad t \in \Big([0,1] \setminus \mathcal{M}_\mu\Big) \cup \{0,1\}$$

(where $\mathcal{M}_\mu$ denotes the set of atoms of μ), and

$$y_i \longrightarrow y \qquad \textit{strongly in} \;\; C([0,1];\Re^n).$$

Proof. See [18, Theorem 5.1]. □

Consider now the dynamic equation of the problem (P) (the optimal control problem introduced in section 1)

$$\left\{ \begin{array}{lcl} dx(t) & = & f(t,x(t),u(t))dt + g(t,x(t))\mu(dt), \\ x(0) & = & x_0. \end{array} \right. \tag{3.7}$$

For a fixed control policy (u,μ) this is just an example of MDI (3.1) (set $F_1(t,x) = \{f(t,x,u(t))\}$ and $f_2 = \{g\}$). We may therefore speak of robust solutions of the dynamic equation (3.4) (corresponding to a control policy (u,μ)). Henceforth, a process is taken to be a triple of elements (x,u,μ), in which (u,μ) is a control policy (see section 1) and x is a robust solution of the dynamical equation (3.7). Notice that, under the hypotheses imposed below there will be a unique robust solution to (3.7) for given (u,μ). This follows from the characterization of robust solutions provided by Proposition (3.1) since, for the MDI associated with (3.7), the differential inclusion (3.2) reduces to a differential equation which is known to have a unique solution.

4. A maximum principle. Our aim is to obtain optimality conditions for problem (P) in the form of a maximum principle. These will follow from conditions on processes (as defined is section 3) which generate boundary points of some "reachable set" of the control system with dynamic equations:

$$\begin{array}{c} dx(t) = f(t,x(t),u(t))dt + g(t,x(t))\mu(dt), \quad t \in [0,1], \\ u(t) \in U_t \quad \text{a.e. } t \in [\,0,1\,]\,. \end{array}$$

Take a locally Lipschitz continuous function $\psi : \Re^n \to \Re^k$ and a closed set $D \subset \Re^n$. We define the (ψ, D)-reachable set $\mathcal{R}_{\psi,D}$ to be

$$\mathcal{R}_{\psi,D} := \{\psi(x(1)) : (x,u,\mu) \text{ is a process and } x(0) \in D\}.$$

The following hypotheses will be invoked:

(H1) There exists a constant $K_f(\cdot) \in L^1$ such that

$$|f(t,x,u) - f(t,y,u)| \;\leq K_f(t)\;|x-y| \;\text{ for }\; (x,u) \in \Re^n \times \Re^m \text{ and } t \in [\,0,1\,].$$

(H2) $f(\cdot,x,\cdot)$ is $\mathcal{L} \times \mathcal{B}$ -measurable.

(H3) $g(\cdot,\cdot)$ is continuous and there exists a constant K_g such that

$$|g(t,x) - g(t,y)| \;\leq K_g\;|x-y| \quad \forall\; x,\; y \in \Re^n,\; t \in [\,0,1\,]\,.$$

(H4) $U \in \Re^{1+m}$ is a Borel set.

THEOREM 4.1. *Let $\{\bar{x}(\cdot),\bar{u}(\cdot),\bar{\mu}(\cdot)\}$ be a process for which $x(0) \in D$ and $\psi(\bar{x}(1))$ is a boundary point of $\mathcal{R}_{\psi,D}$. Assume that hypotheses* (H1)–(H4) *are satisfied. Then*

there exist a function $p \in BV^+([0,1];\Re^n)$ *and a unit vector* $d \in \Re^k$ *such that* $(\bar{x}(\cdot), p(\cdot))$ *is a robust solution of the MDI*

$$d\begin{bmatrix}\bar{x}(t)\\ p(t)\end{bmatrix} \in \begin{bmatrix} f(t,\bar{x}(t),\bar{u}(t))\\ -p(t)\cdot co\partial_x f(t,\bar{x}(t),\bar{u}(t))\end{bmatrix}dt + \begin{bmatrix} g(t,\bar{x}(t))\\ -p(t)\cdot \overline{\partial_x} g(t,\bar{x}(t))\end{bmatrix}\bar{\mu}(dt). \tag{4.1}$$

Furthermore,

$$-p(1) \in d\cdot\partial\psi(\bar{x}(1)), \tag{4.2}$$

$$p(0) \in N_D(\bar{x}(0)), \tag{4.3}$$

$$p(t)\cdot f(t,\bar{x}(t),\bar{u}(t)) = \max_{u\in U_t}\{p(t)\cdot f(t,\bar{x}(t),u)\} \quad a.e.\ t\in[0,1], \tag{4.4}$$

$$p(t)\cdot g(t,\bar{x}(t)) \le 0 \quad \forall\ t\in(0,1), \tag{4.5}$$

$$p(t)\cdot g(t,\bar{x}(t)) \ge 0, \quad \bar{\mu}-a.e.\ t\in[0,1],$$

and corresponding to every atom t *of* $\bar{\mu}$*, there exists a solution* (ξ_t,α_t) *to*

$$\begin{bmatrix}\dot{\xi}_t(s)\\ \dot{\alpha}_t(s)\end{bmatrix} \in \bar{\mu}(\{t\})\begin{bmatrix} g(t,\xi_t(s))\\ -\alpha_t(s)\cdot\overline{\partial_x} g(t,\xi_t(s))\end{bmatrix}, \quad t\in[0,1],$$

$$(\xi_t(0),\alpha_t(0)) = (\bar{x}(t^-),p(t^-)), \quad (\xi_t(1),\alpha_t(1)) = (\bar{x}(t),p(t)) \tag{4.6}$$

that satisfies

$$\alpha_t(s)\cdot g(t,\xi_t(s)) \ge 0 \quad \forall\ s\in[0,1]. \tag{4.7}$$

Here (4.6) is interpreted as

$$(\xi_t(0),\alpha_t(0)) := (\bar{x}(0),p(0)), \quad (\xi_t(1),\alpha_t(1)) := (\bar{x}(0^+),p(0^+))$$

if $t=0$. Also $\overline{\partial_x} g(t,x)$ denotes the set

$$\overline{\partial_x} g(t,x) := \left\{\lim_i a_i : a_i \in co\partial_x g(t_i,x_i) \text{ for some } t_i\to t,\ x_i\to x\right\}.$$

A proof is given in section 5.

The transition from Theorem 4.1 to a maximum principle for problem (P) follows the standard lines. Indeed, suppose that $h:\Re^n\times\Re^n\to\Re$ is a locally Lipschitz continuous function and C is a closed subset of $\Re^n\times\Re^n$, and let $(\bar{x},\bar{u},\bar{\mu})$ be a minimizing process for problem (P).

Consider the control system in which state trajectories are triples (x,y,z):

$$d\left(x(t),y(t),z(t)\right) = \left(f(t,x(t),u(t)),0,0\right)dt + \left(g\left(t,x(t)\right),0,0\right)\mu(dt),$$
$$(x(0),y(0),z(0)) \in D.$$

Here

$$D := \{(x,y,z) : (x,y)\in C \text{ and } z\ge h(x,y)\}.$$

It is easy to deduce from the optimality of $(\bar{x},\bar{u},\bar{\mu})$ that $(0,h(\bar{x}(0),\bar{x}(1))$ is a boundary point of $\mathcal{R}_{\psi,D}$, $\psi(x,y,z) := (y-x,z)$.

Applying the earlier theorem to this boundary process we arrive at the following maximum principle for (P).

THEOREM 4.2. *Let $(\bar{x}(\cdot), \bar{u}(\cdot), \bar{\mu}(\cdot))$ be a minimizing process for* (P). *Assume that h is locally Lipschitz continuous, that C is a closed subset, and that hypotheses* (H1)–(H4) *are satisfied.*

Then there exist $\lambda \geq 0$ and $p \in BV^+([0,1]; \Re^n)$ such that $\|p(\cdot)\|_{L^\infty} + \lambda > 0$ and $(\bar{x}(\cdot), p(\cdot))$ is a robust solution of the MDI

$$d\begin{bmatrix} \bar{x}(t) \\ p(t) \end{bmatrix} \in \begin{bmatrix} f(t, \bar{x}(t), \bar{u}(t)) \\ -p(t) \cdot co\partial_x f(t, \bar{x}(t), \bar{u}(t)) \end{bmatrix} dt + \begin{bmatrix} g(t, \bar{x}(t)) \\ -p(t) \cdot \overline{\partial_x} g(t, \bar{x}(t)) \end{bmatrix} \bar{\mu}(dt).$$

Furthermore,

$$\begin{aligned} &(p(0), -p(1)) \in N_C(\bar{x}(0),\ \bar{x}(1)) + \lambda \partial h(\bar{x}(0),\ \bar{x}(1)), \\ &p(t) \cdot f(t, \bar{x}(t), \bar{u}(t)) = \max_{u \in U_t} \{p(t) \cdot f(t, \bar{x}(t), u)\} \quad a.e.\ t \in [0,1], \\ &p(t) \cdot g(t, \bar{x}(t)) \leq 0 \quad \forall\ t \in [0,1], \\ &p(t) \cdot g(t, \bar{x}(t)) = 0, \quad \bar{\mu} - a.e.\ on\ [0,1]. \end{aligned}$$

Corresponding to every atom t of $\bar{\mu}$, there exists a solution $\left[\begin{smallmatrix} \xi_t(\cdot) \\ \alpha_t(\cdot) \end{smallmatrix}\right]$ to

$$\begin{bmatrix} \dot{\xi}_t(s) \\ \dot{\alpha}_t(s) \end{bmatrix} \in \bar{\mu}(\{t\}) \begin{bmatrix} g(t, \xi_t(s)) \\ -\alpha_t(s) \cdot \bar{\partial}_x g(t, \xi_t(s)) \end{bmatrix} \quad on\ [0,1]$$

which satisfies

$$\begin{gathered} (\xi_t(0), \alpha_t(0)) = (\bar{x}(t^-), p(t^-)), \quad (\xi_t(1), \alpha_t(1)) = (\bar{x}(t), p(t)), \\ \alpha_t(s) \cdot g(t, \xi_t(s)) \geq 0 \quad \forall\ s \in [0,1]. \end{gathered}$$

5. Proof of Theorem 4.1. Take a process $(\bar{x}, \bar{u}, \bar{\mu})$ with the stated "boundary" property. The assertions of the theorem are proven first in the special case when an interim hypothesis,

($\bar{\text{H}}$) There exists $\alpha \in L^1$ such that $\sup_{u \in U(t)} |f(t, \bar{x}(t), u)| \leq \alpha(t)$ a.e.

is added to (H1)–(H4). We show how to remove ($\bar{\text{H}}$) in the final stage of the proof. ($\bar{\text{H}}$) has the role of ensuring suitable "linear growth properties" of the data, namely, the following lemma.

LEMMA 5.1. *There exist $\alpha_1,\ \alpha_2 \in L^1$ and $\beta_1 \geq 0,\ \beta_2 \geq 0$ such that*

$$|f(t,x,u)| \ \leq \alpha_1(t)|x| + \alpha_2(t) \quad \forall \quad x \in \Re^n \ and \ \ u \in U_t,\ \mathcal{L}-a.e.$$

and

$$|g(t,x)| \ \leq \beta_1|x| + \beta_2 \quad \forall \quad x \in \Re^n \ and \ \ t \in [0,1].$$

Proof. An appeal to (H1), (H3), and ($\overline{\text{H}}$) and repeated applications of the triangle inequality validate the inequalities with

$$\begin{aligned} \alpha_1(t) &:= K_f(t), \qquad \alpha_2(t) := \alpha(t) + K_f(t)\ \|\bar{x}(\cdot)\|_{L^\infty}, \\ \beta_1 &:= K_g, \quad \beta_2 := \sup_{s \in [0,1]} |g(s, \bar{x}(s))| + K_g\ \|\bar{x}(\cdot)\|_{L^\infty}. \end{aligned}$$

□

Next, we have a lemma describing the continuity properties of state trajectories with respect to control policies and initial states.

LEMMA 5.2. *Given any control policy (u, μ) and initial state x_0 there is a unique robust solution x of the dynamical equation*

$$\left[\begin{array}{l} dx(t) = f\,(t, x(t), u(t))\,dt + g\,(t, x(t))\,\mu(dt), \quad t \in [\,0, 1\,], \\ x(0) = x_0 \end{array}\right. \tag{5.1}$$

and a unique solution y of the reparameterized equation

$$\left[\begin{array}{l} \dot{y}(s) = f\,(\theta(s), y(s), u(\theta(s)))\,\dot{\theta}(s) + g\,(\theta(s), y(s))\,\dot{\gamma}(s), \quad s \in [0, 1], \\ y(0) = x_0. \end{array}\right. \tag{5.2}$$

(Here $(\theta(s), \gamma(s))$ is the graph completion of the measure μ.) If $\{(u_i(\cdot), \mu_i)\}$ and $\{x_0^i\}$ are sequences of control policies and initial states, respectively, such that

$$\mathcal{L}\text{-}meas\{t : u_i(t) \neq u(t)\} \to 0,$$
$$\mu_i \to \mu \quad weakly^*,$$

and

$$x_0^i \to x_0,$$

then $x_i(t) \to x(t)\ \forall\ t \in \mathcal{C}_\mu \cup \{0, 1\}$, $dx_i \to dx$ (weakly), and $y_i(t) \to y(t)$ uniformly, where $\{x_i(\cdot)\}$ and $\{y_i(\cdot)\}$ are the corresponding sequences of state trajectories and reparameterized state trajectories and $\mathcal{C}_\mu$ is the set of continuity points of μ.*

Proof. Notice that the right-hand side

$$\phi(s, y) := f(\theta(s), y, u(\theta(s)))\ \dot{\theta}\ (s) + g(\theta(s), y)\ \dot{\gamma}\ (s)$$

of the reparameterized equation (5.2) satisfies the growth condition,

$$|\phi(s, y)| \leq \alpha_1^{'}(s)|y| + \alpha_2^{'}(s) \quad \text{a.e.} \quad s \in [0, 1] \tag{5.3}$$

$\forall\ y \in \Re^n$, where $\alpha_1^{'}$ and $\alpha_2^{'}$ are the integrable functions

$$\begin{array}{l} \alpha_1^{'}(s) := \alpha_1(\theta(s))\ \dot{\theta}\ (s) + \beta_1(1 + \mu([0, 1])), \\ \alpha_2^{'}(s) := \alpha_2(\theta(s))\ \dot{\theta}\ (s) + \beta_2(1 + \mu([0, 1])), \end{array} \tag{5.4}$$

and also the Lipschitz condition,

$$|\phi(s, y) - \phi(s, z)| \leq k^{'}(s)|y - z|$$

for all $y, z \in \Re^n$, where $k^{'}(s)$ is the integrable function

$$k^{'}(s) := K_f(\theta(s))\ \dot{\theta}\ (s) + K_g(1 + \mu([0, 1])).$$

These are known conditions under which the reparameterized equation (5.2) has a unique solution $y \in AC([0, 1]; \Re^n)$. But then by Proposition 3.1 there is a unique robust solution to (5.1).

An important observation is that, if $c > 0$ is a constant such that $\mu([0, 1]) < c$, then the coefficients in the linear growth inequality (5.3) have L^1 norm bounded above by a number which depends only on c. (This is clear from (5.4) and the fact that $\int \alpha_1^{'}(\theta(s))\ \dot{\theta}(s)ds = \int \alpha_1^{'}(t)dt$.) Since $\{\mu_i\}$ is a weak* convergent sequence, the μ_i's and μ are uniformly bounded in total variation, and the x_0^i's too are uniformly

bounded, it follows from an application of Gronwall's lemma that the solutions y_i to the reparameterized equations, resulting from replacing (u, μ) by (u_i, μ_i) and x_0 by $x_o^{(i)}$, $i = 1, 2, \ldots$, are uniformly bounded in the supremum norm. Since $\|x_i\|_{L^\infty} \leq \|y_i\|_{L^\infty}$, $i = 1, 2, \ldots$, the x_i's are likewise bounded. The y_i's are, in addition, equicontinuous. To show this, we choose constants $r > 0$ such that $\|y_i\|_{L^\infty} \leq r$, $i = 1, 2, \ldots$, and $\rho > 0$ such that $|g(t, y)| \leq \rho$ for all $(t, y) \in [0, 1] \times r\bar{B}$. Fix $\epsilon > 0$. Choose $\delta > 0$ and let $[\sigma_1, \sigma_2] \subset [0, 1]$ be an arbitrarily interval such that $|\sigma_2 - \sigma_1| < \delta$ and i an arbitrary index value. We have

$$\begin{aligned}
|y_i(\sigma_2) - y_i(\sigma_1)| &\leq \left| \int_{\sigma_1}^{\sigma_2} f(\theta_i(s), y_i(s), u_i(\theta_i(s)))\dot{\theta}(s)ds \right| \\
&\quad + \left| \int_{\sigma_1}^{\sigma_2} g(\theta_i(s), y_i(s))\dot{\gamma}(s)ds \right| \\
&\leq \left| \int_{t_1}^{t_2} f(t, x_i(t), u_i(t))dt \right| + \rho(1 + \mu([0, 1]))|\sigma_2 - \sigma_1| \\
&\leq r \int_{t_1}^{t_2} \alpha_1(t)dt + \int_{t_1}^{t_2} \alpha_2(t)dt + \rho(1 + \mu([0, 1]))|\sigma_2 - \sigma_1|.
\end{aligned}$$

Here $t_1 = \theta_i(\sigma_1)$ and $t_2 = \theta_i(\sigma_2)$. Since $|t_2 - t_1| \leq (1 + \mu_i([0, 1]))|\sigma_2 - \sigma_1|$ and the $|\mu_i([0, 1])|$'s are uniformly bounded, this last inequality implies that, by choosing $\delta > 0$ sufficiently small, we can arrange that $|y_i(\sigma_2) - y_i(\sigma_1)| < \epsilon$ independently of our choice of interval $[\sigma_1, \sigma_2]$ and index value i. This confirms the equicontinuity of the y_i's.

The hypotheses are satisfied under which Proposition 3.2 may be applied when $F_1^i(t, x) := \{f(t, x, u_i(t))\}$, $i = 1, 2, \ldots$, and $F_2(t, x) := \{g(t, x)\}$. For arbitrary subsequences, further subsequences may be chosen such that $\{y_i\}$ and $\{x_i\}$ converge as described. Since, however, the limits are unique, the original sequences must converge. □

Since $\psi(\bar{x}(1)) \in \partial\mathcal{R}_{\psi,D}$ there exists a sequence of vectors $\{\xi_j\}$ such that $\xi_j \notin \mathcal{R}_{\psi,D}$ for $j = 1, 2, \ldots$ and

$$\xi_j \to \psi(\bar{x}(1)) \quad \text{as } j \to \infty.$$

We may construct, using standard procedures (see, e.g., [20]), a sequence of nonnegative-valued L^∞ functions $\{\bar{m}_j(\cdot)\}$ such that

$$\bar{m}_j(t)dt \to \bar{\mu}(dt) \qquad \text{weakly*}.$$

Let $\bar{x}_j(\cdot)$ be the state trajectory corresponding to the conventional and measure controls, $\bar{u}(\cdot)$ and $m_j(t)dt$, and initial state $\bar{x}_0$. According to Lemma 5.2,

$$\bar{x}_j \to \bar{x}(t) \qquad \forall\ t \in C_{\bar{\mu}} \cup \{0, 1\}$$

and

$$d\bar{x}_j(t) \to d\bar{x}(t) \qquad \text{weakly*}.$$

Now define $\epsilon_j := |\xi_j - \psi(\bar{x}(t))|^{1/2}$. Note that, by the properties of $\{\xi_j\}$, $\epsilon_j \to 0$ as $i \to \infty$.

For each j, set $r_j := j + \|\bar{m}_j(\cdot)\|_{L^\infty}$ and consider the optimal control problem

$$(\overline{\mathrm{P}}_j) \qquad \begin{cases} \text{Minimize} \quad |\xi_j - \psi(\bar{x}(t))| \\ \text{subject to} \quad \dot{x}(t) = f\,(t, x(t), u(t)) + g(t, x(t))m(t), \quad t \in\ [\,0, 1], \\ (u(t),\ m(t)) \in U_t \times [0, r_j] \ \ \mathcal{L}\text{-a.e.}, \\ x_0 \in D. \end{cases}$$

Since the dynamic constraint defines a unique state trajectory $x_{a,u,m}(\cdot)$ corresponding to a given control function $(u(\cdot), m(\cdot))$ and initial state a, we may regard this as a minimization problem over triples $(a, u(\cdot), m(\cdot))$ such that $a \in D$; $u(\cdot)$ and $m(\cdot)$ are measurable functions; and $(u(t), m(t)) \in U_t \times [0, r_j]$ a.e. Denote this collection of triples by W_j. We provide W_j with the metric

$$\begin{aligned} \rho((a, u, m); (a', u', m')) := |a - a'| \quad &+ \ \mathcal{L}\text{-meas}\{t \in [0, 1] : u(t) \neq u'(t)\} \\ &+ \int_0^1 |m(t) - m'(t)|dt. \end{aligned}$$

The newly reformulated problem can be expressed as

$$\text{minimize}\{\Psi_j(a, u, m) :\ (a, u, m) \in W_j\},$$

in which

$$\Psi_j(a, u, m) := |\xi_j - \psi(x_{a,u,m}(1))|.$$

By Lemma 5.2, W_j is a complete space and Ψ_j is continuous with respect to the above metric. According to Ekeland's theorem [5] then there exists a triple $(a_j, u_j, m_j) \in W_j$ such that

$$\Psi_j(a_j, u_j, m_j) \leq \Psi_j(a, u, m) + \epsilon_j \rho((a, u, m), (a_j, u_j, m_j)) \tag{5.5}$$

for all $(a, u, m) \in W_j$ and

$$\rho(\bar{x}(0), \bar{u}, \bar{m}_j), (a_j, u_j, m_j)) \leq \epsilon_j \tag{5.6}$$

for each j. Let x_j be the trajectory corresponding to (a_j, u_j, m_j). Since $\epsilon_j \to 0$ we deduce from (5.6) that

$$\begin{gathered} x_j(0) \to \overline{x}(0), \\ \mathcal{L} - \text{meas}\{t \in [0, 1] : u_j(t) \neq \overline{u}(t)\} \to 0, \end{gathered}$$

and

$$m_j(t)dt \to \overline{\mu}(dt) \quad \text{ weakly*}.$$

For each j, (5.5) implies that $(x_j(\cdot), u_j(\cdot), m_j(\cdot))$ is a minimizer for the problem

$$\begin{aligned} \text{Minimize} \quad & |\xi_j - \psi(x(1))| + \epsilon_j \left\{ |x(0) - x_j(0)| \right. \\ & \qquad \left. + \int_0^1 |m_j(t) - m(t)|dt + \int_0^1 \chi_j(t, u(t))dt \right\} \\ \text{subject to} \quad & \dot{x}(t) = f(t, x(t), u(t)) + g(t, x(t))m(t) \quad \mathcal{L}\text{-a.e.}, \\ & (u(t), m(t)) \in U_t \times [0, r_j] \quad \mathcal{L}\text{-a.e.}, \end{aligned}$$

in which

$$\chi_j(t,u) := \begin{cases} 0 & \text{if } u = u_j(t), \\ 1 & \text{if } u \neq u_j(t). \end{cases}$$

Now apply the maximum principle to this last problem (see, e.g., [17]). Since $\xi_j \notin \Psi(x_j(1))$, this tells us that there exists an absolutely continuous function p_j and a vector $d_j \in \Re^k$ of unit length such that

$$-\dot{p}_j(t) \in p_j \cdot \mathrm{co}\partial_x f(t, x_j(t), u_j(t)) + p_j \cdot \mathrm{co}\partial_x g(t, x_j(t))m_j, \quad t \in [0,1],$$

$$p_j(0) \in N_D(x_j(0)) + \epsilon_j \partial_x |x - x_0|_{x=x_j(0)}, \tag{5.7}$$

$$-p_j(1) \in d_j \cdot \partial\psi(x_j(1)), \tag{5.8}$$

and

$$H_j(t, u_j(t), m_j(t)) = \max\{H_j(t,u,m) : \ u \in U_t, \ m \in [0, r_j]\} \text{ a.e.} \tag{5.9}$$

Here

$$H_j(t,u,m) := p_j(t) \cdot f(t, x_j(t), u) + p_j(t) \cdot g(t, x_j(t))m - \epsilon_j(|m - m_j(t)| + \chi_j(t,u)).$$

Notice that the maximization of the Hamiltonian condition (5.9) implies that $m_j(t) = r_j$ a.e. on the set $\{t : p_j(t) \cdot g(t, x_j(t) > \epsilon_j\}$. Since $m_j(t)dt \to \bar{\mu}(dt)$ weakly*, the sequence $\{m_j\}$ is bounded in L^1 norm. Observing that $r_j \to \infty$, we deduce

$$\mathcal{L}\text{-meas}\{t : p_j(t) \cdot g(t, x_j(t) \leq \epsilon_j\} \to 1 \quad \text{as } j \to \infty. \tag{5.10}$$

It follows also from (5.9) that $p_j(t) \cdot g(t, x_j(t)) \geq -\epsilon_j$ a.e. (with respect to Lebesgue measure) on $\{t : m_j(t) > 0\}$. This property can be expressed as

$$\int_0^1 \max\{-p_j(t) \cdot g(t, x_j(t)) - \epsilon_j, \ 0\} m_j(t) dt = 0, \quad j = 1, 2, \ldots. \tag{5.11}$$

We can regard x_j and p_j as solutions to the following MDI, corresponding to controls $(u, \mu) = (u_j, \ m_j(t)dt)$ and initial values $(x_j(0), p_j(0))$:

$$d(x(t), p(t)) \in F_1(t, x(t), p(t), u(t))dt + F_2(t, x(t), p(t))\mu(dt), \tag{5.12}$$

in which

$$F_1(t,x,p,u) := \{f(t,x,u)\} \times \{-p \cdot \mathrm{co}\partial_x f(t,x,u)\}$$

and

$$F_2(t,x,p) := \{g(t,x)\} \times \{-p \cdot \overline{\partial_x} g(t,x)\}.$$

Incorporation of the hybrid gradient $\overline{\partial_x g}$ in these relationships ensures that F_2 has the requisite upper semicontinuity properties for application of the convergence results of section 5.

Bearing in mind the sequences $\{x_j\}$ and $\{p_j\}$ are uniformly bounded (in the case of p_j this follows from the uniform bound on the right endpoints, and an application of Gronwall's lemma), we deduce from Proposition 4.2 that (following an extraction

of subsequences) there exist $\overline{x} \in BV^+([0,1],\Re^n)$, $p \in BV^+([0,1],\Re^n)$, and absolutely continuous functions $y(\cdot)$ and $q(\cdot)$ with the following properties:

$$\begin{aligned} &x_j(t) \to \overline{x}(t) \quad \text{and} \quad p_j(t) \to p(t) \quad \forall \quad t \in \mathcal{C}_\mu \cup \{0,1\}, \\ &x_j(\theta_j(s)) \to y(s) \quad \text{and} \quad p_j(\theta_j(s)) \to q(t) \text{ uniformly,} \end{aligned} \tag{5.13}$$

and

$$\bar{x}(t) = y(\eta(t)) \quad \text{and} \quad p\,(t) = q(\eta(t)) \quad \forall \quad t \in [0,1]. \tag{5.14}$$

Here η is the reparameterization function of μ, (θ,γ) is the graph completion of μ, and (θ_j,γ_j) is the graph completion of μ_j for $j = 1,2,\ldots$. Furthermore $(\overline{x}(\cdot),p(\cdot))$ is a robust solution of the MDI (5.12), and y and q satisfy the differential inclusion

$$\begin{aligned} (\dot{y}(s),\dot{q}(s)) \in &\{f(\theta(s),y(s),\overline{u}(\theta(s)))\} \times \{-q(s)\cdot \mathrm{co}\partial_x f(\theta(s),y(s),\overline{u}(\theta(s)))\}\dot{\theta}(s) \\ &+ \{g(\theta(s),y(s))\} \times \{-q(s)\cdot \overline{\partial_x} g(\theta(s),y(s))\}\dot{\gamma}(s), \quad s \in [0,1]. \end{aligned} \tag{5.15}$$

Notice that we are justified in taking the cluster point of the sequence $\{x_j(\cdot)\}$ to be the "boundary" state trajectory $\overline{x}(\cdot)$ because the *original* sequence is known to have converged to $\overline{x}(\cdot)$ on $\mathcal{C}_\mu \cup \{0,1\}$, and two functions in $BV^+([0,1];\Re^n)$ coincide if they have the same values on a dense set including $\{0,1\}$.

We now arrange, by further subsequence extraction, that the sequence of vectors $\{d_j\}$ in (5.8) has a limit:

$$d_j \to d \quad \text{as } j \to \infty$$

for some vector d of unit length.

We have seen that $(\overline{x},p)$ can be interpreted as a robust solution of the combined state and costate equations (5.12). Our goal now is to show that the $p(\cdot)$ and d we have constructed satisfy the remaining conditions in the theorem statement.

For each j let S_j be the set of points $t \in [0,1]$ such that

$$p_k(t)\cdot g(t,x_k(t)) \le \epsilon_k \qquad \forall\ k \ge j,$$

$$u_k(t) = \overline{u}(t) \qquad \forall\ k \ge j,$$

$$H_k(t,u_k(t),m_k(t)) = \max\{H_k(t,u,m) : u \in U_t, m \in [0,r_k]\} \quad \forall\ k \ge j,$$

and

$$x_k \to \overline{x}(t), \quad p_k(t) \to p(t).$$

In view of (5.6), (5.9), (5.10), and (5.13), we can arrange by a further subsequence extraction that $\mathcal{L}$-meas $\{S_j\} \to 1$.

Take any $t \in \cup_j S_j$ and any $u \in U_t$. The above relationships imply that, in the limit,

$$p(t)\cdot f(t,\overline{x}(t),\overline{u}(t)) = \max\{p(t)\cdot f(t,\overline{x}(t),u) :\ u \in U_t\}$$

and

$$p(t)\cdot g(t,\overline{x}(t)) \le 0.$$

We have shown (4.5) on $\cup_j S_j$. But then (4.5) holds on $(0,1)$ because $\cup_j S_j$ is dense and $t \to p(t) \cdot g(t, \overline{x}(t))$ is right continuous on $(0,1)$. The boundary conditions on the costate function $p(\cdot)$ are verified by passage to the limit in (5.7) and (5.8).

Next we examine the consequences of (5.11). Since $\{m_j(\cdot)\}$ is bounded in L^1 norm, we know that

$$\lim_{j\to\infty} \int_0^1 \max\{-p_j(t) \cdot g(t, x_j(t)),\ 0\} m_j(t) dt = 0.$$

Applying the change-of-variables lemma and using the facts that

$$p_j(\theta_j(s)) \cdot g(\theta_j(s), x_j(\theta_j(s))) \to q(s) \cdot g(\theta(s), y(s))$$

uniformly and

$$\dot{\gamma}_j(\cdot) \to \dot{\gamma}(\cdot) \quad \text{weakly in } L^1$$

(see Proposition 2.1), we conclude that

$$\begin{aligned} &\int_0^1 \max\{-q(s) \cdot g(\theta(s), y(s)),\ 0\} \dot{\gamma}(s) ds \\ &= \lim_j \int_0^1 \max\{-p_j(\theta_j(s)) \cdot g(\theta_j(s), x_j(\theta_j(s))),\ 0\} \dot{\gamma}_j(s) ds = 0. \end{aligned}$$

Since the integrand of the expression on the left is nonnegative, it follows that

$$\int_{[0,1]\setminus(\cup_i I_i)} \max\{-p(\theta(s)) \cdot g(\theta(s), \overline{x}(\theta(s))),\ 0\} \dot{\gamma}(s) ds = 0 \tag{5.16}$$

(we have noted that $\overline{x}(\theta(s)) = y(s),\ p(\theta(s)) = q(s)$ on $[0,1] \setminus (\cup_i I_i)$) and, for each i,

$$\int_{[s_i^{'}, s_i^{''}]} \max\{-q(s) \cdot g(t_i, y(s)),\ 0)\} ds = 0 \tag{5.17}$$

(in view of the fact that $\dot{\gamma} \equiv 1 + \mu([0,1])$ on $[s_i^{'}, s_i^{''}]$). Here $\{t_i\}$ is an enumeration of the atoms of μ and

$$[s_i^{'}, s_i^{''}] = I_i = \theta^{-1}(\{t_i\}),\ i = 1, 2, \ldots.$$

The change-of-variables lemma applied to (5.16) gives

$$\int_{[0,1]\setminus\cup\{t_i\}} \max\{-p(t) \cdot g(t, \overline{x}(t)),\ 0\} \overline{\mu}(dt) = 0. \tag{5.18}$$

Since the integrand in (5.17) is continuous, we conclude from this equation that

$$q(s) \cdot g(t_i, y(s)) \geq 0 \quad \forall\ s \in [s_i^{'}, s_i^{''}],\ i = 1, 2, \ldots. \tag{5.19}$$

We note, however, that, by (5.14), the continuity of y and q, and the continuity of $\bar{x}$ from the right,

$$\bar{x}(t_i) = y(s_i^{''}) \quad \text{and} \quad p(t_i) = q(s_i^{''}), \tag{5.20}$$

and

$$x(t_i^-) = y(s_i^{'}) \quad \text{and} \quad p(t_i^-) = q(s_i^{'}). \tag{5.21}$$

Equations (5.19) and (5.20) now give

$$p(t_i) \cdot g(t, \overline{x}(t_i)) = q(s_i^{''}) \cdot g(t_i, y(s_i^{''})) \geq 0$$

for all i. Combining this relationship with (5.18) we conclude that

$$p(t) \cdot g(t, \overline{x}(t) \geq 0 \quad \mu\text{-a.e.}$$

Next, for each i, define $\xi_i : [0,1] \to \Re^n,\ \pi_i : [0,1] \to \Re^n$ to be

$$\xi_i(\sigma) = y(s_i^{'} + \sigma(s_i^{''} - s_i^{'})),\ \pi_i(\sigma) = q(s_i^{'} + \sigma(s_i^{''} - s_i^{'})).$$

By (5.20) and (5.21)

$$(\xi_i(0), \pi_i(0)) = (x(t_i^-), p(t_i^-)), \quad (\xi_i(1), \pi_i(1)) = (x(t_i), p(t_i))$$

for each i. Furthermore, since $\dot{\theta}(s) \equiv 0$ and $\dot{\gamma}(s) \cdot |s_i^{''} - s_i^{'}| \equiv \mu(\{t_i\})$ on $[s_i^{'}, s_i^{''}]$, we deduce from (5.2) that

$$(\dot{\xi}_i(s), \dot{\pi}_i(s)) \in g(t_i, \xi_i(s)) \times (-\xi_i(s) \cdot \overline{\partial_x} g(t_i, \xi_i(s)) \mu(\{t_i\}) \quad \text{a.e.}$$

This concludes the proof of the theorem in the case that ($\bar{\text{H}}$) is added to the hypotheses. It remains to deal with the case when ($\bar{\text{H}}$) is not valid. For each index value k replace $U(t)$ with $\widetilde{U}_k(t)$:

$$\widetilde{U}_k(t) = U(t) \bigcap \{u \in \Re^m : |f(t, \overline{x}(t), u)| \leq |f(t, \overline{x}(t), \overline{u}(t))| + k\}.$$

We see that $(\overline{x}, \overline{u}, \overline{\mu})$ remains a minimizer. Furthermore, ($\bar{\text{H}}$) is now satisfied. By our earlier analysis there exist $(p_k(\cdot), d_k)$ with the properties described in the above theorem (except that now $\widetilde{U}_k(t)$ replaces $U(t)$). Passage to the limit as $k \to \infty$ (the convergence analysis is along the lines of the first half of the proof) yields the assertions of Theorem 4.1 in this case also. □

REFERENCES

[1] A. Bressan, *On differential systems with impulsive controls*, Rend. Sem. Mat. Univ. Padova, 78 (1987), pp. 227–236.

[2] A. Bressan and F. Rampazzo, *On differential systems with vector valued impulsive controls*, Boll. Un. Mat. Ital. B (7), 3 (1988), pp. 641–656.

[3] A. Bressan and F. Rampazzo, *Impulsive control systems with commutative vector fields*, J. Optim. Theory Appl., 71 (1991), pp. 67–83.

[4] A. Bressan and F. Rampazzo, *Impulsive control systems without commutativity assumptions*, J. Optim. Theory Appl., 81 (1994), pp. 435–457.

[5] F. H. Clarke, *Optimization and Nonsmooth Analysis*, John Wiley, New York, 1983.

[6] G. Dal Maso and F. Rampazzo, *On systems of ordinary differential equations with measures as controls*, Differential and Integral Equations, 4 (1991), pp. 739–765.

[7] J. F. C. Kingman and S. J. Taylor, *Introduction to Measure and Probability*, Cambridge Univ. Press, Cambridge, UK, 1966.

[8] D. F. Lawden, *Optimal Trajectories for Space Navigation*, Butterworth, London, 1663.

[9] P. D. Loewen, *Optimal Control Via Nonsmooth Analysis*, CRM Proceedings and Lecture Notes, American Mathematical Society, Providence, RI, 1993.

[10] J. P. Marec, *Optimal Space Trajectories*, Elsevier, New York, 1979.
[11] B. M. Miller, *Optimality conditions in generalized control problems*, I, II, Automat. Remote Control, 53 (1992), pp. 362–370 and pp. 505–513.
[12] B. M. Miller, *Method of discontinuous time change in problems of control for impulse and discrete continuous system*, Automat. Remote Control, 54 (1993), pp. 1727–1750.
[13] B. M. Miller, *The generalized solutions of nonlinear optimization problems with impulse control*, SIAM J. Control Optim., 34 (1996), pp. 1420–1440.
[14] B. S. Mordukhovich, *Generalized differential calculus for nonsmooth and set-valued mappings*, J. Math. Anal. Appl., 183 (1994), pp. 250–288.
[15] M. Motta and F. Rampazzo, *Dynamic programming for nonlinear systems driven by ordinary and impulsive control*, SIAM J. Control Optim., 34 (1996), pp. 199–225.
[16] R. W. Rishel, *An extended Pontryagin principle for control systems whose control laws contain measures*, SIAM J. Control, 3 (1965), pp. 191–205
[17] J. D. L. Rowland and R. B. Vinter, *Dynamic optimization for problems with free time and active state constraints*, SIAM J. Control Optim., 31 (1993), pp. 677–697.
[18] G. N. Silva and R. B. Vinter, *Measure differential inclusions*, J. Math. Anal. Appl., 202 (1996), pp. 727–746.
[19] H. J. Sussmann, *On the gap between deterministic and stochastic ordinary differential equations*, Ann. Probab., 6 (1978), pp. 17–41.
[20] R. B. Vinter and G. Pappas, *A maximum principle for nonsmooth optimal control problems with state constraints*, J. Math. Anal. Appl., 89 (1982), pp. 212–232.
[21] R. B. Vinter and F. M. L. Pereira, *A maximum principle for optimal processes with discontinuous trajectories*, SIAM J. Control Optim., 26 (1988), pp. 205–229.
[22] J. Warga, *Optimal Control of Differential and Functional Equations*, Academic Press, New York, 1972.
[23] J. Yong, *Volterra-Stieltjes evolution equations and related optimal control problems*, SIAM J. Control Optim., 30 (1993), pp. 539–568.

SIAM J. CONTROL OPTIM.
Vol. 35, No. 6, pp. 1847–1885, November 1997

002

DYNAMICS AND APPROXIMATIONS OF A VELOCITY TRACKING PROBLEM FOR THE NAVIER–STOKES FLOWS WITH PIECEWISE DISTRIBUTED CONTROLS*

L. S. HOU† AND Y. YAN‡

Abstract. We study the dynamics of a piecewise (in time) distributed optimal control problem for the Navier–Stokes equations which models velocity tracking over time. We also study the dynamics of semidiscrete and fully discrete approximations of this velocity tracking problem. We prove that the rates of velocity tracking are exponential. Some computational results are presented, which reinforces the theoretical results derived.

Key words. optimal control, Navier–Stokes equations, semidiscrete approximations, fully discrete approximations

AMS subject classifications. 35B40, 35B37, 35Q30, 65M60

PII. S036301299529286X

1. Introduction.

1.1. Motivation. This work is motivated by the desire to steer over time a candidate viscous, incompressible velocity field $\mathbf{u}$ to a target velocity field $\mathbf{U}$ by appropriately controlling (adjusting) the body force. The control of body force can be affected by appropriately adjusting the external magnetic field in the case of electromagneto fluids such as liquid metal. Thus the study of body force control problems is important in its own right. Boundary velocity control is far more common in practical applications than body force control, but there has been a lack of adequate mathematical theories for both the (uncontrolled) boundary value problems and the boundary control problems for viscous, incompressible flows (see [FGH] for some recent progress in this direction). For this reason, the study of body force control problems is also important in that it provides some insight into boundary velocity control problems.

In [HY], a velocity tracking problem on the infinite time interval was formulated as an optimal control problem: find a triplet $(\mathbf{u}, p, \mathbf{f})$ such that the functional

$$\mathcal{J}_{(0,\infty)}(\mathbf{u}, \mathbf{f}) = \frac{\alpha}{2} \int_0^\infty \int_\Omega |\mathbf{u} - \mathbf{U}|^2 \, d\mathbf{x} \, dt + \frac{\beta}{2} \int_0^\infty \int_\Omega |\mathbf{f} - \mathbf{F}|^2 \, d\mathbf{x} \, dt$$

is minimized subject to the two-dimensional Navier–Stokes equations with an initial condition $\mathbf{u}_0$ and the zero boundary condition. Here, Ω is a two-dimensional bounded domain which is convex or of class C^2, and $\partial\Omega$ denotes its boundary; $\mathbf{u}$ and p denote the velocity field and the pressure field, respectively; $\mathbf{f}$ is the body force—the control field; $\mathbf{u}_0$ is the initial velocity field; and $\mathbf{F}$ is a prescribed body force. α and β are two positive parameters that adjust the relative weight of the two terms in the functional. Also, the viscosity $\nu > 0$ appears in the Navier–Stokes equations. This optimal control

*Received by the editors October 6, 1995; accepted for publication (in revised form) July 31, 1996. http://www.siam.org/journals/sicon/35-6/29286.html

†Department of Mathematics and Statistics, York University, North York, ON M3J 1P3, Canada (hou@mathstat.yorku.ca). The research of this author was supported in part by Natural Science and Engineering Research Council of Canada grant OGP-0169786.

‡Department of Mathematics, Virginia Polytechnic Institute and State University, Blacksburg, VA 24061-0531 (yan@math.ut.ued). The research of this author was supported in part by Natural Science Foundation grant DMS-9626154.

problem on the infinite time interval is of both physical and mathematical interest. The dynamics of the solutions for this problem were studied and an optimality system of equations was derived in [HY]. In particular, it was shown that $\|\mathbf{u}(t)-\mathbf{U}(t)\|_{\mathbf{L}^2(\Omega)} \to 0$ and $\|\nabla\mathbf{u}(t) - \nabla\mathbf{U}(t)\|_{\mathbf{L}^2(\Omega)} \to 0$ as $t \to \infty$. However, the computation of the controlled flows based on solving the derived optimality system is difficult due to the fact that the system needs to be resolved on the infinite time interval $(0,\infty)$ and more importantly that it involves a coupled system of the state variables $(\mathbf{u}, p)$ and the adjoint state variables $(\boldsymbol{\mu}, \pi)$ with forward and backward "initial" conditions—in other words, this system has to be solved on the entire time-space cylinder and it cannot be solved by marching in time. Also, although the $\mathbf{L}^2(\Omega)$ norm and $\mathbf{H}^1(\Omega)$ norm of the difference between the controlled flow $\mathbf{u}$ and the desired flow $\mathbf{U}$ both decay to zero in time, the decay rates can be slow. Therefore, there are practical interests and needs to develop other approaches for velocity tracking on the infinite interval that are devoid of these aforementioned difficulties.

The purpose of this paper is to design an appropriate optimal control mechanism—piecewise (in time) distributed controls—which possesses the following features: the physical objective of tracking the velocity field over time is achieved; $\mathbf{u}(t) - \mathbf{U}(t)$ and $\nabla\mathbf{u}(t) - \nabla\mathbf{U}(t)$ decay to zero exponentially as $t \to \infty$; and the optimal control problem can be solved numerically by marching in time. Precisely, we will study the following piecewise (in time) optimal control problem:

- First, choose a sufficiently small $\delta > 0$, choose a sequence $\{t_n\}_{n=0}^{\infty}$ defined by $t_n = n\delta$, and define $\widehat{\mathbf{u}}^{(0)}(0) = \mathbf{u}_0$.
- Then, inductively, for each $n \geq 0$, find a solution $(\widehat{\mathbf{u}}^{(n+1)}, \widehat{p}^{(n+1)}, \widehat{\mathbf{f}}^{(n+1)})$ on the interval (t_n, t_{n+1}) which minimizes the (localized) functional

$$\begin{aligned} &\mathcal{J}_{(t_n,t_{n+1})}(\mathbf{u},\mathbf{f}) \\ &\stackrel{\text{def}}{=\!=} \frac{\alpha}{2}\int_{t_n}^{t_{n+1}}\int_{\Omega}|\mathbf{u}-\mathbf{U}|^2\,d\mathbf{x}\,dt + \frac{\beta}{2}\int_{t_n}^{t_{n+1}}\int_{\Omega}|\mathbf{f}-\mathbf{F}|^2\,d\mathbf{x}\,dt \end{aligned} \tag{1.1}$$

subject to the (localized) two-dimensional Navier–Stokes equations:

$$\partial_t\mathbf{u} - \nu\Delta\mathbf{u} + (\mathbf{u}\cdot\nabla)\mathbf{u} + \nabla p = \mathbf{f} \qquad \text{in } \Omega\times(t_n,t_{n+1})\,, \tag{1.2}$$

$$\operatorname{div}\mathbf{u} = 0 \qquad \text{in } \Omega\times(t_n,t_{n+1})\,, \tag{1.3}$$

$$\mathbf{u} = \mathbf{0} \qquad \text{on } \partial\Omega\times(t_n,t_{n+1}), \tag{1.4}$$

and

$$\mathbf{u}(\cdot,t_n) = \widehat{\mathbf{u}}^{(n)}(t_n) \qquad \text{in } \Omega\,. \tag{1.5}$$

We define a global (in time) solution $(\widehat{\mathbf{u}}, \widehat{p}, \widehat{\mathbf{f}})$ by patching together all the local optimal control solutions $(\widehat{\mathbf{u}}^{(n)}, \widehat{p}^{(n)}, \widehat{\mathbf{f}}^{(n)})$, i.e., $\widehat{\mathbf{u}}|_{(t_n,t_{n+1})} \stackrel{\text{def}}{=\!=} \widehat{\mathbf{u}}^{(n+1)}$, $\widehat{p}|_{(t_n,t_{n+1})} \stackrel{\text{def}}{=\!=} \widehat{p}^{(n+1)}$ and $\widehat{\mathbf{f}}|_{(t_n,t_{n+1})} \stackrel{\text{def}}{=\!=} \widehat{\mathbf{f}}^{(n+1)}$ for all n. Note that this global solution is determined by marching in time with a sequence of local optimal control problems; therefore, a velocity tracking problem on a finite time interval $(0,T)$ with $T < \infty$ can be studied in the same manner by patching together only a finite number of the local optimal control solutions. In the finite time interval case, we terminate the marching in time after a finite number of time steps.

In this paper, we will analyze the long-time behavior of global solutions for the piecewise optimal control problem; define and analyze the dynamics of semidiscrete approximations; and define and analyze the dynamics of fully discrete approximations. The main results of this paper are that the $\mathbf{L}^2(\Omega)$ and $\mathbf{H}^1(\Omega)$ norms of the difference between $\mathbf{U}(t)$ and $\widehat{\mathbf{u}}(t)$ decay to zero exponentially in time, with the rates independent of the interval length δ; the same is true of the $\mathbf{L}^2(\Omega)$ and $\mathbf{H}^1(\Omega)$ norms of the difference between $\mathbf{U}(t_n)$ and $\widehat{\mathbf{u}}^n$ ($\widehat{\mathbf{u}}^n$ denotes the solution of the semidiscrete piecewise optimal control problem), modular the error due to the approximation of the time derivative; the same is also true of the $\mathbf{L}^2(\Omega)$ norm of the difference between $\mathbf{U}(t_n)$ and $\widehat{\mathbf{u}}_h^n$ ($\widehat{\mathbf{u}}_h^n$ denotes the solution of the fully discrete piecewise optimal control problem), modular the error due to the approximation of the time derivative and the error due to spatial discretizations. In short, the piecewise optimal control approach achieves the physical objective of steering $\mathbf{u}$ to $\mathbf{U}$ over time. This analytical result is also supported by the error analysis and by the computational results presented at the end of the paper.

We mention that in [HRY], further computational results of the fully discrete schemes were presented and some implementation issues were discussed. It should also be noted that the fully discrete approximation schemes bear similarities to those used in some other situations, e.g., [CTMK] for the Burgers equation and [Ra] for an electrically conducting fluid, wherein fully discrete schemes were introduced from a viewpoint different from that taken in this paper and wherein no error analysis were given.

1.2. Preliminaries. Throughout this paper, C denotes a generic constant depending only on the physical domain Ω and the viscosity constant ν. We will use the standard notations for the function spaces $L^r(\Omega)$ with the norm denoted by $\|\cdot\|_{L^r(\Omega)}$ and the Sobolev spaces $H^m(\Omega)$ with the norm denoted by $\|\cdot\|_m$. $H^0(\Omega) = L^2(\Omega)$. Also, $H_0^m(\Omega)$ is the closure of $C_0^\infty(\Omega)$ under the $\|\cdot\|_m$ norm. The dual space of $H_0^r(\Omega)$ is denoted by $H^{-r}(\Omega)$, $r > 0$. The vector-valued ($\mathbb{R}^2$-valued) counterparts of these spaces are denoted by $\mathbf{L}^r(\Omega)$, $\mathbf{H}^m(\Omega)$, $\mathbf{H}_0^m(\Omega)$, and $\mathbf{H}^{-m}(\Omega)$. For details, see [Ad] and [GR]. We introduce the solenoidal spaces

$$\mathbf{W} = \{\mathbf{u} \in \mathbf{L}^2(\Omega) \ : \ \nabla \cdot \mathbf{u} = 0, \ (\mathbf{u} \cdot \mathbf{n})|_\Gamma = 0\} \qquad \text{equipped with the } \|\cdot\|_0 \text{ norm}$$

and

$$\mathbf{V} = \{\mathbf{u} \in \mathbf{H}_0^1(\Omega) \ : \ \nabla \cdot \mathbf{u} = 0\} \qquad \text{equipped with the } \|\cdot\|_1 \text{ norm},$$

i.e., the closure of div-free, $\mathbf{C}_0^\infty(\Omega)$-functions under the $\|\cdot\|_0$ and $\|\cdot\|_1$ norms, respectively; see [GR] or [Te] for a discussion of this functional spaces. We identify the dual space of $\mathbf{W}$ with $\mathbf{W}$ itself under the $\mathbf{L}^2(\Omega)$ inner product. We will also need the following subspace of $L^2(\Omega)$:

$$L_0^2(\Omega) \stackrel{\text{def}}{=\!=} \left\{ r \in L^2(\Omega) : \int_\Omega r \, d\mathbf{x} = 0 \right\}.$$

We next introduce the temporal-spatial function spaces, defined on $\Omega \times (T_1, T_2)$ with $(T_1, T_2) \subset (0, \infty)$, $r \in [1, \infty]$, and $s \in (-\infty, \infty)$:

$$L^r\left(T_1, T_2; \mathbf{H}^s(\Omega)\right) \qquad r \in [1, \infty], \ s \in (-\infty, \infty),$$

equipped with the norm

$$\|\mathbf{u}\|_{L^r(T_1,T_2;\mathbf{H}^s(\Omega))} = \left(\int_{T_1}^{T_2} \|\mathbf{u}(t)\|_s^r \, dt \right)^{1/r}$$

and with $s \geq 1$:

$$\mathcal{H}^{(s)}\left(\Omega \times (T_1, T_2)\right) = \left\{ \mathbf{v} \in L^2\left(T_1, T_2; \mathbf{H}^s(\Omega) \cap \mathbf{H}_0^1(\Omega)\right) : \partial_t \mathbf{v} \in L^2\left(T_1, T_2; \mathbf{H}^{s-2}(\Omega)\right) \right\}$$

equipped with the norm

$$\|\mathbf{v}\|^2_{\mathcal{H}^{(s)}(\Omega\times(T_1,T_2))} = \|\mathbf{v}\|^2_{L^2(T_1,T_2;\mathbf{H}^s(\Omega))} + \|\partial_t \mathbf{v}\|^2_{L^2(T_1,T_2;\mathbf{H}^{s-2}(\Omega))}.$$

For a function $\mathbf{u}$ in a temporal-spatial space, we often use the notation

$$\mathbf{u}(t) \stackrel{\text{def}}{=\!=} \mathbf{u}(\cdot, t)$$

to stand for the restriction of $\mathbf{u}$ at time t as a function defined over the spatial domain Ω.

We introduce the simplified norm notations

$$\|\cdot\| \stackrel{\text{def}}{=\!=} \|\cdot\|_{L^2(\Omega)} \qquad \text{or} \qquad \|\cdot\| \stackrel{\text{def}}{=\!=} \|\cdot\|_{\mathbf{L}^2(\Omega)}, \tag{1.6}$$

$$|||\cdot||| \stackrel{\text{def}}{=\!=} \|\cdot\|_{L^\infty(0,\infty;\mathbf{L}^2(\Omega))}, \tag{1.7}$$

and

$$|||\cdot|||_\infty \stackrel{\text{def}}{=\!=} \|\cdot\|_{\mathbf{L}^\infty(\Omega\times(0,\infty))}. \tag{1.8}$$

Also, $(\cdot,\cdot)$ denotes the inner product on $L^2(\Omega)$ or $\mathbf{L}^2(\Omega)$ and $\langle\cdot,\cdot\rangle$ denotes the duality pairing between $\mathbf{H}_0^1(\Omega)$ and $\mathbf{H}^{-1}(\Omega)$.

To define the solution for the Navier–Stokes equations (in a weak sense), we introduce some standard continuous bilinear or trilinear forms:

$$a(\mathbf{u}, \mathbf{v}) = \nu \int_\Omega \nabla \mathbf{u} : \nabla \mathbf{v}\, d\mathbf{x} \qquad \forall\, \mathbf{u}, \mathbf{v} \in \mathbf{H}^1(\Omega),$$

$$b(\mathbf{u}, p) = -\int_\Omega p \operatorname{div} \mathbf{u}\, d\mathbf{x} \qquad \forall\, \mathbf{u} \in \mathbf{H}^1(\Omega); \forall\, p \in L^2(\Omega),$$

and

$$c(\mathbf{u}, \mathbf{v}, \mathbf{w}) = \int_\Omega (\mathbf{u}\cdot\nabla)\mathbf{v}\cdot\mathbf{w}\, d\mathbf{x} \qquad \forall\, \mathbf{u}, \mathbf{v}, \mathbf{w} \in \mathbf{H}^1(\Omega),$$

where the colon notation: denotes the inner product on $\mathbb{R}^{2\times 2}$. We have the following useful properties for the trilinear form $c(\cdot,\cdot,\cdot)$:

$$c(\mathbf{u}, \mathbf{v}, \mathbf{w}) = -c(\mathbf{u}, \mathbf{w}, \mathbf{v}) \qquad \forall\, \mathbf{u} \in \mathbf{V},\ \forall\, \mathbf{v}, \mathbf{w} \in \mathbf{H}^1(\Omega) \tag{1.9}$$

and

$$c(\mathbf{u}, \mathbf{v}, \mathbf{v}) = 0 \qquad \forall\, \mathbf{u} \in \mathbf{V},\ \forall\, \mathbf{v} \in \mathbf{H}^1(\Omega). \tag{1.10}$$

We denote by C_0 the continuity constant for the trilinear form $c(\cdot,\cdot,\cdot)$, i.e.,

$$|c(\mathbf{u}, \mathbf{v}, \mathbf{w})| \leq C_0 \|\nabla \mathbf{u}\|\, \|\nabla \mathbf{v}\|\, \|\mathbf{w}\|_{\mathbf{L}^4(\Omega)} \qquad \forall\, \mathbf{u} \in \mathbf{V},\ \forall\, \mathbf{v}, \mathbf{w} \in \mathbf{H}^1(\Omega). \tag{1.11}$$

We now give the following definitions of a solution for the Navier–Stokes equations on a finite time interval and on the infinite time interval, respectively.

DEFINITION 1.1. *Given* $(T_1, T_2) \subset (0, \infty)$ *with* $T_2 < \infty$, $\mathbf{f} \in L^2(T_1, T_2; \mathbf{H}^{-1}(\Omega))$ *and* $\boldsymbol{\xi} \in \mathbf{W}$, $(\mathbf{u}, p)$ *is said to be a solution of the Navier–Stokes equations on* (T_1, T_2) *iff* $\mathbf{u} \in \mathcal{H}^{(1)}(\Omega \times (T_1, T_2))$, $p \in L^2(T_1, T_2; L^2_0(\Omega))$, *and* $(\mathbf{u}, p)$ *satisfies*

$$\begin{aligned} \langle \partial_t \mathbf{u}(t), \mathbf{w} \rangle + a(\mathbf{u}(t), \mathbf{w}) + c(\mathbf{u}(t), \mathbf{u}(t), \mathbf{w}) + b(\mathbf{w}, p(t)) \\ = \langle \mathbf{f}(t), \mathbf{w} \rangle \qquad \forall\, \mathbf{w} \in \mathbf{H}^1_0(\Omega),\ \textit{a.e. } t \in (T_1, T_2), \end{aligned} \tag{1.12}$$

$$b(\mathbf{u}(t), r) = 0 \qquad \forall\, r \in L^2_0(\Omega),\ \textit{a.e. } t \in (T_1, T_2) \tag{1.13}$$

and

$$\lim_{t \to T_1^+} \mathbf{u}(t) = \boldsymbol{\xi} \qquad \textit{in } \mathbf{W}. \qquad \square \tag{1.14}$$

We point out that $\mathbf{u} \in \mathcal{H}^{(1)}(\Omega \times (T_1, T_2))$ implies $\mathbf{u} \in C([T_1, T_2]; \mathbf{W})$ so that (1.14) makes sense. It is well known that if (T_1, T_2) is a finite subinterval of $(0, \infty)$ and if $\mathbf{f} \in L^2\left(T_1, T_2; \mathbf{L}^2(\Omega)\right)$, then there is indeed a strong solution $\mathbf{u}$ for the Navier–Stokes equations satisfying the regularity result $\mathbf{u} \in \mathbf{L}^2(T_1, T_2; \mathbf{H}^1_0(\Omega)) \cap L^2(T_1 + \varepsilon, T_2; \mathbf{H}^2(\Omega))$ and $\mathbf{u}_t \in L^2(T_1 + \varepsilon, T_2; \mathbf{W})$ for all $\varepsilon > 0$. Furthermore, if $\boldsymbol{\xi} \in \mathbf{V}$, then $\mathbf{u} \in C([T_1, T_2]; \mathbf{V}) \cap L^2(T_1, T_2; \mathbf{H}^2(\Omega))$ (see [CF] and [Te]).

For $T = \infty$, we define a solution for the Navier–Stokes equations as follows.

DEFINITION 1.2. *Given* $\boldsymbol{\xi} \in \mathbf{W}$ *and* $\mathbf{f} \in L^2(0, T'; \mathbf{H}^{-1}(\Omega))$ *for all* $T' > 0$, $(\mathbf{u}, p)$ *is said to be a solution of the Navier–Stokes equations on* $(0, \infty)$ *iff* $\mathbf{u} \in L^\infty(0, \infty; \mathbf{W})$ *and* $(\mathbf{u}, p)$ *is a solution of the Navier–Stokes equations on* $(0, T')$ *for all* $T' > 0$. $\square$

We now turn to the definition of the piecewise optimal control problem. We introduce the (local) functional

$$\mathcal{J}_{(T_1, T_2)}(\mathbf{u}, \mathbf{f}) = \frac{\alpha}{2} \int_{T_1}^{T_2} \int_\Omega |\mathbf{u} - \mathbf{U}|^2 \, d\mathbf{x}\, dt + \frac{\beta}{2} \int_{T_1}^{T_2} \int_\Omega |\mathbf{f} - \mathbf{F}|^2 \, d\mathbf{x}\, dt. \tag{1.15}$$

We choose the fixed body force $\mathbf{F}$ as

$$\mathbf{F} = \mathbf{N}(\mathbf{U}) \stackrel{\text{def}}{=\!=} \partial_t \mathbf{U} - \nu \Delta \mathbf{U} + (\mathbf{U} \cdot \nabla)\mathbf{U}. \tag{1.16}$$

We make the following regularity assumptions on the prescribed data $\mathbf{U}$ and $\mathbf{F}$:

$$\text{(A1)} \qquad \begin{cases} \mathbf{U} = \mathbf{U}(\mathbf{x}, t) \in C\left([0, \infty); \mathbf{H}^2(\Omega) \cap \mathbf{H}^1_0(\Omega)\right) \cap L^\infty\left(0, \infty; \mathbf{H}^2(\Omega)\right), \\ \operatorname{div} \mathbf{U} = 0, \\ \mathbf{F} = \mathbf{N}(\mathbf{U}) \in L^\infty\left(0, \infty; \mathbf{L}^2(\Omega)\right). \end{cases}$$

We define the (local) admissible elements as follows.

DEFINITION 1.3. *Let* $(T_1, T_2) \subset (0, \infty)$ *with* $T_2 < \infty$ *and* $\boldsymbol{\xi} \in \mathbf{W}$ *be given. A triplet* $(\mathbf{u}, p, \mathbf{f})$ *is said to be an admissible element on* (T_1, T_2) *if* $\mathbf{u} \in \mathcal{H}^{(1)}(\Omega \times (T_1, T_2))$, $p \in L^2\left(T_1, T_2; L^2_0(\Omega)\right)$, $\mathbf{f} \in L^2\left(T_1, T_2; \mathbf{L}^2(\Omega)\right)$, *and* $(\mathbf{u}, p, \mathbf{f})$ *satisfies* (1.12)–(1.14). *The set of all admissible elements on* (T_1, T_2) *is denoted by* $\mathcal{V}_{\text{ad}}(T_1, T_2, \boldsymbol{\xi})$. $\square$

We are now prepared to state precisely the piecewise optimal control problem to be studied in this paper.

- Choose $\delta > 0$ sufficiently small, and set $t_n = n\delta$ for $n = 0, 1, 2, \ldots$.
- Set $\widehat{\mathbf{u}}^{(0)}(0) \stackrel{\text{def}}{=\!=} \mathbf{u}_0$.
- For $n = 0, 1, 2, \ldots$, find $(\widehat{\mathbf{u}}^{(n+1)}, \widehat{p}^{(n+1)}, \widehat{\mathbf{f}}^{(n+1)}) \in \mathcal{V}_{\text{ad}}(t_n, t_{n+1}, \widehat{\mathbf{u}}^{(n)}(\cdot, t_n))$ such that

$$\mathcal{J}_{(t_n, t_{n+1})}(\widehat{\mathbf{u}}^{(n+1)}, \widehat{\mathbf{f}}^{(n+1)}) \le \mathcal{J}_{(t_n, t_{n+1})}(\mathbf{u}, \mathbf{f})$$

  for all admissible elements $(\mathbf{u}, p, \mathbf{f}) \in \mathcal{V}_{\text{ad}}(t_n, t_{n+1}, \widehat{\mathbf{u}}^{(n)}(\cdot, t_n))$.

The global optimal solution $(\widehat{\mathbf{u}}, \widehat{p}, \widehat{\mathbf{f}})$ is defined as in section 1.1 by patching together local optimal solutions $(\widehat{\mathbf{u}}^{(n+1)}, \widehat{p}^{(n+1)}, \widehat{\mathbf{f}}^{(n+1)})$.

We terminate this section with some useful inequalities. The first inequality is the Poincaré inequality

$$\|\nabla \mathbf{w}\|^2 \geq \lambda_1 \|\mathbf{w}\|^2 \qquad \forall\, \mathbf{w} \in \mathbf{H}_0^1(\Omega), \tag{1.17}$$

where $\lambda_1 > 0$ is the greatest real number such that (1.17) holds. The second inequality is connected with the Leray operator

$$\Pi : \mathbf{L}^2(\Omega) \to \mathbf{W} \tag{1.18}$$

(i.e., the orthogonal projection with respect to the $\mathbf{L}^2(\Omega)$ norm). It is well known (see [CF]) that there is a constant $C_\Pi > 0$ depending only on Ω such that

$$C_\Pi \|\mathbf{w}\|_2 \leq \|\Pi \Delta \mathbf{w}\| \leq \|\mathbf{w}\|_2 \qquad \forall\, \mathbf{w} \in \mathbf{H}^2(\Omega) \cap \mathbf{V}\,. \tag{1.19}$$

In other words, $\|\Pi\Delta \cdot \|$ is equivalent to the $\mathbf{H}^2(\Omega)$ norm on $\mathbf{H}^2(\Omega) \cap \mathbf{V}$.

2. Dynamics of the piecewise optimal control problem. We now study the dynamics of the piecewise optimal control problem stated in the last section. We point out that for each n, the existence of an optimal solution was established in [Fu1], [Fu2], [Fu3], and [Li] (see also [AT]) for finite time interval and in [HY] for infinite time interval. Recall from our physical objective that we wish the optimal solution $\widehat{\mathbf{u}}$ to match $\mathbf{U}$ over time; i.e., we wish $\|\widehat{\mathbf{u}}(t) - \mathbf{U}(t)\| \to 0$ and $\|\nabla \widehat{\mathbf{u}}(t) - \nabla \mathbf{U}(t)\| \to 0$ as $t \to \infty$. Such decay properties are established in [HY] for the solutions of the infinite-time optimal control problem. We will prove in this paper that these decay properties are true for the solutions of the piecewise optimal control problem, and, more importantly, the decay rates are exponential.

For technical reasons, we will need the following assumption:

$$\|\mathbf{U}\|_{L^\infty(0,\infty;\mathbf{L}^4(\Omega))} \leq \frac{\nu}{2C_0}, \tag{A2}$$

where C_0 is the constant in (1.11).

PROPOSITION 2.1. *Assume that* (A1) *and* (A2) *hold. Then for any* T_1 *and* T_2 *with* $0 \leq T_1 < T_2$ *and any* $\boldsymbol{\xi} \in \mathbf{W}$,

$$\begin{aligned} \inf_{(\mathbf{u},p,\mathbf{f}) \in \mathcal{V}_{\rm ad}(T_1,T_2,\boldsymbol{\xi})} \mathcal{J}_{(T_1,T_2)}(\mathbf{u},\mathbf{f}) &\leq \frac{\alpha \|\boldsymbol{\xi} - \mathbf{U}(T_1)\|^2}{\nu\lambda_1} \left(1 - e^{-\nu\lambda_1 (T_2 - T_1)/2}\right) \\ &\leq \frac{\alpha(T_2 - T_1)}{2} \|\boldsymbol{\xi} - \mathbf{U}(T_1)\|^2, \end{aligned} \tag{2.1}$$

where $\lambda_1 > 0$ *is the Poincaré constant defined by* (1.17).

Proof. Let $(\widetilde{\mathbf{u}}, \widetilde{p}) \in \mathcal{H}^{(1)}((T_1,T_2)\times\Omega) \times L^2(T_1,T_2;L_0^2(\Omega))$ be the solution of

$$\begin{aligned} &\langle \partial_t \widetilde{\mathbf{u}}(t), \mathbf{w}\rangle + a(\widetilde{\mathbf{u}}(t),\mathbf{w}) + c(\widetilde{\mathbf{u}}(t),\widetilde{\mathbf{u}}(t),\mathbf{w}) + b(\mathbf{w},\widetilde{p}(t)) \\ &\qquad\qquad = (\mathbf{F}(t),\mathbf{w}) \qquad \forall\, \mathbf{w} \in \mathbf{H}_0^1(\Omega),\ \text{a.e. } t \in (T_1,T_2)\,, \\ &b(\widetilde{\mathbf{u}}(t), r) = 0 \qquad \forall\, r \in L_0^2(\Omega),\ \text{a.e. } t \in (T_1,T_2), \end{aligned}$$

and

$$\lim_{t \to T_1^+} \widetilde{\mathbf{u}}(t) = \boldsymbol{\xi} \qquad \text{in } \mathbf{W}\,.$$

Let $\widetilde{\mathbf{f}} \stackrel{\text{def}}{=\!=} \mathbf{F}$. Then $(\widetilde{\mathbf{u}}, \widetilde{p}, \widetilde{\mathbf{f}}) \in \mathcal{V}_{ad}(T_1, T_2, \boldsymbol{\xi})$. Setting $\widetilde{\mathbf{v}} = \widetilde{\mathbf{u}} - \mathbf{U}$, we obtain

$$\begin{aligned} &\langle \partial_t \widetilde{\mathbf{v}}(t), \mathbf{w} \rangle + a\left(\widetilde{\mathbf{v}}(t), \mathbf{w}\right) + c\left(\widetilde{\mathbf{v}}(t), \widetilde{\mathbf{v}}(t), \mathbf{w}\right) + c\left(\mathbf{U}(t), \widetilde{\mathbf{v}}(t), \mathbf{w}\right) \\ &\qquad + c\left(\widetilde{\mathbf{v}}(t), \mathbf{U}(t), \mathbf{w}\right) + b\left(\mathbf{w}, \widetilde{q}(t)\right) = 0 \qquad \forall\, \mathbf{w} \in \mathbf{H}_0^1(\Omega), \text{ a.e. } t \in (T_1, T_2)\,, \end{aligned} \tag{2.2}$$

$$b\left(\widetilde{\mathbf{v}}(t), r\right) = 0 \qquad \forall\, r \in L_0^2(\Omega), \text{ a.e. } t \in (T_1, T_2), \tag{2.3}$$

and

$$\lim_{t \to T_1^+} \widetilde{\mathbf{v}}(t) = \boldsymbol{\xi} - \mathbf{U}(\cdot, T_1) \qquad \text{in } \mathbf{W}\,. \tag{2.4}$$

By setting $\mathbf{w} = \widetilde{\mathbf{v}}$ in (2.2), we have that

$$\frac{1}{2}\frac{d}{dt}\left\|\widetilde{\mathbf{v}}(t)\right\|^2 + \nu\left\|\nabla\widetilde{\mathbf{v}}(t)\right\|^2 + c\left(\widetilde{\mathbf{v}}(t), \mathbf{U}(t), \widetilde{\mathbf{v}}(t)\right) = 0. \tag{2.5}$$

By (1.9) and (1.11),

$$\left|c\left(\widetilde{\mathbf{v}}(t), \mathbf{U}(t), \widetilde{\mathbf{v}}(t)\right)\right| \le C_0 \|\mathbf{U}(t)\|_{\mathbf{L}^4(\Omega)} \cdot \left\|\nabla\widetilde{\mathbf{v}}(t)\right\|^2\,. \tag{2.6}$$

Then, from (2.5) and (2.6), we have that

$$\frac{1}{2}\frac{d}{dt}\left\|\widetilde{\mathbf{v}}(t)\right\|^2 + \left(\nu - C_0\|\mathbf{U}(t)\|_{L^\infty(0,\infty;\mathbf{L}^4(\Omega))}\right)\left\|\nabla\widetilde{\mathbf{v}}(t)\right\|^2 \le 0.$$

Since (A2) implies $\nu - C_0\|\mathbf{U}(t)\|_{L^\infty(0,\infty;\mathbf{L}^4(\Omega))} \ge \frac{\nu}{2}$, we see that

$$\frac{d}{dt}\left\|\widetilde{\mathbf{v}}(t)\right\|^2 + \frac{\nu}{2}\left\|\nabla\widetilde{\mathbf{v}}(t)\right\|^2 \le 0.$$

Applying the Poincaré inequality and then Gronwall's inequality, we obtain

$$\left\|\widetilde{\mathbf{v}}(t)\right\|^2 \le \left\|\widetilde{\mathbf{v}}(T_1)\right\|^2 \cdot e^{-\nu\lambda_1(t-T_1)/2} = \left\|\boldsymbol{\xi} - \mathbf{U}(T_1)\right\|^2 \cdot e^{-\nu\lambda_1(t-T_1)/2}. \tag{2.7}$$

Hence,

$$\begin{aligned} \mathcal{J}_{(T_1,T_2)}(\widetilde{\mathbf{u}}, \widetilde{\mathbf{f}}) &= \frac{\alpha}{2}\int_{T_1}^{T_2}\left\|\widetilde{\mathbf{v}}(t)\right\|^2 dt \le \frac{\alpha}{2}\left\|\boldsymbol{\xi} - \mathbf{U}(T_1)\right\|^2 \int_{T_1}^{T_2} e^{-\nu\lambda_1(t-T_1)/2}\, dt \\ &= \frac{\alpha\|\boldsymbol{\xi} - \mathbf{U}(T_1)\|^2}{\nu\lambda_1}\left(1 - e^{-\nu\lambda_1(T_2-T_1)/2}\right) \le \frac{\alpha}{2}(T_2 - T_1)\|\boldsymbol{\xi} - \mathbf{U}(T_1)\|^2, \end{aligned}$$

where we have used the fact that $1 - e^{-y} \le y$ for all $y \in [0, \infty)$. This completes the proof. ∎

LEMMA 2.2. *Let η, σ, and γ be positive numbers satisfying $\gamma < 1$ and $\sigma \le \eta$. Then there is a constant $\delta_0 > 0$ such that*

$$0 < \varphi(s) \stackrel{\text{def}}{=\!=} e^{-\eta s} + \gamma\left(1 - e^{-\sigma s}\right) < 1 \qquad \textit{for } s \in (0, \delta_0).$$

Proof. Note that $\varphi(0) = 1$ and $\varphi'(0) = -\eta + \gamma\sigma < 0$. Thus the lemma follows by elementary calculus. ∎

We are now prepared to show that $\|\mathbf{u}(t_n) - \mathbf{U}(t_n)\|$ is monotonically decreasing.

THEOREM 2.3. *Assume that* (A1) *and* (A2) *hold. Assume further that*

$$\frac{\alpha}{\beta} < \left(\frac{\nu\lambda_1}{2}\right)^2. \tag{A3}$$

Then there are constants

$$\delta_0 = \delta_0(\nu, \Omega) > 0,$$

$$\mu = \mu(\nu, \Omega) \in (0, 1),$$

and

$$M_1 = M_1(\nu, \Omega) > 0$$

such that if $T_2 - T_1 \le \delta_0$ *and* $(\mathbf{u}, p, \mathbf{f}) \in \mathcal{V}_{ad}(T_1, T_2, \boldsymbol{\xi})$ *satisfies*

$$\mathcal{J}_{(T_1,T_2)}(\mathbf{u}, \mathbf{f}) \le \mathcal{J}_{(T_1,T_2)}(\mathbf{w}, \mathbf{k}) \qquad \forall\, (\mathbf{w}, r, \mathbf{k}) \in \mathcal{V}_{\mathrm{ad}}(T_1, T_2; \boldsymbol{\xi})$$

(*i.e.,* $(\mathbf{u}, p, \mathbf{f})$ *is an optimal solution on* (T_1, T_2))*, then the terminal state* $\mathbf{u}(T_2)$ *satisfies*

$$\|\mathbf{u}(T_2) - \mathbf{U}(T_2)\|^2 \le \mu \|\mathbf{u}(T_1) - \mathbf{U}(T_1)\|^2$$

and the intermediate state $\mathbf{u}(t)$ *satisfies*

$$\|\mathbf{u}(t) - \mathbf{U}(t)\|^2 \le M_1 \|\mathbf{u}(T_1) - \mathbf{U}(T_1)\|^2 \qquad \forall\, t \in [T_1, T_2].$$

Proof. As usual we set $\mathbf{v} = \mathbf{u} - \mathbf{U}$ and $\mathbf{g} = \mathbf{f} - \mathbf{F}$; then $(\mathbf{v}, p, \mathbf{g})$ satisfies

$$\begin{aligned} &\langle \partial_t \mathbf{v}, \mathbf{w}\rangle + c(\mathbf{v}, \mathbf{v}, \mathbf{w}) + c(\mathbf{U}, \mathbf{v}, \mathbf{w}) + c(\mathbf{v}, \mathbf{U}, \mathbf{w}) \\ &\qquad + a(\mathbf{v}, \mathbf{w}) + b(\mathbf{w}, p) = (\mathbf{g}, \mathbf{w}) \qquad \forall \mathbf{w} \in \mathbf{H}_0^1(\Omega), \quad \text{a.e. } t \in (T_1, T_2) \end{aligned} \tag{2.2a}$$

and (2.3), (2.4). Setting $\mathbf{w} = \mathbf{v}$ in (2.2a) and using (A2), (1.11), and Young's inequality, we have

$$\begin{aligned} \frac{1}{2}\frac{d}{dt}\|\mathbf{v}(t)\|^2 + \nu \|\nabla \mathbf{v}(t)\|^2 &\le C_0 \|\mathbf{U}(t)\|_{\mathbf{L}^4(\Omega)} \cdot \|\nabla \mathbf{v}(t)\|^2 + \|\mathbf{v}(t)\| \cdot \|\mathbf{g}(t)\| \\ &\le \frac{\nu}{2}\|\nabla \mathbf{v}(t)\|^2 + \frac{1}{2\sqrt{\alpha\beta}}\left(\alpha \|\mathbf{v}(t)\|^2 + \beta \|\mathbf{g}(t)\|^2\right), \end{aligned}$$

so that by applying the Poincaré inequality,

$$\begin{aligned} \frac{d}{dt}\|\mathbf{v}(t)\|^2 + \nu\lambda_1 \|\mathbf{v}(t)\|^2 &\le \frac{d}{dt}\|\mathbf{v}(t)\|^2 + \nu \|\nabla \mathbf{v}(t)\|^2 \\ &\le \frac{1}{\sqrt{\alpha\beta}}\left(\alpha \|\mathbf{v}(t)\|^2 + \beta \|\mathbf{g}(t)\|^2\right). \end{aligned} \tag{2.8}$$

Multiplying both sides of the last equation by $e^{\nu\lambda_1(t-T_1)}$ and integrating over (T_1, t),

$$\|\mathbf{v}(t)\|^2 e^{\nu\lambda_1 t} - \|\mathbf{v}(T_1)\|^2 e^{\nu\lambda_1 T_1} \le \frac{1}{\sqrt{\alpha\beta}} \int_{T_1}^{t} \left(\alpha \|\mathbf{v}(s)\|^2 + \beta \|\mathbf{g}(s)\|^2\right) e^{\nu\lambda_1 s} ds$$

or

$$\begin{aligned} &\|\mathbf{v}(t)\|^2 - \|\mathbf{v}(T_1)\|^2 e^{-\nu\lambda_1(t-T_1)} \\ &\qquad \le \frac{1}{\sqrt{\alpha\beta}} \int_{T_1}^{t} \left(\alpha \|\mathbf{v}(s)\|^2 + \beta \|\mathbf{g}(s)\|^2\right) e^{-\nu\lambda_1(t-s)} ds \\ &\qquad \le \frac{1}{\sqrt{\alpha\beta}} \int_{T_1}^{t} \left(\alpha \|\mathbf{v}(s)\|^2 + \beta \|\mathbf{g}(s)\|^2\right) ds \le \frac{2}{\sqrt{\alpha\beta}} \mathcal{J}_{(T_1,T_2)}(\mathbf{u}, \mathbf{f}). \end{aligned} \tag{2.9}$$

We obtain, upon applying Proposition 2.1,

$$\begin{aligned}\|\mathbf{v}(t)\|^2 &\le \|\mathbf{v}(T_1)\|^2 e^{-\nu\lambda_1(t-T_1)} + \frac{2}{\sqrt{\alpha\beta}}\cdot \mathcal{J}_{(T_1,T_2)}(\mathbf{u},\mathbf{f}) \\ &\le \|\mathbf{v}(T_1)\|^2 \left(e^{-\nu\lambda_1(t-T_1)} + \frac{2}{\nu\lambda_1}\sqrt{\frac{\alpha}{\beta}}\cdot\left(1-e^{-\nu\lambda_1(T_2-T_1)/2}\right)\right).\end{aligned} \tag{2.10}$$

Set

$$\eta \stackrel{\text{def}}{=\!=} \nu\lambda_1, \quad \sigma \stackrel{\text{def}}{=\!=} \nu\lambda_1/2, \quad \gamma \stackrel{\text{def}}{=\!=} \frac{2}{\nu\lambda_1}\sqrt{\frac{\alpha}{\beta}},$$
$$\delta \stackrel{\text{def}}{=\!=} T_2 - T_1 \quad \text{and} \quad \varphi(r) \stackrel{\text{def}}{=\!=} e^{-\eta r} + \gamma(1-e^{-\sigma\delta}) \ \ \forall r \in [0,\delta].$$

Then (2.10) can be rewritten as

$$\|\mathbf{v}(t)\|^2 \le \|\mathbf{v}(T_1)\|^2 \varphi(t-T_1) \qquad \forall\, t \in [T_1,T_2]\,.$$

Clearly, $\varphi(r) \le 1+\gamma$ for all $r \in [0,\delta]$ so that by setting $M_1 = 1+\gamma$, $\|\mathbf{v}(t)\|^2 \le M_1\|\mathbf{v}(T_1)\|^2$ for all $t \in [T_1,T_2]$. Moreover, by (A2) and (A3), the conditions of Lemma 2.2 are satisfied. Thus it follows from Lemma 2.2 that there exists a $\delta_0 > 0$ such that $0 < \varphi(\delta) < 1$ if $\delta \in (0,\delta_0]$ and, in particular,

$$\mu \stackrel{\text{def}}{=\!=} \varphi(T_2 - T_1) < 1.$$

Hence $\|\mathbf{v}(T_2)\|^2 \le \mu\|\mathbf{v}(T_1)\|^2$. □

Set $\widehat{\mathbf{u}}^{(0)}(0) = \mathbf{u}_0$ and fix $\delta \in (0,\delta_0]$, where δ_0 is as in Theorem 2.3. For $n = 0,1,2,\dots$, let $t_n = n\delta$ and $(\widehat{\mathbf{u}}^{(n+1)}, \widehat{p}^{(n+1)}, \widehat{\mathbf{f}}^{(n+1)}) \in \mathcal{V}_{\text{ad}}(t_n, t_{n+1}, \widehat{\mathbf{u}}^{(n)}(t_n))$ be the optimizer minimizing $\mathcal{J}_{(t_n,t_{n+1})}(\mathbf{u},\mathbf{f})$ over $\mathcal{V}_{\text{ad}}(t_n,t_{n+1},\widehat{\mathbf{u}}^{(n)}(t_n))$. Define $(\widehat{\mathbf{u}},\widehat{p},\widehat{\mathbf{f}})$ by

$$\begin{aligned}&\widehat{\mathbf{u}}(\mathbf{x},t) = \widehat{\mathbf{u}}^{(n+1)}(\mathbf{x},t), \qquad \widehat{p}(\mathbf{x},t) = \widehat{p}^{(n+1)}(\mathbf{x},t), \\ &\text{and} \quad \widehat{\mathbf{f}}(\mathbf{x},t) = \widehat{\mathbf{f}}^{(n+1)}(\mathbf{x},t) \qquad \text{for } t \in [t_n,t_{n+1})\ \forall\, n\,.\end{aligned} \tag{2.11}$$

We can easily verify that $(\widehat{\mathbf{u}},\widehat{p})$ gives a solution of the Navier–Stokes equations (with a forcing term $\widehat{\mathbf{f}}$) on the infinite time interval $(0,\infty)$ in the sense of Definition 1.2.

PROPOSITION 2.4. *Assume that the hypotheses of Theorem* 2.3 *hold. Let* η, σ, γ, δ_0, *and* $\mu = e^{-\eta\delta} + \gamma(1-e^{-\sigma\delta})$ *be as in Theorem* 2.3 *with* $\delta \le \delta_0$. *Then* $(\widehat{\mathbf{u}},\widehat{p},\widehat{\mathbf{f}})$ *constructed in* (2.11) *satisfies*

$$\widehat{\mathbf{u}} \in C([0,\infty);\mathbf{W}) \tag{2.12}$$

and

$$\mathcal{J}_{(t_n,\infty)}(\widehat{\mathbf{u}},\widehat{\mathbf{f}}) \le \frac{\alpha\delta}{2(1-\mu)}\cdot\|\mathbf{u}_0-\mathbf{U}_0\|^2\cdot\mu^n, \qquad n = 0,1,2,\dots. \tag{2.13}$$

In particular,

$$\mathcal{J}_{(0,\infty)}(\widehat{\mathbf{u}},\widehat{\mathbf{f}}) \le \frac{\alpha\delta}{2(1-\mu)}\cdot\|\mathbf{u}_0-\mathbf{U}_0\|^2$$

and

$$\widehat{\mathbf{u}} - \mathbf{U} \in L^2(0,\infty;\mathbf{W}) \qquad \textit{and} \qquad \widehat{\mathbf{f}} - \mathbf{F} \in L^2(0,\infty;\mathbf{W}).$$

Proof. By the classical results for the Navier–Stokes equations (see [Te]) we have that $\widehat{\mathbf{u}}|_{(t_n,t_{n+1})} = \widehat{\mathbf{u}}^{(n+1)} \in C([t_n, t_{n+1}]; \mathbf{W})$ for each $n \geq 0$. The initial conditions $\widehat{\mathbf{u}}^{(n+1)}(t_n) = \widehat{\mathbf{u}}^{(n)}(t_n)$ guarantee the continuity of $\widehat{\mathbf{u}}$ at each t_n in $\mathbf{W}$. Hence, (2.12) is proven.

Applying Proposition 2.1 and Theorem 2.3, we obtain

$$\begin{aligned}\mathcal{J}_{(t_n,t_{n+1})}(\widehat{\mathbf{u}},\widehat{\mathbf{f}}) &\leq \mathcal{J}_{(t_n,t_{n+1})}(\widetilde{\mathbf{u}}^{(n+1)},\widetilde{\mathbf{f}}^{(n+1)}) \\ &\leq \frac{\alpha}{\nu\lambda_1}\left(1-e^{-\nu\lambda_1\delta/2}\right)\cdot\left\|\widehat{\mathbf{u}}^{(n)}(t_n)-\mathbf{U}(t_n)\right\|^2 \\ &\leq \frac{\alpha}{\nu\lambda_1}\left(1-e^{-\nu\lambda_1\delta/2}\right)\cdot\mu^n\left\|\mathbf{u}_0-\mathbf{U}_0\right\|^2.\end{aligned}$$

Hence

$$\mathcal{J}_{(t_n,\infty)}(\widehat{\mathbf{u}},\widehat{\mathbf{f}}) = \sum_{m=n}^{\infty}\mathcal{J}_{(t_m,t_{m+1})}(\widehat{\mathbf{u}},\widehat{\mathbf{f}}) \leq \frac{\alpha\left(1-e^{-\nu\lambda_1\delta/2}\right)}{\nu\lambda_1(1-\mu)}\cdot\left\|\mathbf{u}_0-\mathbf{U}_0\right\|^2\cdot\mu^n.$$

When $n=0$, this reduces to

$$\mathcal{J}_{(0,\infty)}(\widehat{\mathbf{u}},\widehat{\mathbf{f}}) \leq \frac{\alpha\delta}{2(1-\mu)}\cdot\|\mathbf{u}_0-\mathbf{U}_0\|^2,$$

which in turn implies (from the definition of $\mathcal{J}_{(0,\infty)}$) that

$$\widehat{\mathbf{u}}-\mathbf{U}\in L^2(0,\infty;\mathbf{W}) \qquad \text{and} \qquad \widehat{\mathbf{f}}-\mathbf{F}\in L^2(0,\infty;\mathbf{W}). \qquad \square$$

Remark 2.5. As a consequence of Lemma 2.2 and Theorem 2.3 we note that when $0<\delta\ll 1$, $\mu=\varphi(\delta)=e^{-\eta\delta}+\gamma\left(1-e^{-\sigma\delta}\right)=1-(\eta-\gamma\sigma)\delta+O\left(\delta^2\right)$ so that

$$\mathcal{J}_{(0,\infty)}(\widehat{\mathbf{u}},\widehat{\mathbf{f}}) \leq \frac{\alpha\|\mathbf{u}_0-\mathbf{U}_0\|^2}{2(\eta-\gamma\sigma)}+O(\delta).$$

Hence $\mathcal{J}_{(0,\infty)}(\widehat{\mathbf{u}},\widehat{\mathbf{f}})$ remains bounded as $\delta\to 0^+$. $\square$

LEMMA 2.6. *Suppose that a function $\psi(t)$ defined on $[0,\infty)$ is nonnegative and that there exist constants $\mu\in(0,1)$, $M_1>0$ and $\delta\in(0,\infty)$ such that*

$$\psi(n\delta)\leq\mu\psi\left((n-1)\delta\right), \qquad n=1,2,\ldots$$

and

$$\psi(s)\leq M_1\psi(n\delta) \qquad \forall\, s\in[n\delta,(n+1)\delta),\ n=0,1,2,\ldots.$$

Then

$$\psi(t)\leq M_2\, e^{-\kappa t} \qquad \forall\, t\in[0,\infty),$$

where

$$M_2 \stackrel{\text{def}}{=\!=} \frac{M_1}{\mu}\psi(0)\geq 0 \qquad \textit{and} \qquad \kappa \stackrel{\text{def}}{=\!=} \frac{-\ln\mu}{\delta}>0.$$

Proof. We can verify by induction that $\psi(n\delta)\leq\mu^n\psi(0)$. Then for each $t\in[0,\infty)$, there is a unique integer n such that $n\delta\leq t\leq(n+1)\delta$, which yields

$$\begin{aligned}\psi(t) &\leq M_1\psi(n\delta)\leq M_1\mu^n\psi(0)=M_1\psi(0)e^{n\ln\mu} \\ &= M_1\psi(0)e^{(n\delta-t)(\ln\mu)/\delta}\cdot e^{t(\ln\mu)/\delta}\leq\left[M_1\psi(0)e^{-\ln\mu}\right]\cdot e^{t(\ln\mu)/\delta}=M_2e^{-\kappa t}. \qquad \square\end{aligned}$$

THEOREM 2.7. *Assume that the conditions of Theorem* 2.3 *hold and let* δ_0 *and* μ *be as defined in Theorem* 2.3 *and* $\delta \le \delta_0$. *Then* $\|\widehat{\mathbf{u}}(t) - \mathbf{U}(t)\|$ *decays exponentially:*

$$\|\widehat{\mathbf{u}}(t) - \mathbf{U}(t)\|^2 \le M_3 \|\mathbf{u}_0 - \mathbf{U}_0\|^2 e^{-\kappa t}, \tag{2.14}$$

where

$$M_3 = M_3(\nu, \Omega) \stackrel{\text{def}}{=\!=} \frac{M_1}{\mu} \quad \textit{and} \quad \kappa = M_3(\nu, \Omega) \stackrel{\text{def}}{=\!=} -\frac{\ln \mu}{\delta}.$$

Proof. (2.14) is a direct consequence of Theorem 2.3 and Lemma 2.6. ∎

Remark 2.8. By Theorems 2.3 and 2.7 together with (A2), we see that when $0 < \delta \ll 1$,

$$M_3 = \frac{1+\gamma}{1+O(\delta)} = 1 + \frac{2}{\nu\lambda_1}\sqrt{\frac{\alpha}{\beta}} + O(\delta) \le 2 + O(\delta)$$

and

$$\kappa = \frac{-\ln \mu}{\delta} = \frac{-\ln\left[1 - (\eta - \gamma\sigma)\delta + O\left(\delta^2\right)\right]}{\delta} = (1 - \gamma/2)\nu\lambda_1 + O(\delta);$$

both remain bounded as $\delta \to 0^+$. ∎

To prove the exponential decay property of $\|\nabla\mathbf{u}(t) - \nabla\mathbf{U}(t)\|$, we study the local behavior of the optimizer $(\widehat{\mathbf{u}}^{(n+1)}, \widehat{p}^{(n+1)}, \widehat{\mathbf{f}}^{(n+1)})$ defined by (2.11). We need to assume the following global regularity on the desired flow $\mathbf{U}$:

$$\sup_{t\in(0,\infty)} \|\nabla\mathbf{U}(t)\|_{\mathbf{L}^\infty(\Omega)} < \infty. \tag{A4}$$

Also, it follows from (A1) and the continuous embedding $\mathbf{H}^2(\Omega) \hookrightarrow \mathbf{L}^\infty(\Omega)$ that

$$\sup_{t\in(0,\infty)} \|\mathbf{U}(t)\|_{\mathbf{L}^\infty(\Omega)} < \infty\,.$$

LEMMA 2.9. *Suppose that the conditions of Theorem* 2.7 *and* (A4) *hold. Denote by* $\widehat{\mathbf{v}} = \widehat{\mathbf{u}} - \mathbf{U}$ *and* $\widehat{\mathbf{g}} = \widehat{\mathbf{f}} - \mathbf{F}$. *Then*

$$\nu \int_{t_n}^{\infty} \|\nabla\widehat{\mathbf{v}}(t)\|^2 dt \le \|\widehat{\mathbf{v}}^{(n)}(t_n)\|^2 + \frac{1}{\sqrt{\alpha\beta}}\mathcal{J}_{(t_n,\infty)}(\widehat{\mathbf{u}}, \widehat{\mathbf{f}})\,. \tag{2.15}$$

Also, there exist constants $C_1, C_2 > 0$ *such that for* δ *as in Theorem* 2.7,

$$\begin{aligned} &\|\nabla\widehat{\mathbf{v}}(T_2)\|^2 \exp\left\{-C_2\|\widehat{\mathbf{v}}(t_{n_0})\|^2 \int_{t_{n_0}}^{T_2} \|\nabla\widehat{\mathbf{v}}(s)\|^2 ds\right\} \\ &\qquad - \|\nabla\widehat{\mathbf{v}}(T_1)\|^2 \exp\left\{-C_2\|\widehat{\mathbf{v}}(t_{n_0})\|^2 \int_{t_{n_0}}^{T_1} \|\nabla\widehat{\mathbf{v}}(s)\|^2 ds\right\} \\ &\le C_1 \int_{T_1}^{T_2} \left(\|\nabla\widehat{\mathbf{v}}(s)\|^2 + \|\widehat{\mathbf{g}}(s)\|^2\right) \exp\left\{-C_2\|\widehat{\mathbf{v}}(t_{n_0})\|^2 \int_{t_{n_0}}^{s} \|\nabla\widehat{\mathbf{v}}(\sigma)\|^2 d\sigma\right\} ds \end{aligned} \tag{2.16}$$

for all $[T_1, T_2] \subset [t_n, t_{n+1}]$ *if* $n \ge 1$ *or* $[T_1, T_2] \subset (t_0, t_1]$. *Here, in* (2.16), $t_{n_0} \le T_1$ *is arbitrarily fixed.*

Proof. Note that for $t \in (t_n, t_{n+1})$, $(\widehat{\mathbf{u}}(t), \widehat{\mathbf{f}}(t)) = (\widehat{\mathbf{u}}^{(n+1)}(t), \widehat{\mathbf{f}}^{(n+1)}(t))$ and $(\widehat{\mathbf{v}}^{(n+1)}(t), \widehat{\mathbf{f}}^{(n+1)}(t)) = (\widehat{\mathbf{u}}^{(n+1)}(t) - \mathbf{U}(t), \widehat{\mathbf{f}}^{(n+1)}(t) - \mathbf{F}(t))$. For each $j \geq n$, integrating the second inequality of (2.8), we obtain

$$\|\mathbf{v}(t_{j+1})\|^2 - \|\mathbf{v}(t_j)\|^2 \leq \frac{1}{\sqrt{\alpha\beta}} \int_{t_j}^{t_{j+1}} \left(\alpha \|\mathbf{v}(s)\|^2 + \beta \|\mathbf{g}(s)\|^2\right) ds \qquad \forall j \geq n.$$

Summing over $j \geq n$, we obtain (2.15).

By classical arguments (see [Te] and [CF]), the weak form (2.2a) implies that $(\mathbf{v}, p, \mathbf{g})$ satisfies the strong form

$$\partial_t \mathbf{v} - \Delta \mathbf{v} + (\mathbf{v} \cdot \nabla)\mathbf{v} + (\mathbf{U} \cdot \nabla)\mathbf{v} + (\mathbf{v} \cdot \nabla)\mathbf{U} + \nabla p = \mathbf{g} \qquad \text{a.e. } t \in (T_1, T_2). \tag{2.17}$$

Taking the inner product of (2.17) with $-\Pi\Delta\widehat{\mathbf{v}}$, where Π is the Leray projector defined in (1.18), and applying the Schwarz inequality, we obtain, for $t \in (t_n, t_{n+1})$,

$$\begin{aligned}
&\frac{1}{2}\frac{d}{dt}\|\nabla\widehat{\mathbf{v}}(t)\|^2 + \nu\|\Pi\Delta\widehat{\mathbf{v}}(t)\|^2 \\
&\leq ((\widehat{\mathbf{v}}(t) \cdot \nabla)\widehat{\mathbf{v}}(t), \Pi\Delta\widehat{\mathbf{v}}(t)) + ((\mathbf{U}(t) \cdot \nabla)\widehat{\mathbf{v}}(t), \Pi\Delta\widehat{\mathbf{v}}(t)) \\
&\quad + ((\widehat{\mathbf{v}}(t) \cdot \nabla)\mathbf{U}(t), \Pi\Delta\widehat{\mathbf{v}}(t)) - (\mathbf{g}(t), \Pi\Delta\widehat{\mathbf{v}}(t)) \\
&\leq C\|\widehat{\mathbf{v}}(t)\|_{\mathbf{L}^4(\Omega)} \|\nabla\widehat{\mathbf{v}}(t)\|_{\mathbf{L}^4(\Omega)} \|\Pi\Delta\widehat{\mathbf{v}}(t)\| \\
&\quad + \left(\|\mathbf{U}(t)\|_{\mathbf{L}^\infty(\Omega)} \|\nabla\widehat{\mathbf{v}}(t)\| + \|\nabla\mathbf{U}(t)\|_{\mathbf{L}^\infty(\Omega)} \|\widehat{\mathbf{v}}(t)\| + \|\mathbf{g}(t)\|\right) \cdot \|\Pi\Delta\widehat{\mathbf{v}}(t)\|.
\end{aligned} \tag{2.18}$$

But $\|\mathbf{w}\|_{L^4} \leq C\|\mathbf{w}\|^{1/2}\|\nabla\mathbf{w}\|^{1/2}$ for all $\mathbf{w} \in \mathbf{H}_0^1(\Omega)$, so together with (1.19), we have that $\|\widehat{\mathbf{v}}\|_{L^4}\|\nabla\widehat{\mathbf{v}}\|_{L^4}\|\Pi\Delta\widehat{\mathbf{v}}\| \leq C\|\widehat{\mathbf{v}}\|^{1/2}\|\nabla\widehat{\mathbf{v}}\| \cdot \|\Pi\Delta\widehat{\mathbf{v}}\|^{3/2}$. By applying Young's inequality, (2.18) yields

$$\begin{aligned}
\frac{d}{dt}\|\nabla\widehat{\mathbf{v}}(t)\|^2 + \nu\|\Pi\Delta\widehat{\mathbf{v}}(t)\|^2 &\leq C\|\widehat{\mathbf{v}}(t)\|^2 \|\nabla\widehat{\mathbf{v}}(t)\|^4 \\
&\quad + C(\|\mathbf{U}(t)\|^2_{\mathbf{L}^\infty(\Omega)} + \|\nabla\mathbf{U}(t)\|^2_{\mathbf{L}^\infty(\Omega)} + 1) \cdot (\|\nabla\widehat{\mathbf{v}}(t)\|^2 + \|\widehat{\mathbf{g}}(t)\|^2),
\end{aligned} \tag{2.19}$$

where, and from now on, C is a generic constant depending only on Ω and ν. By (A4),

$$C_1 = C_1(\nu, \Omega, \mathbf{U}) \stackrel{\text{def}}{=\!=} C \sup_{s \in (0,\infty)} \left\{ \|\mathbf{U}(s)\|^2_{\mathbf{L}^\infty(\Omega)} + \|\nabla\mathbf{U}(s)\|^2_{\mathbf{L}^\infty(\Omega)} + 1 \right\} < \infty\,. \tag{2.20}$$

Theorem 2.3 and the fact that $t \geq T_1 \geq t_n \geq t_{n_0}$ imply that

$$C\|\widehat{\mathbf{v}}(t)\|^2 \leq CM_1\mu^{n-n_0}\|\widehat{\mathbf{v}}(t_{n_0})\|^2 \leq CM_1\|\widehat{\mathbf{v}}(t_{n_0})\|^2 \qquad \forall\, t \in [T_1, T_2].$$

Setting $C_2 = C_2(\nu, \Omega) = CM_1$, we obtain from (2.19) that

$$\frac{d}{dt}\|\nabla\widehat{\mathbf{v}}(t)\|^2 - C_2\|\widehat{\mathbf{v}}(t_{n_0})\|^2\|\nabla\widehat{\mathbf{v}}(t)\|^4 \leq C_1(\|\nabla\widehat{\mathbf{v}}(t)\|^2 + \|\widehat{\mathbf{g}}(t)\|^2).$$

Multiplying this inequality by $\exp\{-C_2\|\widehat{\mathbf{v}}(t_{n_0})\|^2 \int_{t_{n_0}}^t \|\nabla\widehat{\mathbf{v}}(s)\|^2 ds\}$ and integrating over $t \in (T_1, T_2)$, we obtain (2.16). ∎

Now we can prove the exponential decay of $\|\nabla\widehat{\mathbf{v}}(t)\| = \|\nabla\widehat{\mathbf{u}}(t) - \nabla\mathbf{U}(t)\|$. We will keep the above notations.

THEOREM 2.10. *Assume that the conditions of Theorem* 2.7 *and* (A4) *hold. Then, for each fixed* $\tau \in (0,1)$,

$$\|\nabla\widehat{\mathbf{v}}(t)\|^2 \le C_3\|\mathbf{u}_0-\mathbf{U}_0\|^2 e^{-\kappa t}\exp\left\{C_4\|\mathbf{u}_0-\mathbf{U}_0\|^4 e^{-2\kappa t}\right\} \qquad \forall\, t\ge\tau,$$

where

$$\begin{aligned} C_3 &= C_3(\nu,\Omega,\mathbf{U}) \\ &\overset{\text{def}}{=\!=} \left(\left(\frac{1}{\tau}+C_1\right)\left(M_3+\frac{\delta}{2(1-\mu)}\sqrt{\frac{\alpha}{\beta}}\right)\frac{e^{\kappa\delta}}{\nu}+\frac{\alpha C_1\delta}{\beta(1-\mu)}\right)e^{\kappa(\tau+\delta)} \end{aligned}$$

and

$$C_4 = C_4(\nu,\Omega,\mathbf{U}) \overset{\text{def}}{=\!=} C_2M_3\left(M_3+\frac{\delta}{2(1-\mu)}\sqrt{\frac{\alpha}{\beta}}\right)e^{\kappa(2\tau+3\delta)}$$

with κ, μ, *and* δ *defined in Theorem* 2.7 *and* C_1 *and* C_2 *defined in Lemma* 2.9.

Proof. Let $\tau\in(0,1)$ and $t>\tau$ be given. Denote by $[s]$ the maximum integer not greater than s. Then $n_1 \overset{\text{def}}{=\!=} [(t-\tau)/\delta]\ge 0$ and $t_{n_1}\le t-\tau$. For any $s_1\in[t-\tau,t]$, by repeatedly applying (2.16) for $(T_1,T_2)=(s_1,t_{[s_1/\delta]+1}),(t_{[s_1/\delta]+1},t_{[s_1/\delta]+2}),\dots,(t_{[t/\delta]},t)$ and taking the summation, we obtain

$$\begin{aligned} &\|\nabla\widehat{\mathbf{v}}(t)\|^2\exp\left\{-C_2\|\widehat{\mathbf{v}}(t_{n_1})\|^2\int_{t_{n_1}}^t\|\nabla\widehat{\mathbf{v}}(s)\|^2ds\right\} \\ &\qquad\qquad -\|\nabla\widehat{\mathbf{v}}(s_1)\|^2\exp\left\{-C_2\|\widehat{\mathbf{v}}(t_{n_1})\|^2\int_{t_{n_1}}^{s_1}\|\nabla\widehat{\mathbf{v}}(s)\|^2ds\right\} \\ &\le C_1\int_{s_1}^t\left(\|\nabla\widehat{\mathbf{v}}(s)\|^2+\|\widehat{\mathbf{g}}(s)\|^2\right)\exp\left\{-C_2\|\widehat{\mathbf{v}}(t_{n_1})\|^2\int_{t_{n_1}}^s\|\nabla\widehat{\mathbf{v}}(\sigma)\|^2d\sigma\right\}ds \end{aligned}$$

or

$$\begin{aligned} \|\nabla\widehat{\mathbf{v}}(t)\|^2 &\le \|\nabla\widehat{\mathbf{v}}(s_1)\|^2\exp\left\{C_2\|\widehat{\mathbf{v}}(t_{n_1})\|^2\int_{s_1}^t\|\nabla\widehat{\mathbf{v}}(s)\|^2ds\right\} \\ &\quad + C_1\int_{s_1}^t\left(\|\nabla\widehat{\mathbf{v}}(s)\|^2+\|\widehat{\mathbf{g}}(s)\|^2\right)\exp\left\{C_2\|\widehat{\mathbf{v}}(t_{n_1})\|^2\int_s^t\|\nabla\widehat{\mathbf{v}}(\sigma)\|^2d\sigma\right\}ds \\ &\le \left(\|\nabla\widehat{\mathbf{v}}(s_1)\|^2+C_1\int_{t-\tau}^t\left(\|\nabla\widehat{\mathbf{v}}(s)\|^2+\|\widehat{\mathbf{g}}(s)\|^2\right)ds\right) \\ &\qquad\qquad\cdot\exp\left\{C_2\|\widehat{\mathbf{v}}(t_{n_1})\|^2\int_{t-\tau}^t\|\nabla\widehat{\mathbf{v}}(s)\|^2ds\right\}. \end{aligned}$$

Integrating this inequality over $s_1\in(t-\tau,t)$, we obtain

$$\begin{aligned} \|\nabla\widehat{\mathbf{v}}(t)\|^2 &\le \left(\left(\frac{1}{\tau}+C_1\right)\int_{t-\tau}^t\|\nabla\widehat{\mathbf{v}}(s)\|^2\,ds + C_1\int_{t-\tau}^t\|\widehat{\mathbf{g}}(s)\|^2\,ds\right) \\ &\qquad\qquad\cdot\exp\left\{C_2\|\widehat{\mathbf{v}}(t_{n_1})\|^2\int_{t-\tau}^t\|\nabla\widehat{\mathbf{v}}(s)\|^2\,ds\right\}. \end{aligned} \tag{2.21}$$

Now we apply Lemma 2.9 and Theorem 2.3 to obtain

$$\nu\int_{t-\tau}^t\|\nabla\widehat{\mathbf{v}}(s)\|^2ds \le \nu\int_{t_{n_1}}^\infty\|\nabla\widehat{\mathbf{v}}(s)\|^2ds \le \|\widehat{\mathbf{v}}(t_{n_1})\|^2+\frac{1}{\sqrt{\alpha\beta}}\mathcal{J}_{(t_{n_1},\infty)}(\widehat{\mathbf{u}},\widehat{\mathbf{f}}).$$

By Theorem 2.7 and Proposition 2.4,

$$\int_{t-\tau}^{t} \|\nabla \widehat{\mathbf{v}}(s)\|^2 ds \le \left(M_3 + \frac{\delta}{2(1-\mu)} \sqrt{\frac{\alpha}{\beta}} \right) \frac{e^{\kappa\delta}}{\nu} \|\mathbf{u}_0 - \mathbf{U}_0\|^2 \mu^{[t-\tau]} e^{-\kappa(t-\tau)},$$

$$\int_{t-\tau}^{t} \|\widehat{\mathbf{g}}(s)\|^2 ds \le \frac{2}{\beta} \mathcal{J}_{(t_{n_1}, t_{n_2})}(\widehat{\mathbf{u}}^{(n)}, \widehat{\mathbf{f}}^{(n)}) \le \frac{\alpha\delta e^{\kappa\delta}}{\beta(1-\mu)} \|\mathbf{u}_0 - \mathbf{U}_0\|^2 e^{-\kappa(t-\tau)},$$

and

$$\|\widehat{\mathbf{v}}(t_{n_1})\|^2 \le M_3 \|\mathbf{u}_0 - \mathbf{U}_0\|^2 e^{-\kappa(t-\tau)}.$$

The theorem is therefore proven when one combines the last three estimates and (2.21). ∎

Remark 2.11. By virtue of Remarks 2.5 and 2.8, C_3 and C_4 in Theorem 2.10 remain bounded as $\delta \to 0^+$. ∎

COROLLARY 2.12. *Assume that the conditions of Theorem* 2.10 *hold. Then there exists a function* $\theta(\delta) = \theta(\delta; \mathbf{u}_0, \mathbf{U}_0, \nu, \Omega)$ *which is continuous in* $\delta \in (0, \delta_0]$ *such that* $\lim_{\delta \to 0^+} \theta(\delta) < \infty$ *and*

$$\|\nabla \widehat{\mathbf{v}}(t)\| \le \theta(\delta) t^{-1/2} \qquad \forall t \le 1.$$

Namely,

$$\|\nabla \widehat{\mathbf{v}}(t)\| = O(t^{-1/2}) \qquad \text{as } t \to 0^+$$

holds uniformly for $\delta > 0$ *small enough.*

Proof. Set $t = 2\tau$ in Theorem 2.10; then the corollary is evident. ∎

3. Semidiscrete approximations of the piecewise optimal control problem. In order to compute the optimal solutions of the piecewise optimal control problem analyzed in section 2, we need to discretize this problem in both time and space. In this section we will discuss only semidiscretizations (i.e., time discretizations).

3.1. Definition of the semidiscrete piecewise optimal control problem. We semidiscretize the functional $\mathcal{J}_{(t_n, t_{n+1})}(\mathbf{u}, \mathbf{f})$ by the right-endpoint rectangle rule $\int_{t_n}^{t_{n+1}} \varphi(t)\, dt \approx \delta\varphi(t_{n+1})$ so that the semidiscretized functional becomes

$$\mathcal{J}^{n+1}(\mathbf{u}, \mathbf{f}) = \frac{\delta\alpha}{2} \|\mathbf{u} - \mathbf{U}^{n+1}\|^2 + \frac{\delta\beta}{2} \|\mathbf{f} - \mathbf{F}^{n+1}\|^2 \qquad \forall\, \mathbf{u} \in \mathbf{H}_0^1(\Omega), \forall\, \mathbf{f} \in \mathbf{L}^2(\Omega)\,,$$

where $\mathbf{U}^{n+1} = \mathbf{U}^{n+1}(\mathbf{x}) = \mathbf{U}(\mathbf{x}, t_{n+1})$ and $\mathbf{F}^{n+1} = \mathbf{F}^{n+1}(\mathbf{x}) = \mathbf{F}(\mathbf{x}, t_{n+1})$ with $t_n = n\delta$ for $n = 0, 1, 2, \ldots$. Since δ is fixed, the minimization of $\mathcal{J}^{n+1}(\mathbf{u}, \mathbf{f})$ is equivalent to the minimization of $\mathcal{L}^{n+1}(\mathbf{u}, \mathbf{f}) \stackrel{\text{def}}{=\!=} \delta^{-1} \mathcal{J}^{n+1}(\mathbf{u}, \mathbf{f})$. Thus, instead of the functional $\mathcal{J}^{n+1}(\mathbf{u}, \mathbf{f})$, we will use the functional

$$\mathcal{L}^{n+1}(\mathbf{u}, \mathbf{f}) = \frac{\alpha}{2} \|\mathbf{u} - \mathbf{U}^{n+1}\|^2 + \frac{\beta}{2} \|\mathbf{f} - \mathbf{F}^{n+1}\|^2 \qquad \forall\, \mathbf{u} \in \mathbf{H}_0^1(\Omega), \forall\, \mathbf{f} \in \mathbf{L}^2(\Omega)\,.$$

To semidiscretize the Navier–Stokes equations on (t_n, t_{n+1}), we in principle may choose a time step Δt independent of δ (of course $\Delta t \le \delta$). From the analysis in the last section, the exponential decay rates remain bounded as $\delta \to 0^+$. In other words, the piecewise controlled dynamics eventually remain unchanged for arbitrarily small δ. Therefore it is reasonable to choose $\Delta t = \delta$, i.e., the functional and the Navier–Stokes equations are semidiscretized by the same time stepping. We state the semidiscrete approximation of the piecewise optimal control problem as follows.

- Set $\Delta t = \delta$.
- Define $\widehat{\mathbf{u}}^0 = \mathbf{U}_0$.
- *The $(n+1)$th semidiscrete optimal control problem*: for $n = 0, 1, 2, \ldots$, find $(\widehat{\mathbf{u}}^{n+1}, \widehat{p}^{n+1}, \widehat{\mathbf{f}}^{n+1}) \in \mathbf{H}_0^1(\Omega) \times L_0^2(\Omega) \times \mathbf{L}^2(\Omega)$ such that the functional $\mathcal{L}^{n+1}(\mathbf{u}^{n+1}, \mathbf{f}^{n+1})$ is minimized subject to the semidiscrete Navier–Stokes equations

$$\begin{aligned} &\frac{1}{\Delta t}(\mathbf{u}^{n+1}, \mathbf{w}) + a(\mathbf{u}^{n+1}, \mathbf{w}) + c(\mathbf{u}^{n+1}, \mathbf{u}^{n+1}, \mathbf{w}) + b(\mathbf{w}, p^{n+1}) \\ &\qquad = \frac{1}{\Delta t}(\widehat{\mathbf{u}}^n, \mathbf{w}) + (\mathbf{f}^{n+1}, \mathbf{w}) \qquad \forall\, \mathbf{w} \in \mathbf{H}_0^1(\Omega) \end{aligned} \tag{3.1}$$

and

$$b(\mathbf{u}^{n+1}, r) = 0 \qquad \forall\, r \in L_0^2(\Omega)\,. \tag{3.2}$$

Note that for each $n = 0, 1, 2, \ldots$, the $(n+1)$th optimal control problem is a *steady-state* problem for the state variable pair $(\mathbf{u}^{n+1}, p^{n+1})$ and the control variable $\mathbf{f}^{n+1}$. Using the techniques of [GHS] concerning optimal control problems for the steady-state Navier–Stokes equations, we can show the existence of a solution $(\widehat{\mathbf{u}}^{n+1}, \widehat{p}^{n+1}, \widehat{\mathbf{f}}^{n+1})$ for the $(n+1)$th optimal control problem. The remainder of this section will be devoted to the study of $\widehat{\mathbf{u}}^n$ as $n \to \infty$.

3.2. Dynamics of the semidiscrete solutions of the piecewise optimal control problem. We now study the behavior of the semidiscrete solution $\widehat{\mathbf{u}}^n$ as $n \to \infty$.

By the finite difference approximation formula

$$\partial_t \mathbf{U}(\mathbf{x}, t) = \frac{\mathbf{U}(\mathbf{x}, t + \Delta t) - \mathbf{U}(\mathbf{x}, t)}{\Delta t} + \partial_{tt}^2 \mathbf{U}(\mathbf{x}, t + \theta \Delta t)\Delta t,$$

where $\theta = \theta(\mathbf{x}, t)$ and $|\theta| < 1$, we have that

$$\begin{aligned} &\frac{1}{\Delta t}(\mathbf{U}^{n+1}, \mathbf{w}) + a(\mathbf{U}^{n+1}, \mathbf{w}) + c(\mathbf{U}^{n+1}, \mathbf{U}^{n+1}, \mathbf{w}) \\ &\qquad = \frac{1}{\Delta t}(\mathbf{U}^n, \mathbf{w}) + (\mathbf{F}^{n+1} - \boldsymbol{\tau}^{n+1}, \mathbf{w}) \qquad \forall\, \mathbf{w} \in \mathbf{H}_0^1(\Omega) \end{aligned} \tag{3.3}$$

and

$$b(\mathbf{U}^{n+1}, r) = 0 \qquad \forall\, r \in L_0^2(\Omega)\,, \tag{3.4}$$

where

$$\boldsymbol{\tau}^{n+1} = \Delta t\, \partial_{tt} \mathbf{U}\left(\mathbf{x}, t_n + \theta(\mathbf{x}, t_n)\Delta t\right). \tag{3.5}$$

LEMMA 3.1. *Assume that* (A1)–(A2) *and*

$$\text{(A5)} \quad \partial_t \mathbf{U} \in C\left([0, \infty); \mathbf{H}^1(\Omega)\right), \qquad \partial_{tt} \mathbf{U} \in \mathbf{L}^\infty\left(\Omega \times (0, \infty)\right) \cap C\left([0, \infty); \mathbf{L}^2(\Omega)\right)$$

hold. Assume further that $(\widehat{\mathbf{u}}^{n+1}, \widehat{p}^{n+1}, \widehat{\mathbf{f}}^{n+1})$ is a solution of the $(n+1)$th semidiscrete optimal control problem for $n = 1, 2, \ldots$. Then

$$\mathcal{L}^{n+1}(\widehat{\mathbf{u}}^{n+1}, \widehat{\mathbf{f}}^{n+1}) \le \frac{\alpha}{2}\left(\frac{\|\widehat{\mathbf{u}}^n - \mathbf{U}^n\|^2}{1 + C_5 \Delta t} + \frac{C_6 (\Delta t)^3}{1 + C_5 \Delta t}\right),$$

where

$$C_5 = C_5(\nu,\Omega) \stackrel{\text{def}}{=\!=} \frac{\nu\lambda_1}{2} \qquad \textit{and} \qquad C_6 = C_6(\nu,\Omega,\mathbf{U}) \stackrel{\text{def}}{=\!=} \frac{2\,|||\partial_{tt}\mathbf{U}|||^2}{\nu\lambda_1} \tag{3.6}$$

with the norm $|||\cdot|||$ *defined by* (1.7).

Proof. Let $(\widetilde{\mathbf{u}}^{n+1},\widetilde{p}^{n+1})$ be a solution of the equations

$$\begin{aligned}\frac{1}{\Delta t}(\widetilde{\mathbf{u}}^{n+1},\mathbf{w}) + a(\widetilde{\mathbf{u}}^{n+1},\mathbf{w}) + c(\widetilde{\mathbf{u}}^{n+1},\widetilde{\mathbf{u}}^{n+1},\mathbf{w}) + b(\mathbf{w},\widetilde{p}^{n+1})\\ = \frac{1}{\Delta t}(\widehat{\mathbf{u}}^{n},\mathbf{w}) + (\mathbf{F}^{n+1},\mathbf{w}) \qquad \forall\,\mathbf{w}\in\mathbf{H}_0^1(\Omega)\end{aligned} \tag{3.7}$$

and

$$b(\widetilde{\mathbf{u}}^{n+1},r) = 0 \qquad \forall\, r\in L_0^2(\Omega)\,. \tag{3.8}$$

(The existence of such a $(\widetilde{\mathbf{u}}^{n+1},\widetilde{p}^{n+1})$ can be proved by using the techniques for proving the existence of a solution for the steady-state Navier–Stokes equations.) Set $\widetilde{\mathbf{f}}^{n+1} = \mathbf{F}^{n+1}$; then we see that $(\widetilde{\mathbf{u}}^{n+1},\widetilde{p}^{n+1},\widetilde{\mathbf{f}}^{n+1})$ satisfies the semidiscrete Navier–Stokes equations (3.7), (3.8). Let $\widetilde{\mathbf{v}}^{n+1} = \widetilde{\mathbf{u}}^{n+1} - \mathbf{U}^{n+1}$ and $\widetilde{q}^{n+1} = \widetilde{p}^{n+1}$. Then by subtracting (3.3), (3.4) from (3.7), (3.8), we obtain

$$\begin{aligned}&\frac{1}{\Delta t}(\widetilde{\mathbf{v}}^{n+1},\mathbf{w}) + a(\widetilde{\mathbf{v}}^{n+1},\mathbf{w}) + c(\widetilde{\mathbf{v}}^{n+1},\widetilde{\mathbf{v}}^{n+1},\mathbf{w}) + c(\mathbf{U}^{n+1},\widetilde{\mathbf{v}}^{n+1},\mathbf{w})\\ &+ c(\widetilde{\mathbf{v}}^{n+1},\mathbf{U}^{n+1},\mathbf{w}) + b(\mathbf{w},\widetilde{q}^{n+1}) = \frac{1}{\Delta t}(\widehat{\mathbf{v}}^{n},\mathbf{w}) + (\boldsymbol{\tau}^{n+1},\mathbf{w}) \quad \forall\,\mathbf{w}\in\mathbf{H}_0^1(\Omega)\end{aligned} \tag{3.9}$$

and

$$b(\widetilde{\mathbf{v}}^{n+1},r) = 0 \qquad \forall\, r\in L_0^2(\Omega)\,. \tag{3.10}$$

Setting $\mathbf{w} = \widetilde{\mathbf{v}}^{n+1}$ in (3.9), integrating by parts, and using (1.11), we obtain

$$\begin{aligned}&\frac{1}{2\Delta t}\left(\|\widetilde{\mathbf{v}}^{n+1}\|^2 - \|\widehat{\mathbf{v}}^{n}\|^2 + \|\widetilde{\mathbf{v}}^{n+1}-\widehat{\mathbf{v}}^{n}\|^2\right) + \nu\|\nabla\widetilde{\mathbf{v}}^{n+1}\|^2\\ &= -c(\widetilde{\mathbf{v}}^{n+1},\mathbf{U}^{n+1},\widetilde{\mathbf{v}}^{n+1}) + (\boldsymbol{\tau}^{n+1},\widetilde{\mathbf{v}}^{n+1})\\ &= c(\widetilde{\mathbf{v}}^{n+1},\widetilde{\mathbf{v}}^{n+1},\mathbf{U}^{n+1}) + (\boldsymbol{\tau}^{n+1},\widetilde{\mathbf{v}}^{n+1})\\ &\le C_0\|\mathbf{U}^{n+1}\|_{\mathbf{L}^4(\Omega)}\|\nabla\widetilde{\mathbf{v}}^{n+1}\|^2 + \|\boldsymbol{\tau}^{n+1}\|\,\|\widetilde{\mathbf{v}}^{n+1}\|,\end{aligned}$$

so that by (A2) and Young's inequality

$$\begin{aligned}&\frac{1}{2\Delta t}\left(\|\widetilde{\mathbf{v}}^{n+1}\|^2 - \|\widehat{\mathbf{v}}^{n}\|^2 + \|\widetilde{\mathbf{v}}^{n+1}-\widehat{\mathbf{v}}^{n}\|^2\right) + \nu\|\nabla\widetilde{\mathbf{v}}^{n+1}\|^2\\ &\le \frac{\nu}{2}\|\nabla\widetilde{\mathbf{v}}^{n+1}\|^2 + \frac{1}{\nu\lambda_1}\|\boldsymbol{\tau}^{n+1}\|^2 + \frac{\nu\lambda_1}{4}\|\widetilde{\mathbf{v}}^{n+1}\|^2.\end{aligned} \tag{3.11}$$

Dropping the term $\|\widetilde{\mathbf{v}}^{n+1}-\widehat{\mathbf{v}}^{n}\|^2$ and rearranging, we have

$$\frac{1}{2\Delta t}\left(\|\widetilde{\mathbf{v}}^{n+1}\|^2 - \|\widehat{\mathbf{v}}^{n}\|^2\right) + \frac{\nu}{4}\|\nabla\widetilde{\mathbf{v}}^{n+1}\|^2 \le \frac{1}{\nu\lambda_1}\|\boldsymbol{\tau}^{n+1}\|^2 \tag{3.12}$$

so that using the estimate $\|\boldsymbol{\tau}^{n+1}\| \le \Delta t\,|||\partial_{tt}\mathbf{U}|||$ and the Poincaré inequality, we are led to

$$(1 + C_5\Delta t)\,\|\widetilde{\mathbf{v}}^{n+1}\|^2 \le \|\widehat{\mathbf{v}}^{n}\|^2 + C_6(\Delta t)^3,$$

where C_5 and C_6 are defined by (3.6). Hence, we arrive at

$$\mathcal{L}^{n+1}(\widetilde{\mathbf{u}}^{n+1},\widetilde{\mathbf{f}}^{n+1}) = \frac{\alpha}{2}\|\widetilde{\mathbf{v}}^{n+1}\|^2 \le \frac{\alpha}{2}\left[\frac{\|\widehat{\mathbf{v}}^n\|^2}{1+C_5\Delta t} + \frac{C_6(\Delta t)^3}{1+C_5\Delta t}\right].$$

$(\widehat{\mathbf{u}}^{n+1},\widehat{p}^{n+1},\widehat{\mathbf{f}}^{n+1})$ being a solution for the $(n+1)$th optimal control problem, the desired estimate follows trivially from this last inequality. ∎

THEOREM 3.2. *Assume that* (A1)–(A3) *and* (A5) *hold and* $0 < \Delta t \le 1$. *Then there are positive constants* C_7, C_8, κ_1, *and* ρ_1 *such that*

$$\|\widehat{\mathbf{u}}^{n+1} - \mathbf{U}^{n+1}\|^2 \le (1 - C_7\Delta t)\|\widehat{\mathbf{u}}^n - \mathbf{U}^n\|^2 + C_8(\Delta t)^3$$

with $1 - C_7\Delta t > 0$ *and*

$$\|\widehat{\mathbf{u}}^n - \mathbf{U}^n\|^2 \le \|\mathbf{u}_0 - \mathbf{U}_0\|^2\, e^{-\kappa_1 t_n} + \rho_1\Delta t^2. \tag{3.13}$$

Proof. The optimizer $(\widehat{\mathbf{u}}^{n+1},\widehat{p}^{n+1},\widehat{\mathbf{f}}^{n+1})$ satisfies (3.1), (3.2). By setting $\widehat{\mathbf{v}}^{n+1} = \widehat{\mathbf{u}}^{n+1} - \mathbf{U}^{n+1}$, $\widehat{q}^{n+1} = \widehat{p}^{n+1}$, $\widehat{\mathbf{g}}^{n+1} = \widehat{\mathbf{f}}^{n+1} - \mathbf{F}^{n+1}$, and subtracting (3.3), (3.4) from (3.1), (3.2), we see that

$$\begin{aligned}
&\frac{1}{\Delta t}(\widehat{\mathbf{v}}^{n+1},\mathbf{w}) + a(\widehat{\mathbf{v}}^{n+1},\mathbf{w}) + c(\widehat{\mathbf{v}}^{n+1},\widehat{\mathbf{v}}^{n+1},\mathbf{w}) \\
&\quad + c(\mathbf{U}^{n+1},\widehat{\mathbf{v}}^{n+1},\mathbf{w}) + c(\widehat{\mathbf{v}}^{n+1},\mathbf{U}^{n+1},\mathbf{w}) + b(\mathbf{w},\widehat{q}^{n+1}) \\
&\quad = \frac{1}{\Delta t}(\widehat{\mathbf{v}}^n,\mathbf{w}) + (\widehat{\mathbf{g}}^{n+1} + \boldsymbol{\tau}^{n+1},\mathbf{w}) \qquad \forall\, \mathbf{w} \in \mathbf{H}_0^1(\Omega)
\end{aligned} \tag{3.14}$$

and

$$b(\widehat{\mathbf{v}}^{n+1}, r) = 0 \qquad \forall\, r \in L_0^2(\Omega)\,.$$

Setting $\mathbf{w} = \widehat{\mathbf{v}}^{n+1}$ in (3.14) and using the same techniques as in the proof of (3.11) and (3.12), we obtain

$$\begin{aligned}
&\frac{1}{2\Delta t}\left(\|\widehat{\mathbf{v}}^{n+1}\|^2 - \|\widehat{\mathbf{v}}^n\|^2\right) + \frac{\nu}{4}\|\nabla\widehat{\mathbf{v}}^{n+1}\|^2 \\
&\le \left(\widehat{\mathbf{g}}^{n+1},\widehat{\mathbf{v}}^{n+1}\right) + \frac{1}{\nu\lambda_1}\,\|\boldsymbol{\tau}^{n+1}\|^2 \\
&\le \frac{1}{\sqrt{\alpha\beta}}\left(\frac{\alpha}{2}\|\widehat{\mathbf{v}}^{n+1}\|^2 + \frac{\beta}{2}\|\widehat{\mathbf{g}}^{n+1}\|^2\right) + \frac{C_6(\Delta t)^2}{2}\,.
\end{aligned} \tag{3.15}$$

It thus follows from Lemma 3.1 and the Poincaré inequality that

$$\begin{aligned}
&\left(1 + \frac{\nu\lambda_1}{2}\Delta t\right)\|\widehat{\mathbf{v}}^{n+1}\|^2 \\
&\le \left(1 + \frac{\Delta t}{1+C_5\Delta t}\sqrt{\frac{\alpha}{\beta}}\right)\|\widehat{\mathbf{v}}^n\|^2 + C_6(\Delta t)^3\left(1 + \frac{\Delta t}{1+C_5\Delta t}\sqrt{\frac{\alpha}{\beta}}\right).
\end{aligned}$$

By (A3) and the fact that $\Delta t \le 1$, we can find constants

$$C_7 = C_7(\nu,\Omega) \overset{\text{def}}{=\!=} \left(\sqrt{\frac{\alpha}{\beta}} - \frac{\nu\lambda_1}{2}\right) \Big/ \left(1 + \sqrt{\frac{\alpha}{\beta}}\right) > 0$$

and $C_8 = C_8(\nu, \Omega, \mathbf{U}) \stackrel{\text{def}}{=\!=} C_6(\nu, \Omega, \mathbf{U}) > 0$ such that

$$\|\widehat{\mathbf{v}}^{n+1}\|^2 \le (1 - C_7\Delta t)\|\widehat{\mathbf{v}}^n\|^2 + C_8(\Delta t)^3 \qquad \forall\, n \ge 0.$$

It is straightforward to verify by induction that

$$\begin{aligned}\|\widehat{\mathbf{v}}^n\|^2 &\le (1 - C_7\Delta t)^n\|\widehat{\mathbf{v}}^0\|^2 + C_8(\Delta t)^3\sum_{j=0}^{n-1}(1 - C_7\Delta t)^j \\ &= (1 - C_7\Delta t)^n\|\widehat{\mathbf{v}}^0\|^2 + [1 - (1 - C_7\Delta t)^n]\,\frac{C_8}{C_7}(\Delta t)^2 \\ &\le \left[(1 - C_7\Delta t)^{\frac{1}{C_7\Delta t}}\right]^{C_7\, n\Delta t}\|\widehat{\mathbf{v}}^0\|^2 + \frac{C_8}{C_7}(\Delta t)^2\,.\end{aligned}$$

Applying the inequality $1 - y \le e^{-y}$ for all $y \ge 0$, we are led to

$$\|\widehat{\mathbf{v}}^n\|^2 \le \|\widehat{\mathbf{v}}^0\|^2 e^{-C_7\, t_n} + \frac{C_8}{C_7}(\Delta t)^2.$$

Setting $\kappa_1 = \kappa_1(\nu, \Omega) = C_7$ and $\rho_1 = \rho_1(\nu, \Omega, \mathbf{U}) = C_8/C_7$ completes the proof. ∎

As an easy consequence of Theorem 3.2, we obtain the following estimate for the difference between the continuous and semidiscrete solutions of the piecewise optimal control problem.

COROLLARY 3.3. *Assume the hypotheses of Theorem* 3.2 *hold. Let* $\widehat{\mathbf{u}}(t)$ *denote the global solution of the piecewise optimal control problem as defined in section* 2. *Let* $\widehat{\mathbf{u}}^n$ *denote the semidiscrete solution of the piecewise optimal control problem as defined section* 3.1. *Then there exist positive constants* $K_1 = K_1(\nu, \Omega)$, $K_2 = K_2(\nu, \Omega)$, *and* $K_3 = K_3(\nu, \Omega, \mathbf{U})$ *such that*

$$\|\widehat{\mathbf{u}}(t_n) - \widehat{\mathbf{u}}^n\| \le K_1\|\widehat{\mathbf{u}}_0 - \mathbf{U}_0\|\, e^{-K_2 t_n} + K_3\Delta t, \qquad n = 0, 1, 2, \ldots. \qquad ∎$$

Remark 3.4. In the semidiscretization of the Navier–Stokes equations we used the first-order backward Euler scheme. Therefore, the appearance of the term $O(\Delta t)$ in the last estimate is expected. If we use higher-order approximation schemes (see, e.g., [HR]), we expect to obtain improved estimates. However, the analysis in the context of semidiscrete piecewise optimal control with more sophisticated schemes becomes complicated. ∎

The proof of Theorem 3.2 gives a rough estimate of $\|\nabla\widehat{\mathbf{u}}^n - \nabla\mathbf{U}^n\| = \|\nabla\widehat{\mathbf{v}}^n\|$.

PROPOSITION 3.5. *Assume that the conditions of Theorem* 3.2 *hold. Then*

$$\begin{aligned}\Delta t\,\|\nabla\widehat{\mathbf{u}}^n - \nabla\mathbf{U}^n\|^2 \le{}& \frac{2}{\nu}\left(1 + \sqrt{\frac{\alpha}{\beta}}\right) e^{\kappa_1\delta}\|\mathbf{u}_0 - \mathbf{U}_0\|^2 e^{-\kappa_1 t_n} \\ &+ \frac{2}{\nu}\left(1 + \sqrt{\frac{\alpha}{\beta}}\right)(\rho_1 + C_6)\,(\Delta t)^2.\end{aligned} \tag{3.16}$$

Proof. By (3.15), we have that

$$\Delta t\,\|\nabla\widehat{\mathbf{v}}^n\|^2 \le \frac{2}{\nu}\|\widehat{\mathbf{v}}^{n-1}\|^2 + \frac{4\Delta t}{\nu\sqrt{\alpha\beta}}\,\mathcal{L}^n(\widehat{\mathbf{u}}^n, \widehat{\mathbf{f}}^n) + \frac{4C_6}{\nu}(\Delta t)^3. \tag{3.17}$$

But from Lemma 3.1 and Theorem 3.2, we obtain

$$\mathcal{L}^n(\widehat{\mathbf{u}}^n, \widehat{\mathbf{f}}^n) \le \frac{\alpha}{2}\left(\|\widehat{\mathbf{v}}^{n-1}\|^2 + C_6(\Delta t)^3\right)$$

and

$$\|\widehat{\mathbf{v}}^{n-1}\|^2 \le \|\mathbf{u}_0 - \mathbf{U}_0\|^2 e^{-\kappa t_{n-1}} + \rho_1(\Delta t)^2.$$

By substituting these inequalities into (3.17) and noting that $e^{-\kappa_1 t_{n-1}} = e^{\kappa_1 \delta}\, e^{-\kappa_1 t_n}$, the proposition is proven. ∎

Remark 3.6. Because the left-hand side of (3.16) involves Δt, a bound of the eventual error in $\|\cdot\|_1$ is

$$\limsup_{t\to\infty} \|\nabla\widehat{\mathbf{u}}^n - \nabla\mathbf{U}^n\| = O\left((\Delta t)^{1/2}\right) \qquad \text{as } \Delta t \to 0^+,$$

which is of only half the order of the truncation error for the implicit Euler's scheme. On the other hand, however, the bound of the finite time error is singular:

$$\sup_{t\in[0,T]} \|\nabla\widehat{\mathbf{u}}^n - \nabla\mathbf{U}^n\| = O((\Delta t)^{-1/2}) \qquad \text{as } \Delta t \to 0^+. \qquad ∎$$

We next derive an improved bound for the eventual error in $\mathbf{H}^1(\Omega)$ norm. Such an improvement is not always possible for the full discretization. We first observe the following direct consequence of (3.16).

COROLLARY 3.7. *Assume that the conditions of Theorem* 3.2 *hold. Then for any constant* $\sigma > 0$*, there exist constants* $\varepsilon_0 = \varepsilon_0(\Omega, \nu; \sigma) > 0$ *and* $\widetilde{t} = \widetilde{t}(\Omega, \nu, \|\mathbf{u}_0 - \mathbf{U}_0\|; \sigma) > 0$ *such that*

$$\Delta t \|\nabla\widehat{\mathbf{v}}^n\|^2 \le \sigma \qquad \forall\, t_n \ge \widetilde{t},\ \forall\, \Delta t \in (0, \varepsilon_0). \qquad ∎ \tag{3.18}$$

We also need a stronger version of Proposition 3.5.

PROPOSITION 3.8. *Assume the conditions of Theorem* 3.2 *hold. Then for each* $n \ge 1$*,*

(3.19)

$$\mathcal{L}^{n+1}(\widehat{\mathbf{u}}^{n+1}, \widehat{\mathbf{f}}^{n+1}) \le \frac{\alpha}{2}\left(\|\mathbf{u}_0 - \mathbf{U}_0\|^2 e^{-\kappa_1 t_n} + \rho_1(\Delta t)^3 + C_6(\Delta t)^3\right)/(1 + C_5\Delta t).$$

Moreover, for all $n_2 \ge n_1 \ge 1$*,*

$$\Delta t \sum_{n=n_1+1}^{n_2} \|\nabla\widehat{\mathbf{v}}^n\|^2 \le \|\widehat{\mathbf{v}}^{n_1}\|^2 + C_9(t_{n_2} - t_{n_1})\left(\|\mathbf{u}_0 - \mathbf{U}_0\|^2 e^{-\kappa_1 t_{n_1}} + (\Delta t)^2\right), \tag{3.20}$$

where

$$C_9 = C_9(\nu, \Omega, \mathbf{U}) = (1 + \rho_1 + C_6)\sqrt{\frac{\alpha}{\beta}}.$$

Proof. (3.19) follows from Lemma 3.1 and (3.13). By using (3.15) together with (3.19) and (3.13), we obtain that

$$\begin{aligned} &\|\widehat{\mathbf{v}}^{n+1}\|^2 - \|\widehat{\mathbf{v}}^n\|^2 + \frac{\nu\Delta t}{2}\|\nabla\widehat{\mathbf{v}}^{n+1}\|^2 \\ &\le \sqrt{\frac{\alpha}{\beta}}\|\mathbf{u}_0 - \mathbf{U}_0\|^2\, e^{-\kappa_1 t_{n+1}}\Delta t + (\rho_1 + C_6)\sqrt{\frac{\alpha}{\beta}}\,(\Delta t)^3. \end{aligned} \tag{3.21}$$

Summing up (3.21), (3.20) is proven. ∎

We are ready to derive the improved estimates of the $\mathbf{H}^1(\Omega)$ error $\|\nabla\widehat{\mathbf{u}}^n - \nabla\mathbf{U}^n\|$ in (3.16).

THEOREM 3.9. *Suppose that* (A1)–(A5) *hold. Then there exist constants* $\varepsilon \in (0,1)$ *depending only on* Ω *and* ν *and* $\widetilde{t}$ *depending only on* Ω, ν, *and* $\mathbf{U}$ *such that*

$$\begin{aligned}\|\nabla\widehat{\mathbf{u}}^n - \nabla\mathbf{U}^n\|^2 \le C_{10}\left(\frac{1}{\tau}+1+\tau\right)\left(\|\mathbf{u}_0-\mathbf{U}_0\|^2 e^{-\kappa_1(t_n-\tau)}+(\Delta t)^2\right)\\ \cdot\exp\left\{C_{11}(1+\tau)\left(\|\mathbf{u}_0-\mathbf{U}_0\|^4 e^{-2\kappa_1(t_n-\tau)}+(\Delta t)^4\right)\right\}\end{aligned} \tag{3.22}$$

$\forall\,\Delta t \in (0,\varepsilon)$ *and* $\forall\, t_n \ge \widetilde{t}$, *where* κ_1 *is as in Theorem* 3.2 *and* C_{10} *and* C_{11} *are constants depending only on* Ω, ν, *and* $\mathbf{U}$.

Proof. By the standard theory for the Navier–Stokes equations, $\widehat{\mathbf{v}}^{n+1}$ satisfying the strong form of (3.14),

$$\begin{aligned}&\frac{1}{\Delta t}(\widehat{\mathbf{v}}^{n+1}-\widehat{\mathbf{v}}^n)-\nu\Delta\widehat{\mathbf{v}}^{n+1}+(\widehat{\mathbf{v}}^{n+1}\cdot\nabla)\widehat{\mathbf{v}}^{n+1}\\ &+(\mathbf{U}^{n+1}\cdot\nabla)\widehat{\mathbf{v}}^{n+1}+(\widehat{\mathbf{v}}^{n+1}\cdot\nabla)\mathbf{U}^{n+1}-\nabla\widehat{q}^{n+1}=\widehat{\mathbf{g}}^{n+1}+\boldsymbol{\tau}^{n+1}\end{aligned} \tag{3.23}$$

with $\operatorname{div}\widehat{\mathbf{v}}^{n+1}=0$. Taking the $\mathbf{L}^2(\Omega)$ inner product of (3.23) with $-\Pi\Delta\widehat{\mathbf{v}}^{n+1}$, noting that

$$\begin{aligned}\frac{1}{\Delta t}\left(\widehat{\mathbf{v}}^{n+1}-\widehat{\mathbf{v}}^n\,,\,-\Pi\Delta\widehat{\mathbf{v}}^{n+1}\right)&=\frac{1}{\Delta t}\left(\nabla\widehat{\mathbf{v}}^{n+1}-\nabla\widehat{\mathbf{v}}^n\,,\,\nabla\widehat{\mathbf{v}}^{n+1}\right)\\ &=\frac{1}{2\Delta t}\left(\|\nabla\widehat{\mathbf{v}}^{n+1}\|^2-\|\nabla\widehat{\mathbf{v}}^n\|^2+\|\nabla\widehat{\mathbf{v}}^{n+1}-\nabla\widehat{\mathbf{v}}^n\|^2\right)\end{aligned}$$

and

$$\left(-\nu\Delta\widehat{\mathbf{v}}^{n+1}\,,\,-\Pi\Delta\widehat{\mathbf{v}}^{n+1}\right)=\nu\,\|\Pi\Delta\widehat{\mathbf{v}}^{n+1}\|^2\ge\nu C_\Pi\,\|\nabla\widehat{\mathbf{v}}^{n+1}\|^2$$

with $C_\Pi>0$ defined by (1.19) depending only on Ω, and using the similar treatment in (2.18) and (2.19), we arrive at

$$\begin{aligned}&\frac{1}{\Delta t}\left(\|\nabla\widehat{\mathbf{v}}^{n+1}\|^2-\|\nabla\widehat{\mathbf{v}}^n\|^2+\|\nabla\widehat{\mathbf{v}}^{n+1}-\nabla\widehat{\mathbf{v}}^n\|^2\right)+\nu\|\Pi\Delta\widehat{\mathbf{v}}^{n+1}\|^2\\ &\le C\|\widehat{\mathbf{v}}^{n+1}\|^2\|\nabla\widehat{\mathbf{v}}^{n+1}\|^4+C\left(\|\mathbf{U}^{n+1}\|_{L^\infty}^2+\|\nabla\mathbf{U}^{n+1}\|_{L^\infty}^2+1\right)\\ &\qquad\cdot\left(\|\nabla\widehat{\mathbf{v}}^{n+1}\|^2+\|\widehat{\mathbf{g}}^{n+1}\|^2+\|\boldsymbol{\tau}^{n+1}\|^2\right),\end{aligned} \tag{3.24}$$

where C is a constant depending only on Ω and ν.

Denote $K_4=C(\|\mathbf{U}\|^2_{L^\infty(0,\infty;\mathbf{L}^2(\Omega))}+\|\nabla\mathbf{U}\|^2_{L^\infty(0,\infty;\mathbf{L}^2(\Omega))}+1)$ and $\sigma_n=C\|\widehat{\mathbf{v}}^n\|^2$. Then, by Theorem 3.2, for $n=0,1,2,\ldots$

$$\sigma_n\le C\left(\|\mathbf{u}_0-\mathbf{U}_0\|^2e^{-\kappa_1 t_n}+\rho_1(\Delta t)^2\right). \tag{3.25}$$

Hence σ_n is uniformly bounded in n. Now we apply Corollary 3.7 to choose $\varepsilon>0$ and $\widetilde{t}>0$ such that

$$\sigma_n\cdot\Delta t\|\nabla\widehat{\mathbf{v}}^n\|^2\le y_0\qquad\forall\,\Delta t\in(0,\varepsilon),\,\forall\,t_n\ge\widetilde{t},$$

where the constant $y_0>0$ is chosen such that $1-y\ge e^{-2y}$ for each $y\in[0,y_0]$. Hence (3.24) gives rise to, for all $t_n\ge\widetilde{t}$,

$$\begin{aligned}&\exp\left\{-2\sigma_{n+1}\Delta t\,\|\nabla\widehat{\mathbf{v}}^{n+1}\|^2\right\}\|\nabla\widehat{\mathbf{v}}^{n+1}\|^2\\ &\le\left(1-\sigma_{n+1}\Delta t\,\|\nabla\widehat{\mathbf{v}}^{n+1}\|^2\right)\|\nabla\widehat{\mathbf{v}}^{n+1}\|^2\\ &\le\|\nabla\widehat{\mathbf{v}}^n\|^2+(\Delta t)K_4\left(\|\nabla\widehat{\mathbf{v}}^{n+1}\|^2+\|\widehat{\mathbf{g}}^{n+1}\|^2+\|\boldsymbol{\tau}^{n+1}\|^2\right).\end{aligned} \tag{3.26}$$

Let us fix $\tau > 2\Delta t$ and $n_0 > 0$ such that $t_{n_0} \geq \widetilde{t}$. Let n_2 be the largest integer such that $t_{n_2} \leq t_{n_0} + \tau$. Then for each n_1 satisfying $n_0 \leq n_1 \leq n_2$, we have that, by (3.26) and induction,

$$\begin{aligned}
&\exp\left\{-2\Delta t \sum_{j=n_1+1}^{n_2} \sigma_j \|\nabla \widehat{\mathbf{v}}^j\|^2\right\} \|\nabla \widehat{\mathbf{v}}^{n_2}\|^2 \\
&\leq \exp\left\{-2\Delta t \sum_{j=n_1+1}^{n_2-1} \sigma_j \|\nabla \widehat{\mathbf{v}}^j\|^2\right\} \\
&\qquad \cdot \left(\|\nabla \widehat{\mathbf{v}}^{n_2-1}\|^2 + (\Delta t) K_4 \left(\|\nabla \widehat{\mathbf{v}}^{n_2}\|^2 + \|\widehat{\mathbf{g}}^{n_2}\|^2 + \|\boldsymbol{\tau}^{n_2}\|^2\right)\right) \\
&\leq \exp\left\{-2\Delta t \sum_{j=n_1+1}^{n_2-1} \sigma_j \|\nabla \widehat{\mathbf{v}}^j\|^2\right\} \|\nabla \widehat{\mathbf{v}}^{n_2-1}\|^2 + (\Delta t) K_4 \left(\|\nabla \widehat{\mathbf{v}}^{n_2}\|^2 + \|\widehat{\mathbf{g}}^{n_2}\|^2 + \|\boldsymbol{\tau}^{n_2}\|^2\right) \\
&\leq \cdots \\
&\leq \|\nabla \widehat{\mathbf{v}}^{n_1}\|^2 + (\Delta t) K_4 \sum_{j=n_1+1}^{n_2} \left(\|\nabla \widehat{\mathbf{v}}^j\|^2 + \|\widehat{\mathbf{g}}^j\|^2 + \|\boldsymbol{\tau}^j\|^2\right).
\end{aligned}$$

Multiplying the above inequality by Δt and summing up n_1 over $n_0 \leq n_1 \leq n_2 - 1$, we obtain

$$\begin{aligned}
&(t_{n_2} - t_{n_0}) \exp\left\{-2\Delta t \sum_{j=n_0}^{n_2} \sigma_j \|\nabla \widehat{\mathbf{v}}^j\|^2\right\} \|\nabla \widehat{\mathbf{v}}^{n_2}\|^2 \\
&\leq \Delta t \sum_{j=n_0}^{n_2} \|\nabla \widehat{\mathbf{v}}^j\|^2 + K_4 (t_{n_2} - t_{n_0}) \Delta t \sum_{j=n_0}^{n_2} \left(\|\nabla \widehat{\mathbf{v}}^j\|^2 + \|\widehat{\mathbf{g}}^j\|^2 + \|\boldsymbol{\tau}^j\|^2\right).
\end{aligned}$$

Hence

$$\begin{aligned}
\|\nabla \widehat{\mathbf{v}}^{n_2}\|^2 \leq \exp\left\{2\Delta t \sum_{j=n_0}^{n_2} \sigma_j \|\nabla \widehat{\mathbf{v}}^j\|^2\right\} \cdot \Bigg(&\frac{\Delta t}{t_{n_2} - t_{n_0}} \sum_{j=n_0}^{n_2} \|\nabla \widehat{\mathbf{v}}^j\|^2 \\
&+ K_4 (\Delta t) \sum_{j=n_0}^{n_2} \left(\|\nabla \widehat{\mathbf{v}}^j\|^2 + \|\widehat{\mathbf{g}}^j\|^2 + \|\boldsymbol{\tau}^j\|^2\right)\Bigg).
\end{aligned} \tag{3.27}$$

But $\tau > 2\Delta t$ implies

$$\frac{1}{t_{n_2} - t_{n_0}} \leq \frac{1}{\tau - \Delta t} \leq \frac{2}{\tau}.$$

Moreover, by (3.5),

$$\Delta t \sum_{j=n_0}^{n_2} \|\boldsymbol{\tau}^j\|^2 \leq (\Delta t)^3 \|\|\partial_{tt} \mathbf{U}\|\|_\infty^2 \sum_{j=n_0}^{n_2} 1 \leq \tau \|\|\partial_{tt} \mathbf{U}\|\|_\infty^2 \cdot (\Delta t)^2,$$

and by Proposition 3.8 and Theorem 3.2, there exists a constant $\widetilde{C}$ depending only on Ω and ν such that

$$\Delta t \sum_{j=n_0}^{n_2} \|\widehat{\mathbf{g}}^j\|^2 \leq \frac{2\Delta t}{\beta} \sum_{j=n_0}^{n_2} \mathcal{L}^j(\widehat{\mathbf{u}}^j, \widehat{\mathbf{f}}^j) \leq \widetilde{C}(1+\tau)\left(\|\mathbf{u}_0 - \mathbf{U}_0\|^2 e^{-\kappa_1 t_{n_0}} + (\Delta t)^3\right)$$

and

$$\Delta t \sum_{j=n_0}^{n_2} \|\nabla \widehat{\mathbf{v}}^j\|^2 \le \widetilde{C}(1+\tau)\left(\|\mathbf{u}_0 - \mathbf{U}_0\|^2 e^{\kappa_1 t_{n_0}} + (\Delta t)^2\right).$$

Therefore, by (3.27) and (3.25), we conclude that there exist constants C_{10} and C_{11} depending only on Ω, ν, and $\mathbf{U}$ such that

$$\|\widehat{\mathbf{v}}^{n_2}\|^2 \le C_{10}\left(\frac{1}{\tau} + 1 + \tau\right)\left(\|\mathbf{u}_0 - \mathbf{U}_0\|^2 e^{-\kappa_1 t_{n_0}} + (\Delta t)^2\right) \cdot \exp\left\{C_{11}(1+\tau)\left(\|\mathbf{u}_0 - \mathbf{U}_0\|^4 e^{-2\kappa_1 t_{n_0}} + (\Delta t)^4\right)\right\}. \qquad \square$$

Remark 3.10. According to Theorem 3.9 together with its proof, the exponential decay rate for $\|\nabla\widehat{\mathbf{u}}^n - \nabla\mathbf{U}^n\|$ starts to be effective when $t_n \ge \widetilde{t}$, where $\widetilde{t}$ is independent of Δt. In other words, for each fixed, sufficiently small Δt,

$$\limsup_{n\to\infty} \|\nabla\widehat{\mathbf{u}}^n - \nabla\mathbf{U}^n\| = O(\Delta t),$$

with the bound uniform in $\Delta t \in (0, \varepsilon)$. The error in the $\mathbf{H}^1(\Omega)$ norm is reduced in $t_n = n\Delta t$ exponentially with the exponential rate independent of Δt. We emphasize that $\|\nabla\widehat{\mathbf{u}}^n - \nabla\mathbf{U}^n\|$ remains finite for $t_n \in (0, \widetilde{t})$. But unless $\mathbf{u}_0 - \mathbf{U}_0 \in \mathbf{H}_0^1(\Omega)$, we expect the singular behavior of $\|\nabla\widehat{\mathbf{u}}^n - \nabla\mathbf{U}^n\|$ for small t_n as in Remark 3.6. This type of singular behavior is true even when computing the solutions of the uncontrolled continuous Navier–Stokes equations.

Remark 3.11. We defined the solution for the $(n+1)$th semidiscrete piecewise optimal control problem and analyzed the dynamics of the solutions as $n \to \infty$. However, we said nothing about how to solve the $(n+1)$th optimal control problem. We point out that by introducing a Lagrange multiplier $(\widehat{\boldsymbol{\mu}}^{n+1}, \widehat{\pi}^{n+1}) \in \mathbf{H}_0^1(\Omega) \times L_0^2(\Omega)$, we can convert the problem of solving the $(n+1)$th optimal control problem into the problem of solving the optimality system of equations for $(\widehat{\mathbf{u}}^{n+1}, \widehat{p}^{n+1}, \widehat{\mathbf{f}}^{n+1}, \widehat{\boldsymbol{\mu}}^{n+1}, \widehat{\pi}^{n+1})$ which consists of (3.1), (3.2),

$$\frac{1}{\Delta t}(\boldsymbol{\mu}^{n+1}, \boldsymbol{\omega}) + a(\boldsymbol{\mu}^{n+1}, \boldsymbol{\omega}) + c(\boldsymbol{\omega}, \mathbf{u}^{n+1}, \boldsymbol{\mu}^{n+1}) + c(\mathbf{u}^{n+1}, \boldsymbol{\omega}, \boldsymbol{\mu}^{n+1}) + b(\boldsymbol{\omega}, \pi^{n+1}) = \alpha(\mathbf{u}^{n+1} - \mathbf{U}^{n+1}, \boldsymbol{\omega}) \qquad \forall\, \boldsymbol{\omega} \in \mathbf{H}_0^1(\Omega),$$

$$b(\boldsymbol{\mu}^{n+1}, \tau) = 0 \qquad \forall\, \tau \in L_0^2(\Omega),$$

and

$$(\beta\mathbf{f}^{n+1} - \beta\mathbf{F}^{n+1} + \boldsymbol{\mu}^{n+1}, \mathbf{z}) = 0 \qquad \forall\, \mathbf{z} \in \mathbf{L}^2(\Omega).$$

Using the techniques of [GHS] for the study of optimal control problems for the steady-state Navier–Stokes equations, we can show that the above optimality system of equations indeed has a solution. $\square$

4. Fully discrete approximations of the piecewise optimal control problem. Based on the semidiscrete approximations of the piecewise control problem, we now turn to the study of fully discrete approximations. We will discretize the spatial variables by finite element methods.

4.1. Definition of the fully discrete piecewise optimal control problem. We choose families of finite dimensional subspaces $\mathbf{X}_h \subset \mathbf{H}_0^1(\Omega)$ and $S_h \subset L_0^2(\Omega)$. These families are parameterized by the parameter h that tends to zero; commonly, this parameter is chosen to be some measure of the grid size in a subdivision of Ω into finite elements. One may choose any pair of subspaces $\mathbf{X}_h$ and S_h that can be used for finding finite element solutions of the steady-state Navier–Stokes equations. Thus, concerning these subspaces, we make the following standard assumptions. First, we have the approximation properties: there exist an integer $k \geq 1$ and a constant $C' > 0$, independent of h, $\mathbf{v}$, and q such that

$$\inf_{\mathbf{v}_h \in \mathbf{X}_h} \|\mathbf{v} - \mathbf{v}_h\|_1 \leq C' h^m \|\mathbf{v}\|_{m+1} \quad \forall\, \mathbf{v} \in \mathbf{H}^{m+1}(\Omega) \cap \mathbf{H}_0^1(\Omega)\,,\ 1 \leq m \leq k, \tag{4.1}$$

and

$$\inf_{q_h \in S_h} \|q - q_h\|_0 \leq C' h^m \|q\|_m \quad \forall\, q \in H^m(\Omega) \cap L_0^2(\Omega)\,,\ 1 \leq m \leq k\,. \tag{4.2}$$

Next, we assume the *inf-sup condition*, or *Ladyzhenskaya–Babuska–Brezzi condition*: there exists a constant C'', independent of h, such that

$$\inf_{0 \neq q_h \in S_h} \sup_{\mathbf{0} \neq \mathbf{v}_h \in \mathbf{V}_h} \frac{\int_\Omega q_h \operatorname{div} \mathbf{v}_h \, d\mathbf{x}}{\|\mathbf{v}_h\|_1 \, \|q_h\|_0} \geq C''\,. \tag{4.3}$$

This condition assures the stability of finite element discretizations of the Navier–Stokes equations. For thorough discussions of the approximation properties (4.1), (4.2), see, e.g., [Ci] or [GR], and for like discussions of the stability condition (4.3), see, e.g., [GR]. The reference [GR] may also be consulted for a catalogue of finite element subspaces that meet the requirements of (4.1), (4.2).

For each $n \geq 0$, we define the affine space $\mathbf{Y}_h^{n+1} \stackrel{\text{def}}{=\!=} \{\mathbf{f}_h = \mathbf{y}_h + \mathbf{F}_h^{n+1} : \mathbf{y}_h \in \mathbf{X}_h\}$ for the approximate distributed controls, where $\mathbf{F}_h^{n+1}$ is the L^2 projection of $\mathbf{F}^{n+1}$ onto $\mathbf{X}_h$. Note that $\mathbf{Y}_h^{n+1} \subset \mathbf{X}_h$. To preserve the antisymmetry of the trilinear form $c(\cdot,\cdot,\cdot)$ on the finite element spaces which are in general not divergent-free, we introduce the modified trilinear form (see [Te])

$$\overline{c}(\mathbf{u},\mathbf{v},\mathbf{w}) \stackrel{\text{def}}{=\!=} \frac{1}{2}\{c(\mathbf{u},\mathbf{v},\mathbf{w}) - c(\mathbf{u},\mathbf{w},\mathbf{v})\} \qquad \forall\, \mathbf{u},\mathbf{v},\mathbf{w} \in \mathbf{H}_0^1(\Omega)\,.$$

It can be easily verified that

$$\overline{c}(\mathbf{u},\mathbf{v},\mathbf{w}) = c(\mathbf{u},\mathbf{v},\mathbf{w}) \qquad \forall\, \mathbf{u} \in \mathbf{V},\ \forall\, \mathbf{v},\mathbf{w} \in \mathbf{H}_0^1(\Omega)\,, \tag{4.4}$$

$$\overline{c}(\mathbf{u},\mathbf{v},\mathbf{w}) = -\overline{c}(\mathbf{u},\mathbf{w},\mathbf{v}) \qquad \forall\, \mathbf{u},\mathbf{v},\mathbf{w} \in \mathbf{H}_0^1(\Omega)\,, \tag{4.5}$$

$$\overline{c}(\mathbf{u},\mathbf{v},\mathbf{v}) = 0 \qquad \forall\, \mathbf{u},\mathbf{v} \in \mathbf{H}_0^1(\Omega)\,, \tag{4.6}$$

$$|\overline{c}(\mathbf{u},\mathbf{v},\mathbf{w})| \leq \overline{C}_0 \|\nabla \mathbf{u}\| \, \|\mathbf{v}\|_{\mathbf{L}^4(\Omega)} \, \|\nabla \mathbf{w}\| \qquad \forall\, \mathbf{u},\mathbf{v},\mathbf{w} \in \mathbf{H}_0^1(\Omega)\,, \tag{4.7}$$

$$|\overline{c}(\mathbf{u},\mathbf{v},\mathbf{w})| \leq \overline{C}_1 \|\nabla \mathbf{u}\| \, \|\nabla \mathbf{v}\| \, \|\nabla \mathbf{w}\| \qquad \forall\, \mathbf{u},\mathbf{v},\mathbf{w} \in \mathbf{H}_0^1(\Omega) \tag{4.8}$$

and

$$\left.\begin{array}{l} |\overline{c}(\mathbf{u},\mathbf{v},\mathbf{w})| \leq \overline{C}_2 \|\mathbf{u}\|_2 \, \|\mathbf{v}\| \, \|\nabla \mathbf{w}\| \\ |\overline{c}(\mathbf{v},\mathbf{u},\mathbf{w})| \leq \overline{C}_2 \|\mathbf{u}\|_2 \, \|\mathbf{v}\| \, \|\nabla \mathbf{w}\| \end{array}\right\} \qquad \forall\, \mathbf{u} \in \mathbf{H}^2(\Omega) \cap \mathbf{H}_0^1(\Omega), \forall\, \mathbf{v},\mathbf{w} \in \mathbf{H}_0^1(\Omega)\,. \tag{4.9}$$

Once the finite element spaces $\mathbf{X}_h$ and S_h have been chosen, we define the fully discrete approximations of the piecewise optimal control problem as follows (the time discretization is the same as that for the semidiscrete approximations).

- Set $\Delta t = \delta$.
- Define $\widehat{\mathbf{u}}_h^0 = \mathbf{u}_{0,h}$ where $\mathbf{u}_{0,h}$ is the $\mathbf{L}^2(\Omega)$-projection (or interpolation) of $\mathbf{u}_0$ onto $\mathbf{X}_h$.
- *The $(n+1)$th fully discrete optimal control problem*:
  For $n = 0, 1, 2, \ldots$, find a $(\widehat{\mathbf{u}}_h^{n+1}, \widehat{p}_h^{n+1}, \widehat{\mathbf{f}}_h^{n+1}) \in \mathbf{X}_h \times S_h \times \mathbf{Y}_h^{n+1}$ such that the functional

$$\mathcal{L}_h^{n+1}(\mathbf{u}_h^{n+1}, \mathbf{f}_h^{n+1}) \stackrel{\text{def}}{=\!=} \frac{\alpha}{2}\|\mathbf{u}_h^{n+1} - \mathbf{U}^{n+1}\|^2 + \frac{\beta}{2}\|\mathbf{f}_h^{n+1} - \mathbf{F}^{n+1}\|^2 \qquad \forall\, \mathbf{u}_h^{n+1} \in \mathbf{X}_h,\ \forall\, \mathbf{f}_h^{n+1} \in \mathbf{Y}_h^{n+1}$$

  is minimized subject to the fully discrete Navier–Stokes equations

$$\begin{aligned} &\frac{1}{\Delta t}(\mathbf{u}_h^{n+1}, \mathbf{w}_h) + a(\mathbf{u}_h^{n+1}, \mathbf{w}_h) + \overline{c}(\mathbf{u}_h^{n+1}, \mathbf{u}_h^{n+1}, \mathbf{w}_h) + b(\mathbf{w}_h, p_h^{n+1}) \\ &\qquad = (\mathbf{f}_h^{n+1}, \mathbf{w}_h) + \frac{1}{\Delta t}(\widehat{\mathbf{u}}_h^n, \mathbf{w}_h) \qquad \forall\, \mathbf{w}_h \in \mathbf{X}_h \end{aligned} \tag{4.10}$$

  and

$$b(\mathbf{u}_h^{n+1}, r_h) = 0 \qquad \forall\, r_h \in S_h\,. \tag{4.11}$$

Using the techniques of [GHS] concerning the finite element approximations of optimal control problems for the steady-state Navier–Stokes equations, we can show that for each integer $n \geq 0$, there is a solution $(\widehat{\mathbf{u}}_h^{n+1}, \widehat{\mathbf{f}}_h^{n+1})$ for the $(n+1)$th fully discrete optimal control problem.

4.2. Dynamics of the fully discrete solutions of the piecewise optimal control problem. We now study the behavior of the fully discrete solutions $\widehat{\mathbf{u}}_h^n$ as $n \to \infty$.

For every t, we introduce an auxiliary element $(\mathbf{U}_h(t), P_h(t)) \in \mathbf{X}_h \times S_h$ determined by

$$a\,(\mathbf{U}_h(t), \mathbf{w}_h) + b\,(\mathbf{w}_h, P_h(t)) = a\,(\mathbf{U}(t), \mathbf{w}_h) \qquad \forall\, \mathbf{w}_h \in \mathbf{X}_h \tag{4.12}$$

and

$$b\,(\mathbf{U}_h(t), r_h) = 0 \qquad \forall\, r_h \in S_h\,. \tag{4.13}$$

The existence of such a $(\mathbf{U}_h(t), P_h(t))$ follows from the well-known results for the finite element approximations of the steady-state Stokes equations. Furthermore, under the assumption that there is a $k \geq 1$ such that

$$\mathbf{U} \in L^\infty\left(0, \infty; \mathbf{H}^{k+1}(\Omega)\right) \cap C\left([0, \infty); \mathbf{H}^{k+1}(\Omega)\right)\,, \tag{A6}$$

the following error estimates hold:

$$\begin{aligned} &\|\mathbf{U}_h(t) - \mathbf{U}(t)\|_1 + \|P_h(t)\| \\ &\qquad \leq \overline{C}_3 h^k \|\mathbf{U}(t)\|_{k+1} \leq \overline{C}_3 h^k \|\mathbf{U}\|_{L^\infty(0,\infty;\mathbf{H}^{k+1}(\Omega))} \end{aligned} \tag{4.14}$$

and

$$\|\mathbf{U}_h(t) - \mathbf{U}(t)\| \leq \overline{C}_4 h^{k+1} \|\mathbf{U}(t)\|_{k+1} \leq \overline{C}_4 h^{k+1} \|\mathbf{U}\|_{L^\infty(0,\infty;\mathbf{H}^{k+1}(\Omega))}, \tag{4.15}$$

where $\overline{C}_3$ and $\overline{C}_4$ are constants depending on Ω only; see, e.g., [GR] and [GH]. By differentiating (4.12), (4.13) with respect t, we see that $(\partial_t \mathbf{U}_h(t), \partial_t P_h(t))$ satisfies a system of equations similar to (4.12), (4.13) so that under the assumption

$$\partial_t \mathbf{U} \in L^\infty\left(0,\infty;\mathbf{H}^{k+1}(\Omega)\right) \cap C\left([0,\infty);\mathbf{H}^{k+1}(\Omega)\right) \tag{A7}$$

we have the error estimates

$$\begin{aligned} &\|\partial_t \mathbf{U}_h(t) - \partial_t \mathbf{U}(t)\|_1 + \|\partial_t P_h(t)\| \\ &\quad \le \overline{C}_3 h^k \|\partial_t \mathbf{U}(t)\|_{k+1} \le \overline{C}_3 h^k \|\partial_t \mathbf{U}\|_{L^\infty(0,\infty;\mathbf{H}^{k+1}(\Omega))} \end{aligned} \tag{4.16}$$

and

$$\begin{aligned} &\|\partial_t \mathbf{U}_h(t) - \partial_t \mathbf{U}(t)\| \\ &\quad \le \overline{C}_4 h^{k+1} \|\partial_t \mathbf{U}(t)\|_{k+1} \le \overline{C}_4 h^{k+1} \|\partial_t \mathbf{U}\|_{L^\infty(0,\infty;\mathbf{H}^{k+1}(\Omega))} \qquad \forall\, s \in [0,2]\,. \end{aligned} \tag{4.17}$$

By differentiating (4.12), (4.13) twice with respect t, we see that $(\partial_{tt} \mathbf{U}_h(t), \partial_{tt} P_h(t))$ also satisfies a system of equations similar to (4.12), (4.13) so that under the assumption

$$\partial_{tt} \mathbf{U} \in L^\infty\left(0,\infty;\mathbf{H}^1(\Omega)\right) \cap C\left([0,\infty);\mathbf{H}^1(\Omega)\right)\,, \tag{A8}$$

we have the error estimates for $(\partial_{tt} \mathbf{U}_h(t), \partial_{tt} P_h(t))$:

$$\|\partial_{tt} \mathbf{U}_h(t) - \partial_{tt} \mathbf{U}(t)\|_1 + \|\partial_{tt} P_h(t)\| \le \overline{C}_3 \|\partial_{tt} \mathbf{U}(t)\|_1 \le \overline{C}_3 \|\partial_{tt} \mathbf{U}\|_{L^\infty(0,\infty;\mathbf{H}^1(\Omega))}$$

and

$$\|\partial_{tt} \mathbf{U}_h(t) - \partial_{tt} \mathbf{U}(t)\| \le \overline{C}_4 h^s \|\partial_{tt} \mathbf{U}(t)\|_s \le \overline{C}_4 h^s \|\partial_{tt} \mathbf{U}\|_{L^\infty(0,\infty;\mathbf{H}^s(\Omega))} \qquad \forall\, s \in [0,1]\,;$$

in particular,

$$\|\partial_{tt} \mathbf{U}_h(t) - \partial_{tt} \mathbf{U}(t)\| \le \overline{C}_4 \|\partial_{tt} \mathbf{U}(t)\| \le \overline{C}_4 \|\partial_{tt} \mathbf{U}\|_{L^\infty(0,\infty;\mathbf{H}^1(\Omega))}\,. \tag{4.18}$$

Note that the regularity assumption (A8) for $\partial_{tt}\mathbf{U}$ is weaker than the assumptions (A6) for $\mathbf{U}$ or (A7) for $\partial_t \mathbf{U}$. As a result, the error estimate (4.18) is weaker than (4.15) or (4.17).

LEMMA 4.1. *Assume that hypotheses* (A1), (A2), (A5), (A6), (A7), *and* (A8) *hold. Assume further*

$$\|\mathbf{U}\|_{L^\infty(0,\infty;\mathbf{L}^4(\Omega))} < \frac{\nu}{\overline{C}_0}. \tag{A9}$$

For each integer $n \ge 0$, let $(\widehat{\mathbf{u}}_h^{n+1}, \widehat{p}_h^{n+1}, \widehat{\mathbf{f}}_h^{n+1})$ be a solution of the $(n+1)$th fully discrete optimal control problem. Then there exists an $h_0 > 0$ and constants $\overline{K}_1$, $\overline{K}_2$, and $\overline{K}_3$ such that for all $h \le h_0$ and all n,

$$\begin{aligned} &\mathcal{L}_h^{n+1}(\widehat{\mathbf{u}}_h^{n+1}, \widehat{\mathbf{f}}_h^{n+1}) \\ &\le \alpha\left(\frac{\|\widehat{\mathbf{u}}_h^n - \mathbf{U}_h^n\|^2}{1+\lambda_1 \overline{K}_1 \Delta t} + \frac{\overline{K}_2 h^{2k+2} \Delta t}{1+\lambda_1 \overline{K}_1 \Delta t} + \frac{\overline{K}_3 (\Delta t)^3}{1+\lambda_1 \overline{K}_1 \Delta t}\right) + \alpha \overline{C}_4^2 h^{2k+2} \|\mathbf{U}\|^2_{L^\infty(0,\infty;\mathbf{H}^{k+1}(\Omega))}\,, \end{aligned}$$

where

$$h_0 = h_0(\nu, \Omega, \mathbf{U}) \stackrel{\text{def}}{=\!=} \min\left\{ \left\{ \frac{\nu - \overline{C}_0 \|\mathbf{U}\|_{L^\infty(0,\infty;\mathbf{L}^4(\Omega))}}{2\overline{C}_1 \overline{C}_3 \|\mathbf{U}\|_{L^\infty(0,\infty;\mathbf{H}^{k+1}(\Omega))}} \right\}^{1/k}, 1 \right\}, \tag{4.19}$$

$$\begin{aligned}\overline{K}_1 &= \overline{K}_1(\nu,\Omega,\mathbf{U})\\ &\stackrel{\text{def}}{=\!=} \frac{1}{2}\left(\nu - \overline{C}_0\|\mathbf{U}\|_{L^\infty(0,\infty;\mathbf{L}^4(\Omega))} - \overline{C}_1\overline{C}_3 h_0^k \|\mathbf{U}\|_{L^\infty(0,\infty;\mathbf{H}^{k+1}(\Omega)))}\right),\end{aligned} \tag{4.20}$$

$$\begin{aligned}\overline{K}_2 = \overline{K}_2(\nu,\Omega,\mathbf{U}) \stackrel{\text{def}}{=\!=} \frac{4}{\overline{K}_1}\Bigg(&\left[\overline{C}_1^2\overline{C}_3^4 + 4\overline{C}_2^2\overline{C}_4^2\right]\|\mathbf{U}\|^4_{L^\infty(0,\infty;\mathbf{H}^{k+1}(\Omega))}\\ &+\frac{1}{\lambda_1}\overline{C}_4^2\,\|\partial_t\mathbf{U}\|^2_{L^\infty(0,\infty;\mathbf{H}^{k+1}(\Omega))}\Bigg),\end{aligned} \tag{4.21}$$

and

$$\overline{K}_3 = \overline{K}_3(\nu,\Omega,\mathbf{U}) \stackrel{\text{def}}{=\!=} \frac{8(\overline{C}_4^2+1)}{\lambda_1\overline{K}_1}\,\|\partial_{tt}\mathbf{U}\|^2_{L^\infty(0,\infty;\mathbf{H}^1(\Omega))} \tag{4.22}$$

with the constants $\overline{C}_0$, $\overline{C}_1$, $\overline{C}_2$, $\overline{C}_3$, *and* $\overline{C}_4$ *defined by* (4.7), (4.8), (4.9), (4.14), *and* (4.15), *respectively.*

Proof. Let $(\widetilde{\mathbf{u}}_h^{n+1}, \widetilde{p}_h^{n+1}) \in \mathbf{X}_h \times S_h$ be a solution of the equations

$$\begin{aligned}\frac{1}{\Delta t}(\widetilde{\mathbf{u}}_h^{n+1},\mathbf{w}_h) + a(\widetilde{\mathbf{u}}_h^{n+1},\mathbf{w}_h) + \overline{c}(\widetilde{\mathbf{u}}_h^{n+1},\widetilde{\mathbf{u}}_h^{n+1},\mathbf{w}_h) + b(\mathbf{w},\widetilde{q}_h^{n+1})\\ = (\mathbf{F}_h^{n+1},\mathbf{w}_h) + \frac{1}{\Delta t}(\widehat{\mathbf{u}}_h^n,\mathbf{w}_h) \qquad \forall\,\mathbf{w}_h \in \mathbf{X}_h\end{aligned} \tag{4.23}$$

and

$$b(\widetilde{\mathbf{u}}_h^{n+1}, r_h) = 0 \qquad \forall\, r_h \in S_h\,. \tag{4.24}$$

(The existence of such a $(\widetilde{\mathbf{u}}^{n+1}, \widetilde{p}^{n+1})$ can be proved by using the techniques for proving the existence of finite element solutions for the steady-state Navier–Stokes equations; see, e.g., [GR].) Set $\widetilde{\mathbf{f}}_h^{n+1} = \mathbf{F}_h^{n+1} \in \mathbf{Y}_h$.

From (1.16), (A1), and (4.4), we obtain

$$\begin{aligned}(\partial_t\mathbf{U}(t_{n+1}),\mathbf{w}_h) + a\left(\mathbf{U}(t_{n+1}),\mathbf{w}_h\right) + \overline{c}\left(\mathbf{U}(t_{n+1}),\mathbf{U}(t_{n+1}),\mathbf{w}_h\right)\\ = (\mathbf{F}(t_{n+1}),\mathbf{w}_h) \qquad \forall\,\mathbf{w}_h \in \mathbf{X}_h\end{aligned}$$

and

$$b\left(\mathbf{U}(t_{n+1}), r_h\right) = 0 \qquad \forall\, r_h \in S_h\,.$$

Let $(\mathbf{U}_h(t), P_h(t))$ be defined by (4.12), (4.13). Setting $\mathbf{V}_h(t) = \mathbf{U}_h(t) - \mathbf{U}(t)$ and using (4.12), (4.13) and the fact that $(\mathbf{F}_h^{n+1},\mathbf{w}_h) = (\mathbf{F}^{n+1},\mathbf{w}_h)$ for all $\mathbf{w}_h \in \mathbf{X}_h$, we are led to

$$\begin{aligned}(\partial_t\mathbf{U}_h(t_{n+1}),\mathbf{w}_h) - (\partial_t\mathbf{V}_h(t_{n+1}),\mathbf{w}_h) + a\left(\mathbf{U}_h(t_{n+1}),\mathbf{w}_h\right) + b\left(\mathbf{w}_h, P_h(t_{n+1})\right)\\ + \overline{c}\left(\mathbf{U}(t_{n+1}),\mathbf{U}(t_{n+1}),\mathbf{w}_h\right) = (\mathbf{F}(t_{n+1}),\mathbf{w}_h) \qquad \forall\,\mathbf{w}_h \in \mathbf{X}_h\end{aligned} \tag{4.25}$$

and

$$b\left(\mathbf{U}_h(t_{n+1}), r_h\right) = 0 \qquad \forall\, r_h \in S_h\,. \tag{4.26}$$

We introduce the notations $\mathbf{U}_h^{n+1} = \mathbf{U}_h(t_{n+1})$, $P_h^{n+1} = P_h(t_{n+1})$, $\mathbf{V}_h^{n+1} = \mathbf{V}_h(t_{n+1})$, and $(\partial_t\mathbf{V}_h)^{n+1} = \partial_t\mathbf{V}_h(t_{n+1})$ (also recall $\mathbf{F}^{n+1} = \mathbf{F}(t_{n+1})$). By plugging into (4.25) the finite difference approximation formula

$$\partial_t\mathbf{U}_h(\mathbf{x},t_{n+1}) = \frac{\mathbf{U}_h^{n+1}(\mathbf{x}) - \mathbf{U}_h^n(\mathbf{x})}{\Delta t} + \boldsymbol{\tau}_h^{n+1}(\mathbf{x})\,,$$

we obtain

(4.27)
$$\frac{1}{\Delta t}\left(\mathbf{U}_h^{n+1}-\mathbf{U}_h^n,\mathbf{w}_h\right)+a\left(\mathbf{U}_h^{n+1},\mathbf{w}_h\right)+\overline{c}\left(\mathbf{U}^{n+1},\mathbf{U}^{n+1},\mathbf{w}_h\right)+b\left(\mathbf{w}_h,P_h^{n+1}\right)$$
$$=\left(\mathbf{F}^{n+1},\mathbf{w}_h\right)-\left(\boldsymbol{\tau}_h^{n+1},\mathbf{w}_h\right)+\left((\partial_t\mathbf{V}_h)^{n+1},\mathbf{w}_h\right)\qquad\forall\,\mathbf{w}_h\in\mathbf{X}_h$$

and

$$b\left(\mathbf{U}_h^{n+1},r_h\right)=0\qquad\forall\,r_h\in S_h\,. \tag{4.28}$$

We now examine the error term $\boldsymbol{\tau}_h^{n+1}(\mathbf{x})$ in the previous finite difference formula. Since

$$\begin{aligned}\boldsymbol{\tau}_h^{n+1}(\mathbf{x})&=\partial_t\mathbf{U}(\mathbf{x},t_{n+1})-\frac{\mathbf{U}_h(\mathbf{x},t_{n+1})-\mathbf{U}_h(\mathbf{x},t_n)}{\Delta t}\\&=\frac{1}{\Delta t}\int_{t_n}^{t_{n+1}}\int_s^{t_{n+1}}\partial_{tt}\mathbf{U}_h(\mathbf{x},\sigma)\,d\sigma\,ds,\end{aligned}$$

we see that

$$\begin{aligned}\|\boldsymbol{\tau}_h^{n+1}\|^2=(\boldsymbol{\tau}_h^{n+1},\boldsymbol{\tau}_h^{n+1})&=\frac{1}{\Delta t}\int_{t_n}^{t_{n+1}}\int_s^{t_{n+1}}\int_\Omega\partial_{tt}\mathbf{U}_h(\mathbf{x},\sigma)\cdot\boldsymbol{\tau}_h^{n+1}(\mathbf{x})\,d\mathbf{x}\,d\sigma\,ds\\&\le\frac{1}{\Delta t}\int_{t_n}^{t_{n+1}}\int_s^{t_{n+1}}\|\partial_{tt}\mathbf{U}_h(\sigma)\|\,\|\boldsymbol{\tau}_h^{n+1}\|\,d\sigma\,ds\,.\end{aligned}$$

Hence

$$\begin{aligned}\|\boldsymbol{\tau}_h^{n+1}\|&\le\|\partial_{tt}\mathbf{U}_h\|_{L^\infty(0,\infty;\mathbf{L}^2(\Omega))}\cdot\frac{1}{\Delta t}\int_{t_n}^{t_{n+1}}\int_s^{t_{n+1}}d\sigma\,ds\\&\le\Delta t\,\|\partial_{tt}\mathbf{U}_h\|_{L^\infty(0,\infty;\mathbf{L}^2(\Omega))}.\end{aligned}$$

We note that by (4.18),

$$\begin{aligned}\|\boldsymbol{\tau}_h^{n+1}\|^2&\le2(\Delta t)^2\left(\|\partial_{tt}\mathbf{V}_h\|^2_{L^\infty(0,\infty;\mathbf{L}^2(\Omega))}+\|\partial_{tt}\mathbf{U}\|^2_{L^\infty(0,\infty;\mathbf{L}^2(\Omega))}\right)\\&\le2(\Delta t)^2\left(\overline{C}_4^2+1\right)\|\partial_{tt}\mathbf{U}\|^2_{L^\infty(0,\infty;\mathbf{H}^1(\Omega))}\,.\end{aligned} \tag{4.29}$$

Now, let $\widetilde{\mathbf{v}}_h^{n+1}=\widetilde{\mathbf{u}}_h^{n+1}-\mathbf{U}_h^{n+1}$, $\widetilde{q}_h^{n+1}=\widetilde{p}_h^{n+1}-P_h^{n+1}$, and $\widehat{\mathbf{v}}_h^n=\widehat{\mathbf{u}}_h^n-\mathbf{U}_h^n$. Then by subtracting (4.27), (4.28) from (4.23), (4.24) and again by using $(\mathbf{F}_h^{n+1},\mathbf{w}_h)=(\mathbf{F}^{n+1},\mathbf{w}_h)$ for all $\mathbf{w}_h\in\mathbf{X}_h$, we obtain

$$\begin{aligned}&\frac{1}{\Delta t}(\widetilde{\mathbf{v}}_h^{n+1},\mathbf{w}_h)-\frac{1}{\Delta t}(\widehat{\mathbf{v}}_h^n,\mathbf{w}_h)+a(\widetilde{\mathbf{v}}_h^{n+1},\mathbf{w}_h)+\overline{c}(\widetilde{\mathbf{v}}_h^{n+1},\widetilde{\mathbf{v}}_h^{n+1},\mathbf{w}_h)+\overline{c}(\widetilde{\mathbf{v}}_h^{n+1},\mathbf{V}_h^{n+1},\mathbf{w}_h)\\&\quad+\overline{c}(\mathbf{V}_h^{n+1},\widetilde{\mathbf{v}}_h^{n+1},\mathbf{w}_h)+\overline{c}(\widetilde{\mathbf{v}}_h^{n+1},\mathbf{U}^{n+1},\mathbf{w}_h)+\overline{c}(\mathbf{U}^{n+1},\widetilde{\mathbf{v}}_h^{n+1},\mathbf{w}_h)\\&\quad+\overline{c}(\mathbf{V}_h^{n+1},\mathbf{V}_h^{n+1},\mathbf{w}_h)+\overline{c}(\mathbf{V}_h^{n+1},\mathbf{U}^{n+1},\mathbf{w}_h)+\overline{c}(\mathbf{U}^{n+1},\mathbf{V}_h^{n+1},\mathbf{w}_h)\\&\quad+b(\mathbf{w}_h,\widetilde{q}_h^{n+1})=(\boldsymbol{\tau}_h^{n+1},\mathbf{w}_h)-\left((\partial_t\mathbf{V}_h)^{n+1},\mathbf{w}_h\right)\qquad\forall\,\mathbf{w}_h\in\mathbf{X}_h\end{aligned}$$

and

$$b(\widetilde{\mathbf{v}}_h^{n+1},r_h)=0\qquad\forall\,r_h\in S_h\,.$$

Setting $\mathbf{w}_h = \widetilde{\mathbf{v}}_h^{n+1}$ in the first of the last two equations and using (4.5)–(4.9), we have

$$\begin{aligned}
&\frac{1}{2\Delta t}\left(\|\widetilde{\mathbf{v}}_h^{n+1}\|^2 - \|\widehat{\mathbf{v}}_h^{n}\|^2 + \|\widetilde{\mathbf{v}}_h^{n+1} - \widehat{\mathbf{v}}_h^{n}\|^2\right) + \nu\|\nabla\widetilde{\mathbf{v}}_h^{n+1}\|^2 \\
&= -\overline{c}(\widetilde{\mathbf{v}}_h^{n+1}, \mathbf{V}_h^{n+1}, \widetilde{\mathbf{v}}_h^{n+1}) - \overline{c}(\widetilde{\mathbf{v}}_h^{n+1}, \mathbf{U}^{n+1}, \widetilde{\mathbf{v}}_h^{n+1}) - \overline{c}(\mathbf{V}_h^{n+1}, \mathbf{V}_h^{n+1}, \widetilde{\mathbf{v}}_h^{n+1}) \\
&\quad - \overline{c}(\mathbf{V}_h^{n+1}, \mathbf{U}^{n+1}, \widetilde{\mathbf{v}}_h^{n+1}) - \overline{c}(\mathbf{U}^{n+1}, \mathbf{V}_h^{n+1}, \widetilde{\mathbf{v}}_h^{n+1}) \\
&\quad + (\boldsymbol{\tau}_h^{n+1}, \widetilde{\mathbf{v}}_h^{n+1}) - \left((\partial_t \mathbf{V}_h)^{n+1}, \widetilde{\mathbf{v}}_h^{n+1}\right) \\
&\le \overline{C}_1\|\nabla\widetilde{\mathbf{v}}_h^{n+1}\|^2\,\|\nabla\mathbf{V}_h^{n+1}\| + \overline{C}_0\|\nabla\widetilde{\mathbf{v}}_h^{n+1}\|^2\,\|\mathbf{U}^{n+1}\|_{\mathbf{L}^4(\Omega)} + \overline{C}_1\|\nabla\mathbf{V}_h^{n+1}\|^2\,\|\nabla\widetilde{\mathbf{v}}_h^{n+1}\| \\
&\quad + 2\overline{C}_2\,\|\mathbf{V}_h^{n+1}\|\,\|\mathbf{U}^{n+1}\|_2\,\|\nabla\widetilde{\mathbf{v}}_h^{n+1}\| + \|\boldsymbol{\tau}_h^{n+1}\|\,\|\widetilde{\mathbf{v}}_h^{n+1}\| + \|(\partial_t\mathbf{V}_h)^{n+1}\|\,\|\widetilde{\mathbf{v}}_h^{n+1}\|\,.
\end{aligned}$$

By choosing h_0 as in (4.19) and using (4.14) we see that for all $h \le h_0$,

$$\|\nabla\mathbf{V}_h^{n+1}\| \le \overline{C}_3 h^k \|\mathbf{U}\|_{L^\infty(0,\infty;\mathbf{H}^{k+1}(\Omega))} \le \overline{C}_3 h_0^k \|\mathbf{U}\|_{L^\infty(0,\infty;\mathbf{H}^{k+1}(\Omega))}\,.$$

We define the constant $\overline{K}_1 = \overline{K}_1(\nu, \Omega, \mathbf{U})$ by (4.20) and continue the last inequality involving $\widetilde{\mathbf{v}}_h^{n+1}$ to obtain (by using (4.14) and the inequality $ab \le \epsilon a^2 + \epsilon^{-1}b^2/4$)

$$\begin{aligned}
&\frac{1}{2\Delta t}\left(\|\widetilde{\mathbf{v}}_h^{n+1}\|^2 - \|\widehat{\mathbf{v}}_h^{n}\|^2\right) + \nu\|\nabla\widetilde{\mathbf{v}}_h^{n+1}\|^2 \\
&\le \overline{C}_1\overline{C}_3 h_0^k \|\mathbf{U}\|_{L^\infty(0,\infty;\mathbf{H}^{k+1}(\Omega))}\,\|\nabla\widetilde{\mathbf{v}}_h^{n+1}\|^2 + \overline{C}_0\|\nabla\widetilde{\mathbf{v}}_h^{n+1}\|^2\,\|\mathbf{U}\|_{L^\infty(0,\infty;\mathbf{L}^4(\Omega))} \\
&\quad + \frac{2\overline{C}_1^2}{\overline{K}_1}\|\nabla\mathbf{V}_h^{n+1}\|^4 + \frac{\overline{K}_1}{8}\|\nabla\widetilde{\mathbf{v}}_h^{n+1}\|^2 + \frac{8\overline{C}_2^2}{\overline{K}_1}\|\mathbf{V}_h^{n+1}\|^2\,\|\mathbf{U}^{n+1}\|_2^2 + \frac{\overline{K}_1}{8}\|\nabla\widetilde{\mathbf{v}}_h^{n+1}\|^2 \\
&\quad + \frac{2}{\lambda_1\overline{K}_1}\|\boldsymbol{\tau}_h^{n+1}\|^2 + \frac{\lambda_1\overline{K}_1}{8}\|\widetilde{\mathbf{v}}_h^{n+1}\|^2 + \frac{2}{\overline{K}_1\lambda_1}\|(\partial_t\mathbf{V}_h)^{n+1}\|^2 + \frac{\overline{K}_1\lambda_1}{8}\|\widetilde{\mathbf{v}}_h^{n+1}\|^2\,.
\end{aligned}$$

Using Poincaré inequality for the terms involving $\lambda_1\|\widetilde{\mathbf{v}}_h^{n+1}\|^2$ and simplifying, we obtain

$$\begin{aligned}
&\frac{1}{2\Delta t}\left(\|\widetilde{\mathbf{v}}_h^{n+1}\|^2 - \|\widehat{\mathbf{v}}_h^{n}\|^2\right) + \frac{\overline{K}_1}{2}\|\nabla\widetilde{\mathbf{v}}_h^{n+1}\|^2 \\
&\le \frac{2}{\overline{K}_1}\left(\overline{C}_1^2\,\|\nabla\mathbf{V}_h^{n+1}\|^4 + 4\overline{C}_2^2\,\|\mathbf{V}_h^{n+1}\|^2\,\|\mathbf{U}^{n+1}\|_2^2 + \frac{1}{\lambda_1}\,\|(\partial_t\mathbf{V}_h)^{n+1}\|^2 + \frac{1}{\lambda_1}\,\|\boldsymbol{\tau}_h^{n+1}\|^2\right),
\end{aligned}$$

so that using (4.14)–(4.18) and (4.29), we are led to

$$\begin{aligned}
&\|\widetilde{\mathbf{v}}_h^{n+1}\|^2 + \overline{K}_1\Delta t\,\|\nabla\widetilde{\mathbf{v}}_h^{n+1}\|^2 \\
&\le \|\widehat{\mathbf{v}}_h^{n}\|^2 + \frac{4\Delta t}{\overline{K}_1}\left(\overline{C}_1^2\overline{C}_3^4\,h^{4k}\,\|\mathbf{U}\|^4_{L^\infty(0,\infty;\mathbf{H}^{k+1}(\Omega))} + 4\overline{C}_2^2\overline{C}_4^2\,h^{2k+2}\,\|\mathbf{U}\|^4_{L^\infty(0,\infty;\mathbf{H}^{k+1}(\Omega))}\right. \\
&\qquad\qquad + \frac{1}{\lambda_1}\overline{C}_4^2\,h^{2k+2}\,\|\partial_t\mathbf{U}\|^2_{L^\infty(0,\infty;\mathbf{H}^{k+1}(\Omega))} \\
&\qquad\qquad \left. + \frac{2(\Delta t)^2(\overline{C}_4^2+1)}{\lambda_1}\,\|\partial_{tt}\mathbf{U}\|^2_{L^\infty(0,\infty;\mathbf{H}^{1}(\Omega))}\right).
\end{aligned}$$

Hence, using Poincaré inequality again for the second term on the left-hand side of the last inequality, we obtain

$$\|\widetilde{\mathbf{v}}_h^{n+1}\|^2 \le \frac{\|\widehat{\mathbf{v}}_h^n\|^2}{1+\lambda_1\overline{K}_1\Delta t} + \frac{\overline{K}_2 h^{2k+2}(\Delta t)}{1+\lambda_1\overline{K}_1\Delta t} + \frac{\overline{K}_3(\Delta t)^3}{1+\lambda_1\overline{K}_1\Delta t}\,,$$

where $\overline{K}_2$ and $\overline{K}_3$ are defined by (4.21), (4.22). Using the definition of the functional $\mathcal{L}_h^{n+1}$ and (4.15), we obtain

$$\mathcal{L}_h^{n+1}(\widetilde{\mathbf{u}}_h^{n+1},\widetilde{\mathbf{f}}_h^{n+1}) = \frac{\alpha}{2}\|\widetilde{\mathbf{v}}_h^{n+1}+\mathbf{U}_h^{n+1}-\mathbf{U}^{n+1}\|^2 \le \alpha\,\|\widetilde{\mathbf{v}}_h^{n+1}\|^2 + \alpha\,\|\mathbf{U}_h^{n+1}-\mathbf{U}^{n+1}\|^2$$
$$\le \alpha\left[\frac{\|\widehat{\mathbf{v}}_h^n\|^2}{1+\lambda_1\overline{K}_1\Delta t} + \frac{\overline{K}_2 h^{2k+2}\Delta t}{1+\lambda_1\overline{K}_1\Delta t} + \frac{\overline{K}_3(\Delta t)^3}{1+\lambda_1\overline{K}_1\Delta t}\right] + \alpha\overline{C}_4^2 h^{2k+2}\,\|\mathbf{U}\|^2_{L^\infty(0,\infty;\mathbf{H}^{k+1}(\Omega))}.$$

$(\widehat{\mathbf{u}}_h^{n+1},\widehat{p}_h^{n+1},\widehat{\mathbf{f}}_h^{n+1})$ being a solution for the $(n+1)$th fully discrete optimal control problem, the desired estimate follows trivially from the last inequality. ∎

THEOREM 4.2. *Assume that the hypotheses of Lemma* 4.1 *hold. Assume further that* $\mathbf{u}_0 \in \mathbf{H}^{k+1}(\Omega)$ *and*

$$\frac{\alpha}{\beta} < \frac{(\lambda_1\overline{K}_1)^2}{16}, \tag{A10}$$

where $\overline{K}_1$ *is defined by* (4.20). *Let* h_0 *be defined by* (4.19). *Then there are positive constants* $\delta_0 = \delta_0(\nu,\Omega,\mathbf{U})$, $\overline{K}_4 = \overline{K}_4(\nu,\Omega,\mathbf{U})$, $\overline{K}_5 = \overline{K}_5(\nu,\Omega,\mathbf{U})$, $\overline{K}_6 = \overline{K}_6(\nu,\Omega,\mathbf{U})$, $\overline{\gamma} = \overline{\gamma}(\nu,\Omega,\mathbf{U})$, *and* $\overline{\kappa} = \overline{\kappa}(\nu,\Omega,\mathbf{U})$ *such that for all* $h \le h_0$ *and all* $\Delta t \le \delta_0$,

$$\|\widehat{\mathbf{u}}_h^{n+1}-\mathbf{U}_h^{n+1}\|^2 \le (1-\overline{K}_4\Delta t)\|\widehat{\mathbf{u}}_h^n-\mathbf{U}_h^n\|^2 + \overline{K}_5\Delta t^3 + \overline{K}_6 h^{2k+2}(\Delta t)$$

and

$$\|\widehat{\mathbf{u}}_h^n-\mathbf{U}^n\|^2 \le 3e^{-\overline{\gamma}t_n}\|\widehat{\mathbf{u}}_0-\mathbf{U}^0\|^2 + \overline{\kappa}\left[(\Delta t)^2 + h^{2k+2}\right].$$

Proof. The optimizer $(\widehat{\mathbf{u}}_h^{n+1},\widehat{p}_h^{n+1},\widehat{\mathbf{f}}_h^{n+1})$ satisfies (4.10), (4.11). Let $\mathbf{U}_h^{n+1}$, P_h^{n+1}, $\mathbf{V}_h^{n+1}$, and $(\partial_t\mathbf{V}_h)^{n+1}$ be defined as in the proof of Lemma 4.1 and satisfy (4.25), (4.26). By setting $\widehat{\mathbf{v}}_h^{n+1} = \widehat{\mathbf{u}}_h^{n+1} - \mathbf{U}_h^{n+1}$, $\widehat{q}_h^{n+1} = \widehat{p}_h^{n+1} - \widehat{P}^{n+1}$, $\widehat{\mathbf{g}}_h^{n+1} = \widehat{\mathbf{f}}_h^{n+1} - \mathbf{F}_h^{n+1}$ and subtracting (4.25), (4.26) from (4.10), (4.11) with $(\mathbf{u}_h^{n+1},p_h^{n+1},\mathbf{f}_h^{n+1}) = (\widehat{\mathbf{u}}_h^{n+1},\widehat{p}_h^{n+1},\widehat{\mathbf{f}}_h^{n+1})$, we see that

(4.30)
$$\frac{1}{\Delta t}(\widehat{\mathbf{v}}_h^{n+1},\mathbf{w}_h) + a(\widehat{\mathbf{v}}_h^{n+1},\mathbf{w}_h) + \overline{c}(\widehat{\mathbf{v}}_h^{n+1},\widehat{\mathbf{v}}_h^{n+1},\mathbf{w}_h) + \overline{c}(\widehat{\mathbf{v}}_h^{n+1},\mathbf{V}_h^{n+1},\mathbf{w}_h)$$
$$+ \overline{c}(\mathbf{V}_h^{n+1},\widehat{\mathbf{v}}_h^{n+1},\mathbf{w}_h) + \overline{c}(\widehat{\mathbf{v}}_h^{n+1},\mathbf{U}^{n+1},\mathbf{w}_h) + \overline{c}(\mathbf{U}^{n+1},\widehat{\mathbf{v}}_h^{n+1},\mathbf{w}_h)$$
$$+ \overline{c}(\mathbf{V}_h^{n+1},\mathbf{V}_h^{n+1},\mathbf{w}_h) + \overline{c}(\mathbf{V}_h^{n+1},\mathbf{U}^{n+1},\mathbf{w}_h) + \overline{c}(\mathbf{U}^{n+1},\mathbf{V}_h^{n+1},\mathbf{w}_h) + b(\mathbf{w}_h,\widehat{q}_h^{n+1})$$
$$= (\widehat{\mathbf{g}}_h^{n+1}+\boldsymbol{\tau}_h^{n+1},\mathbf{w}_h) + \frac{1}{\Delta t}(\widehat{\mathbf{v}}_h^n,\mathbf{w}_h) - \left((\partial_t\mathbf{V}_h)^{n+1},\mathbf{w}_h\right) \qquad \forall\,\mathbf{w}_h \in \mathbf{X}_h$$

and

$$b(\widehat{\mathbf{v}}_h^{n+1},r_h) = 0 \qquad \forall\, r_h \in S_h\,. \tag{4.31}$$

Setting $\mathbf{w}_h = \widehat{\mathbf{v}}_h^{n+1}$ in the first of the last two equations and using the same tricks as in the proof of Lemma 4.1 (with $\widehat{\mathbf{u}}_h^{n+1}$ replacing $\widetilde{\mathbf{u}}_h^{n+1}$ and $\widehat{\mathbf{g}}_h^{n+1}+\boldsymbol{\tau}_h^{n+1}$ replacing

$\boldsymbol{\tau}_h^{n+1}$), we obtain

$$\begin{aligned}
&\frac{1}{2\Delta t}\left(\|\widehat{\mathbf{v}}_h^{n+1}\|^2-\|\widehat{\mathbf{v}}_h^{n}\|^2\right)+\frac{\overline{K}_1}{2}\|\nabla\widehat{\mathbf{v}}_h^{n+1}\|^2\\
&\le\frac{2}{\overline{K}_1}\Big(\overline{C}_1^2\,\|\nabla\mathbf{V}_h^{n+1}\|^4+4\overline{C}_2^2\,\|\mathbf{V}_h^{n+1}\|^2\,\|\mathbf{U}^{n+1}\|_2^2\\
&\qquad+\frac{1}{\lambda_1}\,\|(\partial_t\mathbf{V}_h)^{n+1}\|^2+\frac{1}{\lambda_1}\,\|\widehat{\mathbf{g}}_h^{n+1}+\boldsymbol{\tau}_h^{n+1}\|^2\Big)\\
&\le\frac{4\|\widehat{\mathbf{g}}_h^{n+1}\|^2}{\lambda_1\overline{K}_1}+\frac{2}{\overline{K}_1}\Big(\overline{C}_1^2\overline{C}_3^4\,h^{4k}\,\|\mathbf{U}\|^4_{L^\infty(0,\infty;\mathbf{H}^{k+1}(\Omega))}\\
&\qquad\qquad+4\overline{C}_2^2\overline{C}_4^2\,h^{2k+2}\,\|\mathbf{U}\|^4_{L^\infty(0,\infty;\mathbf{H}^{k+1}(\Omega))}\\
&\qquad\qquad+\frac{1}{\lambda_1}\overline{C}_4^2\,h^{2k+2}\,\|\partial_t\mathbf{U}\|^2_{L^\infty(0,\infty;\mathbf{H}^{k+1}(\Omega))}\\
&\qquad\qquad+\frac{4(\Delta t)^2(\overline{C}_4^2+1)}{\lambda_1}\,\|\partial_{tt}\mathbf{U}\|^2_{L^\infty(0,\infty;\mathbf{H}^{1}(\Omega))}\Big)
\end{aligned}$$

so that

$$\|\widehat{\mathbf{v}}_h^{n+1}\|^2+\overline{K}_1\Delta t\,\|\nabla\widehat{\mathbf{v}}_h^{n+1}\|^2\le\|\widehat{\mathbf{v}}_h^{n}\|^2+\overline{K}_2h^{2k+2}(\Delta t)+2\overline{K}_3(\Delta t)^3+\frac{8\Delta t}{\lambda_1\overline{K}_1}\|\widehat{\mathbf{g}}^{n+1}\|^2,$$

where $\overline{K}_1$, $\overline{K}_2$, and $\overline{K}_3$ are defined by (4.20), (4.21), and (4.22), respectively. Using Lemma 4.1 we see that

$$\begin{aligned}
(\beta/2)\|\widehat{\mathbf{g}}_h^{n+1}\|^2&\le\mathcal{L}_h^{n+1}(\widehat{\mathbf{u}}_h^{n+1},\widehat{\mathbf{f}}_h^{n+1})\\
&\le\alpha\,\left(\|\widehat{\mathbf{v}}_h^{n}\|^2+\overline{K}_2h^{2k+2}\Delta t+\overline{K}_3(\Delta t)^3\right)+\alpha\overline{C}_4^2h^{2k+2}\,\|\mathbf{U}\|^2_{L^\infty(0,\infty;\mathbf{H}^{k+1}(\Omega))}.
\end{aligned}$$

By combining the last two inequalities we are led to

$$\begin{aligned}
&\|\widehat{\mathbf{v}}_h^{n+1}\|^2+\overline{K}_1\Delta t\,\|\nabla\widehat{\mathbf{v}}_h^{n+1}\|^2\\
&\le\|\widehat{\mathbf{v}}_h^{n}\|^2+\overline{K}_2h^{2k+2}(\Delta t)+2\overline{K}_3(\Delta t)^3\\
&\qquad+\frac{16\alpha\Delta t}{\beta\lambda_1\overline{K}_1}\left(\|\widehat{\mathbf{v}}_h^{n}\|^2+\overline{K}_2h^{2k+2}(\Delta t)+\overline{K}_3(\Delta t)^3\right)\\
&\qquad+\frac{16\alpha\overline{C}_4^2(\Delta t)h^{2k+2}}{\beta\lambda_1\overline{K}_1}\,\|\mathbf{U}\|^2_{L^\infty(0,\infty;\mathbf{H}^{k+1}(\Omega))}\qquad(4.32)\\
&\le\left[1+16\alpha(\beta\lambda_1\overline{K}_1)^{-1}\Delta t\right]\,\|\widehat{\mathbf{v}}_h^{n}\|^2+\left(2\overline{K}_3+\frac{16\alpha\overline{K}_3\Delta t}{\beta\lambda_1\overline{K}_1}\right)(\Delta t)^3\\
&\qquad+\left(\overline{K}_2+\frac{16\alpha\overline{K}_2\Delta t}{\beta\lambda_1\overline{K}_1}+\frac{16\alpha\overline{C}_4^2}{\beta\lambda_1\overline{K}_1}\,\|\mathbf{U}\|^2_{L^\infty(0,\infty;\mathbf{H}^{k+1}(\Omega))}\right)h^{2k+2}(\Delta t).
\end{aligned}$$

Using (A10), we can find a sufficiently small positive constant $\delta_0=\delta_0(\nu,\Omega,\mathbf{U})$ and a positive constant $\overline{K}_4=\overline{K}_4(\nu,\Omega,\mathbf{U})$ such that

$$0<\frac{\left[1+16\alpha(\beta\lambda_1\overline{K}_1)^{-1}\Delta t\right]}{1+\lambda_1\overline{K}_1\Delta t}\le(1-\overline{K}_4\Delta t)$$

for all $\Delta t \le \delta_0$. Defining

$$\begin{aligned}\overline{K}_5 &\stackrel{\text{def}}{=\!=} \overline{K}_5(\nu,\Omega,\mathbf{U})\\ &= \overline{K}_2 + \frac{16\alpha\overline{K}_2\delta_0}{\beta\lambda_1\overline{K}_1} + \frac{16\alpha\overline{C}_4^2}{\beta\lambda_1\overline{K}_1}\,\|\mathbf{U}\|^2_{L^\infty(0,\infty;\mathbf{H}^{k+1}(\Omega))}\end{aligned}$$

and

$$\overline{K}_6 \stackrel{\text{def}}{=\!=} \overline{K}_6(\nu,\Omega,\mathbf{U}) = 2\overline{K}_3 + \frac{16\alpha\overline{K}_3\delta_0}{\beta\lambda_1\overline{K}_1},$$

we obtain from (4.32) that

$$\|\widehat{\mathbf{v}}_h^{n+1}\|^2 + \overline{K}_1\Delta t\,\|\nabla\widehat{\mathbf{v}}_h^{n+1}\|^2 \le \|\widehat{\mathbf{v}}_h^n\|^2 + \overline{K}_5 h^{2k+2}(\Delta t) + \overline{K}_6(\Delta t)^3\,, \tag{4.33}$$

so by noting

$$\|\widehat{\mathbf{v}}_h^{n+1}\|^2 + \overline{K}_1\Delta t\,\|\nabla\widehat{\mathbf{v}}_h^{n+1}\|^2 \ge (1+\lambda_1\overline{K}_1\Delta t)\,\|\widehat{\mathbf{v}}_h^{n+1}\|^2\,,$$

we are led to

$$\|\widehat{\mathbf{v}}_h^{n+1}\|^2 \le (1-\overline{K}_4\Delta t)\|\widehat{\mathbf{v}}_h^n\|^2 + \overline{K}_5 h^{2k+2}(\Delta t) + \overline{K}_6(\Delta t)^3\,.$$

By induction, we may prove

$$\begin{aligned}\|\widehat{\mathbf{v}}_h^n\|^2 &\le (1-\overline{K}_4\Delta t)^n\|\widehat{\mathbf{v}}_h^0\|^2 + \left(\overline{K}_5 h^{2k+2}(\Delta t) + \overline{K}_6(\Delta t)^3\right)\sum_{j=0}^{n-1}(1-\overline{K}_4\Delta t)^j\\ &= (1-\overline{K}_4\Delta t)^n\|\widehat{\mathbf{u}}_{0,h} - \mathbf{U}_h^0\|^2 + \left(\frac{\overline{K}_5}{\overline{K}_4}h^{2k+2} + \frac{\overline{K}_6}{\overline{K}_4}(\Delta t)^2\right)\left[1-(1-\overline{K}_4\Delta t)^n\right]\\ &\le \left[(1-\overline{K}_4\Delta t)^{\frac{1}{\overline{K}_4\Delta t}}\right]^{\overline{K}_4 n\Delta t}\|\mathbf{u}_{0,h} - \mathbf{U}_h^0\|^2 \ + \frac{\overline{K}_5}{\overline{K}_4}h^{2k+2} + \frac{\overline{K}_6}{\overline{K}_4}(\Delta t)^2\,.\end{aligned}$$

Applying the inequality $1-y \le e^{-y}$ for all $y \ge 0$, we are led to

$$\|\widehat{\mathbf{v}}_h^n\|^2 \le e^{-\overline{K}_4 t_n}\|\mathbf{u}_{0,h} - \mathbf{U}_h^0\|^2 + \frac{\overline{K}_5}{\overline{K}_4}h^{2k+2} + \frac{\overline{K}_6}{\overline{K}_4}(\Delta t)^2\,.$$

Using the triangle inequality we obtain

$$\begin{aligned}\|\widehat{\mathbf{u}}_h^n - \mathbf{U}^n\|^2 &\le 2\|\widehat{\mathbf{u}}_h^n - \mathbf{U}_h^n\|^2 + 2\|\mathbf{U}_h^n - \mathbf{U}^n\|^2\\ &\le e^{-\overline{K}_4 t_n}\|\widehat{\mathbf{u}}_{0,h} - \mathbf{U}_h^0\|^2 + \frac{\overline{K}_5}{\overline{K}_4}h^{2k+2} + \frac{\overline{K}_6}{\overline{K}_4}(\Delta t)^2 + 2\overline{C}_4^2\,h^{2k+2}\,\|\mathbf{U}\|^2_{L^\infty(0,\infty;\mathbf{H}^2(\Omega))}\,.\end{aligned}$$

Using the triangle inequality again we also have

$$\begin{aligned}\|\mathbf{u}_{0,h} - \mathbf{U}_h^0\|^2 &\le 3\|\mathbf{u}_{0,h} - \mathbf{u}_0\|^2 + 3\|\mathbf{u}^0 - \mathbf{U}^0\|^2 + 3\|\mathbf{U}^0 - \mathbf{U}_h^0\|^2\\ &\le 3Ch^{2k+2}\|\mathbf{u}_0\|^2_{\mathbf{H}^{k+1}(\Omega))} + 3\|\widehat{\mathbf{u}}^0 - \mathbf{U}^0\|^2 + 3\overline{C}_4^2 h^{2k+2}\|\mathbf{U}\|^2_{L^\infty(0,\infty;\mathbf{H}^{k+1}(\Omega))}\,.\end{aligned}$$

By combining the last two inequalities we arrive at

$$\|\widehat{\mathbf{u}}_h^n - \mathbf{U}^n\|^2 \le 3e^{-\overline{\gamma}t_n}\|\widehat{\mathbf{u}}_0 - \mathbf{U}^0\|^2 + \overline{\kappa}\left[(\Delta t)^2 + h^{2k+2}\right],$$

where $\overline{\gamma} = \overline{K}_4$ and

$$\overline{\kappa} = \max\left\{\overline{K}_6/\overline{K}_4\,,\, 5\overline{C}_4^2\|\mathbf{U}\|^2_{L^\infty(0,\infty;\mathbf{H}^{k+1}(\Omega))} + 3C\|\mathbf{u}_0\|^2_{\mathbf{H}^{k+1}(\Omega))} + \overline{K}_5/\overline{K}_4\right\}. \qquad \square$$

As an easy consequence of Theorem 4.2, Theorem 2.7, and the inequality

$$\|\widehat{\mathbf{u}}(t_n) - \widehat{\mathbf{u}}_h^n\|^2 \le 2\|\widehat{\mathbf{u}}(t_n) - \mathbf{U}^n\|^2 + 2\|\mathbf{U}^n - \widehat{\mathbf{u}}_h^n\|^2,$$

we obtain the following estimate for the difference between the continuous and fully discrete solutions of the piecewise optimal control problem.

COROLLARY 4.3. *Assume that the hypotheses of Theorems* 4.2 *and* 2.7 *hold. Let* $\widehat{\mathbf{u}}$ *denote the global solution of the piecewise optimal control problem and* $\widehat{\mathbf{u}}_h^n$ *denote the fully discrete solutions of the piecewise optimal control problem. Then there are positive constants* $\overline{M} = \overline{M}(\nu, \Omega, \mathbf{U})$, $\overline{K}_7 = \overline{K}_7(\nu, \Omega, \mathbf{U})$, *and* $\overline{K}_8 = \overline{K}_8(\nu, \Omega, \mathbf{U})$ *such that*

$$\|\widehat{\mathbf{u}}(t_n) - \widehat{\mathbf{u}}_h^n\|^2 \le \overline{M}\, e^{-\overline{K}_7\, t_n}\|\widehat{\mathbf{u}}_0 - \mathbf{U}^0\|^2 + \overline{K}_8\left[(\Delta t)^2 + h^{2k+2}\right]. \qquad \square$$

We now estimate the dynamics in $\mathbf{H}^1(\Omega)$ norm between the fully discrete solution $\mathbf{u}_h^n$ and the desired flow $\mathbf{U}^n$ as $n \to \infty$.

THEOREM 4.4. *Assume that the hypotheses of Lemma* 4.1 *and Theorem* 4.2 *hold. Then there are positive constants* $\overline{K}_9 = \overline{K}_9(\nu, \Omega, \mathbf{U})$ *and* $\overline{\kappa}' = \overline{\kappa}'(\nu, \Omega, \mathbf{U})$ *such that for all* $h \le h_0$ *and* $\Delta t \le \delta_0$,

$$\|\nabla\widehat{\mathbf{u}}_h^n - \nabla\mathbf{U}^n\|^2 \le \frac{\overline{K}_9}{\Delta t}e^{-\overline{\gamma}t_n}\|\mathbf{u}_0 - \mathbf{U}^0\|^2 + \overline{\kappa}'\left((\Delta t) + \frac{h^{2k+2}}{\Delta t}\right), \tag{4.34}$$

where the constants h_0, δ_0, *and* $\overline{\gamma}$ *are as defined in the proof of Theorem* 4.2.

Proof. We use the same notations $(\widehat{\mathbf{u}}_h^{n+1}, \widehat{p}_h^{n+1}, \widehat{\mathbf{f}}_h^{n+1})$, $\widehat{\mathbf{U}}_h^{n+1}$, $\widehat{P}_h^{n+1}$, $(\partial_t\widehat{\mathbf{V}}_h)^{n+1}$, $\widehat{\mathbf{V}}_h^{n+1}$, $\widehat{\mathbf{v}}_h^{n+1}$, $\widehat{q}_h^{n+1}$, and $\widehat{\mathbf{g}}_h^{n+1}$ as in the proof of Theorem 4.2. From the proof of Theorem 4.2, in particular from (4.33), we have

$$\|\widehat{\mathbf{v}}_h^{n+1}\|^2 + \overline{K}_1\Delta t\,\|\nabla\widehat{\mathbf{v}}_h^{n+1}\|^2 \le \|\widehat{\mathbf{v}}_h^n\|^2 + \overline{K}_5 h^{2k+2}(\Delta t) + \overline{K}_6(\Delta t)^3,$$

where the various constants are as defined in the proof of Lemma 4.1 and Theorem 4.2. By dropping the term $\|\widehat{\mathbf{v}}_h^{n+1}\|^2$ and using Theorem 4.2 we obtain

$$\begin{aligned}
&\|\nabla\widehat{\mathbf{v}}_h^{n+1}\|^2 \\
&\le \frac{1}{\overline{K}_1\Delta t}\,\|\widehat{\mathbf{v}}_h^n\|^2 + \overline{K}_5 h^{2k+2} + \overline{K}_6(\Delta t)^2 \\
&\le \frac{1}{\overline{K}_1\Delta t}\left(3e^{-\overline{\gamma}t_n}\|\widehat{\mathbf{u}}_0 - \mathbf{U}^0\|^2 + \overline{\kappa}\left[(\Delta t)^2 + h^{2k+2}\right]\right) + \overline{K}_5 h^{2k+2} + \overline{K}_6(\Delta t)^2 \\
&\le \frac{\overline{K}_9}{\Delta t}e^{-\overline{\gamma}t_n}\|\widehat{\mathbf{u}}_0 - \mathbf{U}^0\|^2 + \overline{\kappa}'\left((\Delta t) + \frac{h^{2k+2}}{\Delta t}\right),
\end{aligned}$$

where $\overline{K}_9 \stackrel{\text{def}}{=\!=} 3/\overline{K}_1$ and $\overline{\kappa}' \stackrel{\text{def}}{=\!=} \max\{\overline{\kappa}, \overline{K}_5, \overline{K}_6\}$. $\square$

Remark 4.5. We note that whenever Δt is fixed, the first term on the right-hand side of (4.34) decays to zero exponentially as $n \to \infty$. If $h = \Delta t^s$ with $s > 0$, then the second term on the right-hand side of (4.34) is of order $O((\Delta t) + (\Delta t)^{2ks+2s-1})$, i.e.,

$$\limsup_{t\to\infty}\|\nabla\widehat{\mathbf{u}}_h^n - \nabla\mathbf{U}^n\| = O((\Delta t)^{1/2} + (\Delta t)^{ks+s-1/2}),$$

which is optimal when $h = (\Delta t)^{1/(k+1)}$. $\square$

Remark 4.6. Naturally one wishes to improve the estimate (4.34) to

$$\|\nabla\widehat{\mathbf{u}}_h^n - \nabla\mathbf{U}^n\|^2 \le \overline{K}_{10}e^{-\overline{\gamma}t_n}\|\mathbf{u}_0 - \mathbf{U}^0\|^2 + \overline{K}_{11}\left((\Delta t) + h^{2k+2}\right). \tag{4.35}$$

It does not appear that one can do so without major modifications in the statement of the problem. We mention one such modification that allows us to obtain the improved estimate (4.35); namely, if we consider the penalized functional

$$\mathcal{P}_h^{n+1}(\mathbf{u}_h^{n+1}, \mathbf{f}_h^{n+1}) \stackrel{\text{def}}{=\!=} \frac{\alpha}{2}\|\mathbf{u}_h^{n+1} - \mathbf{U}^{n+1}\|^2 + \frac{\beta}{2}\|\mathbf{f}_h^{n+1} - \mathbf{F}^{n+1}\|^2 + \frac{\gamma}{2}\left\|\frac{\mathbf{f}_h^{n+1} - \mathbf{f}_h^n}{\Delta t} - \frac{\mathbf{F}^{n+1} - \mathbf{F}^n}{\Delta t}\right\|^2$$

in place of $\mathcal{L}_h^{n+1}(\mathbf{u}_h^{n+1}, \mathbf{f}_h^{n+1})$ and the hypothesis $\partial_{tt}\mathbf{U} \in L^\infty(0,\infty;\mathbf{H}^{k+1}(\Omega)) \cap C([0,\infty);\mathbf{H}^{k+1}(\Omega))$ holds (in addition to the hypotheses of Theorem 4.4, of course), then the estimate (4.35) holds. An outline of a proof for this result is as follows. By virtue of the penalty term in the new functional $\mathcal{P}_h^{n+1}(\mathbf{u}_h^{n+1}, \mathbf{f}_h^{n+1})$, one can estimate the decay property of the discrete time derivative $(\widehat{\mathbf{v}}_h^{n+1} - \widehat{\mathbf{v}}_h^n)/\Delta t$ by taking the difference of the equations for $\widehat{\mathbf{v}}_h^{n+1}$ and $\widehat{\mathbf{v}}_h^n$ and setting $\mathbf{w}_h = (\widehat{\mathbf{v}}_h^{n+1} - \widehat{\mathbf{v}}_h^n)/\Delta t$. Then, (4.35) is proven by estimating $\nu\|\nabla\widehat{\mathbf{v}}_h^{n+1}\|^2$ in (4.30) with $\mathbf{w}_h = \widehat{\mathbf{v}}_h^{n+1}$. We omit the details of the proof because it is very lengthy and, more importantly, by introducing the penalized functional $\mathcal{P}_h^{n+1}(\mathbf{u}_h^{n+1}, \mathbf{f}_h^{n+1})$, we are too far removed from the original control objectives. □

Remark 4.7. In order to solve the $(n+1)$th fully discrete optimal control problem for each n, we need to introduce a Lagrange multiplier $(\widehat{\boldsymbol{\mu}}_h^{n+1}, \widehat{\pi}_h^{n+1})$ to convert the $(n+1)$th fully discrete optimal control problem into a discrete optimality system of equations (similar to the semidiscrete case). A solution for the $(n+1)$th fully discrete optimal control problem can be found by solving the discrete optimality system of equations which consists of (4.10), (4.11),

$$\begin{aligned}\frac{1}{\Delta t}(\boldsymbol{\mu}_h^{n+1}, \boldsymbol{\omega}_h) + a(\boldsymbol{\mu}_h^{n+1}, \boldsymbol{\omega}_h) + \overline{c}(\boldsymbol{\omega}_h, \mathbf{u}_h^{n+1}, \boldsymbol{\mu}_h^{n+1}) + \overline{c}(\mathbf{u}_h^{n+1}, \boldsymbol{\omega}_h, \boldsymbol{\mu}_h^{n+1}) \\ + b(\boldsymbol{\omega}_h, \pi_h^{n+1}) = \alpha(\mathbf{u}_h^{n+1} - \mathbf{U}^{n+1}, \boldsymbol{\omega}_h) \qquad \forall\, \boldsymbol{\omega}_h \in \mathbf{X}_h,\end{aligned} \tag{4.36}$$

$$b(\boldsymbol{\mu}_h^{n+1}, \tau_h) = 0 \qquad \forall\, \tau_h \in S_h, \tag{4.37}$$

and

$$(\boldsymbol{\mu}_h^{n+1} + \beta\mathbf{f}_h^{n+1} - \beta\mathbf{F}^{n+1}, \mathbf{z}_h) = 0 \qquad \forall\, \mathbf{z}_h \in \mathbf{X}_h\,. \tag{4.38}$$

Using the techniques of [GHS], we can show that the above discrete optimality system of equations indeed has a solution $(\widehat{\mathbf{u}}_h^{n+1}, \widehat{p}_h^{n+1}, \widehat{\mathbf{f}}_h^{n+1}, \widehat{\boldsymbol{\mu}}_h^{n+1}, \widehat{\pi}_h^{n+1})$. By eliminating $\mathbf{f}_h^{n+1}$, we arrive at the slightly simplified optimality system of equations for the $(n+1)$th fully discrete optimal control problem, which consists of

$$\begin{aligned}\frac{1}{\Delta t}(\mathbf{u}_h^{n+1}, \mathbf{w}_h) + a(\mathbf{u}_h^{n+1}, \mathbf{w}_h) + \overline{c}(\mathbf{u}_h^{n+1}, \mathbf{u}_h^{n+1}, \mathbf{w}_h) + b(\mathbf{w}_h, p_h^{n+1}) \\ = (\mathbf{F}_h^{n+1} - \beta^{-1}\boldsymbol{\mu}_h^{n+1}, \mathbf{w}_h) + \frac{1}{\Delta t}(\mathbf{u}_h^n, \mathbf{w}_h) \qquad \forall\, \mathbf{w}_h \in \mathbf{X}_h\,,\end{aligned} \tag{4.39}$$

(4.11), (4.36), and (4.37). □

5. An algorithm and a computational example. From the definition of the fully discrete piecewise optimal control problem and Remark 4.6, we summarize an algorithm for solving the fully discrete piecewise optimal control problem.

ALGORITHM 5.1.

- Choose a (sufficiently small) $\delta > 0$ and set $\Delta t = \delta$. Choose a (sufficiently small) h.
- Define $\mathbf{u}_h^0 = \mathbf{U}_h^0$, where $\mathbf{U}_h^0$ is the $\mathbf{L}^2(\Omega)$-projection (or interpolation) of $\mathbf{U}^0$ onto $\mathbf{X}_h$.
- (*Solving the* $(n+1)$*th fully discrete optimal control problem*)
  For $n = 0, 1, 2, \ldots$, find a $(\mathbf{u}_h^{n+1}, p_h^{n+1}, \boldsymbol{\mu}_h^{n+1}, \pi_h^{n+1}) \in \mathbf{X}_h \times S_h \times \mathbf{X}_h \times S_h$ such that

$$\frac{1}{\Delta t}(\mathbf{u}_h^{n+1}, \mathbf{w}_h) + a(\mathbf{u}_h^{n+1}, \mathbf{w}_h) + \overline{c}(\mathbf{u}_h^{n+1}, \mathbf{u}_h^{n+1}, \mathbf{w}_h) + b(\mathbf{w}_h, p_h^{n+1})$$
$$= (\mathbf{F}_h^{n+1} - \beta^{-1}\boldsymbol{\mu}_h^{n+1}, \mathbf{w}_h) + \frac{1}{\Delta t}(\mathbf{u}_h^n, \mathbf{w}_h) \qquad \forall\, \mathbf{w}_h \in \mathbf{X}_h\,,$$

$$b(\mathbf{u}_h^{n+1}, r_h) = 0 \qquad \forall\, r_h \in S_h\,,$$

$$\frac{1}{\Delta t}(\boldsymbol{\mu}_h^{n+1}, \boldsymbol{\omega}_h) + a(\boldsymbol{\mu}_h^{n+1}, \boldsymbol{\omega}_h) + \overline{c}(\boldsymbol{\omega}_h, \mathbf{u}_h^{n+1}, \boldsymbol{\mu}_h^{n+1}) + \overline{c}(\mathbf{u}_h^{n+1}, \boldsymbol{\omega}_h, \boldsymbol{\mu}_h^{n+1})$$
$$+ b(\boldsymbol{\omega}_h, \pi_h^{n+1}) = \alpha(\mathbf{u}_h^{n+1} - \mathbf{U}^{n+1}, \boldsymbol{\omega}_h) \qquad \forall\, \boldsymbol{\omega}_h \in \mathbf{X}_h,$$

  and
$$b(\boldsymbol{\mu}_h^{n+1}, \tau_h) = 0 \qquad \forall\, \tau_h \in S_h\,.$$

- Set $\mathbf{f}_h^{n+1} = \mathbf{F}_h^{n+1} - \beta^{-1}\boldsymbol{\mu}_h^{n+1}$.

We now report some computational results for solving the piecewise optimal control problem by implementing Algorithm 5.1. This example demonstrates that the piecewise optimal control mechanism does a very good job of tracking the velocity field. Also, in the solution process an optimal body force distribution can be obtained. The computational example also reinforces the theoretical results.

Here are some detailed data of the example. We choose the domain $\Omega = (0,1) \times (0,1)$ (i.e., the unit square). The desired velocity field

$$\mathbf{U}(\mathbf{x}, t) = \begin{pmatrix} \varphi_y(x_1, x_2, t) \\ -\varphi_x(x_1, x_2, t) \end{pmatrix}$$

is constructed from the stream function

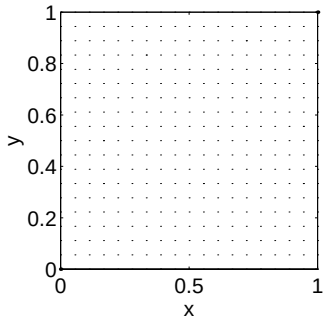
$$\varphi(x_1, x_2, t) = \theta(x_1, t)\theta(x_2, t)$$

with

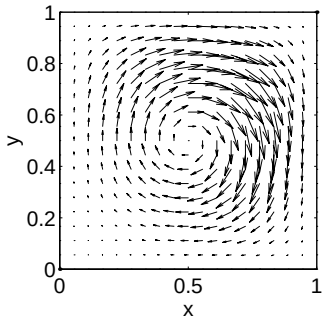
$$\theta(y, t) = (1-y)^2\,(1 - \cos(2k\pi y t))\,, \qquad y \in [0, 1].$$

The integer parameter k ($k = 2$ for the current computation) involved in $\mathbf{U}$ adjusts the number of eddies of circulation presented in the desired flow, thus determines the complexity of the desired flow. Note that $\mathbf{U}$ defined in such a way satisfies the divergence-free condition and the zero boundary conditions. The desired body force is computed by (1.16), namely,

$$\mathbf{F} = \mathbf{u}_t - \nu\Delta\mathbf{U} + (\mathbf{U} \cdot \nabla)\mathbf{U}.$$

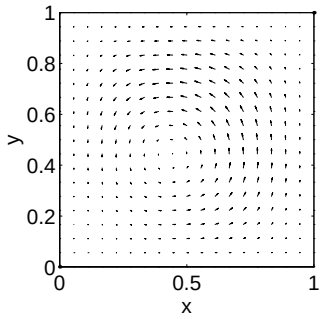


Desired flow: $t = 0$.

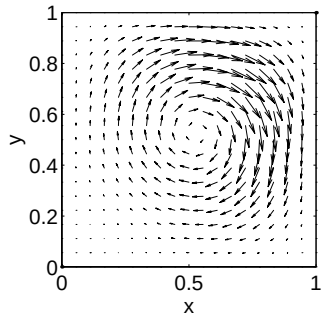


Controlled flow: $t = 0$.

FIG. 1.



Desired flow: $t = 0.3$.



Controlled flow: $t = 0.3$.

FIG. 2.

The uncontrolled initial state is

$$\mathbf{u}_0(\mathbf{x}) = e \begin{pmatrix} (\cos(2\pi x_1) - 1)\sin(2\pi x_2) \\ \sin(2\pi x_1)(1 - \cos(2\pi x_2)) \end{pmatrix}.$$

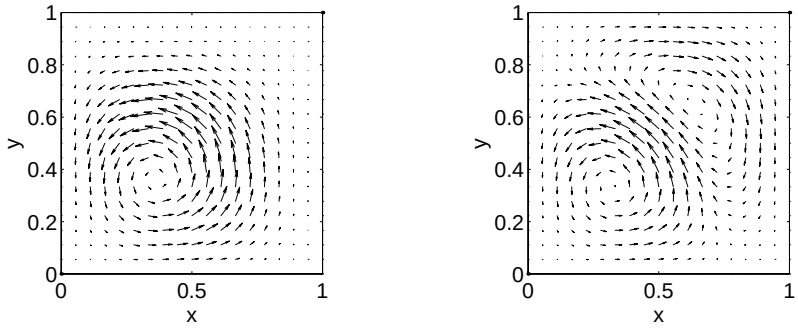
We choose the viscosity constant $\nu = 0.1$ and the time step $\Delta t = \delta = 0.1$ for the computation. The finite elements are chosen to be the Taylor–Hood elements; i.e., the finite element space $\mathbf{X}_h$ is chosen to be piecewise quadratic elements (for $\mathbf{u}_h$ and $\boldsymbol{\mu}_h$) and S_h is chosen to be piecewise linear elements (p_h and π_h). Thus we have $k = 2$ in (4.1), (4.2). The finite-dimensional nonlinear optimality system of equations in Algorithm 5.1 needs to be solved by a nonlinear solver. Newton's method is used to solve the finite-dimensional nonlinear system of equations. Note that our finite element spaces are nonconforming in the sense that the discrete divergence-free condition $b(\mathbf{u}_h^n, r_h) = 0$ does not imply the continuous divergence-free condition, i.e., $\nabla \cdot \mathbf{u}_h^n \neq 0$.

The computational results obtained by implementing Algorithm 5.1 with the above data are presented in graphical form with Figures 1–10.

For better graphics purposes, we plot

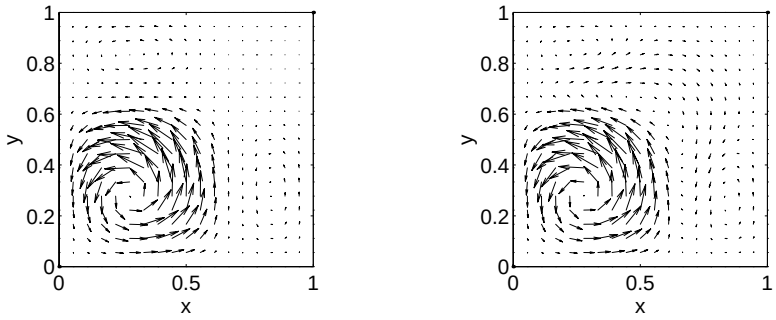
$$\sigma(1 + 100(x_1^2 + x_2^2))\mathbf{u}(\mathbf{x}, t),$$

instead of $\mathbf{u}$, where $\sigma > 0$ is a scaling factor. The purpose of the scaling is to magnify the small structure of the desired flow $\mathbf{U}(\mathbf{x}, t)$ close to the upper right corner of the physical domain Ω.



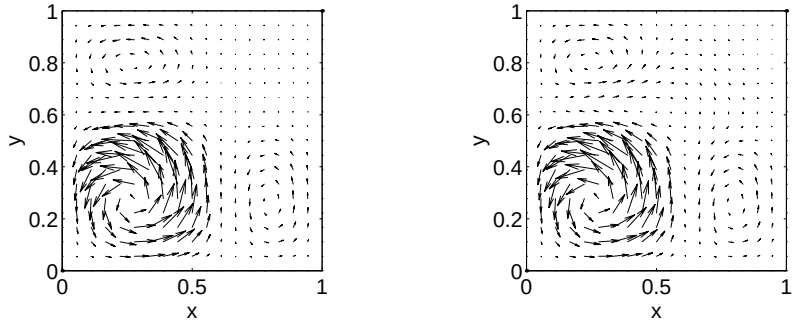
Desired flow: $t = 0.5$. *Controlled flow:* $t = 0.5$.

FIG. 3.



Desired flow: $t = 0.8$. *Controlled flow:* $t = 0.8$.

FIG. 4.



Desired flow: $t = 0.9$. *Controlled flow:* $t = 0.9$.

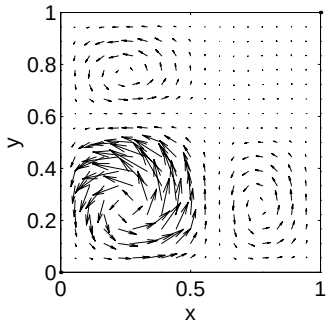
FIG. 5.

Figure 1 shows the initial states $\mathbf{U}_0$ and $\mathbf{u}_0$ of the desired and controlled flows with $\sigma = 0.0005$. Note that $\mathbf{u}_0$ is far away from the zero vector field $\mathbf{U}_0$.

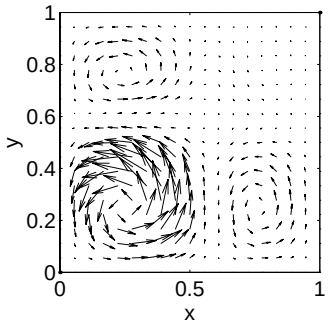
When $t > 0$ becomes larger, the magnitude of the desired flow $\mathbf{U}(t)$ increases but that of the controlled flow $\mathbf{u}(t)$ decreases significantly at the first few steps.

Figures 2–9 are plotted with $\sigma = 0.002$.

During $0 \le t \le 0.7$, the control is in the transient stage. At $t = 0.3$ (Figure 2), neither the position of the eddy nor the magnitude of the velocity field $\mathbf{u}$ matches well with $\mathbf{U}$. At $t = 0.5$, the large eddy of the controlled flow $\mathbf{u}$ splits into two

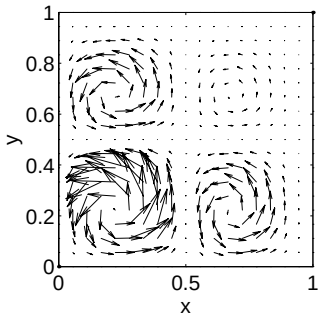


Desired flow: $t = 1.0$.

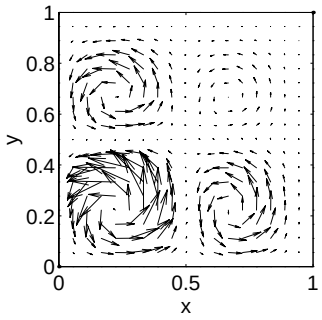


Controlled flow: $t = 1.0$.

FIG. 6.

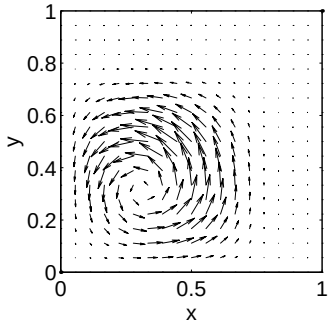


Desired flow: $t = 1.5$.

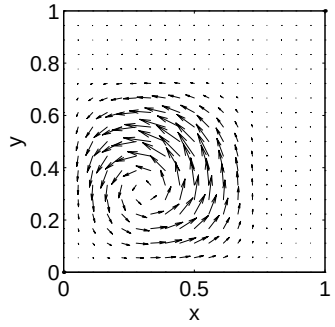


Controlled flow: $t = 1.5$.

FIG. 7.



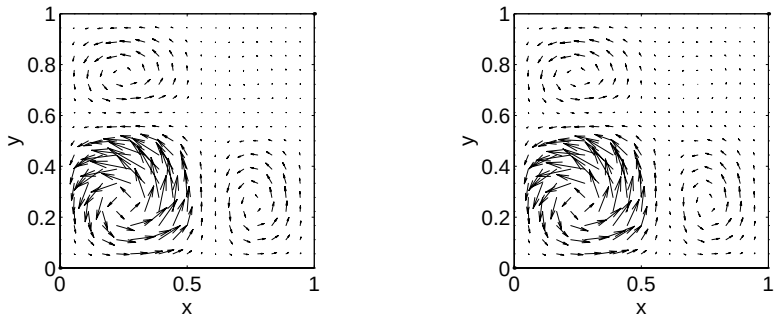
Desired flow: $t = 3.8$.



Controlled flow: $t = 3.8$.

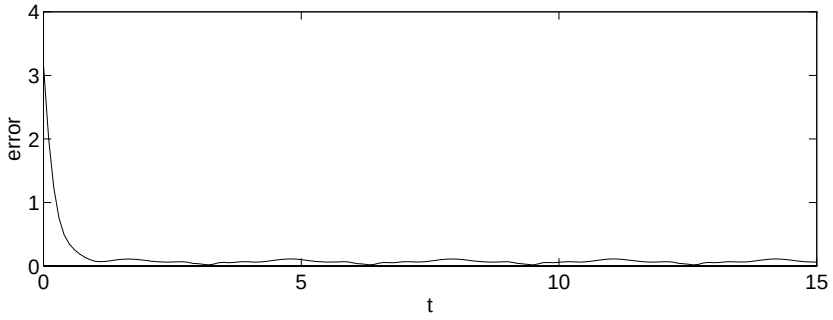
FIG. 8.

eddies, with one at the right position and the other fading as t increases. Starting from $t = 0.8$, the controlled flow matches the desired flow so well that we can hardly distinguish one from the other with naked eyes (Figures 4–6 for $t = 0.8, 0.9, 1.0$). Of course, $\mathbf{u}$ is not identical to $\mathbf{U}$ at this stage, but the difference is at a very small scale. The controlled flow fine-tunes itself to match the desired flow quantitatively. As the distributed control continues to be applied, "perfect matching" (to our naked eyes) is preserved. Some pictures for long time control are included in Figures 7–9 for $t = 1.5, 3.8, 14.7$.



Desired flow: $t = 14.7$. *Controlled flow:* $t = 14.7$.

FIG. 9.



Error reduction in t.

FIG. 10.

In summary, in very short time (starting from $t \sim 0.8$), the controlled velocity field is in nice agreement with the target field. These results reinforce the theoretical exponential decay estimate of $\|\mathbf{u}_h^n - \mathbf{U}^n\|$ (see Corollary 4.3).

On the other hand, in Figure 10, we see that the error $\|\mathbf{u}_h^n - \mathbf{U}^n\|$ is quickly reduced initially and then oscillates between 0 and 0.11 rather than being further reduced. This is due to the discretization error $O(h^3 + \Delta t)$ in Corollary 4.3. By choosing smaller h and Δt, we can reduce the eventual error in the velocity tracking. For more computational details, see [HRY].

Acknowledgments. The authors wish to thank Professor George R. Sell of the University of Minnesota for supporting the access to supercomputer resources. The authors also wish to thank Minnesota Supercomputer Center for granting Cray computer time. All computations in this paper were carried out on the CrayC90 there.

REFERENCES

[Ad] R. ADAMS, *Sobolev Spaces*, Academic Press, New York, 1975.

[AT] F. ABERGEL AND R. TEMAM, *On some control problems in fluid mechanics*, Theoret. Comput. Fluid Dynamics, 1 (1990) pp. 303–325.

[CTMK] H. CHOI, R. TEMAM, P. MOIN, AND J. KIM, *Feedback control for unsteady flow and its application to the stochastic Burgers equation*, J. Fluid Mech., 253 (1993), pp. 509–543.

[Ci] P. Ciarlet, *Finite Element Methods for Elliptic Problems*, North-Holland, Amsterdam, 1978.

[CF] P. Constantin and C. Foias, *Navier-Stokes Equations*, University of Chicago, Chicago, 1988.

[Fu1] A. Fursikov, *On some control problems and results concerning the unique solvability of a mixed boundary value problems for the three-dimensional Navier-Stokes and Euler systems*, Soviet Math. Dokl., 3 (1980), pp. 889–893.

[Fu2] A. Fursikov, *Control problems and theorems concerning the unique solvability of a mixed boundary value problems for the three-dimensional Navier-Stokes and Euler equations*, Math USSR Sb., 43 (1982), pp. 281–307.

[Fu3] A. Fursikov, *Properties of solutions of some extremal problems connected with the Navier-Stokes system*, Math USSR Sb., 46 (1983), pp. 323–351.

[FGH] A. Fursicov, M. Gunzburger, and L. Hou, *Boundary value problems and optimal boundary control for the Navier-Stokes systems: The two-dimensional case*, SIAM J. Control Optim., 36 (1998), to appear.

[GH] M. Gunzburger and L. Hou, *Treating inhomogeneous essential boundary conditions in finite element methods and the calculation of boundary stresses*, SIAM J. Numer. Anal., 29 (1992), pp. 390–424.

[GHS] M. Gunzburger, L. Hou, and T. Svobodny, *Analysis and finite element approximation of optimal control problems for the stationary Navier-Stokes equations with distributed and Neumann controls*, Math. Comp., 57 (1991), pp. 123–151.

[GR] V. Girault and P. Raviart, *Finite Element Methods for Navier-Stokes Equations*, Springer-Verlag, Berlin, 1986.

[HR] J. Heywood and R. Ranacker, *Finite element approximations of the nonstationary Navier-Stokes problem,* IV*, Error analysis for second order time discretization*, SIAM J. Numer. Anal., 27 (1990), pp. 353–384.

[HRY] L. Hou, S. Ravindrin, and Y. Yan, *Numerical solutions of optimal distributed control problems for incompressible flows*, Internat. J. Comput. Fluid Dynamics, 8 (1997), pp. 99–114.

[HY] L. Hou and Y. Yan, *Dynamics for controlled Navier-Stokes systems with distributed control*, SIAM J. Control Optim., 35 (1997), pp. 654–677.

[Li] J.-L. Lions, *Control of Distributed Singular Systems*, Bordas, Paris, 1985.

[Ra] S.S. Ravindran, *A Computational Study of Some Control Problems in Electrically Conducting Flows*, Ph.D. thesis, Simon Fraser University, Burnaby, BC, Canada, 1994.

[Te] R. Temam, *Navier-Stokes Equations, Theory and Numerical Methods*, North-Holland, Amsterdam, 1983.

SIAM J. CONTROL OPTIM.
Vol. 35, No. 6, pp. 1886–1907, November 1997

003

ANNEALING OF ITERATIVE STOCHASTIC SCHEMES*

HAITAO FANG†, GUANGLU GONG‡, AND MINPING QIAN†

Abstract. Annealing of general iterative stochastic schemes is studied by using Wentzell's large deviation theory. The convergence of this algorithm is shown under some conditions. The closed form of the critical constant is specified in terms of its potential function of the mean model.

Key words. global minimum, iterative stochastic schemes, large deviation, potential function, simulated annealing

AMS subject classifications. 60J20, 93E35, 93E99

PII. S0363012995293670

1. Introduction.

Adaptive algorithms with stochastics appear frequently in various applications, such as learning algorithms [17], optimization (simulated annealing [16], genetic algorithms [9], evolution computation [15]), neural networks [12], system identification, adaptive controls, signal modeling, filtering, and transmission [12], [21]. The function of these algorithms is to adjust a state vector (or monitored parameter vector) X_n for specifying the system considered, where n refers to the time of observation of the system. In most cases of applications mentioned above, the rule used to update $X.$ will typically be of the form

$$X_{n+1} = X_n + r_n b\left(X_n, \xi_n\right), \tag{1.1}$$

where $\{r_n\}$ is a sequence of small gains and ξ_n is the input of the system at time n, either deterministic or stochastic. These algorithms have been studied (see [1] and references therein). It is shown that they do not always converge to the desired result. For instance, when a cost (or, say, potential) function $U(x)$ is given, the algorithm will be constructed by the method in [1]; however, it isn't always successful in finding an element in the set $\underline{S} = \{z \in R^d,\, U(z) = \min_{x \in R^d} U(x)\}$. To avoid getting trapped in local minima, stochastic perturbation is added as follows:

$$X_{n+1} = X_n + r_n b\left(X_n, \xi_n\right) + h_n \zeta_n, \tag{1.2}$$

where ζ_n is a sequence of independently and identically distributed (i.i.d.) random variables. Borrowing the idea of annealing process in statistical mechanics, it is called the simulated annealing algorithm [16].

In various special cases, the asymptotic behavior of (1.2) as $n \to \infty$ has been studied in [18], [6], [7], [19], [20]. In [18], the global optimization result is achieved for the case of $b\left(X_n, \xi_n\right) = -\nabla U\left(X_n\right)$, $h_n = r_n$ and $\zeta_n = W_n$ being i.i.d. Gaussian random variables. In [19], using the result of Dubrushin for inhomogeneous Markov chains, the convergence result is achieved, in which the appropriate choice of $\{r_n\}$ depends on the distribution of ξ_n. One of the cases which can be represented explicitly is that ξ_n is also Gaussian; in this case, $\{r_n\}$ is the solution of the equation $cr_n / \ln r_n =$

*Received by the editors October 20, 1995; accepted for publication (in revised form) August 1, 1996. This research was supported by CNSF, Doctoral Program Foundation of the Institute of Higher Education, and Project 863 of CN.

http://www.siam.org/journals/sicon/35-6/29367.html

†Department of Probability and Statistics and Center of Nonlinear Science, Peking University, Beijing, China (htfang@iss03.iss.ac.cn, qianmp@sxx0.math.pku.edu.cn).

‡Department of Applied Mathematics, Tsinghua University, Beijing, China.

$1/\ln n$ and $h_n = -1/\ln r_n$ with a large constant c. The model with $\zeta_n = W_n$, $b(X_n, \xi_n) = -\nabla U(x) + \xi_n$ and $r_n = A/n$, $h_n = \sqrt{B/(n \ln\ln n)}$ is studied in [6], [7], and based on the result of [3], the following result is obtained: If $U(x)$, $\{\xi_n\}$ are suitably behaved and $B/A > \Lambda$ (a constant depending only on $U(x)$), then $X_n \to \underline{S}$ in probability as $n \to +\infty$.

Similar to (1.2), the stochastic differential equation in continuous time

$$dX_t = -\nabla U(X_t)\, dt + c_t dW_t \tag{1.3}$$

is studied, which is a rather familiar model in probability theory called a Langevin-type Markov diffusion. Especially, [14] gives a comprehensive observation of the large time behavior of (1.3), where Hwang and Sheu's main result can be roughly stated as follows: Under some mild conditions on $U(x)$ and $c_t = \sqrt{c/\ln t}$ for large t, there exist Λ, $d^* > 0$ such that, when $c > \Lambda$, for $f \in C_b(R^d)$,

$$E_{0,y}(f(X_t)) \to \int f(x)\, \mu(dx) \qquad \text{as } t \to +\infty$$

uniformly for y in a compact set, where $\mu(dx)$ is the weak limit of the measures $\frac{1}{Z_\epsilon} \exp\{-\frac{U(x)}{\epsilon}\}dx$ as $\epsilon \to 0 (Z_\epsilon = \int \exp\{-\frac{U(x)}{\epsilon}\}dx)$; and when $c > d^*$,

$$P_{0,y}\{X_t \in B(\underline{S},\ l)\} \to 1 \qquad \text{as } t \to +\infty \tag{1.4}$$

uniformly for y in any compact set, where $B(\underline{S},\ l)$ is the l-neighborhood of $\underline{S}$, while for $c < d^*$, (1.4) fails. The proof of these results is based on the large deviation estimates of small Gaussian perturbations of the dynamical system [5].

The model of (1.2) in such a general form studied in the present article provides more flexibility for application. In fact, in many cases, the mechanism of systems is not as simple as those in [18], [19], where the increments of X at time n are only sums of ξ_n and a function of X_n. It often appears as a complicated multivariable function of ξ_n and X_n, e.g., in Kohonen's algorithm and many other algorithms. Hence a consideration for model (1.2) is not only of theoretical generality but also of practical significance. The work of [6], [7] is based on this general form, while $\{r_n\}$ and $\{h_n\}$ therein are too restricted and there is a lack of information on the convergence rate of the algorithm. To obtain the corresponding results of asymptotic behavior in [14] for the model (1.2), the methods used in [6], [7], [19], [14], etc., seem not to be enough. Based on the idea of using large deviations theory to study simulated annealing for a finite number of states in [24], we borrow the generalized large deviation theory of Wentzell [25] and then give a uniform treatment and convergence theorem which includes results in [6], [7], [19].

In [14], the critical constants Λ, $d^*(c^*,\ d^*$ in [14]) are specified by a sequence of definitions connected with cycles, which become rather complicated to reveal their meaning through $U(x)$. In [13], Λ is identified by

$$\Lambda = 2 \sup_{x,y \in R^d} \inf_{\varphi} \sup_{0 \le t \le 1} \left(U(\varphi(t)) - U(x) - U(y) + U_0\right),$$

where φ belongs to the set of continuous routes which link x to y.

In the present article, under the conditions

(1) $b(x, y)$ is a bounded measurable function such that $\bar{b}(x) = Eb(x, \xi_n)$ is uniformly continuous,

(2) ζ_n is an i.i.d. random sequence and $E\zeta_n = 0$,

(3) $\{r_n\}$ and $\{h_n\}$ satisfy assumptions (A1)–(A3) in section 2 and (A4) (or (A4′), (A4″)) in section 3 (we will show that these assumptions do include a fairly wide class of $\{r_n\}$ and $\{h_n\}$),

the large deviation principle for the family of $\{X_n\}$ as $n \to +\infty$ is observed in section 3, which is used mainly as a tool to reach our goal.

The main result of this article is exploited in section 5: If $\bar{b}(x)$ has a potential function $U(x)$, then under some restrictions on the behavior of $\bar{b}(x)$ and $b(x, \xi_n)$ at infinity ((B1)–(B3)), for $\{r_n\}$ in a wide class, one can always choose $\{h_n\}$ such that

$$P_{0,y}\{X_n \in B(\underline{S}, l)\} \to 1 \qquad \text{as } n \to +\infty \tag{1.5}$$

uniformly for y in an arbitrary compact set F. Moreover, for the convergence of X_n in the sense of (1.5), h_n has to be chosen decreasing slowly enough. How slow this is depends on the order of r_n. For instance, when $r_n = 1/n^\gamma + o(1/n^\gamma)(\gamma < 1)$, the critical rate of h_n is $c/\left((1-\gamma) n^{\gamma/2} \ln n\right)$ for c large enough. Furthermore, we identify the closed form of the lower bound d^* of c for convergence:

$$d^* = 2 \sup_{x \in S \setminus \underline{S}} \inf_{y \in \underline{S}} \inf_{\varphi} \sup_{0 \le t \le 1} (U(\varphi(t)) - U(x)) \vee 0,$$

which is analogous to that obtained for finite state processes in [11], where $S = \left\{x \in R^d, \nabla U(x) = 0\right\}$ and $\varphi \in C\left([0, 1], R^d\right)$, $\varphi(0) = x$, $\varphi(1) = y$. We also show that the convergence rate of the algorithm is not the same for different $\{r_n\}$ and $\{h_n\}$. Moreover, changing $\{r_n\}$, $\{h_n\}$ given in [6], [7] will accelerate the convergence speed of the algorithm.

The materials in the appendix are devoted to exploiting the complicated definition of cycles.

2. Preliminaries. We consider the following stochastic approximation model:

$$X_{n+1} = X_n + r_n b(X_n, \xi_n) + \sqrt{r_n h_n}\zeta_n, \tag{2.1}$$

where $\{r_n\}$ is a sequence of small gains monotonically decreasing to zero, satisfying $\sum_n r_n = +\infty$, and $\{h_n\}$ also tends to zero monotonically. We also assume that $\{\xi_n\}$ and $\{\zeta_n\}$ are i.i.d. random vectors respectively, and $\{\xi_n\}$ and $\{\zeta_n\}$ are mutually independent.

Let us make some assumptions on $\{r_n\}$ and $\{h_n\}$. Define t_n and $m(t)$ by

$$t_n = \sum_{i=1}^{n} r_i,$$

$$m(t) = \inf\{n > 0, t_n > t\}.$$

Assume that there are continuous functions $k(s)$ and $t(s)$ such that

$$\frac{r_{m(t_N+s)}}{r_N} \to k(s) \qquad \text{as } N \to +\infty, \tag{A1}$$

$$\frac{h_{m(t_N+s)}}{h_N} \to 1 \qquad \text{as } N \to +\infty, \tag{A2}$$

$$k\left(t_{[s/r_N+N]} - t_N\right) \to q(s) \qquad \text{as } N \to +\infty, \tag{A3}$$

where $[\cdot]$ is the function of taking the integer part.

(A1)–(A3) are too abstract for application. The following propositions and corollary show that in most cases we encounter, these assumptions are satisfied.

PROPOSITION 2.1. *If* $\{r_n\}$ *satisfies following conditions*

(1) *there exists* $\epsilon > 0$, *such that* $\chi \cdot r_{[\chi n]} \geq r_n$ *for large enough* n *and any* $\chi \in (1, 1+\epsilon)$,

(2) $\liminf_{n\to+\infty} r_n \frac{n}{(\ln n)^\alpha} > 0$ *for some* $\alpha > 0$,

then the assumption (A1) *holds and* $k(s) = 1$, *which implies assumption* (A3) *automatically.*

Proof. For the case of $r_n = (\ln n)^\alpha / n$, we have $t_n \sim (\ln n)^{1+\alpha}$. Consider

$$(\ln(n + n/\ln\ln n))^{1+\alpha} - (\ln n)^{1+\alpha},$$

which is equal to

$$\left(\ln n + \ln\left(1 + \frac{1}{\ln\ln n}\right)\right)^{1+\alpha} - (\ln n)^{1+\alpha}$$

$$= (\ln n)^{1+\alpha}\left(1 + \frac{1+\alpha}{\ln n \ln\ln n} + o\left(\frac{1}{\ln n \ln\ln n}\right)\right) - (\ln n)^{1+\alpha} \to \infty, \qquad n \to \infty. \tag{2.2}$$

For any $s > 0$, there exists N such that $r_{m(t_n+s)} \geq r_{n+n/\ln\ln n}$ for $n > N$. Hence

$$1 \geq \limsup_{n\to+\infty} \frac{r_{m(t_n+s)}}{r_n} \geq \liminf_{n\to+\infty} \frac{r_{m(t_n+s)}}{r_n} \geq \liminf_{n\to+\infty} \frac{\left(\ln\left(n + \frac{n}{\ln\ln n}\right)\right)^\alpha}{n + \frac{n}{\ln\ln n}} \frac{n}{(\ln n)^\alpha} = 1, \tag{2.3}$$

i.e.,

$$k(s) = \lim_{n\to+\infty} \frac{r_{m(t_n+s)}}{r_n} = 1.$$

For general $\{r_n\}$ satisfying (1) and (2), there exist N and $c' > 0$ such that for $n > N$,

$$t_{[n+n/\ln\ln n]} - t_n \geq c'\left(\left(\ln\left(n + \frac{n}{\ln\ln n}\right)\right)^{1+\alpha} - (\ln n)^{1+\alpha}\right) \to +\infty, \qquad n \to +\infty,$$

holds. Similarly to (2.3), we obtain

$$1 \geq \limsup_{n\to+\infty} \frac{r_{m(t_n+s)}}{r_n} \geq \liminf_{n\to+\infty} \frac{r_{m(t_n+s)}}{r_n} \geq \liminf_{n\to+\infty} \frac{r_{\left[n+\frac{n}{\ln\ln n}\right]}}{r_n},$$

which implies $k(s) = 1$, since by (1) we have

$$\frac{r_{\left[n+\frac{n}{\ln\ln n}\right]}}{r_n} \geq \frac{r_n}{\left(1 + \frac{1}{\ln\ln n}\right) r_n} \to 1, \qquad n \to +\infty.$$

The proof is complete. □

Remark. $r_n = c/(\ln\ln n)^\alpha (\alpha > 0)$, $r_n = c/(\ln n)^\alpha (\alpha > 0)$, $r_n = c(\ln n)^\beta / n^\alpha (0 < \alpha < 1, \beta \in (-\infty, +\infty))$, and $r_n = c(\ln n)^\alpha / n (\alpha > 0)$ are typical examples satisfying the conditions of Proposition 2.1.

COROLLARY 2.2. *If* $r_n = \hat{r}_n + o(\hat{r}_n)$, *where* $\{\hat{r}_n\}$ *satisfies the conditions of Proposition* 2.1, *then the results of Proposition* 2.1 *hold.*

Proof. The proof is immediate. □

PROPOSITION 2.3. *If* $r_n = c/n + o(1/n)$, *then assumptions* (A1) *and* (A3) *are satisfied with* $k(s) = e^{-s/c}$ *and* $q(s) = c/(s + c)$.

Proof. For any $\epsilon > 0$, there exists N such that for $n > N$,

$$(c-\epsilon)\ln\frac{n+1}{N+1} \le t_n - t_N \le (c+\epsilon)\ln\frac{n}{N}. \tag{2.4}$$

Thus

$$\left[e^{s/(c+\epsilon)}N\right] \le m(t_N+s) \le \left[e^{s/(c-\epsilon)}N+1\right].$$

It follows that

$$\liminf_{n\to+\infty}\frac{r_{m(t_N+s)}}{r_N} \ge e^{-s/(c-\epsilon)} \quad \text{and} \quad \limsup_{n\to+\infty}\frac{r_{m(t_N+s)}}{r_N} \le e^{-s/(c+\epsilon)}$$

for any ϵ and

$$\frac{r_{m(t_N+s)}}{r_N} \to e^{-s/c}, \qquad N \to +\infty.$$

Then we have

$$\begin{aligned}\left(\frac{N}{[sN/(c+o(1))+N]}\right)^{c/(c+\epsilon)} &\le k\left(t_{[sN/(c+o(1))+N]} - t_N\right)\\ &\le \left(\frac{N+1}{[sN/(c+o(1))+N]+1}\right)^{c/(c-\epsilon)}\end{aligned}$$

by (2.4), which shows

$$k\left(t_{[s/r_N+N]} - t_N\right) \to \frac{c}{s+c} \qquad \text{as } N\to+\infty \text{ and } \epsilon\to 0, \tag{2.5}$$

i.e., $q(s) = c/(s+c)$. The proposition is proven. □

Remark. If r_n is given as Corollary 2.2, then assumption (A2) on $\{h_n\}$ can be also specialized as conditions of Corollary 2.2, and if $r_n = c/n + o(1/n)$, then for (A2) h_n can be taken as $c/(\ln\ln n)^\alpha (\alpha > 0)$, or $c/(\ln n)^\alpha (\alpha > 0)$.

3. Large deviation principle. In order to obtain the convergence properties of the model (2.1) under different choices of $\{r_n\}$ and $\{h_n\}$, we investigate the Freidlin–Wentzell-type large deviation law of the following model:

$$X_{n+1}^N = \begin{cases} X_n^N + r_{N+n}b(X_n^N,\xi_{N+n}) + \sqrt{r_{N+n}h_{N+n}}\zeta_{N+n} & \text{if } n < G(N),\\ X_n^N + r_N q(n\cdot r_N)b(X_n^N,\xi_{N+n}) + \sqrt{r_N h_N q(n\cdot r_N)}\zeta_{N+n} & \text{otherwise,}\end{cases} \tag{3.1}$$

where we take $G(N)$ such that $G(N)\cdot r_N \to +\infty$ and let it make

$$\frac{\tilde r_{[s/r_N]}^N}{r_N} \to q(s) \qquad \text{uniformly in } s \text{ as } N\to+\infty \tag{3.2}$$

and

$$\frac{\tilde h_n^N}{h_N} \to 1 \qquad \text{uniformly in } n \text{ as } N\to+\infty, \tag{3.3}$$

where

$$\tilde r_n^N := \begin{cases} r_{n+N} & \text{if } n < G(N),\\ r_N q(n\cdot r_N) & \text{otherwise}\end{cases} \tag{3.4}$$

and

$$\tilde{h}_n^N := \begin{cases} h_{n+N} & \text{if } n < G(N), \\ h_N & \text{otherwise.} \end{cases} \tag{3.5}$$

In the cases of $\{r_n\}$ and $\{h_n\}$ satisfying the conditions of Corollary 2.2, we can take $G(N) = N/\ln\ln N$. When $r_n = c/n + o(1/n)$ and $h_n \approx d/\ln\ln n$, $G(N)$ may be N^k for any $k > 1$.

Remark. The important role of $G(N)$ is clear if one realizes that for $n < G(N)$, we have $X_{n+N} = X_n^N$, which is crucial for our argument in order to make use of Wentzell's powerful estimation of the large deviation principle [25].

Let $\mu(du)$ be the probability measure of ζ_1 and $\nu(dv)$ be the probability measure of ξ_1. We denote

$$A_{ij} = \int_{R^n} u^i u^j \mu(du), \tag{3.6}$$

$$H_0(s,x,y) = \frac{1}{2}\sum_{i,j=1}^{n} (q(s))^{-1} A^{ij} \left(y^i - q(s)\bar{b}^i(x)\right) \cdot \left(y^j - q(s)\bar{b}^j(x)\right), \tag{3.7}$$

$$S_{0,T}(\phi) = \int_0^T H_0\left(s, \phi(s); \dot{\phi}(s)\right) ds, \tag{3.8}$$

where $\left(A^{ij}\right)$ is the inverse of the matrix (A_{ij}) and $\bar{b}(x) = \int_{R^n} b(x,v)\nu(dv)$.

For later use, we cite the following result.

LEMMA 3.1 (see [25, Theorem 3.3.2′]). *Let the following conditions be satisfied:*

(1)

$$\lim_{N\to+\infty} h_N \ln\left(r_N^{-1} \sup_{t\in\{kr_N\},\, x\in R^d} P_{t,x}^N\left\{\left(x, X_{k+1}^N\right) \notin V\right\}\right) = -\infty,$$

where $V = \left\{(x,y) \in R^d \times R^d,\ |x-y| < 1\right\}$.

(2) *For all* $t \in [0,T]$, $x \in R^d$, *and* $|z| \le z_0$,

$$r_N^{-1} h_N G_V^N\left(t,x; h_N^{-1} z\right) - G_0(t,x;z) \to 0 \qquad \textit{uniformly as } N \to +\infty,$$

where

$$G_V^N(z,x;z) = \ln\left[1 + \iint_{\tilde{V}} \left(\exp\left\{Q_{t,x}^N(z,u,v)\right\} - 1\right)\mu(du)\nu(dv)\right], \tag{3.9}$$

$$Q_{t,x}^N(z,u,v) = \tilde{r}_{[t/r_N]}^N z \cdot b(x,v) + \sqrt{\tilde{r}_{[t/r_N]}^N \tilde{h}_{[t/r_N]}^N}\, z \cdot u, \tag{3.10}$$

$$\tilde{V} = \left\{(u,v) \in R^d \times R^d, \left|\tilde{r}_{[t/r_N]}^N b(x,v) + \sqrt{\tilde{r}_{[t/r_N]}^N \tilde{h}_{[t/r_N]}^N}\, u\right| < 1\right\} \tag{3.11}$$

$$G_0(t,x;z) = q(t)\bar{b}(x)\cdot z + \frac{q(t)}{2}\sum_{i,j=1}^{d} A_{ij} z^i z^j. \tag{3.12}$$

(3) *For all* $t \in [0,T]$, $x \in R^d$, *and* $|z| < z_0$,

$$\nabla_z\left(r_N^{-1} h_N G_V^N\left(t,x; h_N^{-1} z\right) - G_0(t,x;z)\right) \to 0 \qquad \textit{uniformly as } N \to +\infty.$$

(4) *For* $|z| \le z_0 < +\infty$, *sufficiently large* N, *and all* t, x,

$$\left| \frac{\partial^2}{\partial z_i \partial z_j} \left(r_N^{-1} h_N G_V^N \left(t, x; h_N^{-1} z \right) \right) \right| \le c < +\infty.$$

Then $\left\{ X_n^N \right\}$ *possesses the large deviation principle as* $N \to +\infty$ *and* $S_{0,T}\{\phi\}$ *is the action functional uniformly with respect to the initial point.*

In the following theorem, the time scale kr_N is used, which means that for any $\phi(t)$, the large deviation of $\left\{ X_k^N - \phi(kr_N) \right\}_k$ is under consideration.

THEOREM 3.2. *Let* $\int_{R^d} u\mu(du) = 0$, $\mu\{u : |u| \ge y\} \le \exp\left\{-ly^\beta\right\}$ *for large enough* y *with* $0 < \beta < 1$, $l > 0$. *For bounded* $b(x, v)$, *uniformly continuous* $\bar{b}(x)$, *and continuous* $q(s)$, *if* $\{r_n\}$ *and* $\{h_n\}$ *satisfy assumptions* (A1)–(A3) *and*

$$r_N^\beta h_N^{\beta-2} \to 0 \qquad \text{as } N \to +\infty, \tag{A4}$$

then $\left\{ X_n^N \right\}$ *possesses the large deviation principle as* $N \to +\infty$ *and* $S_{0,T}\{\phi\}$ *is the action functional uniformly with respect to the initial point.*

Proof. We need only to prove that conditions (1)–(4) of Lemma 3.1 are satisfied.

For any $k > 0$,

$$P_{kr_N,x}^N \left\{ \left(x, X_{k+1}^N \right) \notin V \right\} = \iint_{R^n \times R^n} 1_{\left\{ \left| \tilde{r}_k^N b(x,v) + \sqrt{\tilde{r}_k^N \tilde{h}_k^N} u \right| > 1 \right\}} \nu(dv)\, \mu(du).$$

When N is large enough, it is less than

$$\int_{R^d} 1_{\left\{ \left| 4\sqrt{\tilde{r}_k^N \tilde{h}_k^N} u \right| > \frac{1}{2} \right\}} \mu(du) \le \exp\left\{ -l \left(4\tilde{r}_k^N \tilde{h}_k^N \right)^{-\beta/2} \right\} \le \exp\left\{ -l \left(4 r_N h_N \right)^{-\beta/2} \right\},$$

since $\mu\{u : |u| \ge y\} \le \exp\left\{-ly^\beta\right\}$, $b(x, v)$ is bounded, and $\tilde{r}_k^N \le r_N$ and $\tilde{h}_k^N \le h_N$ for any $k > 0$. Then condition (1) follows.

To prove condition (2), we need to consider the limit of the term

$$r_N^{-1} h_N \iint_{\tilde{V}} \left(\exp\left\{ Q_{t,x}^N \left(h_N^{-1} z, u; v \right) \right\} - 1 \right) \mu(du)\, \nu(dv),$$

which is equal to

$$\begin{aligned} & r_N^{-1} h_N \iint_{\tilde{V}} \left(\exp\left\{ Q_{t,x}^N \left(h_N^{-1} z, u; v \right) \right\} - 1 - Q_{t,x}^N \left(h_N^{-1} z, u; v \right) \right) \mu(du)\, \nu(dv) \\ & + r_N^{-1} h_N \iint_{\tilde{V}} Q_{t,x}^N \left(h_N^{-1} z, u; v \right) \mu(du)\, \nu(dv). \end{aligned} \tag{3.13}$$

The second term can be separated into the following two parts:

$$r_N^{-1} \iint_{\tilde{V}} \tilde{r}_{[t/r_N]}^N z b(x, v)\, \mu(du)\, \nu(dv) + r_N^{-1} \iint_{\tilde{V}} \sqrt{\tilde{r}_{[t/r_N]}^N \tilde{h}_{[t/r_N]}^N} z \cdot u\mu(du)\, \nu(dv). \tag{3.14}$$

By the Hölder inequality and $\int_{R^n} u\mu(du) = 0$, the second term of (3.14) tends to zero uniformly in t, x and $|z| < z_0$ as $N \to +\infty$, and the first part of (3.14) is

$$\frac{\tilde{r}_{[t/r_N]}^N}{r_N} \iint_{R^n \times R^n} z \cdot b(x, v)\, \mu(du)\, \nu(dv) - \frac{\tilde{r}_{[t/r_N]}^N}{r_N} \iint_{\tilde{V}^c} z \cdot b(x, v)\, \mu(du)\, \nu(dv),$$

of which the first part approaches $q(t)\bar{b}(x)\cdot z$ uniformly in t, x and $|z|<z_0$ as $N\to+\infty$, and the second one is bounded by

$$\frac{\tilde{r}^N_{[t/r_N]}}{r_N}z_0 M\mu\left\{u:\left|\sqrt{\tilde{r}^N_{[t/r_N]}\tilde{h}^N_{[t/r_N]}}u\right|>\frac{1}{2}\right\};$$

the latter also goes to zero uniformly in t,x and $|z|<z_0$ as $N\to+\infty$. Now we consider the first term of (3.13). Since

$$\begin{aligned}&\frac{1}{2}r_N^{-1}h_N\iint_{\tilde{V}}\left\{Q^N_{t,x}\left(h_N^{-1}z,u;v\right)\right\}^2\mu(du)\nu(dv)\\&=\frac{1}{2}\iint_{\tilde{V}}\left[\frac{\tilde{r}^N_{[t/r_N]}}{r_N}(z\cdot b(x,v))^2+\frac{\left(\tilde{r}^N_{[t/r_N]}\right)^{1\frac{1}{2}}\left(\tilde{h}^N_{[t/r_N]}\right)^{\frac{1}{2}}}{r_N h_N}(z\cdot b(x,v))(z\cdot u)\right.\\&\qquad\left.+\frac{\tilde{r}^N_{[t/r_N]}\tilde{h}^N_{[t/r_N]}}{r_N h_N}|z\cdot u|^2\right]d\mu d\nu,\end{aligned}$$

it is not difficult to see that the first two terms tend to zero uniformly in t, x, and $|z|<z_0$ as $N\to\infty$, and by (3.1) and (3.2), the limit of the third term is

$$\frac{1}{2}q(t)\sum_{i,j=1}^{d}\left(\int_{R^n}u^iu^j\mu(du)\right)z^iz^j$$

uniformly in t, x, $|z|<z_0$ by a method similar to that above.

Now, we verify

$$r_N^{-1}h_N\iint_{\tilde{V}}\left\{\exp\left\{Q^N_{t,x}\right\}-1-Q^N_{t,x}-\frac{1}{2}\left\{Q^N_{t,x}\right\}^2\right\}\mu(du)\nu(dv)\to 0 \tag{3.15}$$

uniformly in t, x, $|z|<z_0$. Separating $\tilde{V}$ into two parts,

$$\begin{aligned}\tilde{V}_1&=\left\{(u,v)\in R^n\times R^d,\left|\tilde{r}^N_{[t/r_N]}b(x,v)+\sqrt{\tilde{r}^N_{[t/r_N]}\tilde{h}^N_{[t/r_N]}}u\right|<\frac{h_N}{z_0}\right\},\\\tilde{V}_2&=\tilde{V}\backslash\tilde{V}_1,\end{aligned}$$

(3.15) is written as

$$\begin{aligned}&r_N^{-1}h_N\iint_{\tilde{V}_1}\left\{\exp\left\{Q^N_{t,x}\right\}-1-Q^N_{t,x}-\frac{1}{2}\left\{Q^N_{t,x}\right\}^2\right\}\mu(du)\nu(dv)\\&+r_N^{-1}h_N\iint_{\tilde{V}_2}\left\{\exp\left\{Q^N_{t,x}\right\}-1-Q^N_{t,x}-\frac{1}{2}\left\{Q^N_{t,x}\right\}^2\right\}\mu(du)\nu(dv).\end{aligned}\tag{3.16}$$

The first term of (3.16) is less than

$$r_N^{-1}h_N\iint_{\tilde{V}_1}\left\{\left\{Q^N_{t,x}\left(h_N^{-1}z,u;v\right)\right\}^2\right\}\mu(du)\nu(dv),$$

which converges to zero uniformly in t, x, $|z| < z_0$. Let us consider the second term of (3.16). It is less than

$$\left| r_N^{-1} h_N \int_{\frac{h_N}{2z_0}}^{2} \exp\left\{2h_N^{-1} z_0 y\right\} d\mu \left\{ u : |u| \geq \frac{y}{\sqrt{\tilde{r}^N_{[t/r_N]} \tilde{h}^N_{[t/r_N]}}} \right\} \right|$$

$$\leq r_N^{-1} h_N \exp\left\{4h_N^{-1} z_0\right\} \exp\left\{ -l \left(\frac{2}{r_N h_N} \right)^{\beta/2} \right\} + 3 r_N^{-1} h_N \exp\left\{ -l \left(\frac{h_N}{4 z_0^2 r_N} \right)^{\beta/2} \right\}$$

$$+ 2 r_N^{-1} z_0 \int_{\frac{h_N}{2z_0}}^{2} \exp\left\{2h_N^{-1} z_0 y\right\} \exp\left\{ -l \left(\frac{y^2}{r_N h_N} \right)^{\beta/2} \right\} dy,$$

since $\left|Q^N_{t,x}\left(h_N^{-1} z, u; v\right)\right| \leq h_N^{-1} z_0 \left| \sqrt{\tilde{r}^N_{[t/r_N]} \tilde{h}^N_{[t/r_N]}}\, u \right|$ for large enough N, and $e^x - 1 - x - \frac{1}{2}x^2 \leq e^{|x|}$. In terms of $r_N^{-\beta/2} h_N^{1-\beta/2} \to \infty$, it converges to zero. Therefore, condition (2) is checked.

We can verify conditions (3) and (4) similarly. The theorem is then complete. □

Remark. If $Ee^{\lambda \zeta_n} < +\infty$ for some $\lambda > 0$, (A4) may be weakened to

$$r_N h_N^{-1} \to 0 \qquad \text{as } N \to +\infty. \tag{A4$'$}$$

If $\mu\{u : |u| \geq y\} \leq c y^{-\beta}$, $\beta > 4$, $c > 0$, (A4) should be more restricted by assuming

$$h_N^{-1} = o\left(-\ln\left(r_N\right)\right). \tag{A4$''$}$$

But according to the time scale kr_N, the action functional isn't invariant under time shift. To achieve a better form of the action functional, we consider another time scale. Let $s(t) = \int_0^t q(u)\, du$; then we have

$$\int_0^T \frac{1}{2} \sum_{i,j=1}^{d} \left(q(u)\right)^{-1} A^{ij} \left(\dot{\phi}^i(u) - q(u) \bar{b}^i\left(\phi(u)\right) \right) \cdot \left(\dot{\phi}^j(u) - q(u) \bar{b}^j\left(\phi(u)\right) \right) du$$

$$= \int_0^{s(T)} \frac{1}{2} \sum_{i,j=1}^{d} A^{ij} \left(\frac{d\phi_s^i}{du}(u) - \bar{b}^i\left(\phi_s(u)\right) \right) \cdot \left(\frac{d\phi_s^j}{du}(u) - \bar{b}^j\left(\phi_s(u)\right) \right) du,$$

where $\phi_s = \phi \circ s^{-1}$. Denote

$$\tilde{S}_{0,T}(\phi) = \frac{1}{2} \int_0^T \sum_{i,j=1}^{d} A^{ij} \left(\phi^i(v) - \bar{b}^i\left(\phi(v)\right) \right) \cdot \left(\phi^j(v) - \bar{b}^j\left(\phi(v)\right) \right) dv. \tag{3.17}$$

Mapping $\phi(v) \longmapsto \left(\phi \circ s^{-1}\right)(v)$ is bijective in $B(R, R^d)$, the set of Borel mappings from R to R^d; then we have the following.

COROLLARY 3.3. *Denote by $\Phi_{x,T}(\gamma)$ the set of all functions $\phi(t)$, $0 \leq t \leq T$, such that $\phi(0) = x$ and $\tilde{S}_{0,T}(\phi) \leq \gamma$. Then*

(1) *the functional $\tilde{S}_{0,T}(\phi)$ is lower semicontinuous, and $\Phi_{x,T}(\gamma)$ is compact;*

(2) *for any $\delta > 0$, $l > 0$, $\phi \in \Phi_{x,T}(\gamma)$, and sufficiently large N,*

$$P^N_{0,x} \left\{ \sup_{0 \leq s(nr_N) - T_1 \leq T} \left| X^N_n - \phi\left(s(nr_N) - T_1\right) \right| < \delta \right\} \geq \exp\left\{ -h_N^{-1} \left(\tilde{S}_{0,T}(\phi) + l \right) \right\};$$

(3) *for any* $\delta > 0$, $l > 0$, *and sufficiently large* N,

$$P_{0,x}^{N}\left\{\inf_{\phi\in\Phi_{x,T}(\gamma)}\sup_{0\le s(nr_N)-T_1\le T}\left|X_n^N-\phi(s(nr_N)-T_1)\right|\ge\delta\right\}\le\exp\{-h_N^{-1}(\gamma-l)\}.$$

Define

$$V(x,y):=\inf\left\{\tilde{S}_{0,T}(\phi):\phi_0=x,\phi_T=y,\phi\in C([0,T]),T\ge 0\right\}$$

for later use.

4. Lemmas. As in [5], we will give some lemmas to show the relations between the stochastic process $\{X_n^N\}$ and the trajectory $X^0(\cdot)$ of the dynamical system

$$\frac{dX^0(t)}{dt}=\bar{b}\left(X^0(t)\right). \tag{4.1}$$

We have the following lemma.

LEMMA 4.1. *Let* K *be a compact set not entirely containing any* ω*-limit sets of the dynamical system* (4.1) *and* $\tau=\inf\{n>0,X_n^N\notin K\}$. *Assume that the assumptions* (A1)–(A3) *and* (A4) (*or* (A4′), (A4″)) *hold. Then there exist* c, T_0, *and* N_0 *such that*

$$P_x\{s(\tau\cdot r_N)>T\}\le\exp\left\{-c(T-T_0)h_N^{-1}\right\} \tag{4.2}$$

for $N\le N_0$, $T>T_0$.

Proof. The lemma can be proven by using a method similar to that of [5]. □

LEMMA 4.2. *Let* F *be a compact set and the assumptions* (A1)–(A3) *and* (A4) (*or* (A4′), (A4″)) *hold. Then for any* $c>0$, $\delta>0$ *there is* $R>0$ *such that*

$$P_x\left\{s(\theta\cdot r_N)\le\exp\left(ch_N^{-1}\right)\right\}\le\exp\left\{-\delta h_N^{-1}\right\} \tag{4.3}$$

for all $x\in F$, *where* $\theta=\inf\{n>0,|X_n^N|\ge R\}$.

Proof. Choose $R_1>0$, $R_2>0$, and $R_1>R_2$ such that $\{x,|x|\le R_1\}$ contains F and all ω-limit sets of (4.1). Assume $\inf\{V(x,y),|x|\le R_1,|y|\ge R_2\}=d_1>0$, and for any $R>R_2$,

$$d_2(R)=\inf\{V(x,y),|x|\le R_2,|y|>R\}>0.$$

Suppose that $\{\theta_n\}$, $\{\rho_n\}$ are two sequences of stopping times which satisfy

$$\begin{aligned}
\theta_1&=\inf\left\{n\ge 0,\left|X_n^N\right|\ge R_2\right\},\\
\rho_1&=\inf\left\{n\ge\theta_1,\left|X_n^N\right|\le R_1\right\},\\
&\vdots\\
\theta_n&=\inf\left\{n\ge\rho_{n-1},\left|X_n^N\right|\ge R_2\right\},\\
\rho_n&=\inf\left\{n\ge\theta_n,\left|X_n^N\right|\le R_1\right\},\\
&\vdots
\end{aligned}$$

Then for any M,

$$\begin{aligned}&P_x\left\{s\left(\theta\cdot r_N\right)<\exp\left\{ch_N^{-1}\right\}\right\}\\&\le P_x\left\{s\left(\theta_M\cdot r_N\right)\ge s\left(\theta\cdot r_N\right)\right\}+P_x\left\{s\left(\theta_M\cdot r_N\right)<\exp\left\{ch_N^{-1}\right\}\right\}.\end{aligned}$$

To estimate the first term, for sufficiently small ϵ, and $x\in\{R_2\le|x|\le R_2+\epsilon\}$, we have

$$\begin{aligned}&P_x\left\{s\left(\theta_M\cdot r_N\right)<s\left(\theta\cdot r_N\right)\right\}\\&\ge P_x\left\{s\left(\theta_M\cdot r_N\right)<s\left(\theta\cdot r_N\right),R_2\le\left|X_{\theta_M}^N\right|\le R_2+\epsilon\right\}\\&\ge E_x1_{\left\{\tau>\theta_{M-1},R_2\le\left|X_{\theta_{M-1}}^N\right|\le R_2+\epsilon,P_{X_{\theta_{M-1}}^N}\left\{\theta_1<\theta,R_2\le\left|X_{\theta_1}^N\right|\le R_2+\epsilon\right\}\right\}}\\&\ge\left(\inf_{R_2\le|y|\le R_2+\epsilon}P_y\left\{\theta_1<\theta,R_2\le\left|X_{\theta_1}^N\right|\le R_2+\epsilon\right\}\right)^M.\end{aligned}\tag{4.4}$$

Notice that

$$P_x\left\{\theta\le\theta_1\right\}\le\exp\left\{-\frac{d_2\left(R\right)}{2}h_N^{-1}\right\},\tag{4.5}$$

and for N large enough, by Theorem 3.2, there is $\beta>d_2\left(R\right)$ such that

$$P_x\left\{\theta\ge\theta_1,\left|X_{\theta_1}^N\right|>R_2+\epsilon\right\}\le\sup_{t\in(kr_N),x\in R^d}P_{t,x}^N\left\{\left(x,X_{k+1}^N\right)\notin V^\epsilon\right\}\le\exp\left\{-\frac{\beta}{2}h_N^{-1}\right\},\tag{4.6}$$

where $V^\epsilon=\left\{(x,y)\in R^d\times R^d,|x-y|<\epsilon\right\}$. Choosing $M=\exp\{\frac{d_2(R)}{3}h_N^{-1}\}$, by (4.5), (4.6), and

$$P_x\left\{\theta\le\theta_1\text{ or }\left|X_{\theta_1}^N\right|>R_2+\epsilon\right\}\le P_x\left\{\theta\le\theta_1\right\}+P_x\left\{\theta\ge\theta_1,\left|X_{\theta_1}^N\right|>R_2+\epsilon\right\},\tag{4.7}$$

we conclude that (4.4) is bigger than

$$\begin{aligned}&\left(1-\exp\left\{-\frac{d_2\left(R\right)}{2}h_N^{-1}\right\}-\exp\left\{-\frac{\beta}{2}h_N^{-1}\right\}\right)^M\\&=1-\exp\left\{-\frac{d_2\left(R\right)}{6}h_N^{-1}\right\}+o\left(\exp\left\{-\frac{d_2\left(R\right)}{6}h_N^{-1}\right\}\right).\end{aligned}$$

The remainder of the proof is similar to [5, p. 128]. □

Now we consider the following conditions:

$$\limsup_{|y|\to+\infty}\frac{\left|\bar{b}\left(y\right)\right|}{|y|}\le M_1<+\infty,\tag{B1}$$

$$\limsup_{|y|\to+\infty}\frac{\left\langle\bar{b}\left(y\right),y\right\rangle}{|y|^2}\le-c<0,\tag{B2}$$

$$\limsup_{|y|\to+\infty}\frac{E\left(b\left(y,\xi_n\right)-\bar{b}\left(y\right)\right)^2}{|y|^2}\le M_2<+\infty.\tag{B3}$$

LEMMA 4.3. *Suppose that the stochastic approximation model* (2.1) *satisfies conditions* (B1), (B2), *and* (B3). *Then for any compact set* $F \subset R^d$ *and fixed* N, *the family of probability measures*

$$\{P_{N,y}\{X_n \in \cdot\}; N < n, y \in F\}$$

is tight, where $P_{N,y}$ *is the probability starting at* y *at time* N.

Proof. To prove this lemma, we need only to check $E_{N,y}|X_n|^2 \le M$ for any $n > N$ and $y \in F$.

Denote

$$V_N := |y|^2 \text{ and } V_n := \exp\left\{l\sum_{i=N}^{n-1} r_i\right\}\int_{R^d}|x|^2 P_{n,z}^N(dx),$$

where $P_{n,z}^N(dx)$ is the transition probability of $\{X_n\}$ from N to n. Picking $l \le c/4$, we have

$$\begin{aligned}
&V_{n+1} - V_n\\
&= \exp\left\{l\sum_{i=N}^{n} r_i\right\}\int_{R^{2d+n}}\left|y + r_n b(y,v) + \sqrt{r_n h_n}u\right|^2 \nu(dv)\mu(du)P_{n,z}^N(dy)\\
&\quad - \exp\left\{l\sum_{i=N}^{n-1} r_i\right\}\int_{R^d}|y|^2 P_{n,z}^N(dy)\\
&= \exp\left\{l\sum_{i=N}^{n-1} r_i\right\}\left[\int_{R^d}\left[(e^{lr_n}-1)|y|^2 + e^{lr_n}\left(r_n^2|\bar b(y)|^2 + 2r_n\langle y, \bar b(y)\rangle\right)\right]P_{n,z}^N(dy)\right.\\
&\quad \left. + e^{lr_n}\left(\int_{R^{n+d}} r_n^2\left|b(y,v) - \bar b(y)\right|^2\nu(dv)P_{n,z}^N(dy) + \int_{R^{2d}} r_n h_n|u|^2\mu(du)P_{n,z}^N(dy)\right)\right].
\end{aligned}$$

By (B1)–(B3), there exists R such that $\left|\bar b(y)\right| \le (M_1+1)|y|$, $\left\langle \bar b(y), y\right\rangle \le -\frac{2c}{3}|y|^2$, and $E\left(b(y,\xi_n) - \bar b(y)\right)^2 \le (M_2+1)|y|^2$ for all $|y| \ge R$. Then

$$\begin{aligned}
&\int_{\{|y|>R\}}\left[(e^{lr_n}-1)|y|^2 + e^{lr_n}r_n\left[2\langle y, \bar b(y)\rangle + r_n E\left|b(y,\xi_n) - \bar b(y)\right|^2 + h_n E\zeta_n^2\right]\right]P_{n,z}^N(dy)\\
&\le \int_{\{|y|>R\}}\left[-\frac{5c}{6}r_n|y|^2 + 2(M_2+2)r_n^2|y|^2 + 2r_n h_n E\zeta_n^2\right]P_{n,z}^N(dy) \le 0
\end{aligned}$$

for large enough n, and the integral taking over $|y| \le R$ is less than

$$\begin{aligned}
&r_n\int_{\{|y|\le R\}}\left[\frac{c}{2}|y|^2 + 4\left|\left\langle y, \bar b(y)\right\rangle\right| + 2r_n E|b(y,\xi_n)|^2 + h_n E|\zeta_n|^2\right]P_{n,z}^N(dy)\\
&\le Ar_n
\end{aligned}$$

with a constant A. Thus, we obtain

$$V_{n+1} - V_n \le Ar_n\exp\left\{l\sum_{i=N}^{n-1} r_i\right\}$$

and then

$$V_n \le \sum_{i=N+1}^{n-1} Ar_n \exp\left\{l\sum_{i=N}^{i-1} r_i\right\} + V_N \le A\int_0^{\sum_{j=N}^{n-1} r_j} e^{lx}dx + |y|^2,$$

which implies

$$E_{N,y}|X_n|^2 \le \frac{A}{l}\left(1-\exp\left\{-l\sum_{i=N}^{n-1} r_i\right\}\right) + |y|^2 \exp\left\{-l\sum_{i=N}^{n-1} r_i\right\}.$$

This shows that $E_{N,y}|X_n|^2$ are bounded for all $n > N$ and $y \in F$. The lemma is proven. □

Suppose that there exists a continuous function $U(x)$ such that $\bar{b}(x) = -\nabla U(x)$, $A_{ij} = \delta_{ij}$, and $S := \{x \in R^d, \nabla U(x) = 0\}$. If $\liminf_{|x|\to+\infty} |\nabla U(x)| = +\infty$, then S is compact. We assume that S consists of a finite number of connected compact sets $\{K_1, K_2, \ldots, K_L\}$.

LEMMA 4.4. *Let F be a compact set, and suppose that the assumptions* (A1)–(A3), (A4) (*or* (A4′), (A4″)) *hold. Then for given $l > 0$ and $M_0 > 0$, there exists $T^* > 0$ such that for sufficiently large N and $x \in F$,*

$$P_x\{s(\rho \cdot r_N) > T^*\} \le \exp\{-M_0 h_N^{-1}\}, \tag{4.8}$$

with $\rho = \inf\{n \ge 0, X_n^N \in B(S, l)\}$, where $B(\cdot, l)$ stands for the l-neighborhood of a set.

Proof. If $x \in F \cap B(S, l)$, the lemma is obvious. Then we assume $F \cap B(S, l) = \emptyset$. By Lemma 4.1, there is T^* such that

$$P_x\{s(\tau \cdot r_N) < T^*\} \le \frac{1}{2}\exp\{-M_0 h_N^{-1}\}.$$

Choose sufficiently large R such that

$$P_x\{s(\theta \cdot r_N) \le T^*\} \le \exp\{-M_0 h_N^{-1}\},$$

where θ and τ are stopping times defined in Lemmas 4.1 and 4.2, and

$$K := \{x \in R^d, |x| \le R\} \setminus B(S, l).$$

Now we have

$$\begin{aligned} P_x\{s(\tau \cdot r_N) > T^*\} &\ge P_x\{s(\rho \cdot r_N) > T^*, s(\theta \cdot r_N) > T^*\} \\ &\ge P_x\{s(\rho \cdot r_N) > T^*\} - P_x\{s(\theta \cdot r_N) < T^*\}, \end{aligned}$$

since $\tau = \rho \wedge \theta$. Thus the lemma holds. □

For $x, y \in R^d$, suppose that $\mathcal{B}_{x,y}$ is the set of paths linking x and y, i.e., the set of continuous maps $\varphi : [0,1] \to R^d$ such that $\varphi(0) = x$ and $\varphi(1) = y$. For any path $\varphi \in \mathcal{B}_{x,y}$, define

$$e(\varphi) = 2\left(\max_{0\le t\le 1} U(\varphi(t)) - U(x)\right) \text{ and } I(x,y) = \inf_{\varphi\in\mathcal{B}_{x,y}} e(\varphi).$$

By the definition of S, $I(x,y)$ is a constant for fixed y and any $x \in K_1 (\subset S)$. Similarly, $I(x,y)$ is a constant for fixed x and any $y \in K_2 (\subset S)$, $I(x,y)$ is the same for any $x \in$

$K_1\,(\subset S)$ and $y \in K_2\,(\subset S)$. We denote them by $I(K_1, y)$, $I(x, K_2)$, and $I(K_1, K_2)$ correspondingly.

For any $c \geq 0$ and $K \in S$, we define

$$\pi_K^c := \{x \in S, I\,(y, x) \leq c, y \in K\}\,, \quad I\,(\pi_K^c) := \max_{y \in \pi_K^c} \min_{x \in S \backslash \pi_K^c} I\,(y, x)\,,$$

and

$$\underline{S} = \left\{x \in R^d, U\,(x) = \min_{y \in R^d} U\,(y)\right\},$$
$$\underline{S}_c = \left\{x \in S, \inf_{y \in \underline{S}} I\,(y, x) \leq c\right\}.$$

LEMMA 4.5. *For any π_K^c, $\alpha_0 > 0$ and $\delta < \alpha_0$, under the assumptions* (A1)–(A3) *and* (A4) (*or* (A4′), (A4″)) *there exist N_0, l_0 such that*

$$P_x\left\{s\left(\tau_{\pi_k^c} \cdot r_N\right) < \exp\left\{\left(I\left(\pi_K^c\right) - \alpha\right) h_N^{-1}\right\}\right\} \leq \exp\left\{-\delta h_N^{-1}\right\},$$
$$P_x\left\{s\left(\tau_{\pi_k^c} \cdot r_N\right) > \exp\left\{\left(I\left(\pi_K^c\right) + \alpha\right) h_N^{-1}\right\}\right\} \leq \exp\left\{-\delta h_N^{-1}\right\}$$

for $l \leq l_0$, $N > N_0$, and $\alpha > \alpha_0$, where $\tau_{\pi_K^c} = \inf\left\{n > 0, X_n^N \in B\left(S \backslash \pi_K^c, l\right)\right\}$, $x \in B\left(\pi_K^c, l\right)$.

Proof. By Lemma A.4 and Lemma 4.2 of [14], it is obvious. □

LEMMA 4.6. *For any $\alpha_0 > 0$, $c < c'$, under assumptions* (A1)–(A3) *and* (A4) (*or* (A4′), (A4″)), *there is l_0 such that for $l \leq l_0$, $K_i \subset \underline{S}_{c'} \backslash \underline{S}_c$, and $x \in B\left(K_i, l\right)$, we have*

$$E_x\left\{s\left(\tau_{\underline{S}_c} \cdot r_N\right) \middle| \tau_{\underline{S}_{c'}} \leq \rho_{\underline{S}_{c'}}\right\} \leq \exp\left\{-(\bar{d}_{c'}^c + \alpha_0) h_N^{-1}\right\},$$

where

$$\tau_{\underline{S}_c} = \inf\left\{n > 0, X_n^N \in B\left(\underline{S}_c, l\right)\right\},$$
$$\rho_{\underline{S}_{c'}} = \inf\left\{n > 0, X_n^N \notin B\left(\underline{S}_{c'}, l\right)\right\},$$
$$\bar{d}_{c'}^c = \max_{y \in \underline{S}_{c'} \backslash \underline{S}_c} \min_{x \in \underline{S}_c} I\,(y, x)\,.$$

Proof. It is equivalent to the statement that the Markov chain is contained in $\underline{S}_l$. By Lemma 1.6 of [14] and Lemma A.5, the result is obvious. □

LEMMA 4.7. *Under the same assumptions as in Lemma* 4.6, *we have*

$$P_x\left\{s\left(\tau_{\underline{S}_c} \cdot r_N\right) \geq \exp\left\{\left(\bar{d}_{c'}^c + \bar{\alpha}\right) h_N^{-1}\right\} \middle| \tau_{\underline{S}_c} \leq \rho_{\underline{S}_{c'}}\right\} \leq \exp\left\{-\delta h_N^{-1}\right\}$$

for any $\bar{\alpha} > \alpha_0 + \delta$.

Proof. It can be obtained directly by using Lemma 4.6 and the Chebyshev inequality. □

5. Main results. Recall $\underline{S} = \{x \in R^d, U\,(x) = \min_{y \in R^d} U\,(y)\}$, and define $d^* := \max_{x \in S \backslash \underline{S}} \min_{y \in \underline{S}} I\,(x, y)$. Then we have the following lemma.

LEMMA 5.1. *For any compact set $F_0 \in R^d$ and $\alpha > 0$, under assumptions* (B1)–(B3), *if $\{r_n\}$ and $\{h_n\}$ satisfy assumptions* (A1)–(A3), (A4) ((A4′) *or* (A4″)), *and*

$$r_N^{-1} s^{-1}\left(\exp\left\{(d^* + \alpha^*) h_N^{-1}\right\}\right) < G\,(N) \ \ (\textit{defined in } (3.1))$$

for large enough N, then there are $\delta > 0$ and $N_0 > 0$ such that

$$P_{N,y}\left\{X_{N+r_N^{-1}s^{-1}\left(\exp\left\{(d^*+\alpha)h_N^{-1}\right\}\right)} \in B\left(\underline{S},l\right)\right\} \geq 1-\exp\left\{-\delta h_N^{-1}\right\} \tag{5.1}$$

for all $N > N_0$, $\alpha^*/2 \leq \alpha \leq \alpha^*$ and $y \in F_0$.

Proof. For all $n \leq r_N^{-1}s^{-1}\left(\exp\left\{\left(d^*+\alpha\right)h_N^{-1}\right\}\right) < G\left(N\right)$, $X_{N+n} = X_n^N$; therefore we can use all results valid for X_n^N.

For simplicity, we denote

$$T\left(N,c\right) = r_N^{-1}s^{-1}\left(\exp\left\{ch_N^{-1}\right\}\right).$$

Let $\sigma = \inf\left\{n > N, X_n \in B\left(S,l\right)\right\}$. By Lemma 4.4, there is $T^* > 0$ such that for $y \in F_0$ and sufficiently large N,

$$P_{N,y}\left\{\sigma > N + r_N^{-1}s^{-1}\left(T^*\right)\right\} \leq \exp\left\{-M_0h_N^{-1}\right\}$$

for some $M_0 > 0$.

Now for fixed $y \in F_0$,

$$\begin{aligned}
&P_{N,y}\left\{X_{N+T(N,d^*+\alpha)} \in B\left(\underline{S},l\right)\right\}\\
&\geq E_{N,y}\left\{\sigma < N + r_N^{-1}s^{-1}\left(T^*\right), P_{\sigma,X_\sigma}\left\{X_{N+T(N,d^*+\alpha)} \in B\left(\underline{S},l\right)\right\}\right\}\\
&\geq \inf_{\substack{N \leq N_1 \leq N + r_N^{-1}s^{-1}(T^*)\\ y_1 \in B(S,l)}} P_{N_1,y_1}\left\{X_{N+T(N,d^*+\alpha)} \in B(\underline{S},l)\right\} \times \left(1-\exp\{-M_0h_N^{-1}\}\right).
\end{aligned} \tag{5.2}$$

Recall $\underline{S}_{d^*} = \{x \in S, \inf_{y\in\underline{S}} I\left(y,x\right) \leq d^*\}$. Clearly, we have $S\backslash\underline{S}_{d^*} \neq \emptyset$. Fix N_1 with $N \leq N_1 \leq N + r_N^{-1}s^{-1}\left(T^*\right)$ and $y_1 \in B\left(S,l\right)$, and define the stopping time θ by

$$\theta = \inf\left\{n \geq N, X_n \in B\left(\underline{S}_{d^*},l\right)\right\}.$$

Then

$$\begin{aligned}
&P_{N_1,y_1}\left\{X_{N+T(N,d^*+\alpha)} \in B\left(\underline{S},l\right)\right\}\\
&\geq E_{N_1,y_1}\left\{\theta < N + T\left(N,d^*+\frac{\alpha}{2}\right), P_{\theta,X_\theta}\left\{X_{N+T(N,d^*+\alpha)} \in B\left(\underline{S},l\right)\right\}\right\}\\
&\geq P_{N_1,y_1}\left\{\theta < N + T\left(N,d^*+\frac{\alpha}{2}\right)\right\}\\
&\times \inf_{\substack{N \leq N_2 \leq N + T\left(N,\,d^*+\frac{\alpha}{2}\right)\\ y_2 \in B\left(\underline{S}_{d^*},\,l\right)}} P_{N_2,y_2}\left\{X_{N+T(N,d^*+\alpha)} \in B\left(\underline{S},l\right)\right\}.
\end{aligned} \tag{5.3}$$

Now let us estimate

$$P_{N_1,y_1}\left\{\theta < N + T\left(N,d^*+\frac{\alpha}{2}\right)\right\} \text{ and } P_{N_2,y_2}\left\{X_{N+T(N,d^*+\alpha)} \in B\left(\underline{S},l\right)\right\}.$$

For the first estimate, for $y_1 \in B\left(S\backslash\underline{S}_{d^*},l\right)$, there is $\min_{x\in\underline{S}_{d^*}} I\left(y_1,x\right) \leq d^*$, since $\underline{S} \subset \underline{S}_{d^*}$. And by Lemma 4.7, we have

$$P_{N_1,y_1}\left\{\theta \geq N + T\left(N,d^*+\frac{\alpha}{2}\right)\right\} \leq \exp\left\{-\delta h_N^{-1}\right\}; \tag{5.4}$$

otherwise, $y_1 \in B\left(\underline{S}_{d^*},l\right)$. In this case,

$$P_{N_1,y_1}\left\{\theta < N + T\left(N,d^*+\frac{\alpha}{2}\right)\right\} = 1. \tag{5.5}$$

For $P_{N_2,y_2}\left\{X_{N+T(N,d^*+\alpha)}\in B\left(\underline{S},l\right)\right\}$, we assume $y_2\in B\left(\pi_K^{d^*},l\right)$, where $N\le N_2\le N+T\left(N,d^*+\frac{\alpha}{2}\right)$ and $K\subset\underline{S}$, and define

$$d\left(\pi_K^{d^*}\right):=\max_{x\in\pi_K^{d^*}\setminus\underline{S}}\min_{y\in\pi_K^{d^*}\cap\underline{S}}I\left(x,y\right).$$

Fix $y_2\in B\left(\pi_K^{d^*},l\right)$, and define

$$\begin{aligned}\bar{N}_0&=N+T\left(N,d^*+\alpha\right)-T\left(N,d\left(\pi_K^{d^*}\right)+\alpha\right),\\ \bar{N}_1&=N+T\left(N,d^*+\alpha\right)-T\left(N,d\left(\pi_K^{d^*}\right)+\frac{\alpha}{2}\right),\\ \rho&=\inf\left\{n\ge\bar{N}_0,X_n\in B\left(\pi_K^{d^*},l\right)\right\}.\end{aligned}$$

Then

$$\begin{aligned}&P_{N_2,y_2}\left\{X_{N+T(N,d^*+\alpha)}\in B\left(\underline{S},l\right)\right\}\\ &\ge P_{N_2,y_2}\left\{P_{\rho,X_\rho}\left\{X_{N+T(N,d^*+\alpha)}\in B\left(\underline{S},l\right)\right\},\rho\le\bar{N}_1\right\}\\ &\ge\inf_{\substack{\bar{N}\in[\bar{N}_0,\bar{N}_1]\\ y\in B\left(\pi_K^{d^*},l\right)}}P_{\bar{N},y}\left\{X_{N+T(N,d^*+\alpha)}\in B\left(\underline{S},l\right)\right\}\cdot P_{N_2,y_2}\left\{\rho\le\bar{N}_1\right\}.\end{aligned}\tag{5.6}$$

Notice

$$\begin{aligned}P_{N_2,y_2}\left\{\rho>\bar{N}_1\right\}&=P_{N_2,y_2}\left\{X_n\notin B\left(\pi_K^{d^*},l\right)\text{ for all }n\in\left[\bar{N}_0,\bar{N}_1\right]\right\}\\ &\le P_{N_2,y_2}\left\{X_n\notin B\left(S,l\right)\text{ for all }n\in\left[\bar{N}_0,\bar{N}_1\right]\right\}+P_{N_2,y_2}\left\{\tau_1\le\bar{N}_1\right\},\end{aligned}$$

where $\tau_1=\inf\left\{n>N_2,X_n\in B\left(S\backslash\pi_K^{d^*},l\right)\right\}$. Clearly, $I\left(\pi_K^{d^*}\right)>d^*$. Therefore, Lemma 4.5 implies that there is $\delta_1>0$ such that for large N and α small enough,

$$P_{N_2,y_2}\left\{\tau_1\le\bar{N}_1\right\}\le\exp\left\{-\delta_1h_N^{-1}\right\}.$$

On the other hand, for large N, by Lemma 4.1 and 4.2, we obtain

$$P_{N_2,y_2}\left\{X_n\notin S\text{ for all }n\in\left[\bar{N}_0,\bar{N}_1\right]\right\}\le\exp\left\{-M_1h_N^{-1}\right\}.$$

Thus there is $\delta_2>0$ such that

$$P_{N_2,y_2}\left\{\rho\le\bar{N}_1\right\}\ge1-2\exp\left\{-\delta_2h_N^{-1}\right\}.\tag{5.7}$$

Combining (5.3), (5.4), (5.5), (5.6), (5.7), for $N\le N_1\le N+r_N^{-1}s^{-1}\left(T^*\right)$, $y_1\in B\left(S,l\right)$, any $M>0$, and large N, we have

$$\begin{aligned}&P_{N_1,y_1}\left\{X_{N+T(N,d^*+\alpha)}\in B\left(\underline{S},l\right)\right\}\\ &\ge\left(1-\exp\left\{-M_1h_N^{-1}\right\}\right)\left(1-2\exp\left\{-\delta_1h_N^{-1}\right\}\right)\\ &\quad\times\inf_{\substack{N_2\ge N\\ y_2\in B\left(\pi_K^{d^*},l\right)\\ \pi_K^{d^*}\subset\underline{S}_{d^*}}}P_{N_2,y_2}\left\{X_{N_2+T\left(N_2,d\left(\pi_K^{d^*}\right)+\alpha\right)}\in B\left(\underline{S},l\right)\right\}.\end{aligned}$$

Analogously, we also have

$$\inf_{\substack{N \le N_1 \le N + r_N s^{-1}(T^*) \\ y_1 \in B(S,l)}} P_{N_1,y_1}\left\{X_{N+T(N,d^*+\alpha)} \in B(\underline{S},l)\right\} \ge \left(1 - \exp\left\{-\delta h_N^{-1}\right\}\right)^{L+1}.$$

Together with (5.2), we have proved the lemma. □

THEOREM 5.2. *Suppose that $\bar{b}$, the gradient of the potential function $-U(x)$, and $b(x,\xi)$ satisfy* (B1)–(B3). *If $\{r_n\}$ and $\{h_n\}$ meet assumptions* (A1)–(A3) *and* (A4) ((A4′) *or* (A4″)) *and*

$$r_n^{-1} s^{-1}\left(\exp\left\{(d^*+\alpha) h_n^{-1}\right\}\right) < G(n) \,(\textit{defined in } (3.1))$$

for n large enough and some $\alpha > 0$, then

$$P_{0,y}\{X_n \in B(\underline{S},l)\} \to 1 \qquad \textit{as } n \to +\infty$$

uniformly for y in an arbitrary compact set F.

Proof. For any $\epsilon > 0$, by Lemma 4.3, there is a compact set F_0 such that

$$P_{0,y}\{X_n \in F_0\} \ge 1 - \frac{\epsilon}{2} \tag{5.8}$$

for all $y \in F$ and $n > 0$. Note that $T(N,c) := N + r_N^{-1} s^{-1}\left(\exp\left\{c h_N^{-1}\right\}\right)$ is monotonically increasing to infinity as $N \to +\infty$. For any n, there is a unique N_n such that

$$T\left(N_n, d^* + \frac{\alpha^*}{2}\right) \le n < T\left(N_n + 1, d^* + \frac{\alpha^*}{2}\right),$$

where $N_n \to +\infty$, as $n \to +\infty$. By $h_{N+1}/h_N \to 1$ and $r_{N+1}/r_N \to 1$, as $N \to +\infty$, we have $\alpha^*/2 \le \alpha \le \alpha^*$ such that $n = T(N_n, d^* + \alpha)$ for n large enough. In terms of Lemma 5.1, it implies

$$P_{N_n,z}\{X_n \in B(\underline{S},l)\} \ge 1 - \exp\left\{-\delta h_{N_n}^{-1}\right\} > 1 - \frac{\epsilon}{2} \tag{5.9}$$

for all $z \in F_0$ and n large enough. Then the result follows immediately. □

Remarks. (1) If $\{r_n\}$ satisfies the condition of Corollary 2.2, we can choose $G(n) = n/\ln\ln n$. By $s(t) = 1$, we see that $r_n^{-1} s^{-1}\left(\exp\left\{(d^*+\alpha) h_n^{-1}\right\}\right) < n/\ln\ln n$ is equal to

$$h_n^{-1} < \frac{1}{d^* + \alpha}\left(\ln n + \ln r_n - \ln\ln n\right).$$

When $r_n \le 1/\ln^{\alpha} n$, $\alpha > 1$, assumptions (A1)–(A3) and (A4′) are satisfied. To require (A4) or (A4″), we need more restricted conditions on $\{r_n\}$, $\{h_n\}$ such as $r_n \ge 1/n^{\gamma}$, $\gamma < 1$ to meet (A4), etc.

(2) If $r_n = c/n + o(1/n)$,

$$G(n) = n^l \text{ for any } l > 1,$$
$$s(t) = \ln t,$$

then the theorem is valid when $h_n \ge (d+\alpha)/\ln\ln n$.

(3) For given $\epsilon > 0$, in the case of $r_n = 1/n^\gamma$, $\gamma < 1$, and $h_n = (d^* + \alpha)/(1-\gamma)\ln n$, we have $P_{0,x}\{X_N \in \underline{S}\} \geq 1 - \epsilon$, if $N > \exp\left\{-\frac{d^*+\alpha}{\delta(1-\gamma)}\ln(2\epsilon)\right\}$. When $r_n = c/n$, $h_n = (d^*+\alpha)/\ln\ln n$, to obtain the same result above, we need $N > \exp\left\{\exp\left\{-\frac{d^*+\alpha}{\delta}\ln(2\epsilon)\right\}\right\}$, where δ is a constant given in Lemma 5.1.

Here we point out that the $\{h_n\}$ given above is the best possibility for annealing in many cases. In this direction, we consider only when $r_n = c/n^\gamma + o(1/n^\gamma)$, $\gamma < 1$, and $r_n = c/n + o(1/n)$. As in [11], we have the following propositions.

PROPOSITION 5.3. *If $r_n = c/n^\gamma + o(1/n^\gamma)$ and $h_n \leq (d^* - \beta)/((1-\gamma)\ln n)$ for some $\beta > 0$, then there exists a compact set $K \in S\backslash\underline{S}$ satisfying $\pi_K^{d^*-\beta} \cap \underline{S} = \emptyset$, and for any $l > 0$, there is $c_1 > 0$ such that for large N and $y \in B(K, l)$,*

$$P_{N,y}\{\tau < +\infty\} \leq c_1 N^{-\beta(1-\gamma)/(2(d^*-\beta))},$$

where $\tau = \inf\{n > N, X_n \in B(S\backslash\pi_K^{d^*-\beta}, l)\}$.

Proof. Since $d^* = \max_{x\in S\backslash\underline{S}} \min_{y\in\underline{S}} I(x,y)$, we can find a $K \in S\backslash\underline{S}$, which satisfies $\min_{y\in\underline{S}} I(K,y)$. At this time, $\pi_K^{d^*-\beta} \cap \underline{S} = \emptyset$. Now we will prove the remaining part of the proposition. For simplicity, we assume that there is a $\hat{d} > 0$ such that

$$h_n > \frac{\hat{d}}{(1-\gamma)\ln n}.$$

Define

$$N_0 = N^{1/(1-\alpha)}, \quad N_n = (n+N)^{1/(1-\alpha)},\ \alpha < 1, \text{ and } \quad N_n^* = \left(n + N - \frac{1}{2}\right)^{1/(1-\alpha)}.$$

Then

$$N_{n+1} - N_n = cN_n^\alpha + o(N_n^\alpha).$$

Fixing R_0 large enough and considering

$$\tau_0 = \inf\{n > N, |X_n| > R_0\} \wedge \tau,$$

we have

$$P_{N_0,y}\{\tau_0 < N_n\} = \sum_{i=1}^{n} P_{N_0,y}\{N_{i-1} \leq \tau_0 \leq N_i\}.$$

Let ρ be the stopping time defined by

$$\rho = \inf\left\{n > N_{i-1}^*, X_n \in B(S,l)\right\}.$$

Then

$$\begin{aligned} P_{N_0,y}\{N_{i-1} \leq \tau_0 \leq N_i\} &= P_{N_0,y}\left\{N_{i-1} \leq \tau_0 \leq N_i, \rho \leq N_{i-1}^* + r_{N_{i-1}^*}^{-1} T^*\right\} \\ &\quad + P_{N_0,y}\left\{N_{i-1} \leq \tau_0 < N_i, \rho > N_{i-1}^* + r_{N_{i-1}^*}^{-1} T^*\right\}. \end{aligned}$$

Thus for fixed large T^* and by Lemma 4.5, we achieve the upper bound for the second term on the right,

$$\begin{aligned} &P_{N_0,y}\left\{\left|X_{N_{i-1}^*}\right| \leq R_0, \rho > N_{i-1}^* + r_{N_{i-1}^*}^{-1} T^*\right\} \leq \exp\left\{-Mh_{N_{i-1}^*}^{-1}\right\} \\ &\leq \left(N_{i-1}^*\right)^{-\frac{1-\gamma}{d^*-\beta}M}, \end{aligned}$$

where M can be chosen as large as we want if T^* is large enough. The first term is less than

$$P_{N_0,y}\left\{0 \le \tau_0 - N_{i-1} < N_{i-1}^{\alpha} \exp\left\{\left(\frac{\alpha-\gamma}{1-\gamma}\left(d^*-\beta\right) h_{N_{i-1}}^{-1}\right)\right\}, \rho \le N_{i-1}^* + r_{N_{i-1}^*}^{-1} T^*\right\}$$
$$\le \exp\left\{-\delta h_{N_{i-1}}^{-1}\right\} \le \left(N_{i-1}\right)^{-\frac{\delta(1-\alpha)}{d^*-\beta}},$$

where we choose $\gamma < \alpha < \frac{\hat{d}}{d^*-\beta}(1-\gamma)+\gamma$. By Lemma 4.5, there exists δ such that $0 < \delta < \frac{1-\alpha}{1-\gamma}\left(d^*-\beta\right)+\beta$ and satisfying the above inequality. Hence we may take $\delta = \frac{1-\alpha}{1-\gamma}\left(d^*-\beta\right)+\beta/2$ and come to

$$\begin{aligned} P_{N_0,y}\left\{\tau_0 < N_n\right\} &\le \sum_{i=1}^{n} \left(i-1+N\right)^{-\frac{1}{1-\alpha}\left(1-\alpha+\frac{\beta(1-\gamma)}{2(d^*-\beta)}\right)} \\ &= \sum_{i=1}^{n} \left(i-1+N\right)^{-1-\frac{\beta(1-\gamma)}{2(1-\alpha)(d^*-\beta)}} \\ &\le c_1 \left(N-1\right)^{-\frac{\beta(1-\gamma)}{2(1-\alpha)(d^*-\beta)}}. \end{aligned}$$

Let $n \to +\infty$, we complete the proof of the proposition. □

PROPOSITION 5.4. *When $r_n = c/n + o\left(1/n\right)$, if there is $\beta > 0$ such that $h_n \le \left(d^*-\beta\right)/\ln\ln n$, then there exists a set $K \in S\backslash\underline{S}(\pi_K^{d^*-\beta} \cap \underline{S} = \emptyset)$, for any $l > 0$, there is $c_1 > 0$ such that for large N and $y \in B\left(K,l\right)$,*

$$P_{N,y}\left\{\tau < +\infty\right\} \le c_1 \left(\ln N\right)^{-\beta/2(d^*-\beta)},$$

where $\tau = \inf\{n > N, X_n \in B(S\backslash\pi_K^{d^*-\beta}, l)\}$.

Proof. The proof is similar to that of Proposition 5.3. □

Appendix A. In this section, some properties of $I(x,y)$ are studied. Lemma A.1 shows the connection of $I(x,y)$ and $U(\cdot)$; other lemmas are devoted to show the connection between $I(x,y)$ and the cycle given in [14] and reveal the information including the complex definition of cycle.

For any $x \in K \subset S$, $U\left(x\right)$ is a constant, which we also denote by $U(K)$.

LEMMA A.1. *If x and y belong to the same domain of attraction, then $I\left(x,y\right) = 2\left(U\left(y\right)-U\left(x\right)\right) \vee 0$. Otherwise, if $I\left(x,y\right) \ne 2\left(U\left(y\right)-U\left(x\right)\right) \vee 0$, then there exists $K \in S$ such that for any $\epsilon > 0$, we have $\varphi \in \mathcal{B}_{x,y}$ satisfying $e\left(\varphi\right) < I\left(x,y\right)+\epsilon$, which passes through one point of K, and $I(x,K) = I(x,y) = 2(U(K)-U(x))$.*

Proof. The first part is obvious. To prove the second part of the lemma, we define

$$A_n := \left\{z \in R^d,\, U\left(z\right) \le U\left(x\right) + \frac{I\left(x,y\right)}{2} + \frac{1}{n}\right\}.$$

By the continuity of $U(x)$ and the assumption of S consisting of a finite number of components, so has A_n. We also use A_n to denote the component containing x. By the definition of $I(x,y)$, y belongs to A_n, too. Thus,

$$A := \cap_{n=1}^{\infty} A_n \subseteq \left\{z \in R^d,\, U\left(z\right) \le U\left(x\right) + I\left(x,y\right)/2\right\}$$

is closed and connected. Denote

$$\tilde{A} := A \cap \left\{z \in R^d,\, U\left(z\right) = U\left(x\right) + \frac{I\left(x,y\right)}{2}\right\}.$$

We claim that there exists $z \in \tilde{A}$, $\nabla U(x) = 0$. In fact, if on the contrary, for all $z \in \tilde{A}$, $\nabla U(x) \neq 0$, then by the continuity of $\nabla U(x)$ and $\tilde{A}$ being a level set, for any $z_0 \in \tilde{A}$, there exists $l > 0$ such that for any $z \in B(z_0, l) \cap \tilde{A}$ and δ less than some δ_0, it holds that $U(z + \delta \cdot \nabla U(z)) > U(x) + I(x, y)$ and $U(z - \delta \cdot \nabla U(z)) < U(x) + I(x, y)$. These imply that there exists an $\tilde{\delta} > 0$ such that $B(z_0, \tilde{\delta}) \cap A$ is homeomorphic to $R^{d-1} \times [0, +\infty)$, which is path connected. Moreover, A is locally path connected, since $A \backslash \tilde{A}$ is open. Consequently, A is path connected. Therefore, there is a $\varphi(\subset A) \in \mathcal{B}_{x,y}$. Denote

$$B := \left\{ z \in R^d, U(z) < U(x) + \frac{I(x, y)}{2} \right\}.$$

Then B consists of finite number of connected open domains and x and y are in different ones. Let B_x be the component containing x and B_y be that containing y. Assume that if $\varphi(t)$, $0 \leq t \leq 1$, is the path satisfying $I(x, y) = 2(\max_{0 \leq t \leq 1} U(\varphi(t)) - U(x))$. Let $t_1 = \inf \left\{ t > 0, \varphi(t) \notin \bar{B}_x \right\}$. Obviously, $\varphi(t_1) = z_0 \in \partial \bar{B}_x$. We assert $\nabla U(z_0) = 0$ (which leads to contradiction). If $\nabla U(z_0) \neq 0$, as we show above, there is a $\delta > 0$ such that $U(z) > U(x) + I(x, y)$ for $z \in B(z_0, \delta) \backslash \bar{B}_x$. By $\varphi(t) \leq U(x) + I(y, x)$ there exists t_0 for any $t < t_1 + t_0$, $\varphi(t) \in \bar{B}_x$. This contradicts to the assumption of t_1. Hence $\nabla U(z_0) = 0$.

For any $\epsilon > 0$, let A_ϵ be one of the components of

$$\left\{ z \in R^d, U(z) < U(x) + I(x, y)/2 + \epsilon \right\}$$

containing x and y. Then A_ϵ is path connected. By the assertion above, there are $z_0 \in K \subset S$, $z_0 \in \tilde{A}$, and a path $\varphi \in A_\epsilon$ with $\varphi(0) = x$, $\varphi\left(\frac{1}{2}\right) = z_0$, $\varphi(1) = y$. Hence $I(x, K) \leq I(x, y)$. Because of $K \subset \tilde{A}$, we have $I(x, K) \geq I(x, y)$, i.e., $I(x, K) = I(x, y)$. The proof is complete. □

Here we will recall the definition of the hierarchy of cycles and some related notations which are given in [14], and show their connection with $I(x, y)$.

To go back to $V(x, y)$, we define

$$V(K_i, K_j) := \inf \left\{ \tilde{S}_{0,T}(\phi); \phi(0) \in K_i, \phi(T) \in K_j, \phi(t) \notin K_l, l \neq i, j, 0 < t < T \right\}.$$

In the same way, we define $V(x, K_j)$, $V(K_i, y)$. Let $V(K_i) := \min_{j \neq i} V(K_i, K_j)$ and denote $V(K_i, K_j) = V(K_i)$ by $K_i \Longrightarrow K_j$. All of these K_i's with relation "$\Longrightarrow$" define a graph which is still denoted by S. We say K_i is connected to K_j if there are $i_1, \ldots, i_n$ such that $i = i_1$, $j = i_n$, and $K_{i_1} \Longrightarrow K_{i_2} \Longrightarrow \cdots \Longrightarrow K_{i_n}$.

DEFINITION A.2. *A cycle π in S is a subgraph of S satisfying*

(1) *$K \in \pi$ and $K \Longrightarrow K'$ imply $K' \in \pi$;*

(2) *for any $K \neq K'$ in π, K is connected to K' in π.*

DEFINITION A.3. *Let*

$$S^1 = \{\pi : \pi \text{ is a cycle in } S\} \cup \{K : K \in S \text{ and } K \text{ is not in any cycle}\}.$$

An element of S^1 is called a 1-cycle(in S). We use π^1 to denote a 1-cycle.

For each $\pi \in S^1$, let $\hat{V}(\pi) = \max\{V(K); K \in \pi\}$. For $\pi_1, \pi_2 \in S^1$, and $\pi_1 \neq \pi_2$, define

$$V(\pi_1, \pi_2) = \hat{V}(\pi_1) + \min\{V(K_1, K_2) - V(K_1); K_1 \in \pi_1, K_2 \in \pi_2\},$$
$$V(\pi_1) = \min\{V(\pi_1, \pi_2); \pi_1 \neq \pi_2\},$$

and $\pi_1 \Longrightarrow \pi_2$ if $V(\pi_1, \pi_2) = V(\pi_1)$.

To use S^1 instead of S, we can define S^2 and 2-cycles and use the notation π^2 for a specified one. Inductively, if we have defined up to m-cycles π^m in S_m, we will give the following for $m+1$:

$$\hat{V}(\pi^m) = \max\left\{V\left(\pi^{m-1}\right); \pi^{m-1} \in \pi^m\right\},$$
$$V(\pi_1^m, \pi_2^m) = \hat{V}(\pi_1^m) + \min\left\{V(\pi_1^{m-1}, \pi_2^{m-1}) - V(\pi_1^{m-1}); \pi_1^{m-1} \in \pi_1^m, \pi_2^{m-1} \in \pi_2^m\right\},$$
$$V(\pi^m) = \min\left\{V(\pi^m, \hat{\pi}^m); \pi^m \neq \hat{\pi}^m, \hat{\pi}^m \in S^m\right\},$$

and $\pi^m \Longrightarrow \hat{\pi}^m$ if $V(\pi^m) = V(\pi^m, \hat{\pi}^m)$. Then the $(m+1)$-cycle is constructed in the same way as in the 1-cycle case.

For convenience, π^k stands for

$$\cup_{\pi^{k-1}\in\pi^k}\pi^{k-1} \text{ or } \cup_{\pi^{k-1}\in\pi^k}\left(\cup\pi_{k-2} \in \pi_{k-1}\left(\cdots\left(\cup_{\pi^1\in\pi^2}\pi^1\right)\cdots\right)\right).$$

Set $I\left(\pi_1^k, \pi_1^k\right) = \sup_{x\in\pi_1^k}\inf_{y\in\pi_2^k} I(x,y)$ and $I\left(\pi_1^k\right) = \sup_{x\in\pi_1^k}\inf_{y\in S\setminus\pi_1^k} I(x,y)$. Recalling the definition of π_K^c and $I(\pi_K^c)$ in section 4, we have the following lemma, which can be verified by using Lemma A.1 and the cycle decomposition of [23].

LEMMA A.4. *For any $K \in S$, $c > 0$. there exist k and j such that $\pi_K^c = \pi_k^j$ and $V(\pi_k^j) = I(\pi_K^c)$.*

Let $G_{\underline{S}_l}(\underline{S}_c)$, $c < l$, denote the set of all $\underline{S}_c$-graphs in $\underline{S}_l$ (cf. [5], [14]) and

$$V(g) = \sum_{(m\to n)\in g} V(K_m, K_n), \qquad g \in G_{\underline{S}_l}(\underline{S}_c),$$

where $m \to n$ denotes a directed edge of g. (If $\underline{S}_l \backslash \underline{S}_c = \emptyset, V(g) := 0$.)

The following lemma is used to prove Lemma 4.6. To prove it, we need Lemmas A.6 and A.7. However, the detail is omitted; readers are also referred to [23].

LEMMA A.5. *For any $K_i \in \underline{S}_l \backslash \underline{S}_c$,*

$$\begin{aligned}&\min\left\{V(g); g \in G_{\underline{S}_l}(\underline{S}_c)\right\} - \min\left\{V(g); g \in G_{\underline{S}_l}(\underline{S}_c \cup \{K_i\})\right.\\ &\left.\text{or } g \in G_{\underline{S}_l\{K_iK_j\}}(\underline{S}_c \cup \{K_j\}), K_i \neq K_j, K_j \in \underline{S}_l \backslash \underline{S}_c\right\}\\ &\leq \max_{x\in\underline{S}_l\backslash\underline{S}_c}\min_{y\in\underline{S}_c} I(x,y),\end{aligned}$$

where $G_{\underline{S}_l\{K_iK_j\}}(\underline{S}_c \cup \{K_j\})$ denotes the set of $\underline{S}_c \cup \{K_j\}$-graphs in $\underline{S}_l$, in which the sequence of arrows leading from K_i into $\underline{S}_c \cup \{K_j\}$ ends at K_j.

LEMMA A.6. *For any $g_0 \in G_{\underline{S}_l}(\underline{S}_c)$, there is $g \in G_{\underline{S}_l}(\underline{S}_c)$ such that $V(g) \leq V(g_0)$, and for any $\pi^1 \in \underline{S}_l \backslash \underline{S}_c$, we can find $K_0 \in \pi^1$ such that there is a graph h satisfying $h \in G_{\pi^1}(K_0), h \subset g$.*

For $A \in S^k \cap \underline{S}_l$, denote by $G_{\underline{S}_l}^A(\underline{S}_l)$ a new $\underline{S}_c$-graph in $\underline{S}_l$, in which A is considered globally as a node in $\underline{S}_l$; then we have the following lemma.

LEMMA A.7. *For any $K_i \in \underline{S}_l \backslash \underline{S}_c$, if $K_i \in A$, then*

$$\begin{aligned}&\min\left\{V(g); g \in G_{\underline{S}_l}(\underline{S}_c)\right\} - \min\left\{V(g); g \in G_{\underline{S}_l}(\underline{S}_c \cup \{K_i\})\right.\\ &\qquad\left.\text{or } g \in G_{\underline{S}_l\{K_iK_j\}}(\underline{S}_c \cup \{K_j\}), K_i \neq K_j, K_j \in \underline{S}_l \backslash \underline{S}_c\right\}\\ &= \min\left\{V(g); g \in G_{\underline{S}_l}^A(\underline{S}_c)\right\} - \min\left\{V(g); g \in G_{\underline{S}_l}^A(\underline{S}_c \cup \{A\})\right.\\ &\qquad\left.\text{or } g \in G_{\underline{S}_l\{AK_j\}}^A(\underline{S}_c \cup \{K_j\}), K_j \notin A, K_j \in \underline{S}_l \backslash \underline{S}_c\right\};\end{aligned} \tag{A.1}$$

otherwise,

$$\textit{left-hand side} = \min\left\{V(g); g \in G^A_{\underline{S}_l}(\underline{S}_c)\right\} - \min\left\{V(g); g \in G^A_{\underline{S}_l}(\underline{S}_c \cup \{K_i\}) \textit{ or}\right.$$

$$\left. g \in G^A_{\underline{S}_l\{K_iK_j\}}(\underline{S}_c \cup \{K_j\}), K_j \notin \underline{S}_c \cup A \cup \{K_i\} \textit{ or } g \in G^A_{\underline{S}_l\{K_iA\}}(\underline{S}_c \cup A)\right\}. \quad \text{(A.2)}$$

Acknowledgments. The authors are indebted to the referee whose suggestions helped make some technical parts more readable, correct a number of misprints, and complete the references.

REFERENCES

[1] A. Benveniste, M. Métivier, and P. Priouret, *Adaptive Algorithms and Stochastic Approximations*, Springer-Verlag, Berlin, 1990.
[2] O. Catoni, *Rough large deviation estimates for simulated annealing, application to exponential schedules*, Ann. Probab., 20 (1992), pp. 1109–1146.
[3] T. S. Chiang, C. R. Hwang, and S. J. Sheu, *Diffusions for global optimization in R^d*, SIAM J. Control Optim., 25 (1987), pp. 737–752.
[4] J. D. Deuschel and C. Mazza, *L^2 convergence of time non-homogeneous Markov processes: I Spectral estimates*, Ann. Appl. Probab., 4 (1994), pp. 1012–1056.
[5] M. I. Freidlin and A. D. Wentzell, *Random Perturbation of Dynamical Systems*, Springer-Verlag, Berlin, 1984.
[6] S. B. Gelfand and S. K. Mitter, *Metropolis-type annealing algorithms for global optimization in R^d*, SIAM J. Control Optim., 31 (1993), pp. 111–131.
[7] S. B. Gelfand and S. K. Mitter, *Recursive stochastic algorithm for global optimization in R^d*, SIAM J. Control Optim., 29 (1991), pp. 999–1018.
[8] S. Geman and C. R. Hwang, *Diffusions for global optimization*, SIAM J. Control Optim., 24 (1987), pp. 1031–1043.
[9] D. E. Goldberg, *Genetic Algorithms in Search, Optimization and Machine Learning*, Addison-Wesley, Reading, MA, 1989.
[10] F. Götze, *Rate of Convergence of Simulated Annealing Processes*, University of Bielefeld, Germany, 1992, preprint.
[11] B. Hajek, *Cooling schedules for optimal annealing*, Math. Oper. Res., 13 (1988), pp. 311–329.
[12] J. Hertz, A. Krogh, and G. R. Palmar, *Introduction to the Theory of Neural Computation*, Santa Fe Inst. Studies in the Science of Complexity, Addison-Wesley, Reading, MA, 1991.
[13] R. Holley, S. Kusuoka, and D. Strook, *Asymptotic of the spectral gaps with application to the theory of simulated annealing*, J. Funct. Anal., 83 (1989), pp. 333–347.
[14] C. R. Hwang and S. J. Sheu, *Large-time behavior of perturbed diffusion Markov processes with applications to the second eigenvalue problem for Fokker-Planck operators and simulated annealing*, Acta Appl. Math., 19 (1990), pp. 253–295.
[15] *IEEE Transactions on Neural Networks*, 5, 1 (special issue on evolutionary computation), 1994.
[16] S. Kirpatrick, C. D. Gelatt, and M. P. Vecchi, *Optimization by simulated annealing*, Science, 220 (1983), pp. 671–680.
[17] K. Kohonen, *Self-Organization and Associative Memory*, Springer-Verlag, Berlin, 1988.
[18] H. J. Kushner, *Asymptotic global behavior for stochastic approximation and diffusions with slowly decreasing noise effects: Global minimization via Monte Carlo*, SIAM J. Appl. Math., 47 (1987), pp. 169–185.
[19] L. Ljung, G. Pflug, and H. Walk, *Stochastic Approximation and Optimization of Random System*, Birkhäuser-Verlag, Basel, 1992.
[20] D. Márquez and M. Pelletier, *Sur la vitesse de convergence en loi du recuit simulé*, C. R. Acad. Sci. Paris Sér. Math. I, 322 (1996), pp. 391–394.
[21] Z. Michalwicz, *Genetic Algorithms + Datastructures = Evolution Programs*, Springer-Verlag, Berlin, 1992.
[22] L. Miclo, *Recuit simulé sans potentiel sur un ensemble fini*, in Seminaire de Probabilités, vol. 26, Springer-Verlag, New York, 1992, pp. 47–60.
[23] A. Trouve, *Cycle decompositions and simulated annealing*, SIAM J. Control Optim., 34 (1996), pp. 966–986.
[24] J. N. Tsitsiklis, *Markov chains with rare transitions and simulated annealing*, Math. Oper. Res., 14 (1989), pp. 70–90.
[25] A. D. Wentzell, *Limit Theorems on Large Deviations for Markov Stochastic Processes*, Kluwer, Dordrecht, The Netherlands, 1990.

SIAM J. CONTROL OPTIM.
Vol. 35, No. 6, pp. 1908–1923, November 1997

004

EXACT FINITE-DIMENSIONAL FILTERS FOR MAXIMUM LIKELIHOOD PARAMETER ESTIMATION OF CONTINUOUS-TIME LINEAR GAUSSIAN SYSTEMS*

ROBERT J. ELLIOTT† AND VIKRAM KRISHNAMURTHY‡

Abstract. In this paper, we derive a new class of finite-dimensional filters for integrals and stochastic integrals of moments of the state for continuous-time linear Gaussian systems. Apart from being of significant mathematical interest, these new filters can be used with the expectation maximization (EM) algorithm to yield maximum likelihood estimates of the model parameters.

Key words. stochastic systems, finite-dimensional filters, Kalman filter, expectation maximization algorithm, parameter estimation

AMS subject classifications. 93E11, 93E12, 93E10, 60G35

PII. S036301299529255X

1. Introduction.

The Kalman filter is widely used in engineering, economics, and other fields. It considers linear Gaussian dynamics for the state x and observation y processes of the form

$$dx_t = A_t\, x_t\, dt + B_t\, dw_t, \quad x_0 \in \mathbb{R}^m, \tag{1.1}$$

$$dy_t = C_t\, x_t\, dt + D_t\, dv_t, \quad y_0 = 0 \in \mathbb{R}^n. \tag{1.2}$$

The Kalman filter determines the conditional density of the unobserved signal x_t given the observations to time t, $\mathcal{Y}_t = \sigma\{y_s : s \le t\}$, as a Gaussian random variable with mean m_t and variance Σ_t. Here m_t is the conditional mean of x_t given $\mathcal{Y}_t$, and Σ_t is the conditional variance of x_t, although Σ_t turns out to be deterministic and given by the Riccati equation.

To apply the Kalman filter, however, the parameters of the model, that is, the entries in the matrices A, B, C, and D, need to be known. Maximum likelihood estimation of these parameters via the expectation maximization (EM) algorithm has been studied in discrete time in [1], [2], [3] and in continuous time in [9]. In continuous time, the EM algorithm requires computation of the filtered estimates of quantities such as $\int_0^t x_s\, dx'_s$, $\int_0^t x_s\, dy'_s$, $\int_0^t x_s\, x'_s\, ds$.

In all the existing literature on parameter estimation of linear Gaussian models via the EM algorithm, filtered estimates of the above quantities are computed via Kalman smoothing, which requires large memory in any numerical implementation. The main result of this paper, Theorem 3.10, provides *finite-dimensional filters for (the components of) such integral processes.* In fact, as pointed out in section 4, finite-dimensional filters exist for integrals and stochastic integrals of moments of all orders

*Received by the editors September 29, 1995; accepted for publication (in revised form) August 2, 1996.

http://www.siam.org/journals/sicon/35-6/29255.html

†Department of Mathematical Sciences, University of Alberta, Edmonton, T6G 2G1, AB, Canada (relliott@gpu.srv.ualberta.ca). The research of this author was supported in part by NSERC grant A7964, the University of Melbourne, and the Co-operative Research Center for Sensor Signal and Information Processing.

‡Department of Electrical and Electronic Engineering, University of Melbourne, Parkville, Victoria 3052, Australia (vikram@mullian.ee.mu.oz.au). The research of this author was supported by an ARC grant, ATERB, and the Co-operative Research Center for Sensor Signal and Information Processing.

of the state process. In [18] and [19], finite-dimensional filters are described using Lie algebra methods, for time integrals of powers of the state variable. However, our techniques are quite different; we also obtain results for stochastic integrals of powers of the process, and we apply our results to maximum likelihood parameter estimation. Few finite-dimensional filters are known, so our results were certainly a surprise.

The analogous results in discrete time are derived in a companion paper [17]. The continuous-time methods and results in this paper are different from, and not just adaptations of, the discrete-time results in [17]. As described in [17], estimation of the state-space parameters arises in several areas of signal processing and control, including multisensor signal enhancement and speech coding, and also in econometrics.

Parameter estimation for diffusion processes is nicely discussed in [9]; this work also includes a discussion of the EM algorithm. Section 5 describes this application of the results of section 4. The techniques of this paper extend those for Markov chains presented in [5] and the recent book, [6].

The paper is organized as follows: in section 2 we present a standard measure change that simplifies the derivation of our filters. In section 3 the new finite-dimensional filters are derived. Section 4 derives a finite-dimensional filter for higher order moments of the state. In section 5, the filters in section 3 are used to implement a filter-based EM algorithm for maximum likelihood parameter estimation.

2. Dynamics. Consider the classical linear Gaussian model for the signal model and observation processes. That is, the signal $\{x_t\}$, $t \geq 0$, is described by the equation

$$dx_t = A_t\, x_t\, dt + B_t\, dw_t, \qquad x_0 \in \mathbb{R}^m, \tag{2.1}$$

and the observation process $\{y_t\}$, $t \geq 0$, is described as

$$dy_t = C_t\, x_t\, dt + D_t\, dv_t, \qquad y_0 = 0 \in \mathbb{R}^n. \tag{2.2}$$

Here w and v are independent r-dimensional and n-dimensional Brownian motions, respectively, defined on a probability space $(\Omega, \mathcal{F}, P)$. Further, w and v are independent of x_0. We assume that x_0 is a random variable with normal density $\pi_0(x)$, which is $N(\hat{x}_0, P_0)$.

The matrix functions $A_t \in \mathbb{R}^{m\times m}$, $B_t \in \mathbb{R}^{m\times r}$, $C_t \in \mathbb{R}^{n\times m}$, and $D_t \in \mathbb{R}^{n\times n}$ are measurable functions of t. We assume D_t is a positive definite matrix.

We model the above dynamics by supposing that initially we have a probability space $(\Omega, \mathcal{F}, \bar{P})$ such that under $\bar{P}$

1. w is r-dimensional Brownian motion and $\{x_t\}$ is defined by (2.1).
2. $\{y_t\}$ is n-dimensional Brownian motion, independent of w and x_0, and having quadratic variation $\langle y\rangle_t = D_t > 0$; i.e., D_t is a positive definite matrix.

Notation. Consider the complete right continuous filtrations

$$\mathcal{G}_t = \sigma\{x_s, y_s : s \leq t\},$$
$$\mathcal{Y}_t = \sigma\{y_s : s \leq t\}.$$

Vectors x, y will be considered as column vectors. Write

$$\nabla = \left(\frac{\partial}{\partial x_1}, \frac{\partial}{\partial x_2}, \ldots, \frac{\partial}{\partial x_n}\right)'. \tag{2.3}$$

For any function $g : \mathbb{R}^m \to \mathbb{R}$, write

$$\nabla^2 g = \begin{bmatrix} \frac{\partial^2 g}{\partial x_1^2} & \cdots & \frac{\partial^2 g}{\partial x_1 \partial x_m} \\ \vdots & \vdots & \vdots \\ \frac{\partial^2 g}{\partial x_m \partial x_1} & \cdots & \frac{\partial^2 g}{\partial x_m^2} \end{bmatrix}.$$

For a vector field $g(x) = [g_1(x)\ g_2(x)\ \cdots\ g_m(x)]'$ defined on $\mathbb{R}^m$, define

$$\operatorname{div}(g) = \frac{\partial g_1}{\partial x_1} + \frac{\partial g_2}{\partial x_2} + \cdots + \frac{\partial g_m}{\partial x_m}.$$

Consider the exponential

$$\Lambda_t = \exp\left(\int_0^t (C_s\, x_s)' \left(D_s^{-1}\right)' D_s^{-1}\, dy_s - \frac{1}{2} \int_0^t x_s'\, C_s' \left(D_s^{-1}\right)' D_s^{-1}\, C_s\, x_s\, ds \right). \tag{2.4}$$

Then

$$d\Lambda_t = \Lambda_t\, x_t'\, C_t' \left(D_s^{-1}\right)' D_s^{-1}\, dy_t \tag{2.5}$$

and $\bar{\mathbf{E}}\{\Lambda_t\} = 1$, where $\bar{\mathbf{E}}$ denotes expectation under $\bar{P}$.

If we define a measure P in terms of $\bar{P}$ by setting

$$\left.\frac{dP}{d\bar{P}}\right|_{\mathcal{G}_t} = \Lambda_t,$$

then Girsanov's theorem implies that under P, v_t is a standard n-dimensional Brownian motion if we define

$$dv_t = D_t^{-1}\,(dy_t - C_t\, x_t\, dt), \quad v_0 = 0.$$

That is, under P,

$$dy_t = C_t\, x_t\, dt + D_t\, dv_t.$$

Note that under P, the process $\{x_t\}$ still satisfies (2.1). Consequently, under P the processes $\{x_t\}$ and $\{y_t\}$ satisfy the real world dynamics (2.1) and (2.2). However, $\bar{P}$ is a more convenient measure with which to work.

Below, we assume that $g : \mathbb{R}^m \to \mathbb{R}$ is an arbitrary "test" function, which is in $\mathbb{C}^2$ and has compact support.

Filtering is concerned with estimates of the form $\mathbf{E}\{g(x_t)|\mathcal{Y}_t\}$. Using a version of Bayes's theorem [6], we have

$$\mathbf{E}\{g(x_t)|\mathcal{Y}_t\} = \frac{\bar{\mathbf{E}}\{\Lambda_t\, g(x_t)|\mathcal{Y}_t\}}{\bar{\mathbf{E}}\{\Lambda_t|\mathcal{Y}_t\}}. \tag{2.6}$$

Write $\sigma(g)_t = \bar{\mathbf{E}}\{\Lambda_t\, g(x_t)|\mathcal{Y}_t\}$ so that

$$\mathbf{E}\{g(x_t)|\mathcal{Y}_t\} = \frac{\sigma(g)_t}{\sigma(1)_t}.$$

Consequently, $\sigma(g)_t$ is a measure-valued process; it is an unnormalized conditional expectation of $g(x_t)$ given $\mathcal{Y}_t$.

Suppose that $\sigma(\cdot)_t$ has a density $\alpha_t(x)$; then

$$\sigma(g)_t = \bar{\mathbf{E}}\{\Lambda_t\, g(x_t)|\mathcal{Y}_t\} = \int_{\mathbb{R}^m} g(x)\,\alpha_t(x)\,dx.$$

We now give a proof of the well-known Zakai equation for α.

THEOREM 2.1.

$$\begin{aligned}\alpha_t(x) = \alpha_0(x) &- \int_0^t \operatorname{div}\left(A_s\, x\, \alpha_s(x)\right)\, ds + \frac{1}{2}\int_0^t \operatorname{Tr}\left(\nabla^2\alpha_s(x)\, B_s\, B_s'\right)\, ds \\ &+ \int_0^t \alpha_s(x)\, x'\, C_s'\left(D_s^{-1}\right)'\, D_s^{-1}\, dy_s,\end{aligned} \tag{2.7}$$

where $\alpha_0(x) = \pi_o(x)$, the density of x_0.

Proof. From Ito's rule

$$\begin{aligned}g(x_t) = g(x_0) &+ \int_0^t \left(\nabla g(x_s)\right)'\, A_s\, x_s\, ds + \int_0^t \left(\nabla g(x_s)\right)'\, B_s\, dw_s \\ &+ \frac{1}{2}\int_0^t \operatorname{Tr}\left(\nabla^2 g(x_s)\, B_s\, B_s'\right)\, ds.\end{aligned} \tag{2.8}$$

From (2.5) and (2.8)

$$\begin{aligned}\Lambda_t\, g(x_t) = g(x_0) &+ \int_0^t \Lambda_s\, \left(\nabla g(x_s)\right)'\, A_s\, x_s\, ds + \int_0^t \Lambda_s\, \left(\nabla g(x_s)\right)'\, B_s\, dw_s \\ &+ \frac{1}{2}\int_0^t \Lambda_s \operatorname{Tr}\left(\nabla^2 g(x_s) B_s\, B_s'\right) ds + \int_0^t \Lambda_s\, g(x_s)\, x_s'\, C_s'\left(D_s^{-1}\right)'\, D_s^{-1}\, dy_s.\end{aligned} \tag{2.9}$$

Conditioning each side of (2.9) on $\mathcal{Y}_t$ (see Lemma 3.2, p. 261, of [13]) we have

$$\begin{aligned}\sigma(g)_t = \sigma(g)_0 &+ \int_0^t \sigma\left(\left(\nabla g\right)'\, A\, x\right)_s\, ds + \frac{1}{2}\int_0^t \sigma\left(\operatorname{Tr}(\nabla^2 g) B_s\, B_s'\right)\, ds \\ &+ \int_0^t \sigma(g\, x')_s\, C_s'\left(D_s^{-1}\right)'\, D_s^{-1}\, dy_s.\end{aligned} \tag{2.10}$$

Integrating each term by parts as on p. 277 of [7] gives (2.7). □

Remarks. The Zakai equation (2.7) is a stochastic partial differential equation for the unnormalized conditional density of x_t given $\mathcal{Y}_t$. In general, the solution of this equation is a conditional density function, evolving stochastically in time. For the linear, Gaussian dynamics (2.1) and (2.2), however, $\alpha_t(x)$ has a simple form. In fact (see [10] or [11]),

$$\alpha_t(x) = \frac{\nu_t}{(2\,\pi)^{m/2}\,|\Sigma_t|^{1/2}}\, \exp\left(-\frac{1}{2}(x - m_t)'\,\Sigma_t^{-1}\,(x - m_t)\right). \tag{2.11}$$

Here $m_t = \mathbf{E}\{x_t|\mathcal{Y}_t\}$, $m_0 = \hat{x}_0$, $\Sigma_t = \mathbf{E}\{(x_t - m_t)(x_t - m_t)'|\mathcal{Y}_t\}$, $\Sigma_0 = P_0$, and ν_t is a normalizing factor.

It is well known that m_t and Σ_t are given by the Kalman filter equations

$$dm_t = A_t\, m_t\, dt + \Sigma_t\, C_t'\left(D_t^{-1}\right)'\, D_t^{-1}\left(dy_t - C_t\, m_t\, dt\right), \tag{2.12}$$

$$\dot{\Sigma}_t = \Sigma_t\, A_t' + A_t\, \Sigma_t + B_t\, B_t' - \Sigma_t\, C_t'\left(D_t^{-1}\right)'\, D_t^{-1}\, C_t\, \Sigma_t. \tag{2.13}$$

Note that Σ_t is deterministic and can be computed off-line. Also,

(2.14)
$$\nu_t = (2\,\pi)^{m/2} \exp\left(\int_0^t m_s'\, C_s' \big(D_s^{-1}\big)'\, D_s^{-1}\, dy_s - \frac{1}{2} \int_0^t m_s'\, C_s'\, \big(D_s^{-1}\big)'\, D_s^{-1}\, C_s\, m_s\, ds \right).$$

3. Finite-dimensional filters. Let e_i, $e_j \in \mathbb{R}^m$ denote unit vectors with 1 in the ith and jth positions, respectively. Write

$$H_t^{ij} = \int_0^t \langle x_s, e_i \rangle \, \langle e_j, x_s \rangle \, ds = \int_0^t x_s' \, (e_i \, e_j') \, x_s \, ds, \tag{3.1}$$

$$L_t^{ij} = \int_0^t \langle x_s, e_i \rangle \, \langle e_j, dx_s \rangle = \int_0^t x_s' \, (e_i \, e_j') \, dx_s; \tag{3.2}$$

here $\langle \cdot, \cdot \rangle$ denotes the scalar product.

Also let $f_j \in \mathbb{R}^n$ denote the unit vector with 1 in the jth position. Write

$$J_t^{ij} = \int_0^t \langle x_s, e_i \rangle \, \langle f_j, dy_s \rangle = \int_0^t x_s' \, (e_i \, f_j') \, dy_s. \tag{3.3}$$

The parameters of our model are the entries in the matrices A_t, B_t, C_t, and D_t. To estimate these parameters using the EM algorithm (see section 5), it is necessary to obtain filtered estimates of these processes. That is, we wish to obtain expressions for

$$\mathbf{E}\{H_t^{ij} | \mathcal{Y}_t\}, \quad \mathbf{E}\{L_t^{ij} | \mathcal{Y}_t\}, \quad \mathbf{E}\{J_t^{ij} | \mathcal{Y}_t\}.$$

Previously, these estimates have been obtained by smoothing procedures. For example,

$$\mathbf{E}\{H_t^{ij} | \mathcal{Y}_t\} = \int_0^t \mathbf{E}\{x_s' \, (e_i \, e_j) \, x_s | \mathcal{Y}_t\} \, ds.$$

However, this involves a considerable memory requirement. In this section we prove the remarkable result that the filtered estimates for H_t^{ij}, L_t^{ij}, and J_t^{ij} can be described in terms of a finite number of statistics.

Motivated by the techniques in [6], we define a measure associated with H_t^{ij} as follows.

DEFINITION 3.1. *For any test function $g : \mathbb{R}^m \to \mathbb{R}$, define a measure-valued process $\bar{\mathbf{E}}\{\Lambda_t \, H_t^{ij} \, g(x_t) | \mathcal{Y}_t\}$. This has a density $\beta_t^{ij}(x)$ so that*

$$\bar{\mathbf{E}}\{\Lambda_t \, H_t^{ij} \, g(x_t) | \mathcal{Y}_t\} = \int_{\mathbb{R}^m} \beta_t^{ij}(x) \, g(x) \, dx.$$

The existence of the density $\beta_t^{ij}(x)$ follows from the existence and uniqueness of solutions of stochastic partial differential equations such as (3.9). This is established in section 4.2 of [10] and on page 140 of [12].

The following theorem shows the surprising result that we can describe the measure $\beta_t^{ij}(x)$ exactly as a quadratic in x multiplying the $\alpha_t(x)$ of (2.7).

THEOREM 3.2. *At time t, the density $\beta_t^{ij}(x)$ is completely described by the five statistics a_t^{ij}, b_t^{ij}, c_t^{ij}, Σ_t, and m_t as follows:*

$$\beta_t^{ij}(x) = \left(a_t^{ij} + x' \, b_t^{ij} + x' \, c_t^{ij} \, x \right) \alpha_t(x). \tag{3.4}$$

Here $a_t^{ij} \in \mathbb{R}$, $b_t^{ij} \in \mathbb{R}^m$, and $c_t^{ij} \in L_s(\mathbb{R}^m, \mathbb{R})$, the space of symmetric $m \times m$ matrices. Further,

$$\frac{da_t^{ij}}{dt} = \operatorname{Tr}\left(c_t^{ij}\, B_t\, B_t'\right) + b_t^{ij\prime}\, B_t\, B_t'\, \Sigma_t^{-1}\, m_t, \quad a_0^{ij} = 0 \in \mathbb{R}, \tag{3.5}$$

$$\frac{db_t^{ij}}{dt} = -\left(A_t' + \Sigma_t^{-1}\, B_t\, B_t'\right) b_t^{ij} + 2\, c_t^{ij}\, B_t\, B_t'\, \Sigma_t^{-1}\, m_t, \quad b_0^{ij} = 0 \in \mathbb{R}^m, \tag{3.6}$$

$$\frac{dc_t^{ij}}{dt} = -\left(A_t' + \Sigma_t^{-1}\, B_t\, B_t'\right) c_t^{ij} - c_t^{ij}\left(A_t + B_t\, B_t'\, \Sigma_t^{-1}\right) + \frac{1}{2}(e_j\, e_i' + e_i\, e_j'),$$

$$c_0^{ij} = 0 \in L_s(\mathbb{R}^m, \mathbb{R}). \tag{3.7}$$

Proof. First note that for any test function g, applying Ito's rule to (2.1) and (2.5),

$$\begin{aligned}\Lambda_t\, H_t^{ij}\, g(x_t) = {} & \int_0^t \Lambda_s\, H_s^{ij}\ (\nabla g(x_s))'\, A_s\, x_s\, ds + \int_0^t \Lambda_s\, H_s^{ij}\ (\nabla g(x_s))'\, B_s\, dw_s \\ & + \frac{1}{2}\int_0^t \Lambda_s\, H_s^{ij} \operatorname{Tr}\left(\nabla^2 g(x_s)\, B_s\, B_s'\right) ds \\ & + \int_0^t \Lambda_s\, H_s^{ij}\, g(x_s)\, x_s'\, C_s'\left(D_s^{-1}\right)' D_s^{-1}\, dy_s \\ & + \int_0^t \Lambda_s\, g(x_s)\, x_s'\, (e_i\, e_j')\, x_s\, ds. \end{aligned} \tag{3.8}$$

Conditioning on $\mathcal{Y}_t$ under $\bar{P}$ (see Lemma 3.2, p. 261, of [13]) we have

$$\begin{aligned}\bar{\mathbf{E}}\{\Lambda_t\, H_t^{ij}\, g(x_t)|\mathcal{Y}_t\} = {} & \int_0^t \bar{\mathbf{E}}\{\Lambda_s\, H_s^{ij}\ (\nabla g(x_s)'\, A_s\, x_s)\,|\mathcal{Y}_s\}\, ds \\ & + \frac{1}{2}\int_0^t \bar{\mathbf{E}}\{\Lambda_s\, H_s^{ij} \operatorname{Tr}\left(\nabla^2 g(x_s) B_s\, B_s'\right)|\mathcal{Y}_s\}\, ds \\ & + \int_0^t \bar{\mathbf{E}}\{\Lambda_s\, H_s^{ij}\, g(x_s)\, x_s'\, C_s'\left(D_s^{-1}\right)' D_s^{-1}|\mathcal{Y}_s\}\, dy_s \\ & + \int_0^t \bar{\mathbf{E}}\{\Lambda_s\, g(x_s)\, x_s'\, (e_i\, e_j')\, x_s|\mathcal{Y}_s\}\, ds. \end{aligned}$$

That is, in terms of the densities $\beta_t^{ij}(x)$ and $\alpha_t(x)$,

$$\begin{aligned}\int_{\mathbb{R}^m} \beta_t^{ij}(x)\, g(x)\, dx = {} & \int_0^t \int_{\mathbb{R}^m} \beta_s^{ij}(x)\ (\nabla g(x))'\ A_s\, x\, dx\, ds \\ & + \frac{1}{2}\int_0^t \int_{\mathbb{R}^m} \beta_s^{ij}(x) \operatorname{Tr}\left(\nabla^2 g(x)\, B_s\, B_s'\right) dx\, ds \\ & + \int_0^t \int_{\mathbb{R}^m} \beta_s^{ij}(x)\, g(x)\, x'\, C_s'\left(D_s^{-1}\right)' D_s^{-1}\, dx\, dy_s \\ & + \int_0^t \int_{\mathbb{R}^m} \alpha_s(x)\, g(x)\, x'\, (e_i\, e_j')\, x\, dx\, ds. \end{aligned}$$

Integrating by parts in x, because this equation holds for all test functions g, we see

that $\beta_t^{ij}(x)$ must satisfy the stochastic partial differential equation:

$$\begin{aligned}\beta_t^{ij}(x) = &-\int_0^t \operatorname{div}\left(\beta_s^{ij}(x)\, A_s\, x\right)\, ds + \frac{1}{2}\int_0^t \operatorname{Tr}\left(\nabla^2\beta_s^{ij}(x)\, B_s\, B_s'\right)\, ds \\ &+\int_0^t \beta_s^{ij}(x)\, x'\, C_s'\left(D_s^{-1}\right)' D_s^{-1}\, dy_s + \int_0^t \alpha_s(x)\, x'\,(e_i\, e_j')\, x\, ds. \end{aligned} \tag{3.9}$$

We look for a solution of (3.9) of the form

$$\bar\beta_s(x) = \left(a_s^{ij} + x' b_s^{ij} + x'\, c_s^{ij}\, x\right)\alpha_s(x). \tag{3.10}$$

As noted just after Definition 3.1, if such a solution exists, it is unique.

To simplify notation we drop the superscripts i, j on a, b, and c. Then

$$\begin{aligned}\operatorname{div}\left(\bar\beta_s(x)\, A_s\, x\right) &= \operatorname{div}\left((a_s + b_s'\, x + x'\, c_s\, x)\,\alpha_s(x)\, A_s\, x\right) \\ &= (b_s + 2\, c_s\, x)'\, A_s\, x\, \alpha_s(x) + (a_s + b_s'\, x + x'\, c_s\, x)\operatorname{div}\left(\alpha_s(x)\, A_s\, x\right), \\ \nabla\bar\beta_s(x) &= \nabla\left((a_s + b_s'\, x + x'\, c_s\, x)\,\alpha_s(x)\right) \\ &= (b_s + 2\, c_s\, x)\,\alpha_s(x) + (a_s + b_s'\, x + x'\, c_s\, x)\,\nabla\alpha_s(x), \\ \nabla^2\bar\beta_s(x) &= 2\, c_s\,\alpha_s(x) + 2\,(b_s + 2\, c_s\, x)(\nabla\alpha_s(x))' \\ &\quad + (a_s + b_s'\, x + x'\, c_s\, x)\,\nabla^2\alpha_s(x), \\ \operatorname{Tr}\left(\nabla^2\bar\beta_s(x)\, B_s\, B_s'\right) &= 2\,\alpha_s(x)\operatorname{Tr}\left(c_s\, B_s\, B_s'\right) + 2\,(b_s + 2\, c_s\, x)'\, B_s\, B_s'\,\nabla\alpha_s(x) \\ &\quad + (a_s + b_s'\, x + x'\, c_s\, x)\,\operatorname{Tr}\left(\nabla^2\alpha_s(x)\, B_s\, B_s'\right). \end{aligned}$$

Now from (2.11)

$$\nabla\alpha_s(x) = -\Sigma_s^{-1}(x - m_s)\,\alpha_s(x).$$

Consequently, if we substitute $\bar\beta_t(x)$, given by (3.10) in the differential form of the right-hand side of (3.9), we obtain

$$\begin{aligned}&-(b_s + 2\, c_s\, x)'\, A_s\, x\,\alpha_s(x)\, ds - (a_s + b_s'\, x + x'\, c_s\, x)\operatorname{div}\left(\alpha_s(x)\, A_s\, x\right)\, ds \\ &\quad +\alpha_s(x)\operatorname{Tr}\left(c_s\, B_s\, B_s'\right)\, ds - (b_s + 2\, c_s\, x)' B_s\, B_s'\,\Sigma_s^{-1}\,(x - m_s)\,\alpha_s(x)\, ds \\ &\quad +\frac{1}{2}(a_s + b_s'\, x + x'\, c_s\, x)\operatorname{Tr}\left(\nabla^2\alpha_s(x)\, B_s'\, B_s\right)\, ds \\ &\quad +(a_s + b_s'\, x + x'\, c_s\, x)\,\alpha_s(x)\, x'\, C_s'\left(D_s^{-1}\right)' D_s^{-1}\, dy_s \\ &\quad +\alpha_s(x)\, x'\,(e_i\, e_j')\, x\, ds. \end{aligned} \tag{3.11}$$

Also,

$$d\bar\beta_s(x) = (da_s + db_s'\, x + x'\, dc_s\, x)\,\alpha_s(x) + (a_s + b_s'\, x + x'\, c_s\, x)\, d\alpha_s(x). \tag{3.12}$$

Consequently, $\bar\beta_s(x)$, given by (3.10), is a solution of (3.9) if (3.11) equals (3.12). However, $\alpha_s(x)$ solves the Zakai equation (2.7), so

$$d\alpha_s(x) = -\operatorname{div}\left(\alpha_s(x)\, A_s\, x\right)\, ds + \frac{1}{2}\operatorname{Tr}\left(\nabla^2\alpha_s(x)\, B_s\, B_s'\right)\, ds + \alpha_s(x)\, x'\, C_s'\left(D_s^{-1}\right)' D_s^{-1}\, dy_s.$$

Therefore, substituting the above expression for $d\alpha_s(x)$ into (3.12) yields

$$\begin{aligned}d\bar\beta_s(x) = &(da_s + db_s'\, x + x'\, dc_s\, x)\,\alpha_s(x) \\ &-(a_s + b_s'\, x + x'\, c_s\, x)\operatorname{div}\left(\alpha_s(x)\, A_s\, x\right)\, ds \\ &+\frac{1}{2}(a_s + b_s' + x'\, c_s\, x)\operatorname{Tr}\left(\nabla^2\alpha_s(x)\, B_s\, B_s'\right)\, ds \\ &+(a_s + b_s'\, x + x'\, c_s\, x)\, x'\, C_s'\left(D_s^{-1}\right)' D_s^{-1}\,\alpha_s(x)\, dy_s. \end{aligned} \tag{3.13}$$

Finally equating the coefficients of x, x', x and the constants in (3.11) and (3.13), we see that the result holds if (3.5), (3.6), and (3.7) hold. □

We now explicitly solve the ordinary differential equations (3.6) and (3.7).

Write G_t for the matrix solution of

$$\frac{dG_t}{dt} = -(A_t' + \Sigma_t^{-1} B_t B_t')G_t, \quad G_0 = I_{m\times m}. \tag{3.14}$$

Note that G_t is deterministic and can be calculated off-line. Also, as an exponential matrix, G_t has an inverse G_t^{-1}.

LEMMA 3.3. *The explicit solutions of* (3.6) *and* (3.7) *are*

$$b_t^{ij} = 2\,G_t \left(\int_0^t G_s^{-1}\, c_s^{ij}\, B_s\, B_s'\, \Sigma_s^{-1}\, m_s\, ds\right), \tag{3.15}$$

$$c_t^{ij} = \frac{1}{2}\, G_t \left(\int_0^t G_s^{-1}\, (e_j\, e_i' + e_i\, e_j')\, (G_s')^{-1}\, ds\right) G_t'. \tag{3.16}$$

Proof. The above equations follow using variation of constants. □

Remark. We proceed similarly with the process J_t^{ij} and L_t^{ij}, omitting details.

DEFINITION 3.4. *For any test function $g \in \mathbb{C}_0^2(\mathbb{R}^m)$ define the measure-valued process $\bar{\mathbf{E}}\{\Lambda_t\, J_t^{ij}\, g(x_t)|\mathcal{Y}_t\}$. From the results of* [10] *and* [12]*, this has a density $\gamma_t^{ij}(x)$ so that*

$$\bar{\mathbf{E}}\{\Lambda_t\, J_t^{ij}\, g(x_t)|\mathcal{Y}_t\} = \int_{\mathbb{R}^m} \gamma_t^{ij}(x)\, g(x)\, dx.$$

THEOREM 3.5. *At time t, the density $\gamma_t^{ij}(x)$ is completely described by the five statistics $\bar{a}_t^{ij}$, $\bar{b}_t^{ij}$ $\bar{c}_t^{ij}$, Σ_t, and m_t as follows:*

$$\gamma_t^{ij}(x) = \left(\bar{a}_t^{ij} + x'\,\bar{b}_t^{ij} + x'\,\bar{c}_t^{ij}\,x\right)\,\alpha_t(x). \tag{3.17}$$

Here, $\bar{a}_t^{ij} \in \mathbb{R}$, $\bar{b}_t^{ij} \in \mathbb{R}^m$, and $\bar{c}_t^{ij} \in L_s(\mathbb{R}^m, \mathbb{R}^m)$. Further,

$$\frac{d\bar{a}_t^{ij}}{dt} = \mathrm{Tr}\left(\bar{c}_t^{ij}\, B_t\, B_t'\right) + \bar{b}_t^{ij\prime}\, B_t\, B_t'\, \Sigma_t^{-1}\, m_t, \quad \bar{a}_0^{ij} = 0 \in \mathbb{R}, \tag{3.18}$$

$$d\bar{b}_t^{ij} = \left[-(A_t' + \Sigma_t^{-1}\, B_t\, B_t')\bar{b}_t^{ij} + 2\bar{c}_t^{ij}\, B_t\, B_t'\, \Sigma_t^{-1}\, m_t\right]\, dt + dy_t'\, f_j\, e_i, \quad b_0^{ij} = 0 \in \mathbb{R}^m, \tag{3.19}$$

$$\frac{d\bar{c}_t^{ij}}{dt} = -\left(A_t' + \Sigma_t^{-1}\, B_t\, B_t'\right)\, \bar{c}_t^{ij} - \bar{c}_t^{ij}\, \left(A_t + B_t\, B_t'\, \Sigma_t^{-1}\right) + \frac{1}{2}\left(e_i\, f_j'\, C_t + C_t'\, f_j\, e_i'\right), \tag{3.20}$$

$$\bar{c}_0^{ij} = 0 \in L_s(\mathbb{R}^m, \mathbb{R}^m).$$

Proof. The product $\Lambda_t\, J_t^{ij}\, g(x_t)$ is calculated and each side conditioned on $\mathcal{Y}_t$. After integration by parts, the following stochastic partial differential equation is obtained for $\gamma_t^{ij}(x)$:

$$\begin{aligned} d\gamma_t^{ij}(x) = &-\mathrm{div}\left(\gamma_t^{ij}(x)\, A_t\, x\right)\, dt + \frac{1}{2}\mathrm{Tr}\left(\nabla^2\gamma_t^{ij}(x)\, B_t\, B_t'\right)\, dt + \gamma_t^{ij}\, x'\, C_t'\, (D_t')^{-1}\, D_t^{-1}\, dy_t \\ &+ \alpha_t(x)\, x'\, e_i\, f_j'\, dy_t + \alpha_t(x)\, (x'\, C_t'\, f_j\, e_i'\, x)\, dt \end{aligned} \tag{3.21}$$

Recalling that $\alpha_t(x)$ satisfies the Zakai equation (2.7), we see that

$$\left(\bar{a}_t^{ij} + x'\,\bar{b}_t^{ij}\,x + x'\,\bar{c}_t^{ij}\,x\right)\,\alpha_t(x)$$

is a solution of (3.21) if $\bar{a}_t^{ij}$, $\bar{b}_t^{ij}$, and $\bar{c}_t^{ij}$ satisfy (3.18), (3.19), and (3.20), respectively. □

We now obtain explicit solutions to the above equations. Note that $f_j'\,dy_t = dy_t'\,f_j = dy_t^j$, where y_t^j denotes the jth component of y_t.

LEMMA 3.6. *The explicit solutions of* (3.19) *and* (3.20) *are*

$$\bar{b}_t^{ij} = 2\,G_t\,\left(\int_0^t G_s^{-1}\bar{c}_s^{ij}\,B_s\,B_s'\,\Sigma_s^{-1}\,m_s\,ds + \int_0^t G_s^{-1}\,e_i\,dy_s^j\right) \tag{3.22}$$

and

$$\bar{c}_t^{ij} = \frac{1}{2}\,G_t\,\left(\int_0^t G_s^{-1}(e_i\,f_j'\,C_s + C_s'\,f_j\,e_i')\,\left(G_s^{-1}\right)'\,ds\right)\,G_t'. \tag{3.23}$$

DEFINITION 3.7. *For any test function $g \in \mathbb{C}_0^2(\mathbb{R}^m)$ define the measure-valued process $\bar{\mathbf{E}}\{\Lambda_t\,L_t^{ij}\,g(x_t)|\mathcal{Y}_t\}$. From the results of* [10] *and* [12]*, this has a density $\lambda_t^{ij}(x)$ so that*

$$\bar{\mathbf{E}}\{\Lambda_t\,L_t^{ij}\,g(x_t)|\mathcal{Y}_t\} = \int_{\mathbb{R}^m}\lambda_t^{ij}(x)\,g(x)\,dx.$$

THEOREM 3.8. *At time t, the density $\lambda_t^{ij}(x)$ is completely characterized by the five statistics r_t^{ij}, s_t^{ij}, and u_t^{ij} as follows:*

$$\lambda_t^{ij}(x) = \left(r_t^{ij} + x'\,s_t^{ij}\,x + x'\,u_t^{ij}\,x\right)\,\alpha_t(x). \tag{3.24}$$

Here, $r_t^{ij} \in \mathbb{R}$, $s_t^{ij} \in \mathbb{R}^m$, and $u_t^{ij} \in L_s(\mathbb{R}^n, \mathbb{R}^m)$. Further,

$$\frac{dr_t^{ij}}{dt} = \mathrm{Tr}\left(u_t^{ij}\,B_t\,B_t'\right) + s_t^{ij\prime}\,B_t\,B_t'\,\Sigma_t^{-1}\,m_t - \mathrm{Tr}\left(B_t\,B_t'\,e_i\,e_j'\right), \qquad r_0^{ij} = 0 \in \mathbb{R}, \tag{3.25}$$

$$\begin{aligned}\frac{ds_t^{ij}}{dt} &= -\left(A_t' + \Sigma_t^{-1}\,B_t\,B_t'\right)\,s_t^{ij} + 2\,u_t^{ij}\,B_t\,B_t'\,\Sigma_t^{-1}\,m_t \\ &\quad - (e_j\,e_i')\;B_t\,B_t'\,\Sigma_t^{-1}\,m_t, \qquad s_0^{ij} = 0 \in \mathbb{R}^m,\end{aligned} \tag{3.26}$$

$$\begin{aligned}\frac{du^{ij}}{dt} &= -\left(A_t' + \Sigma_t^{-1}\,B_t\,B_t'\right)\,u_t^{ij} - u_t^{ij}\,\left(A_t + B_t\,B_t'\,\Sigma_t^{-1}\right) \\ &\quad + \frac{1}{2}\left(e_i\,e_j'\,\left(A_t + B_t\,B_t'\,\Sigma_t^{-1}\right) + \left(A_t' + \Sigma_t^{-1}\,B_t\,B_t'\right)\,e_j\,e_i'\right), \\ &\quad u_0^{ij} = 0 \in L_s(\mathbb{R}^m, \mathbb{R}^m).\end{aligned} \tag{3.27}$$

Proof. The product $\Lambda_t\,L_t^{ij}\,g(x_t)$ is calculated and each side conditioned on $\mathcal{Y}_t$. After integration by parts, the following stochastic partial differential equation is obtained for $\lambda_t^{ij}(x)$:

$$\begin{aligned}d\lambda_t^{ij}(x) &= -\mathrm{div}\left(\lambda_t^{ij}(x)\,A_t\,x\right)\,dt \\ &\quad + \frac{1}{2}\,\mathrm{Tr}\left(\nabla^2\lambda_t^{ij}(x)\,B_t\,B_t'\right)\,dt + \lambda_t^{ij}(x)\,x'\,C_t'\,\left(D_t^{-1}\right)'\,D_t^{-1}\,dy_t \\ &\quad + \alpha_t(x)\,x'\,e_i\,e_j'\,A_t\,x\,dt - \mathrm{div}\left(x'\,e_i\,e_j'\,B_t\,B_t'\,\alpha_t(x)\right)\,dt.\end{aligned} \tag{3.28}$$

We see that

$$\left(r_t^{ij} + x'\, s_t^{ij}\, x + x'\, u_t^{ij}\, x\right)\, \alpha_t(x)$$

solves (3.28) if r_t^{ij}, s_t^{ij}, and u_t^{ij} satisfy (3.25), (3.26), and (3.27) . □

Again, (3.26) and (3.27) can be solved by variation of constants. We summarize this in the following lemma.

LEMMA 3.9. *The explicit solutions of* (3.26) *and* (3.27) *are*

$$s_t^{ij} = G_t \left(\int_0^t G_s^{-1} (2\, u_s^{ij} - e_j\, e_i')\, \Sigma_s^{-1}\, B_s\, B_s'\, m_s\, ds \right), \tag{3.29}$$

$$u_t^{ij} = \frac{1}{2}\, G_t \left(\int_0^t \left[G_s^{-1} (e_i\, e_j' \left(A_s + B_s\, B_s'\, \Sigma_s^{-1}\right) + (A_s' + \Sigma_s^{-1}\, B_s\, B_s') e_j\, e_i' \right) \left(G_s^{-1}\right)' \right] ds \right) G_t. \tag{3.30}$$

Remark. We observe from the definition of G_t, (3.14), that the integrand in (3.30) includes only half of the four terms in the derivative of $G_t^{-1}(e_i\, e_j' + e_j\, e_i') \left(G_t^{-1}\right)'$, and so the integral cannot be evaluated in closed form.

THEOREM 3.10. *Finite-dimensional filters for* H_t^{ij}, J_t^{ij}, *and* L_t^{ij} *defined in* (3.1), (3.3), *and* (3.2) *are given by*

$$\mathbf{E}\{H_t^{ij} | \mathcal{Y}_t\} = a_t^{ij} + m_t'\, b_t^{ij} + \sum_{p=1}^m \sum_{q=1}^m c_t^{ij}(p,q)\, \Sigma_t(p,q) + m_t'\, c_t^{ij}\, m_t, \tag{3.31}$$

$$\mathbf{E}\{J_t^{ij} | \mathcal{Y}_t\} = \bar{a}_t^{ij} + m_t'\, b_t^{ij} + \sum_{p=1}^m \sum_{q=1}^m \bar{c}_t^{ij}(p,q)\, \Sigma_t(p,q) + m_t'\, \bar{c}_t^{ij}\, m_t, \tag{3.32}$$

$$\mathbf{E}\{L_t^{ij} | \mathcal{Y}_t\} = r_t^{ij} + m_t' s_t^{ij} + \sum_{p=1}^m \sum_{q=1}^m u_t^{ij}(p,q)\, \Sigma_t(p,q) + m_t'\, u_t^{ij}\, m_t. \tag{3.33}$$

Proof. Recall from (2.11) that α_t is an unnormalized Gaussian density with mean m_t and variance Σ_t. Therefore,

$$\int_{\mathbb{R}^m} \alpha_t(x)\, dx = \nu_t.$$

Note that for any $k \in \mathbb{R}^m$

$$\int_{\mathbb{R}^m} k'\, x\, \alpha_t(x)\, dx = (k'\, m_t)\nu_t.$$

Also for any matrix $M \in L(\mathbb{R}^m, \mathbb{R}^m)$ with entries $M(p,q)$, $1 \le p, q \le m$,

$$\begin{aligned} \int_{\mathbb{R}^m} x'\, M\, x\, \alpha_t(x)\, dx &= \int_{\mathbb{R}^m} (x - m_t)'\, M\, (x - m_t)\, \alpha_t(x)\, dx + m_t'\, M\, m_t \int_{\mathbb{R}^m} \alpha_t(x)\, dx \\ &= \left(\sum_{p=1}^M \sum_{q=1}^M M(p,q)\, \Sigma_t(p,q) + m_t'\, M\, m_t \right) \nu_t. \end{aligned}$$

Now from Bayes's theorem (2.6), we have

$$\begin{aligned}\mathbf{E}\{H_t^{ij}|\mathcal{Y}_t\} &= \frac{\bar{\mathbf{E}}\{\Lambda_t\, H_t^{ij}|\mathcal{Y}_t\}}{\bar{\mathbf{E}}\{\Lambda_t|\mathcal{Y}_t\}} \\ &= \frac{\int_{\mathbb{R}^m} \beta_t^{ij}(x)\,dx}{\int_{\mathbb{R}^m} \alpha_t(x)\,dx} \\ &= a_t^{ij} + m_t'\, b_t^{ij} + \sum_{p=1}^{m}\sum_{q=1}^{m} c_t^{ij}(p,q)\,\Sigma_t(p,q) + m_t'\, c_t^{ij}\, m_t\end{aligned}$$

by (3.4) and because the factors ν_t cancel. The proofs of equations (3.32) and (3.33) are similar. □

4. Finite-dimensional filter for higher order moments. The techniques of the previous section can be generalized to show that integrals and stochastic integrals of higher moments of the state variables have filtered estimates which can be expressed in terms of a finite number of statistics. The results also hold for other functions of the state and will be investigated in a subsequent paper.

Assumption 4.1. For notational simplicity, in this section we assume that the state and observation processes are scalar valued, i.e., that $m = n = 1$ in (2.1) and (2.2).

Let Γ_t be the process defined as

$$\Gamma_t = \int_0^t x_s^p\, ds, \quad p \in \mathbb{Z}^+.$$

Suppose $g : \mathbb{R} \to \mathbb{R}$ is any test function and suppose for a density-valued process $\mu_t(x)$,

$$\bar{\mathbf{E}}\{\Lambda_t\, \Gamma_t\, g(x_t)|\mathcal{Y}_t\} = \int_{\mathbb{R}} \mu_t(x)\, g(x)\, dx.$$

We now show that $\mathbf{E}\{\Gamma_t|\mathcal{Y}_t\}$ can be computed via a finite-dimensional filter.

THEOREM 4.2. *At time t, the density μ_t is completely characterized by the $p+3$ statistics, $a_t(0), a_t(1), \ldots, a_t(p)$, Σ_t, and m_t as follows:*

$$\mu_t(x) = \left[\sum_{i=0}^{p} a_t(i)\, x^p\right]\, \alpha_t(x), \tag{4.1}$$

where $a_0(i) = 0$, $i = 1, \ldots, p$, and

$$\begin{aligned}\frac{da_t(p)}{dt} &= -p\left(A_t + \Sigma_t^{-1}\, B_t^2\right)\, a_t(p) + 1, \\ \frac{da_t(p-1)}{dt} &= -(p-1)\left(A_t + \Sigma_t^{-1}\, B_t^2\right)\, a_t(p-1) + p\, a_t(p)\, \Sigma_t^{-1}\, B_t^2\, m_t, \\ \frac{da_t(i)}{dt} &= -i\left(A_t + \Sigma_t^{-1}\, B_t^2\right)\, a_t(i) + \frac{1}{2}\,(i+1)\,(i+2)\, a_t(i+2) \\ &\quad + (i+1)\, a_t(i+1)\, \Sigma_t^{-1}\, B_t^2, \\ &\qquad\qquad\qquad\qquad\qquad\qquad i = 1, \ldots, p-2, \\ \frac{da_t(0)}{dt} &= B_t^2\, a_t(2) + \Sigma_t^{-1}\, B_t^2\, a_t(1)\, m_t. \end{aligned} \tag{4.2}$$

Proof.

$$\Lambda_t\,\Gamma_t\,g(x_t) = \int_0^t \Lambda_s\,\Gamma_s\,\nabla g(x_s)\,A_s\,x_s\,ds + \int_0^t \Lambda_s\,\Gamma_s\,\nabla g(x_s)\,B_s\,dw_s + \frac{1}{2}\int_0^t \Lambda_s\,\Gamma_s\,\nabla^2 g(x_s)\,B_s^2\,ds + \int_0^t \Lambda_s\,\Gamma_s\,g(x_s)\,x_s\,C_s\,D_s^{-2}\,dy_s + \int_0^t \Lambda_s\,g(x_s)\,x_s^p\,ds.$$

Conditioning on $\mathcal{Y}_t$, as in [13],

$$\int_{\mathbb{R}} \mu_t(x)\,g(x)\,dx = \int_0^t \int_{\mathbb{R}} \mu_s(x)\,\nabla g(x)\,A_s\,x\,dx\,ds + \frac{1}{2}\int_0^t \int_{\mathbb{R}} \mu_s(x)\nabla^2 g(x)\,B_s^2\,dx\,ds + \int_0^t \int_{\mathbb{R}} \mu_s(x)\,g(x)\,x\,C_s\,D_s^{-2}\,dx\,dy_s + \int_0^t \int_{\mathbb{R}} x^p\,g(x)\,\alpha_s(x)\,dx\,ds.$$

Integrating by parts in x, we see that $\mu_t(\cdot)$ satisfies the stochastic partial differential equation

$$\mu_t(x) = -\int_0^t \frac{d}{dx}\left(\mu_s(x)\,A_s\,x\right)\,ds + \frac{1}{2}\int_0^t \frac{d^2}{dx^2}\mu_s(x)\,B_s^2\,ds + \int_0^t \mu_s(x)\,x\,C_s\,D_s^{-2}\,dy_s + \int_0^t x^p\,\alpha_s(x)\,ds.$$

It can then be verified that (4.1) is a solution to the above equation if the time-varying coefficients $a_t(0),\ldots,a_t(p)$ satisfy the ordinary differential equations (4.2). □

Remark. The ordinary differential equations (4.2) can be solved explicitly by variation of constants.

Finally, we note that a similar derivation to that given in this section yields finite-dimensional filters for $\int_0^t x_s^p\,dx_s$ and $\int_0^t x_s^p\,dy_s$, $p = 1, 2, 3, \ldots$.

5. Filtered EM algorithm for Gaussian state-space models. The aim of this section is to derive a filter-based EM algorithm for computing maximum-likelihood (ML) parameter estimates of a linear Gaussian state-space system. The finite-dimensional filters of section 3 are used in implementing the E-step of the EM algorithm, resulting in a filter-based EM algorithm.

Consider the time-invariant version of the state-space model (2.1), (2.2):

$$dx_t = A\,x_t\,dt + B\,dw_t, \tag{5.1}$$

$$dy_t = C\,x_t\,dt + D\,dv_t. \tag{5.2}$$

Our aim is to compute ML estimates of the parameters $\theta = (A, C)$ given the observations $\mathcal{Y}_t = \sigma\{y_s : s \le t\}$ and assuming B, D are known. We do this via the EM algorithm.

Remark. Unlike the discrete-time case, in continuous time, it is not possible to obtain ML estimates of the variance terms B and D because measures corresponding to Wiener processes with different variances are not absolutely continuous (see Chapter 6.1 in [13]). At the end of this section, we give estimates for B and D in terms of the quadratic variations of the state and observation processes.

EM ALGORITHM. Suppose that we have observation history $\mathcal{Y}_t$ available. Let $\{P_\theta\,,\,\theta\in\Theta\}$ be a family of probability measures on $(\Omega,\mathcal{F})$, all absolutely continuous with respect to a fixed probability measure P_0. The log likelihood function for computing an estimate of the parameter θ based on the information available in $\mathcal{Y}_t$ is

$$\mathcal{L}(\theta)=\mathbf{E}_0\left\{\log\frac{dP_\theta}{dP_0}\mid\mathcal{Y}_t\right\},$$

and the maximum likelihood estimate (MLE) is defined by

$$\widehat{\theta}\in\operatorname*{argmax}_{\theta\in\Theta}\mathcal{L}(\theta).$$

The EM algorithm is an iterative numerical method for computing the MLE. Let $\hat{\theta}_0$ be the initial parameter estimate. The EM algorithm generates a sequence of parameter estimates as follows.

Each iteration of the EM algorithm consists of two steps:

Step 1 (E-step). Set $\tilde{\theta}=\widehat{\theta}_j$ and compute $Q(\cdot,\tilde{\theta})$, where

$$Q(\theta,\tilde{\theta})=\mathbf{E}_{\tilde{\theta}}\left\{\log\frac{dP_\theta}{dP_{\tilde{\theta}}}\mid\mathcal{Y}_t\right\}.$$

Step 2 (M-step). Find $\widehat{\theta}_{j+1}\in\operatorname*{argmax}_{\theta\in\Theta}Q(\theta,\theta_j)$.

The sequence-generated $\{\widehat{\theta}_j\,,\,j\geq 0\}$ gives nondecreasing values of $\mathcal{L}(\hat{\theta}_j)$ with equality if and only if $\widehat{\theta}_{j+1}=\widehat{\theta}_j$.

It is shown in the appendix that

(5.3)

$$\begin{aligned}Q(\theta,\tilde{\theta})=\mathbf{E}&\left\{\int_0^t x_s'\,A'\,[BB']^{\#}\,dx_s-\frac{1}{2}\int_0^t x_s'\,A'\,[B\,B']^{\#}\,A\,x_s\,ds|\mathcal{Y}_t\right\}\\&+\mathbf{E}\left\{\int_0^t x_s'\,C'\,(D\,D')^{-1}\,dy_s-\frac{1}{2}\int_0^t x_s'\,C'(DD')^{-1}\,C\,x_s\,ds|\mathcal{Y}_t\right\}+\mathbf{E}\{R(\tilde{\theta}|\mathcal{Y}_t\},\end{aligned}$$

where # denotes the pseudoinverse and $R(\tilde{\theta})$ does not involve θ.

To implement the M-step we set the derivatives $\partial Q/\partial\theta=0$. This yields

$$A=\mathbf{E}\left\{\int_0^t dx_s\,x_s'|\mathcal{Y}_t\right\}\left(\mathbf{E}\left\{\int_0^t x_s\,x_s'\,ds|\mathcal{Y}_t\right\}\right)^{-1}=\hat{L}_t'\,\hat{H}_t^{-1},\tag{5.4}$$

$$C=\mathbf{E}\left\{\int_0^t dy_s\,x_s'|\mathcal{Y}_t\right\}\left(\mathbf{E}\left\{\int_0^t x_s\,x_s'\,ds|\mathcal{Y}_t\right\}\right)^{-1}=\hat{J}_t'\,\hat{H}_t^{-1},\tag{5.5}$$

where $\hat{H}_t$ and $\hat{L}_t\in\mathbb{R}^{m\times m}$ denote matrices with elements $\hat{H}_t^{ij}\triangleq\mathbf{E}\{H_t^{ij}|\mathcal{Y}_t\}$ and $\hat{L}_t^{ij}\triangleq\mathbf{E}\{L_t^{ij}|\mathcal{Y}_t\}$, $i,j\in\{1,\dots,m\}$. Also, $\hat{J}_t\in\mathbb{R}^{m\times n}$ denotes the matrix with elements $\hat{J}_t^{ij}\triangleq\mathbf{E}\{J_t^{ij}|\mathcal{Y}_t\}$, $i=1,\dots,m$, $j=1,\dots,n$. The terms $\hat{H}_t^{ij}$, $\hat{L}^{ij}$, and $\hat{J}_t^{ij}$ are computed using Theorems 3.2, 3.8, and 3.5 together with the filters in Theorem 3.10. Thus we have a filter-based EM algorithm.

Remark. We have presented the EM algorithm for updating the parameters of a general state-space model. However, we have not addressed identifiability issues of our state-space model.

Indeed, identifiability and consistency of the ML estimator have been studied in special cases of our model (where our filter-based algorithm also applies). For example, in [9], A is of a known structure but parametrized by an unknown vector θ. Consistency of the ML estimators is proved in [16].

Estimation of B and D. First consider the tensor product of x_t with itself:

$$x_t\, x_t' = x_0\, x_0' + \int_0^t x_s\, dx_s' + \int_0^t dx_s\, x_s' + \int_0^t B\, B'\, ds. \tag{5.6}$$

Conditioning both sides of (5.6) on $\mathcal{Y}_t$, we have

$$\Sigma_t = \mathbf{E}\{x_0\, x_0' | \mathcal{Y}_t\} + \mathbf{E}\left\{ \int_0^t x_s\, dx_s' + \int_0^t dx_s\, x_s' \,\middle|\, \mathcal{Y}_t \right\} + B\, B'\, t. \tag{5.7}$$

$\mathbf{E}\{x_0\, x_0' | \mathcal{Y}_t\}$ in (5.7) is the smoothed second moment and is given in terms of finite-dimensional statistics; see Theorem 12.11, section 12.4 in [8]. The components of the conditioned stochastic integrals in (5.7) are given by the filtered estimates of L_t^{ij}. Consequently, we have a procedure for estimating the matrix $B\, B'$.

Similarly, consider the tensor product of y_t with itself:

$$y_t\, y_t' = \int_0^t y_s\, dy_s' + \int_0^t dy_s\, y_s' + D\, D'\, t. \tag{5.8}$$

This expression simply amounts to evaluating DD' in terms of the quadratic variation of y.

6. Conclusion and extensions. For linear Gaussian dynamics, new finite-dimensional filters have been derived which estimate integrals and stochastic integrals of the moments of the state variable. Used in the EM algorithm, these quantities provide maximum likelihood estimates of the parameters in the dynamics of the Kalman filter.

We mention two extensions of the results. First, similar filters can be derived for nonlinear systems with Benes-type nonlinearity. Details are presented in [21]. Second, techniques similar to those developed in this paper can be applied in the reconstruction of doubly stochastic autoregressive (AR) processes, as we now outline.

Assuming the scalar version ($m = n = 1$) of the state-space model (2.1), (2.2), consider the continuous-time doubly stochastic AR process $z_t \in \mathbb{R}$ defined as

$$dz_t = f(x_t)\, z_t\, dt + du_t, \quad z_0 = 1. \tag{6.1}$$

Here u is a scalar Brownian process independent of v and w, and $f : \mathbb{R} \to \mathbb{R}$ is a polynomial function. Suppose our aim is to derive a filter for computing $\mathbf{E}\{z_t | \mathcal{Y}_t\}$.

In (6.1), z_t can be viewed as a continuous-time AR process with random coefficient $f(x_t)$, where x_t itself evolves according to another AR process. In [4] we derived finite-dimensional filters for such problems in discrete time. The above model can be viewed as a continuous-time version of [4]. The problem of reconstructing z_t when x_t is a finite state Markov chain (instead of a continuous-valued Gaussian process) has recently been studied in the context of maneuvering target tracking [20]. Therefore, (6.1) can be viewed as the continuous-valued analogue of the model in [20].

Let us compute $\mathbf{E}\{z_t | \mathcal{Y}_t\}$. Note that we can solve (6.1) explicitly for z_t; this yields

$$z_t = \exp\left(\int_0^t f(x_s)\, ds \right) \left[1 + \int_0^t \exp\left(- \int_0^s f(x_r) dr \right) du_s \right].$$

Since u is independent of x and y, our aim can be re-expressed as follows: compute the filtered estimate

$$\mathbf{E}\{z_t|\mathcal{Y}_t\} = \mathbf{E}\left\{\exp\left(\int_0^t f(x_s)\,ds\right)|\mathcal{Y}_t\right\}.$$

Using techniques similar to those presented in this paper, it is possible to derive a finite-dimensional filter for $\mathbf{E}\{z_t|\mathcal{Y}_t\}$. Details are presented in [22].

Appendix A. Derivation of $Q(\theta, \tilde{\theta})$. To update the estimate from $\tilde{A}$ to A, we employ Girsanov's theorem and introduce the density

$$\left.\frac{dP(A)}{dP(\tilde{A})}\right|_{\mathcal{G}_t} = \exp\left(\int_0^t x_s'(A'-\tilde{A}')(BB')^{\#}\,dx_s \right.$$
$$\left. -\frac{1}{2}\int_0^t x_s'(A'-\tilde{A}')(BB')^{\#}(A+\tilde{A})x_s\,ds\right),$$

where # denotes the pseudoinverse. Then

(A.1)
$$\mathbf{E}\left\{\log\left.\frac{dP(A)}{dP(\tilde{A})}\right|_{\mathcal{G}_t}|\mathcal{Y}_t\right\} = \mathbf{E}\left\{\int_0^t x_s'\,A'\,(BB')^{\#}\,dx_s - \frac{1}{2}\int_0^t x_s'\,A'\,(BB')^{\#}\,A\,x_s\,ds|\mathcal{Y}_t\right\}$$
$$+R(\tilde{A}),$$

where $R(\tilde{A})$ does not involve A.

Similarly, to update the estimate from $\tilde{C}$ to C, we again apply Girsanov's theorem and introduce the density

$$\left.\frac{dP(C)}{dP(\tilde{C})}\right|_{\mathcal{G}_t} = \exp\left(\int_0^t x_s'(C'-\tilde{C}')(DD')^{-1}\,dy_s \right.$$
$$\left. -\frac{1}{2}\int_0^t x_s'(C'-\tilde{C}')(DD')^{-1}(C+\tilde{C})x_s\,ds\right).$$

Consequently,

(A.2)
$$\mathbf{E}\left\{\log\left.\frac{dP(C)}{dP(\tilde{C})}\right|_{\mathcal{G}_t}|\mathcal{Y}_t\right\} = \mathbf{E}\left\{\int_0^t x_s'C'\,(DD')^{-1}\,dy_s \right.$$
$$\left. -\frac{1}{2}\int_0^t x_s'\,C'\,(DD')^{-1}\,Cx_s\,ds|\mathcal{Y}_t\right\} + S(\tilde{C}),$$

where $S(\tilde{C})$ does not involve C.

Adding (A.1) and (A.2) yields (5.3).

REFERENCES

[1] R.H. Shumway and D.S. Stoffer, *An approach to time series smoothing and forecasting using the EM algorithm*, J. Time Series Anal., 3 (1982), pp. 253–264.

[2] D. Ghosh, *Maximum likelihood estimation of the dynamic shock-error model*, J. Econometrics, 41 (1989), pp. 121–143.

[3] V. Krishnamurthy, *On-line estimation of dynamic shock-error models*, IEEE Trans. Automat. Control, 35 (1994), pp. 1129–1134.

[4] V. Krishnamurthy and R.J. Elliott, *Exact finite dimensional filters for doubly stochastic auto-regressive processes*, IEEE Trans. Automat. Control, June 1997 (to appear).
[5] R.J. Elliott, *Exact adaptive filters for Markov chains observed in Gaussian noise*, Automatica, 30 (1994), pp. 1399–1408.
[6] R.J. Elliott, L. Aggoun, and J.B. Moore, *Hidden Markov Models: Estimation and Control*, Applications of Mathematics, Springer-Verlag, New York, 1995.
[7] R.J. Elliott, *Stochastic Calculus and Applications*, Applications of Mathematics 18, Springer-Verlag, New York, 1982.
[8] R.S. Liptser and A.N. Shiryayev, *Statistics of Random Processes* 2, Springer-Verlag, Berlin, Heidelberg, New York, 1977.
[9] A. Dembo and O. Zeitouni, *Parameter estimation of partially observed continuous-time stochastic processes via the EM algorithm*, Stochastic Process. Appl., 23 (1986), pp. 91–113.
[10] A. Bensoussan, *Stochastic Control of Partially Observable Systems*, Cambridge University Press, Cambridge, UK, 1992.
[11] A. Bensoussan and R.J. Elliott, *General finite dimensional risk sensitive problems and small noise limits*, IEEE Trans. Automat. Control, 41 (1996), pp. 210–215.
[12] E. Pardoux, *Stochastic partial differential equations and filtering of diffusion processes*, Stochastics, 3 (1979), pp. 127–167.
[13] E. Wong and B. Hajek, *Stochastic Processes in Engineering Systems*, Springer-Verlag, New York, 1985.
[14] A.P. Dempster, N.M. Laird, and D.B. Rubin, *Maximum likelihood from incomplete data via the EM algorithm*, J. Royal Statist. Soc. B, 39 (1977), pp. 1–38.
[15] C.F.J. Wu, *On the convergence properties of the EM algorithm*, Ann. Statist., 11 (1983), pp. 95–103.
[16] E.J. Hannan and M. Deistler, *The Statistical Theory of Linear Systems*, Wiley, New York, 1988.
[17] R.J. Elliott and V. Krishnamurthy, *New finite dimensional filters for estimation of discrete-time linear Gaussian models*, IEEE Trans. Automat. Control, submitted.
[18] D.L. Ocone, J.S. Baras, and S.I. Marcus, *Explicit filters for diffusions with certain nonlinear drifts*, Stochastics, 8 (1982), pp. 1–16.
[19] S.I. Marcus and A.S. Willsky, *Algebraic structure and finite dimensional nonlinear estimation*, SIAM J. Math. Anal., 9 (1978), pp. 312–327.
[20] F. Dufour and P. Bertrand, *An image based filter for discrete-time Markovian jump linear systems*, Automatica, 32 (1996), pp. 241–247.
[21] R.J. Elliott, V. Krishamurthy, and H.V. Poor, *Exact filters for certain moments and stochastic integrals of the state of systems with Benes nonlinearity*, IEEE Trans. Automat. Control, submitted.
[22] V. Krishamurthy and R.J. Elliott, *Exact finite dimensional filters for certain exponential functionals of Gaussian state-space processes*, IEEE Trans. Automat. Control, submitted.

SIAM J. CONTROL OPTIM.
Vol. 35, No. 6, pp. 1924–1951, November 1997

005

SYSTEM IDENTIFICATION BY DYNAMIC FACTOR MODELS*

C. HEIJ†, W. SCHERRER‡, AND M. DEISTLER‡

Abstract. This paper concerns the modeling of stochastic processes by means of dynamic factor models. In such models the observed process is decomposed into a structured part called the latent process and a remainder that is called noise. The observed variables are treated in a symmetric way so that no distinction between inputs and outputs is required. This motivates the additional condition that the prior assumptions on the noise are symmetric in nature. One of the central questions in this paper is how uncertainty about the noise structure translates into nonuniqueness of the possible underlying latent processes. We investigate several possible noise specifications and analyze properties of the resulting class of observationally equivalent factor models. This concerns in particular the characterization of optimal models and properties of continuity and consistency.

Key words. linear systems, stationary processes, identification, factor analysis, errors in variables, least squares, consistency

AMS subject classifications. 93A30, 93B11, 93E03, 93E12

PII. S0363012995282127

1. Introduction.

In this paper we are concerned with the identification of linear systems. The most commonly used models in system identification are ARMA and ARMAX models; we refer to [17], [4], and [13]. An ARMA model is symmetric and nonopen in the sense that all observed variables are treated in a symmetric way and they are completely described by the model. On the other hand, ARMAX models are nonsymmetric and open, as a distinction is made between inputs and outputs and the noise is added to the outputs, and the inputs are not modeled.

We will consider linear factor models where the noise model is symmetric and where we have a deterministic, symmetric, and open system model. In a sense, these models combine the symmetry which is inherent in, for example, ARMA models, with the flexibility of models that leave certain process aspects unexplained, as, for example, in input-output models.

Of course, the classical ARMA and ARMAX models are appropriate in a great number of cases. For instance, if we are interested in predicting the outputs from the inputs, then the ARMAX setting is appropriate. On the other hand, there are also situations where this approach can not be justified and may lead to prejudiced results.

- A prediction-based error model is not appropriate, for example, if we are interested in the "true" underlying system and there is noise on the inputs and the outputs.
- There may be uncertainty about the number of system equations or about the classification of the system variables into inputs and outputs. In this case

Received by the editors February 24, 1995; accepted for publication (in revised form) August 15, 1996. This research was supported by Austrian "Fonds zur Förderung der wissenschaftlichen Forschung" Projekt P9176-PHY and the EC-SCIENCE "System Identification" project SC1-CT92-0779.

http://www.siam.org/journals/sicon/35-6/28212.html

†Econometric Institute, Erasmus University Rotterdam, P.O. Box 1738, 3000 DR Rotterdam, The Netherlands (heij@ect.few.eur.nl).

‡Institut für Ökonometrie, Operations Research und Systemtheorie, Technische Universität Wien, Argentinierstrasse 8, A-1040 Wien, Austria (W.Scherrer@tuwien.ac.at, M.Deistler@tuwien.ac.at).

we have to perform a more symmetric way of system modeling, which in turn demands a symmetric noise model.

- In multivariate time series analysis one is confronted with the so-called curse of dimensionality. One method of reducing the dimension of the parameter space for the system model is dynamic factor analysis, which is an essential aspect of the approach described here.

Factor models have been used in statistics, psychometrics, and econometrics for a long time; see [9], [1], [10]. The theory is most well developed for the case of static models. Most applications are also reported within this framework, although there are also contributions on the identification of dynamic factor models; see [11], [8], [5]. Within the area of systems and control there has recently been increasing interest in symmetric modeling. We mention the introduction of the behavioral approach in systems theory in [24], [26], the attention paid to the Frisch problem (see [18], [23], [2]), and low-noise modeling, as proposed in [15]. Most contributions on factor models in this area deal either with the mathematical structure of dynamic models or with data modeling by means of static models. In a certain sense, nonparametric-framework, results on the identification of dynamic factor models within a stochastic setting have been presented in [6], [7]. Procedures for symmetric time series modeling within a deterministic behavioral framework have been proposed in [25], [14], and [21].

In this paper we try to integrate the above two frameworks, i.e., stochastic factor models and deterministic behavioral modeling. The model class consists of stochastic dynamic factor models where the latent process satisfies deterministic behavioral laws. This means that stochastic structure is added to the deterministic behavioral framework, which provides additional tools of analysis. On the other hand, our approach allows for an analysis of dynamic factor models in terms of finite-dimensional systems, as opposed to the nonparametric results that were previously obtained.

We consider a situation that is idealized insofar as we commence from the population second moments of the data. In other words, we analyze the relation between the spectral density of the observed process and the corresponding factor models. Nevertheless, this is done from the point of view of requirements connected with the identification from observed data, and we will indicate how the results of this paper can be used for this purpose. A detailed analysis of procedures for the identification of dynamic factor models from observed time series falls beyond the scope of this paper and will be investigated elsewhere.

One of the issues studied in this paper is the nonuniqueness of the behaviour for given second-order moments. This means that uncertainty about the precise noise structure leads to a corresponding nonuniqueness of the possible factor models that are compatible with the observed process. As is well known, in the mainstream approach of modeling with exogenous inputs, the population second moments of the observations determine, under very general conditions, the transfer function of the underlying system uniquely. This is due to the assumption that the noise is uncorrelated with the inputs. Uniqueness in general does not hold true in cases when all the variables may be corrupted by noise. This means that the set of observationally equivalent models, that is, the set of all models compatible with the population second moments, will in general not be a singleton. Of course, by imposing sufficiently strong conditions, uniqueness can be achieved, but in many cases it may be hard to justify such assumptions. The question then becomes how the lack of knowledge about the error structure translates into nonuniqueness of the resulting model. This is a kind of uncertainty about the underlying system that can not be removed, even in an infinite sample.

We now give an outline of the topics treated in this paper. A dynamic factor model is of the form

$$w = \hat{w} + \tilde{w}, \tag{1}$$

where w is the observed process, $\hat{w}$ is a (in general unobserved) latent process satisfying exact linear dynamic equations, and $\tilde{w}$ is the noise process. The restrictions on $\hat{w}$ can be expressed in terms of deterministic system behaviors as introduced in [24], [26]. The processes $(w, \hat{w}, \tilde{w})$ are assumed to be jointly stationary, and in this case the latent process has a singular spectrum. The noise process represents the error resulting from the approximation of the observations w by the latent process $\hat{w}$.

The central question considered in this paper is how to obtain the restrictions satisfied by the latent process from the observations. Without imposing further conditions, no solutions can be excluded from the knowledge of the observed process alone. This means that we have to impose additional assumptions on the noise structure in order to make meaningful statements about the underlying system. The main topics of this paper can be summarized as follows.

(i) The formulation of noise assumptions and an analysis of their effect on the class of observationally equivalent models. We consider in particular the assumptions of orthogonality (the latent process and the noise process are mutually uncorrelated), observability (the latent process can be expressed as a linear function of the observed process), and bounded noise (the noise process satisfies an a priori specified bound).

(ii) An analysis of the structural properties of identification procedures corresponding to different noise assumptions. This involves an analysis of the mapping relating an observed process to the class of observationally equivalent system models. Continuity of this mapping is related to consistency in case of modeling from observed time series.

(iii) An analysis of the complexity and goodness of fit of factor models, with special attention for optimal models of restricted complexity.

This paper has the following structure. In section 2 we define the dynamic factor model. For this purpose we review the behavioral approach in linear system theory. Factor models are characterized on the behavioral level and also in terms of spectral properties, and we define the complexity and goodness of fit of factor models. The general framework is illustrated by the special case of a white noise process and nondynamic system equations, and it is shown that in this case our set-up coincides with the classical formulation of static factor models. Section 3 is concerned with optimal models, in the sense of minimizing the noise under restrictions on the complexity of the latent process. Section 4 investigates structural properties of the corresponding identification problem, with special attention paid to continuity and consistency. Section 5 contains concluding remarks. Some technical proofs are collected in section 6, the appendix.

2. Dynamic factor models.

2.1. Linear systems. For the formulation of dynamic factor models it is convenient to use the behavioral approach as developed by Willems in [24], [26]. Since this approach may not be well known to the reader, we discuss in this section those aspects that are relevant for our purposes. Readers with an interest in further details and proofs are referred to [24], [26].

In this subsection $\hat{w} : \mathbf{Z} \to \mathbf{R}^q$ denotes a trajectory rather than a process; that is, it is a q-variate time series observed in discrete time. The behaviour of a deterministic

system is defined as the set of all trajectories $\hat{w}$ that may arise within the restrictions imposed by the system. So a behavior is a subset $\mathcal{B}$ of $(\mathbf{R}^q)^{\mathbf{Z}}$. Of special interest are behaviors that are linear, time invariant, and complete. This means that $\mathcal{B} \subset (\mathbf{R}^q)^{\mathbf{Z}}$ is a linear subspace that is invariant under the shift operator σ, defined by $(\sigma\,\hat{w})(t) := \hat{w}(t+1)$, and that the behaviour is in addition closed in the topology of pointwise convergence. The last condition means that for a sequence $\hat{w}_n \in \mathcal{B}$ which converges pointwise (in $\mathbf{R}^q$) to $\hat{w}_0 \in (\mathbf{R}^q)^{\mathbf{Z}}$, it holds that also $\hat{w}_0 \in \mathcal{B}$. These conditions imply that the behavior corresponds to a linear, time invariant, finite-dimensional system. Below, we will simply use the term linear system to refer to a linear, time invariant, complete behaviour $\mathcal{B} \subset (\mathbf{R}^q)^{\mathbf{Z}}$.

Linear systems can be represented in several ways. Here we discuss representations in terms of polynomial equations, state space models with driving variables, and corresponding transfer functions.

Every linear system can be represented in polynomial form as the solution set of the polynomial equations

$$R(\sigma, \sigma^{-1})\,\hat{w} = 0 \tag{2}$$

Here R is a polynomial matrix in the forward and backward shifts. The representation of a given system by a polynomial matrix is highly nonunique. Without loss of generality we could have restricted ourselves to polynomials in either σ or σ^{-1} alone, but (2) is in accordance with [24], [26]. The set of behavioral laws of a linear system $\mathcal{B}$ is defined as the set of all polynomial equations satisfied by the system; that is, it is the module of $1 \times q$ polynomials $\mathcal{L} = \{r; r(\sigma, \sigma^{-1})\,\hat{w} = 0 \text{ for all } \hat{w} \in \mathcal{B}\}$. Every polynomial representation of a given system has the same (polynomial) rank p, which is equal to the dimension of the module $\mathcal{L}$. Full row rank representations are unique up to left multiplication by a unimodular matrix, i.e., a polynomial matrix which has a polynomial inverse. These representations can also be interpreted as input-output systems in polynomial form, where p is the number of outputs and $m := q - p$ is the number of inputs. We denote by n the minimal number of initial conditions required to express future outputs in terms of future inputs, which is equal to the sum of the Kronecker observability indices of the system.

An alternative representation is in terms of state models with driving variables. Every linear system can be represented as

$$\sigma x = Ax + Bv, \quad \hat{w} = Cx + Dv. \tag{3}$$

Here v is an auxiliary vector of unrestricted driving variables and x is a vector of state variables. In contrast with the usual input-state-output model, here all the external variables are described as outputs of a system driven by forces which need not have any external meaning. For a given system this kind of representation is highly nonunique. Minimal representations have n states and m driving variables, and the class of all minimal representations is described by the feedback group $(S(A+BF)S^{-1}, SBR, (C+DF)S^{-1}, DR)$.

Until now no assumptions were made concerning the controllability of systems. For example, if A is a $q \times q$ invertible matrix then the set $\{\hat{w} : \mathbf{Z} \to \mathbf{R}^q; \hat{w}(t+1) = A\,\hat{w}(t), t \in \mathbf{Z}\}$ defines a linear system with autonomous evolution which is clearly not controllable. A system $\mathcal{B}$ is called controllable if every future in $\mathcal{B}$ is attainable from every past in $\mathcal{B}$, that is, if for every $\hat{w}_1, \hat{w}_2 \in \mathcal{B}$ there exist $\hat{w} \in \mathcal{B}$ and $h \geq 0$ such that $\hat{w}(t) = \hat{w}_1(t)$ for $t < 0$ and $\hat{w}(t) = \hat{w}_2(t)$ for $t \geq h$. In terms of the kernel representations (2) this means that $R(z, z^{-1})$ has constant rank over $z \in \mathbf{C} \setminus \{0\}$. In

this case the system can also be represented as the image of a polynomial operator, that is, $\hat{w} \in \mathcal{B}$ is represented as

$$\hat{w} = M(\sigma, \sigma^{-1})f, \tag{4}$$

where f has the interpretation of the underlying generating factors. There is a close connection between the notion of controllability as defined before and the usual notion in terms of state space models, because minimal state models (3) of controllable systems $\mathcal{B}$ are characterized by the property that (A, B) is a controllable pair and (A, C) an observable pair. In this case we can obtain isometric state models (see [21]), that is, representations with the property that

$$\begin{pmatrix} A & B \\ C & D \end{pmatrix}' \begin{pmatrix} A & B \\ C & D \end{pmatrix} = \begin{pmatrix} I_n & O \\ O & I_m \end{pmatrix}, \tag{5}$$

where I_d denotes the d-dimensional identity matrix and Q' denotes the transposed of a matrix Q. If (A, B, C, D) is a minimal isometric state representation of a controllable system, then all such representations are given by (UAU', UBV, CU', DV) with U and V orthogonal matrices.

The model (4) gives a finite impulse response representation of controllable systems. This gives a clear description how to generate all time series belonging to a given system. Alternative descriptions are in terms of transfer functions. For controllable systems we can always choose A to be asymptotically stable, and in this case the square summable time series in the system can be generated as $\hat{w} = G(\sigma^{-1})v$, where v is square summable and G is the causal transfer function defined by $G(z) = D + \sum_{k=1}^{\infty} CA^{k-1}Bz^k$. The rank of the transfer function G is m, and its McMillan degree is n. For an isometric state model this transfer function becomes an isometry, sometimes also called an all-pass transfer function. The driving variables needed to generate a given square summable time series are then obtained by $v = G^*(\sigma^{-1})\,\hat{w}$, where G^* is the adjoint defined by $G^*(\sigma^{-1}) := G'(\sigma)$.

In our analysis we will often make use of isometric representations of linear systems. A state space method for obtaining these models is described in [21]. They can also be obtained from polynomial representations, as follows. Let $\mathcal{B}$ be a controllable linear system with kernel representation $\mathcal{B} = \ker(R) = \{\hat{w}; R(\sigma, \sigma^{-1})\,\hat{w} = 0\}$ and image representation $\mathcal{B} = \mathrm{im}(M) = \{\hat{w}; \hat{w} = M(\sigma, \sigma^{-1})f\}$. If m is the number of inputs of the system, then R can be chosen with $q - m$ rows and M with m columns. Controllability implies that $R(z, z^{-1})$ has constant rank over $\mathbf{C}\backslash\{0\}$, and M can also be chosen of constant rank. In this case the projections $P = M(M^*M)^{-1}M^*$ and $Q = R^*(RR^*)^{-1}R$ are well-defined rational functions with constant rank over the domain $\mathbf{C}\backslash\{0\}$. So there exist causal, miniphase spectral factorizations $P = \hat{G}\,\hat{G}^*$ and $Q = \tilde{G}\,\tilde{G}^*$; see [22, Theorem I.10.1]. These spectral factors are isometric; that is, $\hat{G}^*\,\hat{G} = I_m$ and $\tilde{G}^*\,\tilde{G} = I_{q-m}$. Then the spectral factor $\hat{G}$ is an isometric transfer function for $\mathcal{B}$, and all square summable time series in $\mathcal{B}$ are obtained as the image of $\hat{G}$. Therefore we call this an isometric image representation. Further, all square summable time series in $\mathcal{B}$ are annihilated by $\tilde{G}^*$, and therefore we call $\tilde{G}$ an isometric kernel representation. As R and M describe the same system, it follows that $RM = 0$ so that $\tilde{G}^*\,\hat{G} = 0$. This shows that the $q \times q$ rational matrix $[\hat{G}, \tilde{G}]$ is inner, that is, it is stable and unitary. Conversely, every rational inner matrix $[\hat{G}, \tilde{G}]$ describes a linear system with isometric image representation $\hat{G}$ and isometric kernel representation $\tilde{G}$.

2.2. Factor models and spectra. Let $(\Omega, \mathcal{A}, \mathrm{P})$ denote an underlying probability space and let $\mathbf{L}_2$ be the corresponding Hilbert space of square integrable real-valued

random variables. We assume that the observed process w is a q-dimensional, weakly stationary process, so that $w \in (\mathbf{L}_2^q)^{\mathbf{Z}}$. A dynamic factor model is a process decomposition of the form $w = \hat{w} + \tilde{w}$, where $\tilde{w} \in (\mathbf{L}_2^q)^{\mathbf{Z}}$ is the noise process and $\hat{w} \in (\mathbf{L}_2^q)^{\mathbf{Z}}$ is the latent process that is essentially restricted to a linear system. The behaviour $\mathcal{B}$ of $\hat{w}$ is defined as the smallest linear, time invariant, complete system which contains almost all process realizations, that is, $P\{\hat{w}(\omega) \in \mathcal{B}\} = 1$. The following result states that this definition makes sense.

PROPOSITION 2.1. *For every stochastic process the behavior is well defined.*

Proof. We call a behavior $\mathcal{B}$ compatible with a process $\hat{w}$ if $\mathcal{B}$ contains almost all process realizations. Of course, $(\mathbf{R}^q)^{\mathbf{Z}}$ is always compatible, and countable intersections of compatible behaviors are compatible.

Now let $\mathcal{B}$ be a compatible behavior. If it contains a strictly smaller compatible behavior $\mathcal{B}' \subset \mathcal{B}$, $\mathcal{B}' \neq \mathcal{B}$, then we proceed with $\mathcal{B}'$. This system has either fewer inputs than $\mathcal{B}$ or equal number of inputs and fewer states. Continuing in this way, we end up after a finite number of steps with a compatible behavior $\mathcal{B}^*$ that contains no strictly smaller compatible behavior. This implies that for every compatible $\mathcal{B}$ there holds $\mathcal{B} \cap \mathcal{B}^* = \mathcal{B}^*$, and thus $\mathcal{B}^* \subseteq \mathcal{B}$. This proves that $\mathcal{B}^*$ is the smallest compatible behavior. ∎

We call a behavior nontrivial if $\mathcal{B} \neq (\mathbf{R}^q)^{\mathbf{Z}}$. Dynamic factor models are defined as follows.

DEFINITION 2.2. *A dynamic factor model of a process w is a decomposition $w = \hat{w} + \tilde{w}$, where the latent process $\hat{w}$ has nontrivial behavior $\mathcal{B}$, which is called the behavior of the factor model.*

In this paper we will be mainly concerned with the behavior of factor models, as in many cases this is the main point of interest in system identification. In order to simplify our analysis of dynamic factor models we make some additional assumptions on the processes. Some of these assumptions could be relaxed, but they are imposed to prevent technical complications that could obscure the underlying modeling ideas. To formulate the assumptions we use the following terminology. Let S_t denote the subspace of $\mathbf{L}_2$ spanned by the zero mean random variables $\{w_i(t); i = 1, \ldots, q\}$. Let the Hilbert spaces $\mathbf{H}(w)$ and $\mathbf{H}_t(w)$ be generated by, respectively, $\{S_t; t \in \mathbf{Z}\}$ and $\{S_s; s \leq t\}$, so that $\mathbf{H}(w)$ is generated by the process and $\mathbf{H}_t(w)$ by the past of this process. The process is said to have full rank if the space $\mathbf{H}_t(w) \cap \{\mathbf{H}_{t-1}(w)\}^{\perp}$ has dimension q, that is, if no nontrivial linear combination of the variables $w(t)$ can be predicted without error from the past. It is called purely nondeterministic if $\bigcap_{-\infty}^{\infty} \mathbf{H}_t(w) = \{0\}$, that is, if the prediction of $w(t+h)$ from $\mathbf{H}_t(w)$ converges to zero for $h \to \infty$. As is well known, every purely nondeterministic process can be written as

$$w = T(\sigma^{-1})\varepsilon, \tag{6}$$

that is, $w(t) = \sum_{k=0}^{\infty} T_k \varepsilon(t-k)$, where ε is a white noise process with $\mathrm{E}\{\varepsilon(t)\varepsilon'(t)\} = I_q$ and $\varepsilon(t) \in \mathbf{H}_t(w)$, and where $\sum_{k=0}^{\infty} \|T_k\|_2^2 < \infty$ with $\|\cdot\|_2$ the Frobenius norm of a matrix. This is called a Wold representation of the process; see, e.g., [19]. If $\sum_{k=0}^{\infty} \|T_k\|_2 < \infty$, then this representation is called absolutely summable. In this paper we will always make the following assumptions.

Assumptions.

- A1. The processes w, $\hat{w}$, and $\tilde{w}$ are jointly weakly stationary, with zero mean.
- A2. The observed process w is purely nondeterministic and has full rank.
- A3. The latent process $\hat{w}$ and the noise process $\tilde{w}$ are purely nondeterministic.
- A4. The Wold representations of w, $\hat{w}$, and $\tilde{w}$ are absolutely summable.

The assumption A1 is imposed for convenience, as this means that the usual tools of time series analysis and linear systems theory can be applied. The full rank assumption in A2 implies that the behavior of the observed process is unrestricted, so that it can not be modeled by a factor model without noise. Concerning assumption A3, note that a latent process with nontrivial behavior can not be of full rank. We assume that it is purely nondeterministic, and that the same holds true for the noise. This seems a reasonable requirement in view of assumption A2. Finally, assumption A4 is imposed for technical reasons. It implies that the spectral densities of the processes are continuous functions on the unit circle.

Stated in terms of behaviors, assumption A3 for the latent process means the following.

PROPOSITION 2.3. *The behavior of a purely nondeterministic process is controllable.*

Proof. Let $\hat{w}$ be a purely nondeterministic process. Further let $\mathcal{B}$ be a noncontrollable system with full row rank polynomial representation R, with the property that $R(\sigma, \sigma^{-1})\,\hat{w} = 0$ almost surely. Let $R = UDV$ be the Smith form, with U and V unimodular matrices and with $D = (\Delta, 0)$ where Δ is a diagonal matrix with one-dimensional polynomials unequal to zero on the diagonal.

Define $w^* = V\,\hat{w}$ and let $w^* = (w_1^*, w_2^*)$ be a partitioning corresponding to that of $D = (\Delta, 0)$. Then there holds $\Delta w_1^* = 0$ almost surely. So this process evolves according to an autonomous difference equation and can be predicted without error; that is, w_1^* belongs to $\mathbf{H}_t(\hat{w})$ for all t, the space spanned by the past of $\hat{w}$. As $\hat{w}$ is purely nondeterministic, this means that $w_1^* = 0$. This shows that also $R^*(\sigma, \sigma^{-1})\,\hat{w} = 0$ almost surely, where $R^* = (I, 0)V$. As $R^*(z, z^{-1})$ has constant rank it follows that this defines a controllable system, and of course it defines a system that is strictly smaller than $\mathcal{B}$. So the behavior of $\hat{w}$ is also controllable. ∎

We mention that the converse of this result does not hold true; that is, a latent process with controllable behavior need not be purely nondeterministic. In terms of the representations of controllable systems discussed in section 2.1, the above result means that a factor model can be described as follows:

$$w = M(\sigma, \sigma^{-1})f + \tilde{w}, \tag{7}$$

$$w = Cx + Dv + \tilde{w}, \quad \sigma x = Ax + Bv. \tag{8}$$

Here M is a polynomial matrix and (A, B, C, D) are real-valued matrices. The first representation is a generalization of the static model of classical factor analysis and explains the observed variables in terms of a number of unobserved underlying factors. The second representation gives a more explicit description of the dynamical evolution of the latent process $\hat{w} = Cx + Dv$ in terms of unrestricted factors v and additional factors x that exhibit the memory structure.

Factor models can also be described by means of spectra. In terms of the Wold representation (6), where ε is white noise with unit covariance and where T is an (in general nonrational) causal transfer function with causal inverse, the spectrum of w is given by $\Sigma = TT^*$. The spectra of $\hat{w}$ and $\tilde{w}$ are denoted, respectively, by $\hat{\Sigma}$ and $\tilde{\Sigma}$, and the cross spectrum between $\hat{w}$ and $\tilde{w}$ is denoted by Σ_c. Under assumptions A1–A4, all these spectra are bounded functions on the unit circle. A factor model corresponds to a decomposition

$$\Sigma = \hat{\Sigma} + \tilde{\Sigma} + \Sigma_c + \Sigma_c{}'. \tag{9}$$

By assumption, the behavior of the latent process is nontrivial so that $\hat{\Sigma}$ is singular. The rank of this spectrum corresponds to the number of unrestricted factor components. This is made precise in the following result. Here we denote by $\ker(\hat{\Sigma})$ the set of $1 \times q$ polynomials $r(s, s^{-1})$ for which $r(z, z^{-1})\,\hat{\Sigma}(z) = 0$ on the unit circle. The polynomial rank of $\hat{\Sigma}$ is defined as $q - p$, where p is the dimension of the module $\ker(\hat{\Sigma})$. Further, by $\mathrm{im}(\hat{\Sigma})$ we denote the smallest linear system that contains all time series of the form $\hat{\Sigma}(\sigma)v$, where v is a $q \times 1$ time series with finite support.

THEOREM 2.4.

(i) *A latent process $\hat{w}$ with spectrum $\hat{\Sigma}$ has behavior $\mathcal{B} = \mathrm{im}(\hat{\Sigma})$, and the behavioral laws are given by $\mathcal{L} = \ker(\hat{\Sigma})$.*

(ii) *The number of inputs of the behavior is equal to the polynomial rank of $\hat{\Sigma}$.*

(iii) *A latent process has behavior $\mathcal{B}$ if and only if it can be generated as $\hat{w} = \hat{G}\,v$, where $\hat{G}$ is an isometric image representation of $\mathcal{B}$ and v is a weakly stationary process with zero mean and finite second-order moments that has trivial behavior.*

Proof. (i) Let $\mathcal{B}$ be the behavior of $\hat{w}$ and $\mathcal{L}$ the corresponding set of laws. Then a $1 \times q$ polynomial belongs to $\mathcal{L}$ if and only if $r\,\hat{w} = 0$ holds almost surely, and this is equivalent to the condition $r\,\hat{\Sigma} = 0$, that is, $\mathcal{L} = \ker(\hat{\Sigma})$.

Now let $\mathcal{B}^* = \mathrm{im}(\hat{\Sigma})$ be the smallest linear system that contains all time series of the form $\hat{\Sigma}(\sigma)v$, where v is a $q \times 1$ time series with finite support. Let $\mathcal{L}^*$ denote the set of laws of the system $\mathcal{B}^*$. The system $\mathcal{B}^*$ consists of pointwise limits of time series $\hat{\Sigma}(\sigma)v_n$, $n = 1, 2, \ldots$, where v_n are time series with finite support. If $r \in \mathcal{L}$, then $r\,\hat{\Sigma} = 0$ implies $r(\sigma)\,\hat{\Sigma}(\sigma)v_n = 0$, and the same holds true for the pointwise limit of $\hat{\Sigma}(\sigma)v_n$. This shows that $\mathcal{L} \subseteq \mathcal{L}^*$. Now let r be a $1 \times q$ polynomial with $r\,\hat{\Sigma} \neq 0$, and let $w \in \mathcal{B}^*$ be defined by $w = \hat{\Sigma}(\sigma)v$ where v has Z-transform r'. As $r\,\hat{\Sigma}\,r' \neq 0$, it follows that $r(\sigma)\,\hat{\Sigma}(\sigma)v \neq 0$, so that r does not belong to $\mathcal{L}^*$. This implies that $\mathcal{L}^* \subseteq \mathcal{L}$, so that $\mathcal{L}^* = \mathcal{L}$. As $\mathcal{B}$ and $\mathcal{B}^*$ satisfy the same relations, it follows that $\mathcal{B} = \mathcal{B}^*$.

(ii) The number of inputs of $\mathcal{B}$ is given by $m = q - p$, where p is the dimension of the module $\mathcal{L} = \ker(\hat{\Sigma})$. This was also defined as the polynomial rank of $\hat{\Sigma}$.

(iii) First assume that $\hat{w}$ has behavior $\mathcal{B}$ with m inputs. Let R be a $(q - m) \times q$ polynomial matrix with full rank so that $\mathcal{B} = \ker(R)$, and let $\hat{G}$ be an isometric image representation of $\mathcal{B}$ as defined in section 3.1, so that $R\,\hat{G} = 0$. As $\hat{G}$ is rational it can be written as $p^{-1}Q$, with p a scalar polynomial and Q a $q \times m$ matrix polynomial with full column rank. As $\hat{G}$ is stable, so that it has no poles on the unit circle, it follows that $\hat{v} = \hat{G}^*\,\hat{w}$ is a well-defined stationary process with zero mean and finite second-order moments. As $\hat{G}\,\hat{G}^*$ is the projection onto $\mathcal{B}$ and realizations of the factor process belong almost surely to $\mathcal{B}$, it follows that $\hat{G}\,\hat{v} = \hat{G}\,\hat{G}^*\,\hat{w} = \hat{w}$. It remains to show that $\hat{v}$ has trivial behavior $(\mathbf{R}^m)^{\mathbf{Z}}$. Suppose that this was not the case; then there is a $1 \times m$ polynomial $r \neq 0$ such that $r\,\hat{v} = 0$. As Q has rank m there exists a $1 \times q$ polynomial π so that $\pi Q = r$ and $\pi\,\hat{w} = p^{-1}\pi Q\,\hat{v} = 0$, so that π is a law of the process $\hat{w}$. It then follows that $(R', \pi')'\,\hat{G} = (0, r')'$, where $r \neq 0$. This implies that $(R', \pi')'$ is a polynomial matrix of rank $q - m + 1$ with the property that $(R', \pi')'\,\hat{w} = 0$. This means that the behavior of $\hat{w}$ has fewer than m inputs, but this contradicts (ii).

Second, suppose that $\hat{w} = \hat{G}\,\hat{v}$. As $\hat{v}$ has trivial behavior it follows that r is a behavioral law of $\hat{w}$ if and only if $r\,\hat{G} = 0$, or equivalently $r\,\hat{G}\,\hat{G}^* = rP = 0$ with P the projection operator onto $\mathcal{B}$. This shows that the behavior of $\hat{w}$ is given by $\mathcal{B}$. □

Concerning (ii), note that the polynomial rank of $\hat{\Sigma}$ is $q - p$, where p is the number of independent polynomial relations satisfied by the latent process $\hat{w}$. In

general, the polynomial rank may be larger than the dimension of the innovation space $\mathbf{H}_t(\hat{w}) \cap \{\mathbf{H}_{t-1}(\hat{w})\}^{\perp}$. This dimension is the usual definition of the rank of the process $\hat{w}$, and this is equal to the maximum of $\operatorname{rank}(\hat{\Sigma}(z))$ on the unit circle. This implies that for all $|z| = 1$ the rank of $\hat{\Sigma}(z)$ is smaller than or equal to the polynomial rank of $\hat{\Sigma}$, and if $\hat{w}$ satisfies additional linear relations that are not polynomial, then the rank of $\hat{\Sigma}(z)$ is strictly smaller than the polynomial rank of $\hat{\Sigma}$. As nonpolynomial relations correspond to infinite-dimensional systems, they fall outside the behavioral setting discussed in section 2.1.

2.3. Factor schemes. The basic question considered in this paper concerns the relationship between the spectrum of the observed process and the class of observationally equivalent factor models. Under assumption A2 there exists for every linear system $\mathcal{B}$ a factor model with behavior $\mathcal{B}$, because we can simply define the noise as $\tilde{w} = w - \hat{w}$ for every latent process $\hat{w}$. In the words of Kalman [15], within this setting we can obtain no models without prejudice. So we have to impose additional restrictions on the noise process in order to make meaningful statements about the underlying system. These restrictions should be motivated in each practical situation. Here we consider the following possible specifications, which we call factor schemes.

- The factor model is called orthogonal if the latent process and the noise process are mutually uncorrelated, that is, if $\mathrm{E}\{\hat{w}(t)\,\tilde{w}(s)'\} = 0$ for all t, s. Stated otherwise, there holds $\mathbf{H}(\hat{w}) \perp \mathbf{H}(\tilde{w})$ and $\Sigma_c = 0$.
- The factor model is called observable if $\hat{w}$ is a linear function of w, that is, if $\mathbf{H}(\hat{w}) \subseteq \mathbf{H}(w)$. Stated otherwise, there holds $\hat{\Sigma} = F\,\Sigma\,F^*$, $\tilde{\Sigma} = (I - F)\,\Sigma(I - F)^*$, and $\Sigma_c = F\,\Sigma(I - F)^*$ for some, possibly noncausal, transfer function F.
- The factor model is said to have bounded noise if it satisfies an a priori specified bound in terms of the noise spectrum $\tilde{\Sigma}$.

The quality of factor models is expressed in terms of the complexity and the goodness of fit of the model.

DEFINITION 2.5. *The* complexity *of a dynamic factor model is defined as the pair* (m, n)*, where* m *is the number of driving variables and* n *the number of states of the behavior of the factor model.*

The complexity measures the dimension of the latent process, in the sense that the set of possible realizations $\{\hat{w}(\omega); \omega \in \Omega\}$ on a time interval of length $L \geq n$ is (almost surely) contained in an $(mL+n)$-dimensional subspace of $\mathbf{R}^{qL}$. In parametric terms, the complexity can also be expressed as follows.

PROPOSITION 2.6.

(i) *In terms of a kernel representation* $R(\sigma, \sigma^{-1})\,\hat{w} = 0$*, the complexity is given by* $m = q - \operatorname{rank}(R)$ *and* $n = \sum_{k=1}^{q-m} \nu_k$*, where* $\{\nu_1, \ldots, \nu_{q-m}\}$ *are the Kronecker observability indices.*

(ii) *In terms of an isometric image representation* $\hat{G}$ *of the factor behavior, the complexity is given by the rank* m *and McMillan degree* n *of* $\hat{G}$*.*

Proof. (i) This follows from Theorem 6 in [24].

(ii) This follows from Theorem 4.9 and Lemma 4.10 in Chapter 4 of [14]. ∎

Below, we will sometimes consider another measure of complexity in case the factor model is observable and the spectrum Σ is rational, that is, $w = T(\sigma^{-1})\varepsilon$ in (6), where T is now a rational transfer function. Then a special class of latent processes is obtained by prefiltering the noise, that is, $\hat{w} = T(\sigma^{-1})F(\sigma^{-1})\varepsilon$, where F is a rational, rank deficient transfer function. We define the effective noise space by $\mathcal{N} = \operatorname{im}(F)$, that is, the behavior of the filtered noise process $F(\sigma^{-1})\varepsilon$. In this case the behavior of the latent process is given by $\mathcal{B} = \operatorname{im}(TF)$. As TF is rational

and rank deficient, it follows that $\mathcal{B}$ is a nontrivial linear system. An alternative characterization of complexity is the pair (m', n'), the number of inputs and states of the effective noise behavior $\mathcal{N}$. This measures the complexity of the noise process underlying the latent process.

DEFINITION 2.7. *Let a process with rational Wold representation $w = T(\sigma^{-1})\varepsilon$ and a latent process $\hat{w} = (TF)(\sigma^{-1})\varepsilon$ be given. Then the* noise complexity *of the corresponding factor model is defined as (m', n'), the number of inputs and states of the effective noise space $\mathcal{N} = \mathrm{im}(F)$.*

The two foregoing notions of complexity are not equivalent. If $\mathcal{N}$ is the effective noise space of a factor model with behavior $\mathcal{B}$ and if (m, n) and (m', n') are the complexities of $\mathcal{B}$ and $\mathcal{N}$, respectively, then $m = m'$, but in general $n \neq n'$.

The goodness of fit of factor models is measured in terms of the second moments of the noise process $\tilde{w}$. As is well known, the choice of norms may have an essential effect on the obtained models. Here we will restrict the attention to the mean squares norm and the uniform norm. In the following we use the notation $\tilde{\Sigma}^{1/2}$ for a spectral factor of the noise spectrum $\tilde{\Sigma}$ so that $\tilde{\Sigma} = \tilde{\Sigma}^{1/2}(\tilde{\Sigma}^{1/2})^*$. We define the norm of a $1 \times q$ polynomial $r(\sigma, \sigma^{-1}) = \sum r_k \sigma^k$ by $\|r\|_2^2 := \sum \|r_k'\|^2$, where $\|\cdot\|$ denotes the Euclidean norm on $\mathbf{R}^q$. Further, we define the following norms for spectral factors, where $\lambda_{\max}(Q)$ denotes the spectral radius, that is, the maximum of the absolute values of the eigenvalues of a matrix Q.

$$\| \tilde{\Sigma}^{1/2} \|_2^2 = \frac{1}{2\pi} \int_{-\pi}^{\pi} \mathrm{trace}\{\tilde{\Sigma}(e^{-i\lambda})\} d\lambda, \tag{10}$$

$$\| \tilde{\Sigma}^{1/2} \|_\infty^2 = \sup_{\lambda \in [-\pi, \pi]} \lambda_{\max}\{\tilde{\Sigma}(e^{-i\lambda})\}. \tag{11}$$

DEFINITION 2.8. *For a factor model with noise process $\tilde{w}$ with spectrum $\tilde{\Sigma}$, the mean squares and uniform fit are, respectively, defined by*

$$\| \tilde{w} \|_2 := [\mathrm{E}\{\tilde{w}(t)' \tilde{w}(t)\}]^{1/2} = \| \tilde{\Sigma}^{1/2} \|_2, \tag{12}$$

$$\| \tilde{w} \|_\infty := \sup\{[\mathrm{E}\{(r(\sigma, \sigma^{-1}) \tilde{w})(t)\}^2]^{1/2}; \|r\|_2 = 1\} = \| \tilde{\Sigma}^{1/2} \|_\infty. \tag{13}$$

Because of assumption A3, the noise process is purely nondeterministic, so that the coefficients of $\tilde{\Sigma}^{1/2}$ are square summable and $\|\cdot\|_2$ is well defined, and because of assumption A4, the spectrum is bounded on the unit circle so that $\|\cdot\|_\infty$ is also well defined. The mean squares and uniform norms are monotonic, since they become larger if the spectrum becomes larger in the sense of positive semidefinite matrix functions on the unit circle. Sometimes, when results hold true for both norms, we make no distinction in notation and write $\| \tilde{w} \|$ and $\| \tilde{\Sigma}^{1/2} \|$.

2.4. Illustrations.

2.4.1. Static factor models. As a simple illustration we show that the framework as introduced before is an extension to the dynamic case of the well-known class of static factor models that have been analyzed in [18] and [23], among others. In later sections we will use the static case for further illustration.

Suppose that the observations are uncorrelated over time, so that w is a white noise process. In this subsection we restrict our attention to factor models $w = \hat{w} + \tilde{w}$, where $\hat{w}$ and $\tilde{w}$ are also white noise processes. We further impose the condition that the behavior of the factor model is static in the sense that the state dimension is $n = 0$.

The corresponding linear systems are described by linear nondynamic equations of the form $R\,\hat{w} = 0$, where R is a full row rank $p \times q$ real matrix. Let M be a $q \times (q-p)$ matrix with $\mathrm{im}(M) = \ker(R)$; then the factor model can be written as

$$w(t) = Mf(t) + \tilde{w}(t).$$

This corresponds to the classical static factor model with factors f. If the covariance matrix of f has full rank, then the complexity of this factor model is $(m, 0)$, where $m = q - p$ is the number of factors.

In the literature several possible factor schemes have been proposed. For example, in the principal component analysis of multivariate statistics the aim is to keep the noise process $\tilde{w}$ as small as possible, under a restriction on the number of independent factors m. In the so-called Frisch scheme the aim is to minimize the complexity of the model under the restrictions that the processes $\hat{w}$ and $\tilde{w}$ are orthogonal and that, in addition, the q components of the noise process $\tilde{w}$ are mutually orthogonal.

Our approach resembles principal component analysis. In the next section we will consider minimization of the noise under a restriction of the complexity of the behavior of the latent process.

2.4.2. Dynamic system example. Here we give a simple example of a dynamic factor model. Suppose that the data generating process consists of a single input, single output system where both the input u and the output y are observed under additive noise. That is, we assume that the data $w = (u, y)$ are generated as $w = \hat{w} + \tilde{w}$, with $\tilde{w}$ the noise and $(\hat{u}, \hat{y})$ the latent process with $\hat{y} = g\,\hat{u}$, where g denotes the underlying rational transfer function. For simplicity we assume that the latent input $\hat{u}$ is white noise and that the noise process $\tilde{w}$ is also white noise, all uncorrelated and with unit variance. In this case the spectrum of the data generating process is given by

$$\Sigma = \begin{pmatrix} 2 & g^* \\ g & gg^* + 1 \end{pmatrix}.$$

An obvious factor model for this process is the above decomposition in the latent process $\hat{w}$ and the white noise process $\tilde{w}$. If $g(\sigma, \sigma^{-1}) = r_1(\sigma, \sigma^{-1})/r_2(\sigma, \sigma^{-1})$, then this factor model has a behavior described by the equation $R(\sigma, \sigma^{-1})\,\hat{w} = 0$ where $R = (-r_1, r_2)$. The complexity is $(m, n) = (1, d)$, where d is the maximum of the degrees of the polynomials r_1 and r_2. The mean squares fit is $\| \tilde{w} \|_2 = \sqrt{2}$, and the uniform fit is $\| \tilde{w} \|_\infty = 1$. Because of our assumptions, this factor model is orthogonal but not observable.

Of course, the real question is whether we can identify the underlying transfer function g from the spectrum Σ. This will be investigated in section 3.3.2.

3. Pareto optimal models. The quality of a factor model for an observed process w is measured by its complexity and goodness of fit. In general, the fit can become better if the model is allowed to be more complex. We use a lexicographic ordering of complexities, so that (m_1, n_1) is less complex than (m_2, n_2) if $m_1 < m_2$ or $m_1 = m_2, n_1 < n_2$. A factor model is called Pareto optimal if it satisfies the following two conditions: every less complex model has a strictly worse fit, and no equally complex model has strictly better fit. This means that the fit can not be improved without increasing the complexity and that the complexity can not be reduced without deteriorating the fit.

We characterize Pareto optimal models by optimizing the fit for a given bound on the complexity. This problem is analyzed in three steps. In section 4.1 we investigate

two cases: modeling with a specified behavior and modeling with a restricted number of inputs where the number of states is left free. In section 4.2 we derive Pareto optimal models of restricted complexity, where both the number of inputs and the number of states is limited. The optimality of models depends of course on the specification of the factor scheme, that is, on the choice of norms for the noise and on possible conditions of orthogonality and observability.

3.1. Optimal models of restricted rank. First assume that the behavior of the factor model has been specified a priori, so that the factor equations are given. The aim is to find a model with minimal error that satisfies these equations. Let $\mathcal{B}$ denote the given controllable linear system with polynomial representation $R(\sigma, \sigma^{-1})\,\hat{w} = 0$. The isometric image and kernel representations of the system are denoted, respectively, by $\hat{G}$ and $\tilde{G}$, so that $P_{\mathcal{B}} := \hat{G}\,\hat{G}^* = I - R^*(RR^*)^{-1}R$ is the projection operator onto the system and $\tilde{G}\,\tilde{G}^* = I - P_{\mathcal{B}}$ is the projection onto the set of behavioral equations. The following results hold true both for the mean squares and for the uniform norm.

THEOREM 3.1. *Let w be a process with spectrum Σ and let $\mathcal{B}$ be the required behavior of a factor model.*

(i) *A latent process with optimal fit is given by $\hat{w}_0 := P_{\mathcal{B}} w$, with noise spectrum $\tilde{\Sigma}_0 = (I - P_{\mathcal{B}})\,\Sigma(I - P_{\mathcal{B}}) = \tilde{G}\,\tilde{G}^*\,\Sigma\,\tilde{G}\,\tilde{G}^*$. The corresponding factor model is observable but, in general, not orthogonal.*

(ii) *Among orthogonal models, a latent process with optimal fit is given by $\hat{w}_0 := [I - \Sigma R^*(R\Sigma R^*)^{-1}R]w$, with corresponding noise spectrum $\tilde{\Sigma}_0 = \Sigma R^*(R\Sigma R^*)^{-1}R\Sigma = \Sigma\,\tilde{G}(\tilde{G}^*\,\Sigma\,\tilde{G})^{-1}\,\tilde{G}^*\,\Sigma$.*

Proof.

(i) The relation $\tilde{G}^*\,\hat{w} = 0$ implies for the mean squares norm

$$\begin{aligned} \|\,\tilde{\Sigma}^{1/2}\,\|_2^2 &= \oint_{|z|=1} \operatorname{trace}(\hat{G}^*\,\tilde{\Sigma}\,\hat{G})(z)dz + \oint_{|z|=1} \operatorname{trace}(\tilde{G}^*\,\tilde{\Sigma}\,\tilde{G})(z)dz \\ &\geq \oint_{|z|=1} \operatorname{trace}(\tilde{G}^*\,\Sigma\,\tilde{G})(z)dz. \end{aligned}$$

Therefore the misfit is minimal if and only if $\hat{G}^*\,\tilde{w} = 0$ holds, so that $\hat{w} = (\hat{G}\,\hat{G}^* + \tilde{G}\,\tilde{G}^*)\,\hat{w} = \hat{G}\,\hat{G}^*(w - \tilde{w}) = P_{\mathcal{B}} w$. This model is also optimal for the uniform norm, since

$$\lambda_{\max}(\tilde{\Sigma}(z)) \geq \lambda_{\max}((\tilde{G}\,\tilde{G}^*\,\tilde{\Sigma}\,\tilde{G}\,\tilde{G}^*)(z)) \quad = \quad \lambda_{\max}((\tilde{G}\,\tilde{G}^*\,\Sigma\,\tilde{G}\,\tilde{G}^*)(z))$$

holds for all points z of the unit circle. This optimal model is, in general, not orthogonal since $P_{\mathcal{B}}\,\Sigma(I - P_{\mathcal{B}})$ is not zero in general.

(ii) We show that $\tilde{\Sigma}(z) \geq \tilde{\Sigma}_0(z)$ holds for all points z of the unit circle. For simplicity of notation we omit the argument z in the following. Let $G = [\hat{G}, \tilde{G}]$; then, because of $\tilde{G}^*\,\hat{\Sigma} = 0$, it follows that $\tilde{G}^*\,\Sigma = \tilde{G}^*\,\tilde{\Sigma}$ and hence

$$\begin{aligned} G^*\,\tilde{\Sigma}\,G &= \begin{pmatrix} \hat{G}^*\,\tilde{\Sigma}\,\hat{G} & \hat{G}^*\,\Sigma\,\tilde{G} \\ \tilde{G}^*\,\Sigma\,\hat{G} & \tilde{G}^*\,\Sigma\,\tilde{G} \end{pmatrix} \\ &\geq \begin{pmatrix} (\hat{G}^*\,\Sigma\,\tilde{G})(\tilde{G}^*\,\Sigma\,\tilde{G})^{-1}(\tilde{G}^*\,\Sigma\,\hat{G}) & \hat{G}^*\,\Sigma\,\tilde{G} \\ \tilde{G}^*\,\Sigma\,\hat{G} & \tilde{G}^*\,\Sigma\,\tilde{G} \end{pmatrix} \\ &= G^*\,\Sigma\,\tilde{G}(\tilde{G}^*\,\Sigma\,\tilde{G})^{-1}\,\tilde{G}^*\,\Sigma\,G. \end{aligned}$$

The above inequality is a consequence of the fact that $G^*\,\tilde{\Sigma}\,G \geq 0$. So all orthogonal factor models with behavior $\mathcal{B}$ must satisfy $\tilde{\Sigma} \geq \Sigma\,\tilde{G}(\tilde{G}^*\,\Sigma\,\tilde{G})^{-1}\,\tilde{G}^*\,\Sigma = \tilde{\Sigma}_0$. This

shows the second expression for $\tilde{\Sigma}_0$. The first expression follows from the fact that $\tilde{G} = R^*Q$ where Q is a spectral factor of $(RR^*)^{-1}$, that is, $QQ^* = (RR^*)^{-1}$. □

The optimal factor model is unique in the case of the mean squares norm, but not generally in case of the uniform norm. If we are interested in factor behaviors only, then the above results show that we may restrict our attention to observable models. This leaves four factor schemes of interest: those for the mean squares and the uniform norm, and according to whether orthogonality is imposed or not. We define the distance between a behavior and a spectral density as the fit of the optimal factor model with this behavior. That is, the misfit function is given by

$$\mathrm{d}(\Sigma, \mathcal{B}) = \| \tilde{\Sigma}_0^{1/2} \|, \tag{14}$$

where $\tilde{\Sigma}_0$ is the noise spectrum of the optimal factor models for $\mathcal{B}$ given in Theorem 3.1 and where $\tilde{\Sigma}_0^{1/2}$ denotes a spectral factor of $\tilde{\Sigma}_0$. We use the same notation for the four different factor schemes.

Next we describe optimal models of restricted rank, so that only the number of inputs of the latent process is restricted, but not the number of state variables. Under the assumptions A1–A4 of section 2.2, the observed spectrum Σ is a well-defined matrix function on the unit circle that can be pointwise decomposed in terms of its eigenvalues and eigenvectors as $\Sigma = U\Lambda U^*$. Here U is a $q \times q$ unitary matrix function (i.e., $UU^* = U^*U = I_q$), and Λ is a diagonal matrix of ordered eigenfunctions. For simplicity we assume that the eigenvalues are distinct everywhere on the unit circle, so that $\Lambda = \mathrm{diag}(\lambda_1, \ldots, \lambda_q)$ with $\lambda_1(z) > \lambda_2(z) > \cdots > \lambda_q(z) > 0$ on the unit circle. Let $U = [U_1, U_2]$, where U_1 consists of the first m columns of U and U_2 of the remaining columns, and let $\Lambda = \mathrm{diag}(\Lambda_1, \Lambda_2)$ be a corresponding partitioning. The principal component model of rank m is defined by the factor $\hat{w} = U_1U_1^*w$ and noise $\tilde{w} = U_2U_2^*w$. In terms of the spectra, this gives

$$\hat{\Sigma}_m = U_1\Lambda_1U_1^*, \quad \tilde{\Sigma}_m = U_2\Lambda_2U_2^*, \quad \Sigma_c = 0. \tag{15}$$

Under the above assumptions, this model is well defined and unique (See [3, Theorems 9.3.1, 9.3.2, and 9.3.3]), and it is clearly observable and orthogonal. The latent process spectrum has rank m, but the factor behavior will in general be trivial; that is, it will be $(\mathbf{R}^q)^{\mathbf{Z}}$. This is because, in general, there exist no nontrivial polynomial equations such that $R(z, z^{-1})\,\hat{\Sigma}_m(z) = 0$.

The following result states that the principal component model has optimal fit, and that it can be approximated arbitrarily closely by factor models with complexity (m, n) if the number of state variables n is chosen sufficiently large. The results hold true for all factor schemes, that is, for mean squares and uniform fit and irrespective of whether orthogonality and observability are imposed or not.

THEOREM 3.2.

(i) *No factor model of complexity (m, n) has better fit than the fit $\| \tilde{\Sigma}_m^{1/2} \|$ of the principal component model of rank m.*

(ii) *For every $\varepsilon > 0$ there is a factor model of complexity (m, n), for some finite n, with better fit than $\| \tilde{\Sigma}_m^{1/2} \| + \varepsilon$.*

Proof. Under the assumption $\lambda_1(z) > \lambda_2(z) > \cdots > \lambda_q(z) > 0$ for all $|z| = 1$, the eigenvector matrix $U(z)$ has an absolutely summable Laurent series expansion; see [3, Theorems 9.3.1, 9.3.2, and 9.3.3]. This implies that $\tilde{w}_m = U_2U_2^*w$ and $\hat{w}_m = w - \tilde{w}_m = U_1U_1^*w$ are well-defined processes.

(i) As $\Sigma(z)$ is continuous on the unit circle it follows also that the eigenvalues $\lambda_i(z)$ are continuous functions (See Lemma 6.1 in the appendix), and thus $\| \tilde{\Sigma}_m^{1/2} \|_2^2 =$

$\oint_{|z|=1}\{\lambda_{m+1}(z)+\cdots+\lambda_q(z)\}dz$ and $\|\,\tilde{\Sigma}_m^{1/2}\,\|_\infty^2 = \sup_{|z|=1}\lambda_{m+1}(z)$ are well defined. If $\tilde{G}$ is the isometric kernel representation of a behavior $\mathcal{B}$, then the optimal noise covariance corresponding to $\mathcal{B}$ is, according to Theorem 3.1 given by $\tilde{\Sigma} = \tilde{G}\,\tilde{G}^*\,\Sigma\,\tilde{G}\,\tilde{G}^*$. As $\tilde{G}^*(z)\,\tilde{G}(z) = I$, Lemma 6.1 implies the optimality of the principal component model.

(ii) Since $U_2(\sigma)$ is an absolutely summable filter, we can find a positive integer N and a finite filter $\tilde{G}_N(\sigma) = \sum_{|k|\le N}\tilde{G}_k\,\sigma^k$ such that $\|U_2 - \tilde{G}_N\,\|_\infty$ is arbitrarily small. Thus we can choose N such that $\|I - \tilde{G}_N^*\,\tilde{G}_N\,\|_\infty$ and also $\|U_2U_2^* - \tilde{G}_N(\tilde{G}_N^*\,\tilde{G}_N)^{-1}\,\tilde{G}_N^*\,\|_\infty$ become arbitrarily small. The transfer function $P_N = \tilde{G}_N(\tilde{G}_N^*\,\tilde{G}_N)^{-1}\,\tilde{G}_N^*$ is a rational projection matrix of rank m, so that $(I - P_N)$ is the isometric image representation of a behavior $\mathcal{B}_N$ with m inputs and a finite number of states. Then, analogous to the proof of Proposition 6.3 in the appendix, it follows that $\|\,\tilde{\Sigma}_N - \tilde{\Sigma}_m\,\|_\infty \to 0$, and thus $\|\,\tilde{\Sigma}_N^{1/2}\,\| \to \|\,\tilde{\Sigma}_m^{1/2}\,\|$ by Lemma 6.2. Here U_2 corresponds to $\tilde{G}_0$ in the proof of Proposition 6.3, and this proof can easily be extended to the case where $\tilde{G}_0 = U_2$ is not rational but only absolutely summable. □

So the principal component model gives an optimal reduced rank approximation of the spectrum. Further, this gives a first idea of achievable combinations of complexity and fit. A sufficient condition for the existence of a factor model with fit δ and complexity (m, n), for some finite n, is that $\|\,\tilde{\Sigma}_m^{1/2}\,\| < \delta$, and a necessary condition is that $\|\,\tilde{\Sigma}_m^{1/2}\,\| \le \delta$.

We conclude this section by considering the effect of using weighted norms or, stated otherwise, the effect of prefiltering the observed process. Let Q be a $q \times q$ positive definite matrix function which is bounded on the unit circle. Then the Q-weighted norm is defined as $\|\,\tilde{w}\,\|_Q = \|T^*\,\tilde{w}\,\|$ for a spectral factorization $Q = TT^*$. This norm is well defined, as it does not depend on the choice of the spectral factor.

PROPOSITION 3.3. *Let $\mathcal{B}$ be a controllable linear system of complexity (m, n). Then there is a choice of Q-weights such that $\mathcal{B}$ is the behavior of a factor model that minimizes the Q-weighted norm over the set of all factor models with m inputs.*

Proof. Let $R(\sigma, \sigma^{-1})$ be a full row rank polynomial matrix with rows that form a basis for the set of laws of the behavior $\mathcal{B}$. As $\|\,\tilde{w}\,\|_Q = \|T^*\,\tilde{w}\,\|$ we can use the result of Theorem 3.2 on the transformed data $\bar{w} := T^*w$, with spectrum $T^*\,\Sigma\,T$. The transformed latent process $\hat{\bar{w}} = T^*\,\hat{w}$ satisfies the relation $R(T^*)^{-1}\hat{\bar{w}} = 0$. Thus, by Theorem 3.2, $\mathcal{B}$ is optimal with respect to the weighted norm $\|\,\tilde{w}\,\|_Q$ if $R(T^*)^{-1}$ is a basis of the left eigenspace of $T^*\,\Sigma\,T$ corresponding to the $q-m$ smallest eigenvalues, pointwise on the unit circle. In this case $\bar{w} = \hat{\bar{w}} + \tilde{\bar{w}}$ is the principal component model for the transformed data.

Now let $\bar{S}(\sigma, \sigma^{-1})$ be a full column rank polynomial matrix with columns that form a basis of the right kernel of R (i.e., $R\bar{S} = 0$), and let $S = \Sigma^{-1}\,\bar{S}$ and $Q = \Sigma^{-1} + SS^*$. If $\bar{Q} = \Sigma^{*/2}\,Q\,\Sigma^{1/2}$, then it follows that $R\,\Sigma^{1/2}\,\bar{Q} = R\,\Sigma^{1/2}$ and $S^*\,\Sigma^{1/2}\,\bar{Q} = (I + S^*\,\Sigma\,S)S^*\,\Sigma^{1/2}$. Thus the $q-m$ smallest eigenvalues of $\bar{Q}(z)$ are equal to 1, and $R(z, z^{-1})\,\Sigma^{1/2}(z)$ is a basis of the corresponding left eigenspace. Let $Q = TT^*$ and $\bar{\Sigma} = T^*\,\Sigma\,T$; then there holds $x\,\Sigma^{1/2}\,\bar{Q} = \lambda x\,\Sigma^{1/2}$ if and only if $x(T^*)^{-1}\bar{\Sigma} = \lambda x(T^*)^{-1}$. So the $q-m$ smallest eigenvalues of $\bar{\Sigma} = T^*\,\Sigma\,T$ are equal to one and $U_2 = R(T^*)^{-1}$ is a basis of the corresponding eigenspace. This shows that T is the appropriate transformation and Q the appropriate norm. □

This shows that the choice of norms is decisive for the obtained behaviors. So, in practical applications, it is imperative to take care of appropriate weighting of the data. In our opinion the norms should not be chosen on mathematical grounds alone,

but must be related to the information and objectives of each specific application. Here we will further restrict attention to the unweighted norms, which may be relevant in applications if the observed variables have been transformed appropriately.

3.2. Optimal models of restricted complexity. A straightforward method for determining Pareto optimal models is to fix the complexity and to optimize the fit under this constraint. A model of optimal fit is then Pareto optimal if there are no less complex models of at least equal fit. For complexity (m, n), this can be checked by comparing, first, with the optimal fit of models of complexity $(m, n-1)$ and, second, with the fit achievable by models having less than m inputs. The second comparison is simplified by the result of Theorem 3.2 for the principal component model of rank $m-1$. Because of these considerations, we restrict our attention to the determination of optimally fitting models of given complexity.

The main complication of the corresponding optimization problem is that the set of systems of given complexity (m, n) is not convex and also not compact. We restrict the attention to the mean squares norm and consider both the factor schemes with and without orthogonality. We will not investigate several other questions that are of interest in this context, such as the existence and unicity of optimal models and the case of the uniform norm.

The solution for the mean squares norm is given in terms of the so-called global total least squares algorithm presented in [21]. Let $W = (W_1, \dots, W_r)$ be a square summable $q \times r$ matrix sequence, that is, with $\|W\|_2^2 := \sum_{t=-\infty}^{\infty} \|W(t)\|_2^2 < \infty$, where $\|W(t)\|_2$ denotes the Frobenius norm of the matrix $W(t)$. Further, let the l_2-distance between this sequence and a linear system $\mathcal{B}$ be defined as $\mathrm{d}(W, \mathcal{B}) := \min\{\|W - V\|_2; V = (V_1, \dots, V_r)$ with $V_i \in \mathcal{B}, i = 1, \dots, r\}$. The objective in global total least squares is to determine an optimal model of restricted complexity, that is, one which minimizes the l_2-distance over the set of controllable systems with m inputs and n states. In general the optimal model exists and is unique, but existence and uniqueness may fail to hold true in exceptional cases. For algorithmic details we refer to [21] and [20], where a Gauss–Newton algorithm for the involved projections is described. If $\mathcal{B}$ is the optimal system, then $P_{\mathcal{B}}W$ is called the optimal l_2-approximation of W.

THEOREM 3.4. *Let $w = T\varepsilon$ be a given process with spectrum $\Sigma = TT^*$. For given complexity (m, n), a factor model with optimal mean squares fit is given by $w = \hat{w} + \tilde{w}$, where $\hat{w} = \hat{T}\,\varepsilon$ and $\tilde{w} = (T - \hat{T})\varepsilon$. Here $\hat{T}$ is the optimal l_2-approximation of complexity (m, n) for the spectral factor T. This model is observable but in general not orthogonal.*

Proof. According to Theorem 3.1, it is no restriction of generality if we consider only observable models. So let $\hat{w}(t) = [G_t(\sigma, \sigma^{-1})\varepsilon](t)$; then assumption A1 of joint stationarity of w and $\hat{w}$ implies that G_t is time invariant, say $G_t = G$. This means that we can write $\hat{w} = F(\sigma)w = G(\sigma)\varepsilon$ for some transfer function $G(\sigma) = F(\sigma)T(\sigma)$. As ε has full rank, it follows that the latent process has complexity (m, n), that is, $R\,\hat{w} = 0$ for a polynomial matrix R representing a system $\mathcal{B}$ with complexity (m, n) if and only if $RG = 0$; that is, all columns of G should belong to the system $\mathcal{B}$. The noise $\tilde{w} = (T - G)\varepsilon$ has spectrum $(T - G)(T - G)^*$ and mean squares norm $\|\,\tilde{w}\,\|_2^2 = \frac{1}{2\pi}\int_{-\pi}^{\pi} \mathrm{trace}\{(T - G)(T - G)^*(e^{-i\lambda})\}d\lambda$. But this is precisely equal to $\|T - G\|_2^2$, the l_2-distance between T and G. So this minimization problem is the l_2- approximation problem for T where each of the q columns of G should belong to the same system of complexity (m, n). The optimal choice over this class is by definition given by $\hat{T}$.

It can be shown that the factor filter $\hat{T}$ and the noise filter $\tilde{T} := T - \hat{T}$ satisfy $\hat{T}^*\,\tilde{T} = 0$, but in general $\Sigma_c = \hat{T}\,\tilde{T}^* \neq 0$ so that the processes $\hat{w}$ and $\tilde{w}$ are not orthogonal. ∎

Next we characterize optimal models under the condition of orthogonality. In order to simplify the analysis, we restrict our attention to observed processes with rational spectrum Σ and use the alternative definition of complexity in terms of the effective noise space; see Definition 2.7.

THEOREM 3.5. *Let $w = T\varepsilon$ be a given process with spectrum $\Sigma = TT^*$. For given noise complexity (m, n), an orthogonal factor model with optimal mean squares fit is given by $w = \hat{w} + \tilde{w}$, where $\hat{w} = S\varepsilon$ and $\tilde{w} = (T - S)\varepsilon$. Here S^* is the optimal l_2-approximation of complexity (m, n) for the adjoint T^* of the spectral factor T.*

Proof. Within this setting a latent process is given by $\hat{w} = TF\varepsilon$ where the factor noise space $\mathcal{N} = \mathrm{im}(F)$ has complexity (m, n). The noise is then given by $\tilde{w} = T(I - F)\varepsilon$, and the orthogonality condition is equivalent to requiring $TF(I - F)^*T^* = 0$. As T has full rank everywhere, it follows that $F = F^* = F^2$ is a projection, namely the orthogonal projection onto the system $\mathcal{N}$. The noise has norm $\| \tilde{w} \|_2^2 = \frac{1}{2\pi} \int_{-\pi}^{\pi} \mathrm{trace}\{T(I - F)T^*(e^{-i\lambda})\} d\lambda$. This is equal to $\|(I - F)T^*\|_2^2 = \|T^* - S^*\|_2^2$, where each column of S^* is the optimal l_2-approximation within the system $\mathcal{N}$ of the corresponding column of T^*, as F is the projection onto this system. The optimal choice of the model, that is, of F or equivalently of $\mathcal{N}$ of complexity (m, n), is precisely the optimal l_2-approximation problem of T^*. ∎

3.3. Illustrations.

3.3.1. The static case. The foregoing results can easily be applied to the case of static factor models. Let w be a white noise process, so that the spectrum Σ is a constant function, the covariance matrix of the process. The principal component model is then obtained by the eigenvalue decomposition of the matrix Σ. The optimal latent process with m factors is given by the projection of the observations onto the space spanned by the eigenvectors corresponding to the m leading eigenvalues of Σ. Therefore, in the optimal factor model, both the latent process and the noise are white noise processes. It follows from Theorem 3.2(i) that the principal component model is Pareto optimal among all models of complexity (m, n) for all $n \geq 0$. That is, no gain of fit is possible by allowing for dynamic equations.

For the static case, the result in Proposition 3.3 has also been pointed out in [15] and [16]. In the ordinary least squares scheme the indeterminateness of optimal models is resolved by the assumption that certain variables are noise free, i.e., that a principal submatrix of the noise covariance $\tilde{\Sigma}$ is zero. In terms of the weighting matrix Q this means that certain noise directions are assigned an infinite weight. In our approach, however, we treat all variables in a symmetric way.

3.3.2. Dynamic system example. Next we consider the dynamic errors-in-variables system described in section 2.4.2, and we use the notation introduced there. So let the spectrum Σ be given, and assume that the complexity (m, n) has been specified with $m = 1$ and $n \geq d$. The principal component decomposition for fixed frequency is easily obtained, with eigenvalues $\lambda_1 = 2 + gg^*$ and $\lambda_2 = 1$ and the eigenvector corresponding to λ_2 given by $(-g, 1)^*$. We denote the corresponding latent process by $\hat{w}_* = (\hat{u}_*, \hat{y}_*)$ and the noise process by $\tilde{w}_*$. This shows that the principal component model has a behavior that is finite dimensional, and this model is Pareto optimal among all models of complexity $(1, n)$ with $n \geq d$. The underlying transfer function g has been identified, because $\hat{y}_* = g\,\hat{u}_*$.

Although the underlying behavior has been identified, this is not the case for the true latent process and noise process. This can be seen from the spectral properties of the noise processes. The noise that affects the data has spectrum I_2 of rank 2, whereas the noise $\tilde{w}_*$ has a spectrum of only rank 1. Further, the factor model $w = \hat{w} + \tilde{w}$ has

a mean squares error $\| \tilde{w} \|_2 = \sqrt{2}$, whereas the principal component model has error $\| \tilde{w}_* \|_2 = 1$. We remark that both models are in fact optimal for the uniform norm.

This shows that in this case the Pareto optimal model indeed identifies the latent transfer function g from the observed spectrum Σ, at least when the complexity is not chosen too small. We should remark that this result depends in a crucial way on our assumptions about the way the data are generated. For example, if the observation noise $\tilde{w}$ would not be white, then Pareto optimal models will in general not have transfer function g. In terms of Proposition 3.3 this would require an appropriate prefiltering of the data. In our example, the required filter Q is the identity, that is, our data generating process is such that the unweighted norm is appropriate to identify the underlying transfer function. For practical applications this means that, in order to find good approximations of the underlying system, one should incorporate available information on the noise properties.

4. Consistency.

4.1. System topology. We introduce the topologies on linear systems and spectra that we will use in our analysis of continuity properties of factor models. For linear systems the gap metric is defined in terms of the projections described at the end of section 2.1.

DEFINITION 4.1. *Let $\mathcal{B}_1, \mathcal{B}_2$ be linear systems with isometric image representations $\hat{G}_1$ and $\hat{G}_2$, respectively; then the gap between these systems is defined by*

$$\mathrm{d}(\mathcal{B}_1, \mathcal{B}_2) = \| \hat{G}_1 \hat{G}_1^* - \hat{G}_2 \hat{G}_2^* \|_\infty. \tag{16}$$

This is analogous to the usual definition of the gap between two closed linear subspaces of a Hilbert space as $\|P_1 - P_2\|$, where P_1 and P_2 are the orthogonal projection operators onto the two spaces. Here $\hat{G}_i \hat{G}_i^*$ is the orthogonal projection onto the set of square summable time series in the behavior $\mathcal{B}_i$, $i = 1, 2$.

PROPOSITION 4.2.

(i) *The gap* d *is a metric on the class of controllable linear systems.*

(ii) *In terms of system restrictions, if $\tilde{G}_i$ denotes an isometric kernel representation of $\mathcal{B}_i$, $i = 1, 2$, then* $\mathrm{d}(\mathcal{B}_1, \mathcal{B}_2) = \| \tilde{G}_1 \tilde{G}_1^* - \tilde{G}_2 \tilde{G}_2^* \|_\infty$.

(iii) *If two systems have a different number of inputs, then their gap equals one.*

Proof. (i) This holds true for so-called l_2 systems, and this implies the same result for controllable systems since these are in one-to-one correspondence with l_2 systems. See Corollaries 3-4 and 5-3 of Chapter 4 in [14].

(ii) This follows from the fact that $[\hat{G}, \tilde{G}]$ is inner, so that $\hat{G}\hat{G}^* + \tilde{G}\tilde{G}^* = I$.

(iii) See Proposition 5-5 of Chapter 4 in [14]. ∎

In the following, we denote by $\mathbf{B}(m, n)$ the set of all controllable linear systems with m inputs and n states, by $\overline{\mathbf{B}}(m, n) := \bigcup_{k=1}^{n} \mathbf{B}(m, k)$ the set of all controllable linear systems with m inputs and at most n states, and by $\mathbf{B} := \bigcup_{m=0}^{q} \bigcup_{n=0}^{\infty} \mathbf{B}(m, n)$ the set of all controllable linear systems.

PROPOSITION 4.3.

(i) *For $n > 0$ the set $\mathbf{B}(m, n)$ is neither open nor closed in $\mathbf{B}$.*

(ii) *The set $\overline{\mathbf{B}}(m, n)$ is the closure of $\mathbf{B}(m, n)$ in $\mathbf{B}$.*

(iii) *The sets $\mathbf{B}$ and $\overline{\mathbf{B}}(m, n)$, for $n > 0$, are not compact.*

Proof. (i) For $n = 0$ the only controllable systems are described by the isometric state parameters $(A, B, C, D) = (-, -, -, D)$ with corresponding static projection operator DD'. It follows that $\mathbf{B}(m, 0)$ is a compact set, and this will be described in more detail in section 4.4. We will now consider the case $n > 0$.

In order to show that $\mathbf{B}(m,n)$ is not open it suffices to construct a sequence of systems $\mathcal{B}_k \in \mathbf{B}(m,n+1)$ with $\mathrm{d}(\mathcal{B}_k,\mathcal{B}_0) \to 0$ where $\mathcal{B}_0 \in \mathbf{B}(m,n)$. Let (A_0,B_0,C_0,D_0) be a minimal isometric state representation of $\mathcal{B}_0$ and let $a \in \mathbf{R}, b \in \mathbf{R}^{1\times m}$, and $c \in \mathbf{R}^{q\times 1}$ be such that $A = \left(\begin{smallmatrix} A_0 & 0 \\ 0 & a \end{smallmatrix}\right)$, $B = \left(\begin{smallmatrix} B_0 \\ b \end{smallmatrix}\right)$, $C_k = (C_0, \varepsilon_k c)$ is an observable and controllable quadruple for all $\varepsilon_k > 0$. The system $\mathcal{B}_0$ has transfer function $\hat{G}_0 = D_0 + C_0(zI - A_0)^{-1}B_0$, and let the system $\mathcal{B}_k$ be defined by the transfer function $\hat{G}_k = D_0 + C_k(zI-A)^{-1}B = \hat{G}_0 + \varepsilon_k(z-a)^{-1}cb$ with $\varepsilon_k \to 0$ for $k \to \infty$. Then $\mathcal{B}_k \in \mathbf{B}(m,n+1)$ and clearly $\|G_k - G_0\|_\infty \to 0$ and $\mathrm{d}(\mathcal{B}_k,\mathcal{B}_0) = \| \hat{G}_k(\hat{G}_k^* \hat{G}_k)^{-1} \hat{G}_k^* - \hat{G}_0 \hat{G}_0^* \|_\infty \to 0$ for $k \to \infty$.

That $\mathbf{B}(m,n)$ is not closed follows in a similar way, by constructing a sequence in $\mathbf{B}(m,n)$ that converges to a system in $\mathbf{B}(m,n-1)$.

(ii) Let cl $\mathbf{B}(m,n)$ denote the closure of $\mathbf{B}(m,n)$. Systems with $m' \neq m$ do not belong to this closure, since such systems have gap one with respect to all systems in $\mathbf{B}(m,n)$; see Proposition 4.2(iii). Systems with m inputs and less than n states can be obtained as the limit of sequences of systems in $\mathbf{B}(m,n)$, by similar constructions as in the proof of (i). It remains to prove that systems in $\mathbf{B}(m,n')$ with $n' > n$ do not belong to cl $\mathbf{B}(m,n)$. Let $\mathcal{B} \in \mathbf{B}(m,n')$ with $n' > n$ have isometric image representation $\hat{G}$; then the projection operator $P = \hat{G}\hat{G}^*$ is a rational function with rank m and McMillan degree $2n'$. As projection operators corresponding to systems in $\mathbf{B}(m,n)$ have rank m and McMillan degree $2n$, it follows that such operators can not converge to P, so that $\mathcal{B}$ does not belong to cl $\mathbf{B}(m,n)$.

(iii) As $\mathbf{B}$ is a metric space, it suffices to prove that there exists a sequence of systems $\mathcal{B}_k \in \overline{\mathbf{B}}(m,n)$ which has no convergent subsequence in the set $\mathbf{B}$ of all controllable linear systems. Consider the case $q = 2, m = 1, n = 1$ and the systems described by the isometric state parameters $\left(\begin{smallmatrix} a & \beta\delta C'D \\ \gamma C & \delta D \end{smallmatrix}\right)$, where $0 < a < 1$ is a real number, C and D are 2×1 vectors of unit length, and β, γ, δ are real numbers to obtain an isometric matrix; that is, $\gamma = \sqrt{1-a^2}$, $\beta = -\gamma/a$, and $\delta = \{1 + \beta^2(C'D)^2\}^{-1/2}$. To guarantee minimality it is further assumed that $C'D \neq 0$. The corresponding isometric image representations are given by $\hat{G}(z) = \delta D + \beta\gamma\delta(C'D)C(z-a)^{-1}$, and the projection operators by $P = \hat{G}\hat{G}^*$. If $a \uparrow 1$ then $\gamma \to 0, \beta \to 0$, and $\delta \to 1$, so that the pointwise limit of $\hat{G}(z)$ is D for $z \neq 1$ and $\hat{G}(1)$ converges to $D - 2(C'D)C$. If the corresponding sequence of systems would have a limiting point, say with projection operator P_0, then it should hold that $\|P_0 - P\|_\infty \to 0$ for $a \uparrow 1$. As $P_0(z)$ is continuous on the unit circle the only candidate for P_0 is given by DD', but as $P(z)$ is also continuous and $P(1) \not\to DD'$ for $a \uparrow 1$, it follows that no subsequence can converge to a system in $\mathbf{B}$. □

In our analysis, not only the distance between two systems but also the distance between two sets of systems is of relevance. If $\mathbf{B}_1$ and $\mathbf{B}_2$ are two compact subsets of $\mathbf{B}$, then the Hausdorff distance between these sets is defined as

$$\mathrm{d_H}(\mathbf{B}_1,\mathbf{B}_2) := \max\{\rho(\mathbf{B}_1,\mathbf{B}_2), \rho(\mathbf{B}_2,\mathbf{B}_1)\}, \tag{17}$$

where $\rho(\mathbf{B}_1,\mathbf{B}_2) := \sup_{\mathcal{B}_1\in\mathbf{B}_1} \inf_{\mathcal{B}_2\in\mathbf{B}_2} \mathrm{d}(\mathcal{B}_1,\mathcal{B}_2)$.

In order to investigate continuity properties we also need a topology on the set of spectral densities. We use the metric defined by

$$\mathrm{d}(\Sigma_1,\Sigma_2) = \| \Sigma_1 - \Sigma_2 \|_\infty := \sup_{\lambda\in[-\pi,\pi]} \lambda_{\max}\{\Sigma_1(e^{-i\lambda}) - \Sigma_2(e^{-i\lambda})\}. \tag{18}$$

Under assumption A4 the spectra are bounded on the unit circle, so that this is a well-defined metric.

4.2. Continuity. We consider the relation between observed spectra and identified factor behaviors. For given spectrum Σ, complexity (m, n), and noise bound δ, we denote by $\mathbf{B}(\Sigma; \delta, m, n) \subseteq \mathbf{B}(m, n)$ the set of all behaviors of factor models $w = \hat{w} + \tilde{w}$ satisfying the conditions that the factor behavior has m inputs and n states and that the noise process has norm $\| \tilde{w} \| \leq \delta$. So this corresponds to the factor scheme with bounded noise. The set $\mathbf{B}(\Sigma; \delta, m, n)$ depends of course on the measure of fit and on the possible condition of orthogonality. As the results in this section hold true for all the four corresponding factor schemes, we will make no explicit distinction between them. Systems in $\mathbf{B}(\Sigma; \delta, m, n)$ are called feasible for the data Σ and the specified complexity and fit. The feasibility of a given behavior can be checked by means of the results in Theorem 3.1.

PROPOSITION 4.4.

(i) *The set of feasible systems* $\mathbf{B}(\Sigma; \delta, m, n)$ *depends on whether orthogonality is imposed or not, but it does not depend on whether observability is imposed or not.*

(ii) *The set* $\mathbf{B}(\Sigma; \delta, m, n)$ *is closed in* $\mathbf{B}(m, n)$ *but in general not in* $\mathbf{B}$.

(iii) *If* $\mathcal{B} \in \mathbf{B}(\Sigma; \delta, m, n)$ *has fit strictly better than* δ, *then it is an inner point of* $\mathbf{B}(\Sigma; \delta, m, n)$.

Proof. (i) This follows from Theorem 3.1.

(ii) The set $\mathbf{B}(\Sigma; \delta, m, n)$ is closed in $\mathbf{B}(m, n)$ by Proposition 6.3 in the appendix, but not in $\mathbf{B}$, as follows from Proposition 4.3(i).

(iii) This is immediate from Proposition 6.3. ∎

In order to use the Hausdorff metric (17) we next formulate a sufficient condition for compactness. We call a state dimension n *minimal* for given (Σ, δ, m) if there exists a feasible model of complexity (m, n) but not one of complexity (m, n') with $n' < n$, that is, if $\mathbf{B}(\Sigma; \delta, m, n) \neq \emptyset$ and $\mathbf{B}(\Sigma; \delta, m, n') = \emptyset$ for all $n' < n$. If we are only interested in Pareto optimal models, then this minimality condition can be imposed without loss of generality.

PROPOSITION 4.5. *If* n *is minimal for* (Σ, δ, m), *then the set of feasible systems* $\mathbf{B}(\Sigma; \delta, m, n)$ *is compact.*

Proof. We prove this in terms of isometric state space representations. For this purpose we first describe this parametrization in some more detail. By definition, systems in $\mathbf{B}(m, n)$ are controllable and so can be represented by an isometric state model that satisfies (5). Let $\Pi(m, n) \subset \mathbf{R}^{(n+q)\times(n+m)}$ be the set of all such minimal isometric system matrices and let $\Pi = \bigcup_{m=0}^{q} \bigcup_{n=1}^{\infty} \Pi(m, n)$. On this set we define the metric $\mathrm{d}(\pi_1, \pi_2) = \|\pi_1 - \pi_2\|_\infty$ if $(m_1, n_1) = (m_2, n_2)$ and $\mathrm{d}(\pi_1, \pi_2) = 3$ otherwise. It is easily verified that this is a metric on Π and that $\Pi(m, n)$ is open in Π. That the parametrization of $\mathbf{B}$ by Π is continuous can be seen as follows. Let $\pi_k \to \pi_0$; then for k sufficiently large there holds $(m_k, n_k) = (m_0, n_0)$. As π_0 is a minimal isometric representation, it follows that $A_0 A_0' + C_0 C_0' = I$ with (A_0, C_0) observable, so that A_0 has all its eigenvalues strictly within the unit circle. Then the mapping from (A, B, C, D) to the isometric image representation $\hat{G} = D + C(zI - A)^{-1}B$ is continuous in π_0, and so $\mathrm{d}(\mathcal{B}_k, \mathcal{B}_0) = \| \hat{G}_k \hat{G}_k^* - \hat{G}_0 \hat{G}_0^* \| \to 0$ for $k \to \infty$.

Because the parametrization is continuous, in order to prove that $\mathbf{B}(\Sigma; \delta, m, n)$ is a compact subset of $\mathbf{B}$ it suffices to prove that the corresponding set of parameters denoted by $\Pi_0 \subset \Pi$ is compact. As $\Pi(m, n) \subset \Pi$ is open, it suffices to prove that Π_0 is a compact subset of $\Pi(m, n)$, or also that it is a closed and bounded subset of the Euclidean space $\mathbf{R}^{(n+q)\times(n+m)}$. Because of the isometry condition, boundedness is evident, so it remains to prove the closedness of Π_0. We prove this by contradiction.

Suppose that there is a sequence of systems $\mathcal{B}_k \in \mathbf{B}(\Sigma;\delta,m,n)$ with minimal isometric representations $(A_k,B_k,C_k,D_k) \to (A_0,B_0,C_0,D_0)$ so that the system $\mathcal{B}_0$ corresponding to these limit parameters does not belong to $\mathbf{B}(\Sigma;\delta,m,n)$. Then A_0 has eigenvalues on the unit circle. Indeed, if this were not the case, then the parametrization would be continuous in (A_0,B_0,C_0,D_0) and hence, by Proposition 6.3, it would follow that $\mathrm{d}(\Sigma,\mathcal{B}_0) = \lim \mathrm{d}(\Sigma,\mathcal{B}_k) \le \delta$. As $\mathcal{B}_0$ has m inputs and n is assumed to be minimal for (Σ,δ,m), it would follow that $\mathcal{B}_0 \in \mathbf{B}(\Sigma;\delta,m,n)$, contradicting our assumption. Now, state directions corresponding to eigenvectors of unit eigenvalues of A_0 are not observable because of the isometry condition $A_0'A_0 + C_0'C_0 = I$. So the state space for $\mathcal{B}_0$ can be reduced by deleting such unobservable directions. Let (A,B,C,D_0) be the restriction of (A_0,B_0,C_0,D_0) to the observable subspace, so that A has all its eigenvalues strictly within the unit circle. Because the two representations describe the same system $\mathcal{B}_0$ with the same driving variables, it follows that $G_0(z) := D_0 + C(zI-A)^{-1}B = D_0 + C_0(zI-A_0)^{-1}B_0$ pointwise on the unit circle, with the exception of the eigenvalues $\{e^{-i\lambda_j}; j=1,\dots,r\}$ of A_0. Moreover, as G_0 is the pointwise limit of $G_k = D_k + C_k(zI-A_k)^{-1}B_k$ it follows that G_0 is an isometric image representation of $\mathcal{B}_0$, with m inputs and at most $n-r$ states.

We consider first the factor scheme without orthogonality and with the uniform norm. Using the notation (14), we obtain from Theorem 3.1(i) that the fit of the system in this case is given by $\mathrm{d}(\Sigma,\mathcal{B}_0) = \| \tilde{\Sigma}_0^{1/2} \|_\infty$, where $\tilde{\Sigma}_0 = (I - G_0G_0^*)\,\Sigma(I - G_0G_0^*)$. As n is minimal for (Σ,δ,m) and $\mathcal{B}_0$ has less than n states, it follows that $\sup_{\lambda\in[-\pi,\pi]} \lambda_{\max}\{\tilde{\Sigma}_0(e^{-i\lambda})\} > \delta^2$. Because of the continuity of $G_0(z)$ and $\Sigma(z)$ on the unit circle, there exists an $\varepsilon > 0$ so that also $\sup_{\lambda\in\Lambda} \lambda_{\max}\{\tilde{\Sigma}_0(e^{-i\lambda})\} > \delta^2$ where $\Lambda = \{\lambda \in [-\pi,\pi]; |\lambda - \lambda_j| \ge \varepsilon \text{ for all } j = 1,\dots,r\}$. As G_k converges pointwise to G_0 on the compact set Λ this implies that for k sufficiently large $\{\mathrm{d}(\Sigma,\mathcal{B}_k)\}^2 \ge \sup_{\lambda\in\Lambda} \lambda_{\max}\{(I-G_kG_k^*)\,\Sigma(I-G_kG_k^*)(e^{-i\lambda})\} > \delta^2$, but this contradicts the statement that $\mathcal{B}_k \in \mathbf{B}(\Sigma;\delta,m,n)$. This proves compactness for the factor scheme without orthogonality and with the uniform norm.

The result for the orthogonal factor scheme with uniform norm follows in a similar way by using Theorem 3.1(ii). For the mean squares norm the reasoning is similar. Under the previous assumptions, there would exist an $\varepsilon > 0$ such that $\frac{1}{2\pi}\int_\Lambda \mathrm{trace}\{\tilde{\Sigma}_0(e^{-i\lambda})\}d\lambda > \delta^2$, and as G_k converges uniformly to G_0 on the compact set Λ, this gives a contradiction as before. □

The set of feasible systems does not in general depend in a fully continuous way on the observed spectrum. Therefore, we use the weaker concept of upper semicontinuity. We call the set of feasible systems $\mathbf{B}(\Sigma;\delta,m,n)$ upper semicontinuous in (Σ,δ) if for all $(\Sigma_k,\delta_k) \to (\Sigma,\delta)$ and for $\mathcal{B}_k \in \mathbf{B}(\Sigma_k;\delta_k,m,n)$ with $\mathcal{B}_k \to \mathcal{B}_0$ it holds that $\mathcal{B}_0 \in \mathbf{B}(\Sigma;\delta,m,n)$. As the sets of feasible systems are in general not compact, upper semicontinuity is not equivalent to the condition that $\rho(\mathbf{B}_k,\mathbf{B}_0) = \sup_{\mathcal{B}_k\in\mathbf{B}_k} \inf_{\mathcal{B}_0\in\mathbf{B}_0} \mathrm{d}(\mathcal{B}_k,\mathcal{B}_0) \to 0$, where $\mathbf{B}_k := \mathbf{B}(\Sigma_k;\delta_k,m,n)$ and $\mathbf{B}_0 := \mathbf{B}(\Sigma;\delta,m,n)$. The following continuity results for feasible systems are valid for all factor schemes, that is, for the mean squares and uniform fit and for the cases with and without orthogonality constraint. We use the notation $\overline{\mathbf{B}}(\Sigma;\delta,m,n)$ for the set of all feasible systems for (Σ,δ) with m inputs and at most n states.

PROPOSITION 4.6.

(i) *The set* $\overline{\mathbf{B}}(\Sigma;\delta,m,n)$ *is upper semicontinuous in* (Σ,δ).

(ii) *If* n *is minimal for* (Σ,δ,m), *then* $\mathbf{B}(\Sigma;\delta,m,n)$ *is upper semicontinuous in* (Σ,δ).

(iii) *Let n be minimal for $(\Sigma; \delta+\eta, m, n)$ for some $\eta > 0$ and let $\mathbf{B}(\Sigma; \delta, m, n)$ be nonempty; then $\rho(\mathbf{B}(\Sigma_k; \delta_k, m, n), \mathbf{B}(\Sigma; \delta, m, n)) \to 0$ if $(\Sigma_k, \delta_k) \to (\Sigma, \delta)$.*

(iv) *Under the conditions in* (iii), *$\mathbf{B}(\Sigma; \delta, m, n)$ is continuous from the right in δ.*

Proof. (i) Let $(\Sigma_k, \delta_k) \to (\Sigma, \delta)$ and $\mathcal{B}_k \in \overline{\mathbf{B}}(\Sigma_k; \delta_k, m, n)$ with $\mathcal{B}_k \to \mathcal{B}$; then we have to prove that $\mathcal{B} \in \overline{\mathbf{B}}(\Sigma; \delta, m, n)$. That $\mathcal{B}$ has m inputs and at most n states follows from the fact that $\overline{\mathbf{B}}(m, n)$ is closed; see Proposition 4.3(ii). Further, Proposition 6.3 in the appendix implies that $\mathrm{d}(\Sigma_k, \mathcal{B}_k) \to \mathrm{d}(\Sigma, \mathcal{B})$, and this implies that $\mathrm{d}(\Sigma, \mathcal{B}) \le \delta$ so that $\mathcal{B} \in \overline{\mathbf{B}}(\Sigma; \delta, m, n)$.

(ii) This corresponds to the situation in (i), where now $\mathcal{B}_k$ all have complexity (m, n). If $\mathcal{B}_k \to \mathcal{B}$ then $\mathcal{B} \in \overline{\mathbf{B}}(m, n)$ and $\mathrm{d}(\Sigma, \mathcal{B}) \le \delta$. As n is minimal for (Σ, δ, m), it follows that $\mathcal{B} \in \mathbf{B}(m, n)$, so that $\mathcal{B} \in \mathbf{B}(\Sigma; \delta, m, n)$.

(iii) In a first step we prove that n is minimal for $(\Sigma_k, \delta+\eta, m)$ for all k large enough. If this is not true, then there exist $n' < n$ and infinitely many indices k, so that $\mathbf{B}(\Sigma_k; \delta+\eta, m, n')$ is not empty. For such indices let $\mathcal{B}_k \in \mathbf{B}(\Sigma_k; \delta+\eta, m, n')$ have minimal isometric representation (A_k, B_k, C_k, D_k); then the isometry condition implies that this sequence has a limit point, denoted by (A_0, B_0, C_0, D_0). Let $\mathcal{B}_0$ be the behavior corresponding to these parameters; then $\mathcal{B}_0 \in \mathbf{B}(m, n'')$ with $n'' \le n'$. As in the proof of Proposition 4.5, the isometric kernel representations $\tilde{G}_k$ converge pointwise on the unit circle to the kernel representation $\tilde{G}_0$ of $\mathcal{B}_0$, except for a finite number of points. This implies that $\mathrm{d}(\Sigma_0, \mathcal{B}_0) \le \delta+\eta$, which contradicts the minimality of n for $(\Sigma, \delta+\eta, m)$. So n is minimal for $(\Sigma_k, \delta+\eta, m)$ and therefore $\mathbf{B}(\Sigma_k; \delta_k, m, n)$ is compact for k sufficiently large.

Now suppose that there exists an $\varepsilon > 0$ and a sequence of systems $\mathcal{B}_k \in \mathbf{B}(\Sigma_k; \delta_k, m, n)$ so that $\mathrm{d}(\mathcal{B}_k, \mathcal{B}) \ge \varepsilon$ for all $\mathcal{B} \in \mathbf{B}(\Sigma; \delta, m, n)$. As $\mathrm{d}(\Sigma_k, \mathcal{B}_k) \le \delta_k$ and $(\Sigma_k, \delta_k) \to (\Sigma, \delta)$, it follows from Proposition 6.3 in the appendix that for k sufficiently large, $\mathcal{B}_k \in \mathbf{B}(\Sigma; \delta+\eta, m, n)$. As n is minimal for $(\Sigma, \delta+\eta, m)$, this is according to Proposition 4.5 a compact set, so the sequence $\mathcal{B}_k$ contains a limit point, that we denote by $\mathcal{B}_0 \in \mathbf{B}(\Sigma; \delta+\eta, m, n)$. It follows from Proposition 6.3 that $\mathrm{d}(\Sigma, \mathcal{B}_0) \le \delta$ and thus $\mathcal{B}_0 \in \mathbf{B}(\Sigma; \delta, m, n)$. From the assumption that $\mathrm{d}(\mathcal{B}_k, \mathcal{B}) \ge \varepsilon$ for all $\mathcal{B} \in \mathbf{B}(\Sigma; \delta, m, n)$ this implies that $\mathrm{d}(\mathcal{B}_k, \mathcal{B}_0) \ge \varepsilon$, but this contradicts the fact that $\mathcal{B}_0$ is a limit point of the sequence $\mathcal{B}_k$.

(iv) Let $\delta_k \downarrow \delta$; then, according to Proposition 4.5, the sets $\mathbf{B}(\Sigma; \delta_k, m, n)$ are compact for k sufficiently large. It follows from the result in (iii) that there holds $\rho(\mathbf{B}(\Sigma; \delta_k, m, n), \mathbf{B}(\Sigma; \delta, m, n)) \to 0$, and as $\mathbf{B}(\Sigma; \delta, m, n) \subseteq \mathbf{B}(\Sigma; \delta_k, m, n)$ it is trivial that $\rho(\mathbf{B}(\Sigma; \delta, m, n), \mathbf{B}(\Sigma; \delta_k, m, n)) = 0$. This proves convergence in the Hausdorff metric. ∎

It is also of interest to consider the continuity of Pareto optimal models. Continuity in this respect is connected with robustness in the sense that small perturbations in the data should lead to a small perturbation of optimal models. We analyze this for models that optimize the fit under a complexity constraint. For given spectrum Σ we denote by $\mathbf{B}^*(\Sigma; m, n)$ the set of behaviors of optimally fitting factor models with m inputs and n states, and by $\overline{\mathbf{B}}^*(\Sigma; m, n)$ the set of optimally fitting behaviors with m inputs and at most n states.

PROPOSITION 4.7.

(i) *The set $\overline{\mathbf{B}}^*(\Sigma; m, n)$ is upper semicontinuous in the spectrum Σ.*

(ii) *Let δ^* be the optimal fit in $\mathbf{B}(m, n)$ and let n be minimal for $(\Sigma, \delta^*+\eta, m)$ for some $\eta > 0$; then $\rho(\mathbf{B}^*(\Sigma_k; m, n), \mathbf{B}^*(\Sigma; m, n)) \to 0$ for $\Sigma_k \to \Sigma$.*

Proof. (i) Let $\Sigma_k \to \Sigma$ and let $\mathcal{B}_k$ be an optimal behavior in $\overline{\mathbf{B}}(m, n)$ for Σ_k with $\mathcal{B}_k \to \mathcal{B}$ for $k \to \infty$; then we have to prove that $\mathcal{B}$ is optimal for Σ. As $\overline{\mathbf{B}}(m, n)$

is closed, it follows that $\mathcal{B} \in \overline{\mathbf{B}}(m,n)$, and if this limit system is not optimal, then there exists a system $\mathcal{B}_0 \in \overline{\mathbf{B}}(m,n)$ so that $\mathrm{d}(\Sigma,\mathcal{B}_0) < \mathrm{d}(\Sigma,\mathcal{B})$. It then follows from Proposition 6.3 in the appendix that for k sufficiently large, $\mathrm{d}(\Sigma_k,\mathcal{B}_0) < \mathrm{d}(\Sigma_k,\mathcal{B}_k)$, but this contradicts the optimality of $\mathcal{B}_k$.

(ii) If this is not true, then there exists an $\varepsilon > 0$ and a sequence of systems $\mathcal{B}_k \in \mathbf{B}^*(\Sigma_k;m,n)$ so that for all $\mathcal{B} \in \mathbf{B}^*(\Sigma;m,n)$ there holds $\mathrm{d}(\mathcal{B}_k,\mathcal{B}) \geq \varepsilon$. Now let $\mathcal{B} \in \mathbf{B}^*(\Sigma;m,n)$, so that $\mathrm{d}(\Sigma,\mathcal{B}) = \delta^*$ and $\mathrm{d}(\Sigma_k,\mathcal{B}) \leq \delta^* + \eta_k$ with $\eta_k \downarrow 0$ for $k \to \infty$. It then follows that $\mathrm{d}(\Sigma_k,\mathcal{B}_k) \leq \delta^* + \eta_k$ and hence $\mathrm{d}(\Sigma,\mathcal{B}_k) \leq \delta^* + \eta$ for k sufficiently large. Because n is minimal for $(\Sigma,\delta^*+\eta,m)$ it follows that $\mathbf{B}(\Sigma;\delta^*+\eta,m,n)$ is compact, so that the sequence $\mathcal{B}_k$ has a limit point, say $\mathcal{B}_0 \in \mathbf{B}(m,n)$. As $\mathrm{d}(\mathcal{B}_k,\mathcal{B}) \geq \varepsilon$ for all $\mathcal{B} \in \mathbf{B}^*(\Sigma;m,n)$, the same holds true for $\mathcal{B}_0$, but this contradicts the fact that $\mathrm{d}(\Sigma,\mathcal{B}_0) = \lim \mathrm{d}(\Sigma_k,\mathcal{B}_k) = \delta^*$, so that $\mathcal{B}_0 \in \mathbf{B}^*(\Sigma;m,n)$. ∎

4.3. Consistency. Next we investigate the consistency of dynamic factor models when the spectrum is estimated from observed data. In applications the spectrum of the observed process will in general be unknown. Suppose that, apart from assumptions A1–A4, the available information on the process consists of an observed time series of length T. Let Σ_T denote an estimator of the process spectrum Σ that is based on this time series. In order to simplify the analysis we assume that the estimator is strongly consistent, so that $\mathrm{d}(\Sigma,\Sigma_T) \to 0$ almost surely for $T \to \infty$. A strongly consistent estimator can be obtained, for example, as follows. Let the observed process have spectrum $\Sigma(z) = \sum_{k=-\infty}^{\infty} R(k) z^{-k}$, where $R(k) := \mathrm{E}\{w(t)w'(t-k)\}$ are the process covariances, and let $\hat{R}_T(k) = \frac{1}{T}\sum_{t=k+1}^{T} w(t)w'(t-k)$ be the sample covariances.

PROPOSITION 4.8. *Under weak conditions on the data generating process, a strongly consistent estimator of Σ is given by $\Sigma_T(z) = \sum_{|k|\leq k_T} \hat{R}_T(k) z^{-k}$, where $k_T = \log(T)$.*

Proof. The estimation error is bounded by

$$\|\Sigma(z) - \Sigma_T(z)\|_\infty \leq (2k_T+1) \sup_{|k|\leq k_T} \|R(k) - \hat{R}_T(k)\| + \sum_{|k|>k_T} \|R(k)\|.$$

The second term converges to zero by assumption A4, and the first term converges to zero almost surely under weak conditions. A sufficient condition is that the spectrum Σ is rational, but the result also holds true for a broad class of nonrational spectra. For these results we refer to [13, Theorems 5.3.2 and 7.4.3]. ∎

In the following, let $\mathbf{B}_0 := \mathbf{B}(\Sigma;\delta,m,n)$ be the class of feasible models and $\mathbf{B}_0^* \subset \mathbf{B}(m,n)$ be the set of optimal models of complexity (m,n), that is, with optimal fit in this class. By $\overline{\mathbf{B}}_0$ and $\overline{\mathbf{B}}_0^*$ we denote the sets of feasible and optimal models, respectively, with m inputs and at most n states. Further, let $\mathbf{B}_T := \mathbf{B}(\Sigma_T;\delta,m,n)$ be the set of feasible models and $\mathbf{B}_T^*$ the set of optimal models of complexity (m,n) for the estimated spectrum Σ_T, and let $\overline{\mathbf{B}}_T$ and $\overline{\mathbf{B}}_T^*$ be the sets of feasible and optimal models, respectively, with m inputs and at most n states. These are random sets, since they depend on the observed time series. The next two theorems state consistency properties for feasible and optimal models, where it is assumed that the estimator Σ_T is strongly consistent.

THEOREM 4.9.

(i) *Behaviors with better fit than the noise bound are estimated consistently; that is, if a factor model has behavior $\mathcal{B}$ of complexity (m,n) and fit δ, then for $\delta' > \delta$, it holds almost surely that $\mathcal{B} \in \mathbf{B}(\Sigma_T;\delta',m,n)$ for $T \to \infty$.*

(ii) *The sample estimator of the set of feasible behaviors in $\overline{\mathbf{B}}(m,n)$ is upper semi-consistent, in the sense that $\{\mathcal{B}_T \in \overline{\mathbf{B}}_T, \mathcal{B}_T \to \mathcal{B}_0\} \Rightarrow \{\mathcal{B}_0 \in \overline{\mathbf{B}}_0\}$ holds almost*

surely; that is, the set of data with this convergence property has probability one.

(iii) *If n is minimal for $(\Sigma, \delta+\eta, m)$ for some $\eta > 0$, then the set of feasible sample behaviors in $\mathbf{B}(m,n)$ converges to a subset of the feasible behaviors for the process, in the sense that $\rho(\mathbf{B}_T, \mathbf{B}_0) \to 0$ almost surely for $T \to \infty$.*

Proof. (i) This is evident as $\Sigma_T \to \Sigma$ almost surely and $\mathrm{d}(\Sigma, \mathcal{B})$ is continuous; see Proposition 6.3 in the appendix.

(ii) This follows from Proposition 4.6(i).

(iii) This follows from Proposition 4.6(iii). ∎

THEOREM 4.10.

(i) *The sample estimator of the set of optimal behaviors in $\overline{\mathbf{B}}(m,n)$ is upper semiconsistent in the sense that $\{\mathcal{B}_T \in \overline{\mathbf{B}}^*_T, \mathcal{B}_T \to \mathcal{B}_0\} \Rightarrow \{\mathcal{B}_0 \in \overline{\mathbf{B}}^*_0\}$ almost surely.*

(ii) *If the process spectrum has a unique optimal factor behavior $\mathcal{B}^*_0$ of complexity (m,n) and if the infimum of the fits of models in $\overline{\mathbf{B}}(m,n-1)$ is strictly larger than the fit of $\mathcal{B}^*_0$, then this behavior is estimated consistently in the sense that $\mathrm{d_H}(\mathbf{B}^*_T, \{\mathcal{B}^*_0\}) \to 0$ almost surely for $T \to \infty$.*

Proof. (i) This follows from Proposition 4.7(i).

(ii) As it is given that $\mathbf{B}^*_0 = \{\mathcal{B}^*_0\}$ is a singleton, it follows that $\rho(\{\mathcal{B}^*_0\}, \mathbf{B}^*_T) = \inf_{\mathcal{B}\in\mathbf{B}^*_T} \mathrm{d}(\mathcal{B}^*_0, \mathcal{B}) \le \sup_{\mathcal{B}\in\mathbf{B}^*_T} \mathrm{d}(\mathcal{B}^*_0, \mathcal{B}) = \rho(\mathbf{B}^*_T, \{\mathcal{B}^*_0\})$, so it suffices to prove that the last expression converges to zero. Let the optimal fit for Σ_T among models of complexity (m,n) be given by δ^*_T and let $\delta^*_0 = \mathrm{d}(\Sigma, \mathcal{B}^*_0)$; then it follows from $\mathrm{d}(\Sigma_T, \mathcal{B}^*_0) \to \delta^*_0$ that $\delta^*_T \to \delta^*_0$ almost surely. Further, because of the assumption that $\inf\{\mathrm{d}(\Sigma, \mathcal{B}); \mathcal{B} \in \overline{\mathbf{B}}(m, n-1)\} > \delta^*_0$, it follows that n is minimal for all $(\Sigma, \delta^*_0+\eta, m)$ with $\eta \ge 0$ sufficiently small, and the same then holds true almost surely for $(\Sigma_T, \delta^*_T+\eta, m)$ if $T \to \infty$. Then, for T sufficiently large, $\mathbf{B}^*_T$ is a closed subset of the compact set $\mathbf{B}(\Sigma_T; \delta^*_T+\eta, m, n)$, so that $\mathbf{B}^*_T$ is compact. This means that the Hausdorff distance is well defined. Further, as $(\Sigma_T, \delta^*_T) \to (\Sigma, \delta^*_0)$ almost surely, it follows from Proposition 4.6(iii) that

$$\rho(\mathbf{B}^*_T, \{\mathcal{B}^*_0\}) = \rho(\mathbf{B}(\Sigma_T; \delta^*_T, m, n), \mathbf{B}(\Sigma; \delta^*_0, m, n)) \to 0 \text{ almost surely.} \qquad ∎$$

This means that, under the above conditions, the feasible and optimal finite sample systems are in the limit also feasible and optimal for the data generating process. However, it is possible that not all feasible and optimal systems are identified in this way.

4.4. Low noise consistency. We conclude our analysis by considering another kind of consistency, inspired by the concept of low noise as defined in [15]. This is based on the idea that an identification method which aspires to deal with noisy data must, as a minimal requirement, function well when dealing with data having low noise content. Let the observed process be given by $w = \hat{w}_0 + \tilde{w}_0$, where the latent process $\hat{w}_0$ is fixed and has behavior $\mathcal{B}_0$ of complexity (m_0, n_0) and where the noise process $\tilde{w}_0$ has norm δ_0. Low noise consistency corresponds to the condition that the factor behavior $\mathcal{B}_0$ is identified uniquely if the noise vanishes in the limit. The following result shows that this holds true, provided that the factor scheme is specified correctly.

PROPOSITION 4.11.

(i) *If the factor scheme, noise bound, and complexity have been specified correctly, then the factor behavior is identified; that is, if $\delta \ge \delta_0$, $m = m_0$, and $n = n_0$, then $\mathcal{B}_0 \in \mathbf{B}(\Sigma; \delta, m, n)$. If orthogonality is imposed but the data generating process does not satisfy this property, then the system need not be identified.*

(ii) *Correctly specified factor schemes are low noise consistent; that is, if* $\delta_0 \leq \delta \downarrow 0$, *then the set of feasible behaviors* $\mathbf{B}(\Sigma; \delta, m, n) \to \{\mathcal{B}_0\}$ *(in the sense of the Hausdorff metric) for* $(m, n) = (m_0, n_0)$, *and* $\mathbf{B}(\Sigma; \delta, m, n) \to \emptyset$ *if* $m < m_0$ *or* $m = m_0, n < n_0$. *Consistency is generally lost if orthogonality is imposed but the data generating process does not satisfy this property.*

Proof. (i) This is evident from the definition of $\mathbf{B}(\Sigma; \delta, m, n)$.

(ii) The process decomposition $w = \hat{w}_0 + \tilde{w}_0$ induces a corresponding spectral decomposition $\Sigma = \hat{\Sigma}_0 + \tilde{\Sigma}_0 + \Sigma_c + \Sigma_c{}'$, where $\hat{\Sigma}_0$ is the spectrum of the latent process $\hat{w}_0$, $\tilde{\Sigma}_0$ is the spectrum of the noise $\tilde{w}_0$, and Σ_c is the cross spectrum between $\hat{w}_0$ and $\tilde{w}_0$. As the latent process $\hat{w}_0$ is fixed and the noise converges to zero, it follows that $\| \Sigma - \hat{\Sigma}_0 \|_\infty = \| \tilde{\Sigma}_0 + \Sigma_c + \Sigma_c{}' \|_\infty \to 0$.

First we consider the factor scheme without orthogonality constraint. Then the misfit function $\mathrm{d}(\hat{\Sigma}_0, \mathcal{B})$ is also well defined for the singular spectral density $\hat{\Sigma}_0$; i.e., if P is the projection onto $\mathcal{B}$ and $\tilde{\Sigma} = (I - P)\, \hat{\Sigma}_0 (I - P)$, then $\mathrm{d}(\hat{\Sigma}_0, \mathcal{B}) = \| \tilde{\Sigma}^{1/2} \|$ and $\mathbf{B}(\hat{\Sigma}_0; \delta, m, n) = \{\mathcal{B} \in \mathbf{B}(m, n);\ \mathrm{d}(\hat{\Sigma}_0, \mathcal{B}) \leq \delta\}$. It can easily be shown, along the lines of the proof of Proposition 6.3 in the appendix, that $\mathrm{d}(\Sigma, \mathcal{B}) \to \mathrm{d}(\hat{\Sigma}_0, \mathcal{B}_*)$ if $\Sigma \to \hat{\Sigma}_0$ and $\mathcal{B} \to \mathcal{B}_*$. In addition, it holds that

$$\left| \mathrm{d}^2(\Sigma, \mathcal{B}) - \mathrm{d}^2(\hat{\Sigma}_0, \mathcal{B}) \right| \leq c \| \Sigma - \hat{\Sigma}_0 \|_\infty,$$

where $c = 2$ for the uniform norm and $c = 2\pi q$ for the mean squares norm. The above result follows from the proof of Lemma 6.2 in the appendix and the inequality $\|(I - P)(\Sigma - \hat{\Sigma}_0)(I - P)\|_\infty \leq \| \Sigma - \hat{\Sigma}_0 \|_\infty$.

We now first show that for $m < m_0$ or $m = m_0$, $n < n_0$, the infimum of the misfits $\mathrm{d}(\hat{\Sigma}_0, \mathcal{B})$ over the set of behaviors $\mathbf{B}(m, n)$ is strictly larger than zero. If this were not true, then there would exist a sequence of behaviors $\mathcal{B}_k \in \mathbf{B}(m, n)$ with corresponding projections P_k, such that $\mathrm{d}(\hat{\Sigma}_0, \mathcal{B}_k) \to 0$. As in the proof of Proposition 4.5, it follows that there exists a subsequence $k(l)$ and a behavior $\mathcal{B}_* \in \mathbf{B}(m, n'), n' \leq n$, with a corresponding projection P_*, such that $P_{k(l)}(z) \to P_*(z)$ for $l \to \infty$, pointwise on the unit circle except for a finite number of points. Then $\mathrm{d}(\hat{\Sigma}_0, \mathcal{B}_k) \to 0$ implies that $\mathrm{d}(\hat{\Sigma}_0, \mathcal{B}_*) = 0$, and this means that $\mathcal{B}_0 \subseteq \mathcal{B}_*$. This contradicts the assumption that the complexity (m, n) is smaller than the complexity (m_0, n_0) of $\mathcal{B}_0$. We conclude that the infimum of misfits of models of complexity $m < m_0$ or $m = m_0$, $n < n_0$ is given by a strictly positive number δ_*. Since $\| \Sigma - \hat{\Sigma}_0 \|_\infty$ converges to zero for $\delta \downarrow 0$, there exists a $\delta_+ > 0$ such that $c \| \Sigma - \hat{\Sigma}_0 \|_\infty < \delta_*^2$ for $\delta \leq \delta_+$. By the above considerations and inequalities, it holds for $\delta \leq \delta_+$ that

$$\mathrm{d}^2(\Sigma, \mathcal{B}) \geq \mathrm{d}^2(\hat{\Sigma}_0, \mathcal{B}) - c \| \Sigma - \hat{\Sigma}_0 \|_\infty > \delta_*^2 - \delta_*^2 = 0.$$

This shows that $\mathbf{B}(\Sigma; \delta, m, n)$ is empty for $m < m_0$ and for $m = m_0$, $n < n_0$ if $\delta \leq \delta_+$.

Now suppose that the complexity has been specified correctly. In this case $\mathcal{B}_0 \in \mathbf{B}(\Sigma; \delta, m_0, n_0)$ so that $\rho(\{\mathcal{B}_0\}, \mathbf{B}(\Sigma; \delta, m_0, n_0)) = 0$. Further, from the foregoing, it follows that n_0 is minimal for $(\hat{\Sigma}_0; \delta_+, m_0)$, as $\mathbf{B}(\hat{\Sigma}_0; \delta_+, m_0, n) = \emptyset$ for $n < n_0$ and $\mathbf{B}(\hat{\Sigma}_0; \delta_+, m_0, n_0)$ is not empty, and also $\mathbf{B}(\hat{\Sigma}_0; 0, m_0, n_0) = \{\mathcal{B}_0\}$. It follows from Proposition 4.6(iii) that $\rho(\mathbf{B}(\Sigma; \delta, m_0, n_0), \{\mathcal{B}_0\}) \to 0$.

Next we consider the factor scheme with orthogonality. By imposing the orthogonality constraint the sets $\mathbf{B}(\Sigma; \delta, m, n)$ generally become smaller. Since $\mathcal{B}_0 \in \mathbf{B}(\Sigma; \delta, m_0, n_0)$ for $\delta_0 \leq \delta$, the above results imply that $\mathbf{B}(\Sigma; \delta, m, n) \to \emptyset$ if the complexity (m, n) is smaller than (m_0, n_0) and that $\mathbf{B}(\Sigma; \delta, m_0, n_0) \to \{\mathcal{B}_0\}$.

That consistency is lost if orthogonality is imposed but the data generating process is not orthogonal is evident from Theorem 3.1(i), because this shows that in this

case the misfit δ_0 can in general not be obtained in the class of orthogonal models in $\mathbf{B}(m,n)$. □

4.5. Illustration. We will illustrate the foregoing results for static factor models, as in this case more explicit characterizations can be obtained. We will not further discuss the dynamic system example of sections 2.4 and 3.3, as the consistency analysis for dynamic factor models will be the topic of another paper.

So assume that the observed process w is white noise, and let Σ denote the covariance matrix of w. As we have seen in section 3.3.1, we can without loss of fit restrict ourselves to static relations. The set of all static systems $\mathbf{B}(m,0)$ is isomorphic to the set of all m-dimensional linear subspaces of $\mathbf{R}^q$. Isometric kernel representations of static systems are isometric matrices $\tilde{G} \in \mathbf{R}^{q\times m}$.

It can easily be seen that $\mathcal{B} \in \mathbf{B}(\Sigma;\delta,m,0)$ if and only if the isometry $\tilde{G}$ satisfies the following inequalities: for the nonorthogonal factor scheme, $\operatorname{trace}(\tilde{G}'\,\Sigma\,\tilde{G}) \le \delta^2$ for the mean squares norm and $\tilde{G}'(\Sigma - \delta^2 I)\,\tilde{G} \le 0$ for the uniform norm, and for the orthogonal factor scheme, $\operatorname{trace}(\Sigma\,\tilde{G}(\tilde{G}'\,\Sigma\,\tilde{G})^{-1}\,\tilde{G}'\,\Sigma) \le \delta^2$ and $\tilde{G}'(\Sigma^2 - \delta^2\,\Sigma)\,\tilde{G} \le 0$, respectively. From this characterization it follows that the sets $\mathbf{B}(\Sigma;\delta,m,0)$ of static systems are always compact.

Let $\lambda_1 > \lambda_2 > \cdots > \lambda_q > 0$ denote the eigenvalues of Σ, and let $\Sigma = \hat{\Sigma}_m + \tilde{\Sigma}_m$ be the principal component decomposition of Σ with m factors as in (15). The set $\mathbf{B}(\Sigma;\delta,m,0)$ is nonempty if and only if $\|\,\tilde{\Sigma}_m\,\| \le \delta$, that is, $\lambda_{m+1}^{1/2} \le \delta$ for the uniform norm and $(\lambda_{m+1} + \cdots + \lambda_q)^{1/2} \le \delta$ for the mean squares norm. Furthermore, one can show that the sets $\mathbf{B}(\Sigma;\delta,m,0)$ depend continuously on (Σ,δ) with the exception of points where $\|\,\tilde{\Sigma}_m\,\| = \delta$.

Let Σ_T denote a strongly consistent estimator of Σ. If $\|\,\tilde{\Sigma}_m\,\| < \delta$ then $\mathbf{B}(\Sigma_T;\delta,m,0)$ is a strongly consistent estimator of $\mathbf{B}(\Sigma;\delta,m,0)$. The principal component model of Σ_T is a strongly consistent estimator of the principal component model of Σ, so that the Pareto optimal models are estimated consistently.

5. Conclusion. Dynamic factor models decompose an observed process in terms of an underlying latent component and additional noise. The variables are treated in a completely symmetric way, and no assumptions on inputs and outputs are required. The latent process has a singular spectrum as it satisfies deterministic dynamic relationships. This means that the factor behavior consists of a linear dynamical system. In particular, the latent process has fewer free variables than the observed process. Depending on the chosen factor scheme, several interpretations of the noise process are possible. If the noise can be assumed to be uncorrelated with the latent process, this is called the orthogonal factor scheme. This is the usual assumption in the classical models of factor analysis. In other situations it is more natural to consider the latent process as an approximation of the observed process and to assume that the factor components are constructed from the observations. This is called the observable factor scheme.

Within this framework we investigated the representation of dynamic factor models and defined notions of complexity and goodness of fit. Concerning the identification of factor models we presented characterizations of Pareto optimal models and we derived results on consistency, both in the case of observed data and in the case of low noise.

An advantage of our approach is that it deals explicitly with the symmetric modeling of observed data by means of dynamic stochastic models. Other contributions in symmetric system modeling have been developed in the behavioral identification

of systems and in the structural analysis of factor models. In a sense, our approach can be seen as an extension of these two frameworks. It enriches the deterministic behavioral framework with a stochastic analysis, and it extends the traditionally structure-oriented analysis of factor models to a more empirical modeling setting.

Several questions deserve further investigation. Of special interest is the analysis of identification procedures within this framework. Another issue is the incorporation of prior knowledge, for example, concerning the input-output structure of the model. A further analysis of the probabilistic structure of factor models is needed in order to develop statistical test procedures, for example, to estimate the complexity of factor models from observed data.

6. Appendix.

LEMMA 6.1. *Let $A, B \in \mathbf{C}^{q\times q}$ be two positive semidefinite matrices, and let $\lambda_1(A) \geq \cdots \geq \lambda_q(A) \geq 0$ and $\lambda_1(B) \geq \cdots \geq \lambda_q(B) \geq 0$ be the eigenvalues of A and B, respectively. Then*

(i) $|\lambda_i(A) - \lambda_i(B)| \leq \|A - B\|_\infty$.

(ii) *For every unitary matrix $U \in \mathbf{C}^{q\times m}$, $U^*U = I$, there holds*

$$\begin{aligned} \operatorname{trace}(UU^*AUU^*) &= \operatorname{trace}(UAU^*) \geq \lambda_{m+1}(A) + \cdots + \lambda_q(A), \\ \lambda_{\max}(UU^*AUU^*) &= \lambda_{\max}(UAU^*) \geq \lambda_{m+1}(A). \end{aligned}$$

The lower bound is reached if the columns of U form a basis for the eigenspace of A corresponding to the $q - m$ smallest eigenvalues.

Proof. See [12, Corollary 8.1.3 and Theorem 8.1.2]. □

LEMMA 6.2. *Let Σ_k be a sequence of spectral densities that converges to Σ_0 in the sense that $\| \Sigma_k - \Sigma_0 \|_\infty \to 0$. Then*

(i) $\| \Sigma_k^{1/2} \| \to \| \Sigma_0^{1/2} \|$.

(ii) *if Σ_0 is positive definite, then Σ_k is positive definite for all k sufficiently large and $\| \Sigma_k^{-1} - \Sigma_0^{-1} \|_\infty \to 0$.*

Proof. (i) By Lemma 6.1 $|\lambda_i(\Sigma_k(z)) - \lambda_i(\Sigma_0(z))| \leq \| \Sigma_k - \Sigma_0 \|_\infty$ pointwise on the unit circle, so that

$$\begin{aligned} |\| \Sigma_k^{1/2} \|_2^2 - \| \Sigma_0^{1/2} \|_2^2| &= |\oint_{|z|=1} \operatorname{trace}(\Sigma_k(z) - \Sigma_0(z))dz| \\ &\leq 2\pi q \| \Sigma_k - \Sigma_0 \|_\infty, \\ |\| \Sigma_k^{1/2} \|_\infty^2 - \| \Sigma_0^{1/2} \|_\infty^2| &= |\sup_{|z|=1} \lambda_{\max}(\Sigma_k(z)) - \sup_{|z|=1} \lambda_{\max}(\Sigma_0(z))| \\ &\leq 2\| \Sigma_k - \Sigma_0 \|_\infty. \end{aligned}$$

(ii) By the assumption $\Sigma_0 > 0$ and the result in Lemma 6.1 for the eigenvalues of Σ_k, it follows that $\| \Sigma_0^{-1} \|_\infty = 1/\{\inf_{|z|=1} \lambda_{\min}(\Sigma_0(z))\}$ and $\| \Sigma_k^{-1} \|_\infty$ are bounded. The result then follows from

$$\| \Sigma_k^{-1} - \Sigma_0^{-1} \|_\infty = \| \Sigma_k^{-1}(\Sigma_0 - \Sigma_k) \Sigma_0^{-1} \|_\infty \leq \| \Sigma_k^{-1} \|_\infty \|(\Sigma_0 - \Sigma_k)\|_\infty \| \Sigma_0^{-1} \|_\infty. \qquad \square$$

PROPOSITION 6.3. *The misfit function $\mathrm{d}(\Sigma, \mathcal{B})$ is continuous in $(\Sigma, \mathcal{B})$ for all positive definite spectral densities Σ.*

Proof. Let $\Sigma_k \to \Sigma_0 > 0$ and $\mathcal{B}_k \to \mathcal{B}_0$ be convergent sequences of spectral densities and behaviors, respectively. The corresponding isometric kernel representations of $\mathcal{B}_k$, $\mathcal{B}_0$ are denoted by $\tilde{G}_k$ and $\tilde{G}_0$, respectively. The optimal noise spectra, given in Theorem 3.1, corresponding to the spectral densities Σ_k, Σ_0 and the behaviors $\mathcal{B}_k$, $\mathcal{B}_0$ are denoted by $\tilde{\Sigma}_k$ and $\tilde{\Sigma}_0$, respectively. By Lemma 6.2 it suffices to show that $\| \tilde{\Sigma}_k - \tilde{\Sigma}_0 \|_\infty \to 0$.

For the case without orthogonality the noise spectra are given by

$$\tilde{\Sigma}_k = \tilde{G}_k \, \tilde{G}_k^* \, \Sigma_k \, \tilde{G}_k \, \tilde{G}_k^* \quad \text{and} \quad \tilde{\Sigma}_0 = \tilde{G}_0 \, \tilde{G}_0^* \, \Sigma_0 \, \tilde{G}_0 \, \tilde{G}_0,$$

in which case $\| \tilde{\Sigma}_k - \tilde{\Sigma}_0 \|_\infty \to 0$ is evident.

For the case with orthogonality, let $\bar{G}_k = \tilde{G}_k \, \tilde{G}_k^* \, \tilde{G}_0$; then $\|\bar{G}_k - \tilde{G}_0 \|_\infty \leq \| \tilde{G}_k \, \tilde{G}_k^* - \tilde{G}_0 \, \tilde{G}_0^* \|_\infty \| \tilde{G}_0 \|_\infty \to 0$. The noise spectra for this factor scheme are given by

$$\tilde{\Sigma}_0 = \Sigma_0 \, \tilde{G}_0 (\tilde{G}_0^* \, \Sigma_0 \, \tilde{G}_0)^{-1} \, \tilde{G}_0^* \, \Sigma_0,$$

$$\tilde{\Sigma}_k = \Sigma_k \, \tilde{G}_k (\tilde{G}_k^* \, \Sigma_k \, \tilde{G}_k)^{-1} \, \tilde{G}_k^* \, \Sigma_k = \Sigma_k \, \bar{G}_k (\bar{G}_k^* \, \Sigma_k \, \bar{G}_k)^{-1} \bar{G}_k^* \, \Sigma_k,$$

where the last equality follows from the fact that $\tilde{G}_k^* \, \tilde{G}_0 \to I$, so that this is invertible for k sufficiently large. The result now follows from Lemma 6.2. ∎

REFERENCES

[1] T.W. Anderson and H. Rubin, *Statistical inference in factor analysis*, in Proceedings Third Berkeley Symposium on Mathematical Statistics and Probability, J. Neyman, ed., University of California Press, Berkeley, Los Angeles, 1956, pp. 111–150.

[2] S. Beghelli, R.P. Guidorzi, and U. Soverini, *The Frisch scheme in dynamic system identification*, Automatica, 26 (1990), pp. 171–176.

[3] D.R. Brillinger, *Time Series, Data Analysis and Theory*, Holden-Day, San Francisco, CA, 1981.

[4] P.E. Caines, *Linear Stochastic Systems*, Wiley, New York, 1988.

[5] M. Deistler, *Symmetric modeling in system identification*, in Three Decades of Mathematical System Theory, H. Nijmeijer and J.M. Schumacher, eds., Springer-Verlag, Berlin, 1989, pp. 128–147.

[6] M. Deistler and W. Scherrer, *Identification of linear systems from noisy data*, in New Directions in Time Series Analysis, Part II, D. Brillinger et al., eds., IMA Vol. Math. Appl., 46, Springer-Verlag, 1992, pp. 21–42.

[7] M. Deistler and W. Scherrer, *System identification and errors in the variables*, in Statistical modeling and Latent Variables, K. Haagen, D. J. Bartholomew, and M. Deistler, eds., North-Holland, Amsterdam, 1993, pp. 95–111.

[8] R.F. Engle and M. Watson, *A one-factor multivariate time series model of metropolitan wage rates*, J. Amer. Statist. Assoc., 76 (1981), pp. 774–781.

[9] R. Frisch, *Statistical Confluence Analysis by Means of Complete Regression Systems*, Publ. 5, Economic Institute, University of Oslo, Norway, 1934.

[10] W.A. Fuller, Measurement Error Models, Wiley, New York, 1987.

[11] J.F. Geweke, *The dynamic factor analysis of economic time series models*, in Latent Variables in Socio-economic Models, D.J. Aigner and A.S. Goldberger, eds., North-Holland, Amsterdam, 1977, pp. 365–383.

[12] G.H. Golub and C.F. Van Loan, *Matrix Computations*, The Johns Hopkins University Press, Baltimore, MD, 1989.

[13] E.J. Hannan and M. Deistler, *The Statistical Theory of Linear Systems*, Wiley, New York, 1988.

[14] C. Heij, *Deterministic Identification of Dynamical Systems*, Springer-Verlag, New York, 1989.

[15] R.E. Kalman, *A theory for the identification of linear relations*, in Frontiers in Pure and Applied Mathematics, a Collection of Papers Dedicated to Jacques-Louis Lions on the Occasion of his Sixtieth Birthday, R. Dautray, ed., North-Holland, Amsterdam, 1991, pp. 117–132.

[16] E.L. Leamer, *Errors in variables in linear systems*, Econometrica, 55 (1987), pp. 893-909.

[17] L. Ljung, *System Identification: Theory for the User*, Prentice-Hall, Englewood Cliffs, NJ, 1987.

[18] G. Picci and S. Pinzoni, *Dynamic factor analysis models for stationary processes*, IMA J. Math. Control Inform., 3 (1986), pp. 185–210.

[19] M.B. Priestley, *Spectral Analysis and Time Series*, Academic Press, New York, 1981.

[20] B. Roorda, *Algorithms for global total least squares modeling of finite multivariable time series*, Automatica, 31 (1995), pp. 391–404.

[21] B. Roorda and C. Heij, *Global total least squares modeling of multivariable time series*, IEEE Trans. Automat. Control, 40 (1995), pp. 50–63.
[22] Y.A. Rozanov, *Stationary Random Processes*, Holden-Day, San Francisco, CA, 1967.
[23] J.H. van Schuppen, *Stochastic realization problems*, in Three Decades of Mathematical System Theory, H. Nijmeijer and J.M. Schumacher, eds., Springer-Verlag, Berlin, 1989, pp. 480–523.
[24] J.C. Willems, *From time series to linear system, part* I, Automatica, 22 (1986), pp. 561–580.
[25] J.C. Willems, *From time series to linear system, part* III, Automatica, 23 (1987), pp. 87–115.
[26] J.C. Willems, *Paradigms and puzzles in the theory of dynamical systems*, IEEE Trans. Automat. Control, 36 (1991), pp. 259–294.

SIAM J. CONTROL OPTIM.
Vol. 35, No. 6, pp. 1952–1988, November 1997

006

ERGODIC CONTROL OF SWITCHING DIFFUSIONS*

MRINAL K. GHOSH†, ARISTOTLE ARAPOSTATHIS‡, AND STEVEN I. MARCUS§

Abstract. We study the ergodic control problem of switching diffusions representing a typical hybrid system that arises in numerous applications such as fault-tolerant control systems, flexible manufacturing systems, etc. Under fairly general conditions, we establish the existence of a stable, nonrandomized Markov policy which almost surely minimizes the pathwise long-run average cost. We then study the corresponding Hamilton–Jacobi–Bellman (HJB) equation and establish the existence of a unique solution in a certain class. Using this, we characterize the optimal policy as a minimizing selector of the Hamiltonian associated with the HJB equations. As an example we apply the results to a failure-prone manufacturing system and obtain closed form solutions for the optimal policy.

Key words. switching diffusions, Markov policy, ergodicity, pathwise average cost, Hamilton–Jacobi–Bellman equations

AMS subject classifications. 93E20, 60J60

PII. S0363012996299302

1. Introduction.

We address the problem of controlling switching diffusions by continually monitoring the continuous and discrete component of the state. The objective is to minimize, almost surely (a.s.), the pathwise long-run average (ergodic) cost over all admissible policies. A controlled switching diffusion is a typical example of a hybrid system which arises in numerous applications of systems with multiple modes or failure modes, such as fault-tolerant control systems, multiple target tracking, flexible manufacturing systems, etc. [13], [14], [23]. The state of the system at time t is given by a pair $\big(X(t), S(t)\big) \in \mathbb{R}^d \times \mathcal{S}$, $\mathcal{S} = \{1, 2, \dots, N\}$. The continuous component $X(t)$ is governed by a "controlled diffusion process" with a drift vector which depends on the discrete component $S(t)$. Thus, $X(t)$ switches from one diffusion path to another as the discrete component $S(t)$ jumps from one state to another. On the other hand, the discrete component $S(t)$ is a "controlled Markov chain" with a transition matrix depending on the continuous component. The evolution of the process $\big(X(t), S(t)\big)$ is governed by the following equations:

$$dX(t) = b\big(X(t), S(t), u(t)\big)dt + \sigma\big(X(t), S(t)\big)dW(t), \tag{1.1}$$

$$P\big(S(t+\delta t) = j \mid S(t) = i, X(s), S(s), s \le t\big) = \lambda_{ij}\big(X(t), u(t)\big)\delta t + o(\delta t), \quad i \ne j, \tag{1.2}$$

for $t \ge 0$, $X(0) = X_0$, $S(0) = S_0$, where b, σ, λ are suitable functions, $\lambda_{ij} \ge 0$ for $i \ne j$, $\sum_{j=1}^{N} \lambda_{ij} = 0$, $W(\cdot)$ is a standard Brownian motion, and $u(\cdot)$ is a nonanticipative

*Received by the editors February 26, 1996; accepted for publication (in revised form) August 15, 1996. This research was supported in part by Texas Advanced Research Program (Advanced Technology Program) grant 003658-186, Air Force Office of Scientific Research grants F49620-92-J-0045 and F49620-92-J-0083, and National Science Foundation grants EEC 9402384, NCR-9211343, and NCR-9502582.

http://www.siam.org/journals/sicon/35-6/29930.html

†Department of Mathematics, Indian Institute of Science, Bangalore 560012, India (mkg@math.iisc.ernet.in).

‡Department of Electrical and Computer Engineering, The University of Texas at Austin, Austin, TX 78712 (ari@mail.utexas.edu).

§Electrical Engineering Department and Institute for Systems Research, University of Maryland, College Park, MD 20742 (marcus@src.und.edu).

control process (admissible policy). The latter is called a Markov policy if $u(t) = v\big(X(t), S(t)\big)$ for a suitable function v. Our goal is to minimize a.s. over all admissible policies the functional

$$\limsup_{T\to\infty} \frac{1}{T}\int_0^T c\big(X(t), S(t), u(t)\big)dt\,, \tag{1.3}$$

where c is the running-cost function. Note that in (1.3) there is no expectation; we are minimizing the limiting pathwise average cost. Such a criterion is very important in practical applications since we often deal with a single realization. Under certain conditions, we show that there exists a Markov policy v^* and constant ρ^* such that

$$\lim_{T\to\infty} \frac{1}{T}\int_0^T c\big(X(t), S(t), v^*(X(t), S(t))\big)dt = \rho^* \quad \text{a.s.},$$

and for any other admissible policy $v(\cdot)$

$$\liminf_{T\to\infty} \frac{1}{T}\int_0^T c\big(X(t), S(t), v(t)\big)dt \geq \rho^* \quad \text{a.s.}$$

This establishes that v^* is optimal in a much stronger sense; viz., the most "pessimistic" average cost under v^* is no worse than the most "optimistic" average cost under any other admissible policy. Also, under the conditions assumed in this paper, the optimal pathwise average cost coincides with the optimal expected average cost. So we do not distinguish between these two criteria.

Our paper is organized as follows. In section 2 we present and analyze a motivating example, while in section 3 we introduce a concise mathematical model of the switching diffusion. Section 4 is devoted to the study of recurrence and ergodicity of switching diffusions. The existence of an optimal policy is established in section 5. The HJB equations are studied in section 6. Conclusions are in section 7.

2. A motivating example. The failure-prone manufacturing system presented in [1], [5], [14] is a very good example of the class of systems studied in this paper. This section is devoted to the analysis of this manufacturing model. Results from subsequent sections will be used in this example and thus the reader will have the opportunity to glimpse some of the key developments of the paper.

Suppose that there is one machine producing a single commodity. We assume that the demand rate is a constant $d > 0$. Let the machine state $S(t)$ take values in $\{0, 1\}$, $S(t) = 0$ or 1, according as the machine is down or functional. We model $S(t)$ as a continuous time Markov chain with generator

$$\begin{bmatrix} -\lambda_0 & \lambda_0 \\ \lambda_1 & -\lambda_1 \end{bmatrix},$$

where λ_0 and λ_1 are positive constants corresponding to the infinitesimal rates of repair and failure, respectively. The inventory $X(t)$ is governed by the Ito equation

$$dX(t) = \big(u(t) - d\big)dt + \sigma dW(t)\,, \tag{2.1}$$

where $\sigma > 0$, $u(t)$ is the production rate, and $W(t)$ is a one-dimensional Wiener process independent of $S(t)$. The last term in (2.1) can be interpreted as "sales

return," "inventory spoilage," "sudden demand fluctuations," etc. A negative value of $X(t)$ represents backlogged demand. The production rate is constrained by

$$u(t) \in \begin{cases} \{0\} & \text{if } S(t) = 0, \\ [0, r] & \text{if } S(t) = 1. \end{cases}$$

Let $c : \mathbb{R} \to \mathbb{R}_+$ be the cost function which is assumed to be convex and Lipschitz. Also, $c(x) \geq g(|x|)$ for some increasing function $g : \mathbb{R}_+ \to \mathbb{R}_+$. Thus, c satisfies (5.3), a required condition for the validity of our results. We show later in this section that a certain hedging-point policy is stable. Therefore, by the results of section 5, there exists an a.s. optimal nonrandomized Markov policy with respect to the cost criterion

$$\limsup_{T\to\infty} \frac{1}{T} \int_0^T c(X(t))dt\,.$$

The HJB equations in this case are

$$\begin{aligned} (2.2) \quad & \begin{pmatrix} \frac{\sigma^2}{2}V''(x,0) - dV'(x,0) \\ \frac{\sigma^2}{2}V''(x,1) + \min\limits_{u\in[0,r]}\{(u-d)V'(x,1)\} \end{pmatrix} + \begin{bmatrix} -\lambda_0 & \lambda_0 \\ \lambda_1 & -\lambda_1 \end{bmatrix} \begin{pmatrix} V(x,0) \\ V(x,1) \end{pmatrix} \\ & \qquad + \begin{pmatrix} 1 \\ 1 \end{pmatrix} c(x) = \begin{pmatrix} 1 \\ 1 \end{pmatrix} \rho\,. \end{aligned}$$

The results of section 6 ensure existence of a C^2 solution (V, ρ^*) of (2.2), where ρ^* is the optimal cost. Using the convexity of $c(\cdot)$, it can be shown that $V(\cdot, i)$ is convex for each i. Hence, there exists an x^* such that

$$\begin{aligned} & V'(x,1) \leq 0 \qquad \text{for } x \leq x^*, \\ (2.3) \quad & V'(x,1) \geq 0 \qquad \text{for } x \geq x^*. \end{aligned}$$

It follows from (2.3) that the value of u which minimizes $(u-d)V'(x,1)$ is

$$u = \begin{cases} r & \text{if } x < x^*, \\ 0 & \text{if } x > x^*. \end{cases}$$

Since $V'(x^*,1) = 0$, any $u \in [0,r]$ minimizes $(u-d)V'(x^*,1)$. Therefore, in view of Theorem 6.2, the action $u \in [0,r]$ can be chosen arbitrarily at $x = x^*$. To be specific, we let $u(x^*) = d$, i.e., we produce at the level that meets the demand exactly. Thus, the following stable, nonrandomized Markov policy is optimal:

$$(2.4) \qquad v^*(x,0) = 0, \qquad v^*(x,1) = \begin{cases} r & \text{if } x < x^*, \\ d & \text{if } x = x^*, \\ 0 & \text{if } x > x^*. \end{cases}$$

Note that the stability of the policy (2.4) follows from Theorem 6.3 provided that the set of stable, nonrandomized Markov policies is nonempty. We show next that the

zero-inventory policy v given by

$$v(x,0)=0, \qquad v(x,1)=\begin{cases} r & \text{if } x\le 0,\\ 0 & \text{if } x>0 \end{cases} \tag{2.5}$$

is stable if and only if

$$\frac{(r-d)}{\lambda_1} > \frac{d}{\lambda_0}\,. \tag{2.6}$$

The condition (2.6) is in accord with intuition. Note that λ_0^{-1} and λ_1^{-1} are the mean sojourn times of the chain in states 0 and 1, respectively. In state 0 the mean inventory depletes at a rate d while in state 1 it builds up at a rate $(r-d)$. Thus, if (2.6) is satisfied, one would expect the zero-inventory policy to stabilize the system. Our analysis confirms this intuition. We first show that under v the process $\big(X(\cdot),S(\cdot)\big)$ has an invariant probability measure η_v with a strictly positive density. In view of Lemma 4.1, it then follows from the ergodic theory of Markov processes [25, Chap. 1] that $\big(X(\cdot),S(\cdot)\big)$ is positive recurrent, or equivalently that v is stable.

By Lemma 5.2, the density φ of the invariant probability measure η_v can be obtained by solving the adjoint system

$$(L^v)^*\varphi(x,i)=0\,, \tag{2.7}$$

subject to

$$\varphi(x,i)>0, \qquad \sum_{i\in\{0,1\}}\int_{\mathbb{R}}\varphi(x,i)\,dx=1\,, \tag{2.8}$$

where L^v is the differential generator defined in (3.6)–(3.8). Define

$$\tilde\lambda_0:=\frac{2\lambda_0}{\sigma^2}, \quad \tilde\lambda_1:=\frac{2\lambda_1}{\sigma^2}, \quad \tilde d:=\frac{2d}{\sigma^2}, \quad \text{and} \quad \tilde r:=\frac{2r}{\sigma^2}\,.$$

Then (2.7) is equivalent to

$$\begin{aligned} &\varphi''(x,0)+\tilde d\varphi'(x,0)-\tilde\lambda_0\varphi(x,0)+\tilde\lambda_1\varphi(x,1)=0\\ &\varphi''(x,1)+\tilde d\varphi'(x,1)-\tilde\lambda_1\varphi(x,1)+\tilde\lambda_0\varphi(x,0)=0 \end{aligned} \qquad \text{for } \ x>0, \tag{2.9a}$$

$$\begin{aligned} &\varphi''(x,0)+\tilde d\varphi'(x,0)-\tilde\lambda_0\varphi(x,0)+\tilde\lambda_1\varphi(x,1)=0\\ &\varphi''(x,1)-(\tilde r-\tilde d)\varphi'(x,1)-\tilde\lambda_1\varphi(x,1)+\tilde\lambda_0\varphi(x,0)=0 \end{aligned} \qquad \text{for } \ x<0. \tag{2.9b}$$

A solution of (2.9), subject to the constraint (2.8), exists if and only if (2.6) holds and takes the form

$$\varphi(x)=\begin{pmatrix}\varphi(x,0)\\ \varphi(x,1)\end{pmatrix}=\begin{cases} a_1\begin{pmatrix}\tilde\lambda_1\\ \tilde\lambda_0\end{pmatrix}e^{-s_1x}+a_2\begin{pmatrix}-\tilde\lambda_1\\ \tilde\lambda_1\end{pmatrix}e^{-s_2x} & \text{for } x\ge 0,\\ a_3\begin{pmatrix}\tilde\lambda_1\\ -\psi(s_3)\end{pmatrix}e^{s_3x}+a_4\begin{pmatrix}-\tilde\lambda_1\\ \psi(s_4)\end{pmatrix}e^{s_4x} & \text{for } x<0, \end{cases} \tag{2.10}$$

where $\psi(s) = s^2 + \tilde{d}s - \tilde{\lambda}_0$, $s_1 = \tilde{d}$, $s_2 = \frac{\tilde{d}}{2} + \frac{1}{2}\big[\tilde{d}^2 + 4(\tilde{\lambda}_0 + \tilde{\lambda}_1)\big]^{1/2}$, and s_3, s_4 are the positive roots of the polynomial

$$s^3 - (\tilde{r} - 2\tilde{d})s^2 - \big[(\tilde{r} - \tilde{d})\tilde{d} + \tilde{\lambda}_0 + \tilde{\lambda}_1\big]s + \big[(\tilde{r} - \tilde{d})\tilde{\lambda}_0 - \tilde{d}\tilde{\lambda}_1)\big]\,,$$

ordered by $0 < s_3 < s_4$. Also, the coefficients $\{a_1, a_2, a_3, a_4\}$ are given by

$$\begin{aligned}
a_1 &= \frac{1}{\Delta}\Big\{\frac{(s_4 - s_3)s_2}{\tilde{\lambda}_0 + \tilde{\lambda}_1} + \frac{s_4 + s_2}{s_3 + \tilde{d}} - \frac{s_3 + s_2}{s_4 + \tilde{d}}\Big\},\\
a_2 &= \frac{1}{\Delta}\frac{(s_4 - s_3)s_2}{\tilde{\lambda}_0 + \tilde{\lambda}_1},\\
a_3 &= \frac{1}{\Delta}\frac{s_4 + s_2}{s_3 + \tilde{d}},\\
a_4 &= \frac{1}{\Delta}\frac{s_3 + s_2}{s_4 + \tilde{d}},\\
\Delta &= \frac{(s_4 - s_3)(s_2 - \tilde{d})}{\tilde{d}} + \frac{\tilde{\lambda}_0 + \tilde{\lambda}_1}{\tilde{d}}\Big\{\frac{s_4 + s_2}{s_3} - \frac{s_3 + s_2}{s_4}\Big\}.
\end{aligned} \tag{2.11}$$

Note that if $\varphi_{x^*}(\cdot)$ denotes the density of the invariant measure corresponding to a hedging-point policy as in (2.4), then

$$\varphi_{x^*}(x) = \varphi(x - x^*)\,.$$

Given a convex cost function, the average cost $\rho(x^*)$ corresponding to such a policy can be readily computed and is a convex function of the threshold value x^*.

In [5], Bielecki and Kumar have studied the mean square stability of the piecewise deterministic system, i.e., (2.1) with $\sigma = 0$. They have shown that under (2.6) the policy (2.5) is mean square stable, and have computed the optimal threshold value x^* in (2.4). These results can be easily reproduced here by computing the limiting value of the invariant distribution as $\sigma \to 0$, which we do next. The roots s_2, s_3, and s_4 have the following asymptotic dependence on σ:

$$s_2 = \frac{2d}{\sigma^2} + \mathcal{O}(1), \quad s_3 = \frac{(r-d)\lambda_0 - d\lambda_1}{d(r-d)} + \mathcal{O}(\sigma^2), \quad s_4 = \frac{2(r-d)}{\sigma^2} + \mathcal{O}(1)\,. \tag{2.12}$$

Thus, using (2.11), we obtain

$$\begin{aligned}
a_1, a_2 &= \frac{d\big[(r-d)\lambda_0 - d\lambda_1\big]}{r(\lambda_0 + \lambda_1)^2} + \mathcal{O}(\sigma^2),\\
a_3 &= \frac{\sigma^2}{2}\frac{\big[(r-d)\lambda_0 - d\lambda_1\big]}{d(r-d)(\lambda_0 + \lambda_1)} + \mathcal{O}(\sigma^4),\\
a_4 &= \frac{\sigma^2}{2}\frac{d\big[(r-d)\lambda_0 - d\lambda_1\big]}{r^2(r-d)(\lambda_0 + \lambda_1)} + \mathcal{O}(\sigma^4)\,.
\end{aligned} \tag{2.13}$$

Let

$$\alpha_0 := \frac{(r-d)\lambda_0 - d\lambda_1}{d(r-d)}$$

and $\delta_z(x)$ denote the Dirac measure centered at z. Using (2.12) and (2.13), we can show that as $\sigma \to 0$, $\varphi_{x^*}(\cdot)$ converges weakly to a distribution with "density" $\overline{\varphi}_{x^*}(\cdot)$, given by

$$\overline{\varphi}_{x^*}(x,i) = \begin{cases} \frac{\lambda_1\alpha_0}{\lambda_0+\lambda_1} e^{\alpha_0(x-x^*)} & \text{for } x \le x^*,\ i=0, \\ \frac{d\alpha_0}{\lambda_0+\lambda_1}\delta_{x^*}(x) + \frac{d\lambda_1\alpha_0}{(r-d)(\lambda_0+\lambda_1)} e^{\alpha_0(x-x^*)} & \text{for } x \le x^*,\ i=1, \\ 0 & \text{for } x > x^*. \end{cases}$$

Using, as in [5], a cost of the form

$$c(x) = \frac{c^+ + c^-}{2}|x| + \frac{c^+ - c^-}{2}x\,, \tag{2.14}$$

with c^+ and c^- positive constants, the average cost corresponding to the policy in (2.4) takes the form

$$\begin{aligned} \rho(x^*) &= \sum_{i=0,1} \int_{-\infty}^{x^*} c(x)\overline{\varphi}_{x^*}(x,i)\,dx \\ &= c^+x^* - \frac{c^+r\lambda_1}{(r-d)(\lambda_0+\lambda_1)\alpha_0} + \frac{r\lambda_1(c^+ + c^-)}{(r-d)(\lambda_0+\lambda_1)\alpha_0} e^{-\alpha_0 x^*}. \end{aligned}$$

In this manner, the results in [5] are reproduced exactly. One advantage of our approach is that the class of admissible policies does not have to be restricted as is done in [5], in order to guarantee the existence of solutions. With our method, optimality is obtained with respect to the class of all nonanticipative policies. Furthermore, our analysis shows that the stability of the zero-inventory policy is retained under additive noise in (2.1). Let us also note that conditions for the optimality of the zero-inventory policy under additive noise can be readily obtained for the cost in (2.14) using the density in (2.10).

3. The mathematical model. We first show that the switching diffusion (1.1), (1.2) can be constructed on a given probability space. Our presentation follows [13], [14]; we repeat it here for the sake of clarity and completeness. Let U be a compact metric space, $\mathcal{S} := \{1, 2, \dots, N\}$, and

$$\begin{aligned} &\overline{b} = \big[\overline{b}_1, \dots, \overline{b}_d\big]' : \mathbb{R}^d \times \mathcal{S} \times U \to \mathbb{R}^d, \\ &\sigma = \big[\sigma_{ij}(\cdot,\cdot)\big] : \mathbb{R}^d \times \mathcal{S} \to \mathbb{R}^{d\times d}, \\ &\overline{\lambda}_{ij} : \mathbb{R}^d \times U \to \mathbb{R}, \qquad i,j \in \mathcal{S}\,, \\ &\overline{\lambda}_{ij} \ge 0 \text{ for } i \ne j, \qquad \sum_{j\in\mathcal{S}} \overline{\lambda}_{ij} = 0 \text{ for any } i \in \mathcal{S}. \end{aligned}$$

We also define the matrix $\widetilde{\Lambda} : \mathbb{R}^d \times U \to \mathbb{R}^{N\times N}$ by

$$\big[\widetilde{\Lambda}(x,u)\big]_{ij} = \begin{cases} \overline{\lambda}_{ij}(x,u)\,, & i \ne j, \\ 0\,, & i = j\,. \end{cases}$$

We make the following assumptions which will be in effect throughout the paper.

Assumption 3.1.

(i) The functions $\overline{b}(x,k,u)$, $\sigma_{ij}(x,k)$, and $\overline{\lambda}_{ij}(x,u)$ are continuous and Lipschitz in x, uniformly with respect to u, with a Lipschitz constant γ_0. Let m_0 denote the least upper bound of $\|\overline{b}(0,k,\cdot)\|_\infty$, $|\sigma_{ij}(0,k)|$, and $\|\overline{\lambda}_{ij}(0,\cdot)\|_\infty$.

(ii) $\sigma_{ij}(\cdot,\cdot)$ is uniformly elliptic; i.e., there exists a constant $m>0$ such that $\sigma(\cdot,k)\sigma'(\cdot,k)\ge mI$.

(iii) The matrix $\widetilde{\Lambda}(x,u)$ is irreducible for all $(x,u)\in\mathbb{R}^d\times U$.

For a Polish space Y, $\mathfrak{B}(Y)$ denotes its Borel σ-field and $\boldsymbol{\mathcal{P}}(Y)$ the space of probability measures endowed with the Prohorov topology, i.e., the topology of weak convergence. Let $\mathfrak{M}(Y)$ be the set of all nonnegative, integer-valued, σ-finite measures on $\mathfrak{B}(Y)$. Let $\mathfrak{M}_\sigma(Y)$ be the smallest σ-field on $\mathfrak{M}(Y)$ with respect to which all the maps from $\mathfrak{M}(Y)$ to $\mathbb{N}\bigcup\{\infty\}$ of the form $\mu\mapsto\mu(B)$ with $B\in\mathfrak{B}(Y)$ are measurable. $\mathfrak{M}(Y)$ is assumed to be endowed with this measurability structure. Let $\mathcal{V}=\boldsymbol{\mathcal{P}}(U)$ and $b=\big[b_1,\dots,b_d\big]'\colon\mathbb{R}^d\times\mathcal{S}\times\mathcal{V}\to\mathbb{R}^d$ be defined by

$$b_i(\cdot,\cdot,v)=\int_U \overline{b}_i(\cdot,\cdot,u)v(du)\,. \tag{3.1}$$

Similarly, for $i,j\in\mathcal{S}$ and $v\in\mathcal{V}$, λ_{ij} is defined as

$$\lambda_{ij}(\cdot,v)=\int_U \overline{\lambda}_{ij}(\cdot,u)v(du)\,. \tag{3.2}$$

For $i,j\in\mathcal{S}$, $x\in\mathbb{R}^d$, and $v\in\mathcal{V}$, let $\Delta_{ij}(x,v)$ be consecutive (with respect to the lexicographic ordering on $\mathcal{S}\times\mathcal{S}$), left closed, right open intervals of the real line, each having length $\lambda_{ij}(x,v)$. Define a function $h:\mathbb{R}^d\times\mathcal{S}\times\mathcal{V}\times\mathbb{R}\longrightarrow\mathbb{R}$ by

$$h(x,i,v,z)=\begin{cases} j-i & \text{if } z\in\Delta_{ij}(x,v),\\ 0 & \text{otherwise.}\end{cases} \tag{3.3}$$

Let $\big(X(t),S(t)\big)$ be the $(\mathbb{R}^d\times\mathcal{S})$-valued, controlled, switching diffusion process given by the following stochastic differential equations:

$$\begin{aligned} dX(t)&=b\big(X(t),S(t),v(t)\big)\,dt+\sigma\big(X(t),S(t)\big)dW(t),\\ dS(t)&=\int_{\mathbb{R}} h\big(X(t),S(t-),v(t),z\big)\,\mathfrak{p}(dt,dz)\end{aligned} \tag{3.4}$$

for $t\ge0$ with $X(0)=X_0$, $S(0)=S_0$, where

(i) X_0 is a prescribed $\mathbb{R}^d$-valued random variable.

(ii) S_0 is a prescribed $\mathcal{S}$-valued random variable.

(iii) $W(\cdot)=\big[W_1(\cdot),\dots,W_d(\cdot)\big]'$ is a d-dimensional standard Wiener process.

(iv) $\mathfrak{p}(dt,dz)$ is an $\mathfrak{M}(\mathbb{R}_+\times\mathbb{R})$-valued Poisson random measure with intensity $dt\times m(dz)$, where m is the Lebesgue measure on $\mathbb{R}$.

(v) $\mathfrak{p}(\cdot,\cdot)$, $W(\cdot)$, X_0, and S_0 are independent.

(vi) $v(\cdot)$ is a $\mathcal{V}$-valued process with measurable sample paths satisfying the nonanticipativity property that the σ-fields $\mathfrak{F}_t^v$ and $\mathfrak{F}_{[t,\infty)}^{W,\mathfrak{p}}$ given by

$$\begin{aligned}\mathfrak{F}_t^v&=\sigma\{v(s),s\le t\},\\ \mathfrak{F}_{[t,\infty)}^{W,\mathfrak{p}}&=\sigma\big\{W(s)-W(t),\mathfrak{p}(A,B)\ :\ A\in\mathfrak{B}\big([s,\infty)\big),B\in\mathfrak{B}(\mathbb{R}),s\ge t\big\}\end{aligned}$$

are independent for each $t\in\mathbb{R}$.

A process $v(\cdot)$ satisfying (vi) is called an *admissible (control) policy*. If $v(\cdot)$ is a Dirac measure, i.e., $v(\cdot) = \delta_{u(\cdot)}$, where $u(\cdot)$ is U-valued, then it is called an admissible *nonrandomized* policy. An admissible policy is called *feedback* if $v(\cdot)$ is progressively measurable with respect to the natural filtration $\mathfrak{F}_t = \{X(s), S(s),\ s \le t\}$.

A particular subclass of feedback policies is of special interest. A feedback policy $v(\cdot)$ is called a (homogeneous) Markov policy if $v(t) = \tilde{v}\big(X(t), S(t)\big)$ for a measurable map $\tilde{v} : \mathbb{R}^d \times \mathcal{S} \to \mathcal{V}$. With an abuse in notation the map $\tilde{v}$ itself is called a Markov policy. Let Π, Π_M, and Π_{MD} denote the sets of all admissible, Markov, and nonrandomized Markov policies, respectively.

If $\big(W(\cdot), \mathfrak{p}(\cdot,\cdot), X_0, S_0, v(\cdot)\big)$ satisfying (i)–(vi) above are given on a prescribed probability space $(\Omega, \mathfrak{G}, P)$, then under Assumption 3.1, equation (3.4) admits an almost sure unique strong solution [17, Chap. 3], and $X(\cdot) \in C(\mathbb{R}_+; \mathbb{R}^d)$, $S(\cdot) \in D(\mathbb{R}_+; \mathcal{S})$, where $D(\mathbb{R}_+; \mathcal{S})$ is the space of right continuous functions on $\mathbb{R}_+$ with left limits taking values in $\mathcal{S}$. However, if $v(\cdot)$ is a feedback policy, then there exists a measurable map

$$f : \mathbb{R}_+ \times C(\mathbb{R}_+; \mathbb{R}^d) \times D(\mathbb{R}_+; \mathcal{S}) \longrightarrow \mathcal{V}$$

such that for each $t \ge 0$, $v(t) = f\big(t, X(\cdot), S(\cdot)\big)$ and is progressively measurable with respect to $\{\mathfrak{F}_t\}$. Thus, $v(\cdot)$ cannot be specified a priori in (3.4). Instead, one has to replace $v(t)$ by $f\big(t, X(\cdot), S(\cdot)\big)$, and (3.4) takes the form

$$\begin{aligned} dX(t) &= b\big(X(t), S(t), f(t, X(\cdot), S(\cdot))\big)dt + \sigma\big(X(t), S(t)\big)dW(t), \\ dS(t) &= \int_{\mathbb{R}} h\big(X(t), S(t-), f(t, X(\cdot), S(\cdot)), z\big)\mathfrak{p}(dt, dz)\,, \end{aligned} \tag{3.5}$$

for $t \ge 0$ with $X(0) = X_0$, $S(0) = S_0$. In general, (3.5) does not even admit a weak solution. However, if the feedback policy is Markov, then the existence of a unique strong solution can be established.

If $\boldsymbol{\mathcal{K}}(\mathbb{R}^d)$ is a vector space of real functions over $\mathbb{R}^d$, we adopt the notation $\boldsymbol{\mathcal{K}}(\mathbb{R}^d \times \mathcal{S})$ to indicate the space $\big(\boldsymbol{\mathcal{K}}(\mathbb{R}^d)\big)^N$, endowed with the product topology. For example,

$$L^p(\mathbb{R}^d \times \mathcal{S}) := \Big\{ f : \mathbb{R}^d \times \mathcal{S} \to \mathbb{R} \,:\, f(\cdot, i) \in L^p(\mathbb{R}^d) \text{ for all } i \in \mathcal{S} \Big\},$$

and similarly, we define $C^k(\mathbb{R}^d \times \mathcal{S})$, $W^{k,p}(\mathbb{R}^d \times \mathcal{S})$, etc. For $f \in W^{2,p}_{\ell oc}(\mathbb{R}^d \times \mathcal{S})$ and $u \in U$, we write

$$L^u f(x,k) = L^u_k f(x,k) + \sum_{j \in \mathcal{S}} \overline{\lambda}_{kj}(x,u) f(x,j)\,, \tag{3.6}$$

where

$$L^u_k = \frac{1}{2} \sum_{i,j,\ell=1}^{d} \sigma_{i\ell}(x,k)\sigma_{j\ell}(x,k) \frac{\partial^2}{\partial x_i \partial x_j} + \sum_{j=1}^{d} \overline{b}_j(x,k,u) \frac{\partial}{\partial x_j} \tag{3.7}$$

and, more generally, for $v \in \mathcal{V}$,

$$L^v f(x,k) = \int_U L^u f(x,k) v(du)\,. \tag{3.8}$$

The following result is proved in [14].

THEOREM 3.1. *Under a Markov policy v,* (3.4) *admits an almost sure unique strong solution such that $\big(X(\cdot), S(\cdot)\big)$ is a strong Feller process with differential generator L^v.*

A Markov policy v is called *stable* if the corresponding process $\big(X(\cdot), S(\cdot)\big)$ is positive recurrent. In this case, the process has a unique invariant probability measure, denoted by $\eta_v \in \boldsymbol{\mathcal{P}}(\mathbb{R}^d \times \mathcal{S})$. The uniqueness of η_v is guaranteed by Assumption 3.1. We assume that the set of stable Markov policies is nonempty.

The optimization problem. Let $\overline{c} : \mathbb{R}^d \times \mathcal{S} \times U \to \mathbb{R}_+$ be the cost function. The following assumption on the cost, $\overline{c}$, will be in effect throughout the paper.

Assumption 3.2. For each $i \in \mathcal{S}$, $\overline{c}(\cdot, i, \cdot)$ is continuous.

We define $c : \mathbb{R}^d \times \mathcal{S} \times \mathcal{V} \to \mathbb{R}_+$ by

$$c(x, i, v) = \int_U \overline{c}(x, i, u) v(du)\,. \tag{3.9}$$

Let $v(\cdot)$ be an admissible policy and $\big(X(\cdot), S(\cdot)\big)$ the corresponding process. The pathwise (long-run) average cost incurred under $v(\cdot)$ is

$$\limsup_{T\to\infty} \frac{1}{T} \int_0^T c\big(X(t), S(t), v(t)\big) dt\,. \tag{3.10}$$

We wish to a.s. minimize (3.10) over all admissible policies. Our goal is to establish the existence of a stable Markov policy which is a.s. optimal. In general, this is not the case, as the following simple counterexample shows [6]. Let $\overline{c}(x, i) = \exp(-\|x\|^2)$. Then for every stable Markov policy the average cost is positive a.s., while we can find an unstable Markov policy for which the average cost is a.s. zero, making it an optimal policy. We want to rule out this possibility, as stability is a very desirable property. We carry out our study under two alternate sets of hypotheses: (a) a condition on the cost which penalizes unstable behavior, (b) a blanket stability condition which implies that all Markov policies are stable. We describe these conditions in section 6.

4. Recurrence, ergodicity, and harmonic functions of switching diffusions. Due to the interaction between the continuous and discrete components, the study of recurrence and ergodicity of switching diffusions is quite involved. Let v be a Markov policy which will be fixed throughout this section unless explicitly stated otherwise. Let $P^v : \mathbb{R}_+ \times \mathbb{R}^d \times \mathcal{S} \to \boldsymbol{\mathcal{P}}(\mathbb{R}^d \times \mathcal{S})$ denote the transition function of the corresponding process $\big(X(\cdot), S(\cdot)\big)$. Also $P^v_{x,i}$ and $E^v_{x,i}$ denote the probability measure and the expectation operator, respectively, on the canonical space of the process $\big(X(\cdot), S(\cdot)\big)$ starting at $(x, i) \in \mathbb{R}^d \times \mathcal{S}$. The following result plays a crucial role in recurrence.

LEMMA 4.1. *For any $(t, x, i) \in \mathbb{R}_+ \times \mathbb{R}^d \times \mathcal{S}$, the support of $P^v(t, x, i; \cdot)$ is $\mathbb{R}^d \times \mathcal{S}$.*

Proof. For each $i \in \mathcal{S}$, let τ_i denote the sojourn time of $S(t)$ in state i. Then

$$P^v_{x,i}\big(\tau_i > t\big) = E^v_{x,i}\bigg[\exp\Big(\int_0^t \lambda_{ii}\big(X(s), v(X(s), S(s))\big) ds\Big)\bigg].$$

Let $\lambda^v_{ij}(s) := \lambda_{ij}\big(X(s), v(X(s), S(s))\big)$, $I_{A,j}(s) := I\{X(s) \in A, S(s) = j\}$, and P^v_i be the transition function of the diffusion corresponding to L^v_i, i.e., the diffusion with no

switching and $S(t) \equiv i$. For $A \in \mathfrak{B}(\mathbb{R}^d)$, $i, j \in \mathcal{S}$, and $t > 0$,

(4.1)
$$\begin{aligned} P^v\big(t,x,i,A\times\{j\}\big) &= E^v_{x,i}\big[I_{A,j}(t) \bigm| \tau_i > t\big] P^v_{x,i}\big(\tau_i > t\big) + E^v_{x,i}\big[I_{A,j}(t) I\{\tau_i < t\}\big] \\ &= E^v_{x,i}\biggl[\exp\Big(\int_0^t \lambda^v_{ii}(s)\,ds\Big)\biggr] P^v_i(t,x,A)\delta_{ij} \\ &\quad + E^v_{x,i}\biggl[\int_0^t -\lambda^v_{ii}(s)\exp\Big(\int_0^s \lambda^v_{ii}(s')ds'\Big)ds \\ &\qquad\qquad \int_{\mathbb{R}^d} P^v_i(s,x,dy)\sum_{k\neq i}\lambda^v_{ik}(s)P^v\big(t-s,y,k,A\times\{j\}\big)\biggr] \\ &= E^v_{x,i}\biggl[\exp\Big(\int_0^t \lambda^v_{ii}(s)\,ds\Big)\biggr] P^v_i(t,x,A)\delta_{ij} \\ &\quad + \sum_{k\neq i}\int_0^t E^v_{x,i}\biggl[-\lambda^v_{ii}(s)\lambda^v_{ik}(s)\exp\Big(\int_0^s \lambda^v_{ii}(s')ds'\Big)\biggr] \\ &\qquad\qquad \int_{\mathbb{R}^d} P^v_i(s,x,dy)P^v\big(t-s,y,k,A\times\{j\}\big)ds\,. \end{aligned}$$

Define the transition matrix $\widetilde{\Pi}^v$ by

$$[\widetilde{\Pi}^v(t,x,A)]_{ij} = P^v\big(t,x,i,A\times\{j\}\big)\,.$$

Then we can suitably define the matrix measures

$$\Gamma^v_1, \Gamma^v_2 : \mathbb{R}\times\mathbb{R}^d \to \big(\boldsymbol{\mathcal{P}}(\mathbb{R}^d)\big)^{N\times N}$$

with $\Gamma^v_1(t,x,A)$ positive, diagonal and $\Gamma^v_2(t,x,A)$ nonnegative, irreducible (by Assumption 3.1 (iii)), for all $(t,x,A) \in \mathbb{R}_+ \times \mathbb{R}^d \times \mathfrak{B}(\mathbb{R}^d)$, provided A has positive Lebesgue measure, so as to write (4.1) in the form

$$\widetilde{\Pi}^v(t,x,A) = \Gamma^v_1(t,x,A) + \int_0^t\int_{\mathbb{R}^d} \Gamma^v_2(s,x,dy)\widetilde{\Pi}^v(t-s,y,A)ds\,. \tag{4.2}$$

The desired result follows from (4.2), using the irreducibility of $\Gamma^v_2(t,x,A)$. □

Let τ_{ii}, τ_j be the stopping times defined as follows:

$$\tau_{ii} = \inf\big\{t > 0 \,:\, S(t) = i \text{ and } S(t') \neq i, \text{ for some } 0 < t' < t\big\}, \tag{4.3}$$

$$\tau_j = \inf\big\{t > 0 \,:\, S(t) = j\big\}. \tag{4.4}$$

Let $D \subset \mathbb{R}^d$ be a bounded open set and J a subset of $\mathcal{S}$. Define

$$\tau_{D,J} = \inf\big\{t \geq 0 \,:\, \big(X(t),S(t)\big) \notin D\times J\big\}, \tag{4.5}$$

$$\tau_D = \inf\big\{t \geq 0 \,:\, X(t) \notin D\big\}. \tag{4.6}$$

Using (4.2) and well-known arguments in Markov processes [12, Vol. I, p. 111] the following results can be proved.

LEMMA 4.2. *If τ is a stopping time of the form τ_{ii}, τ_j, $\tau_{D,J}$, or τ_D, as defined in* (4.3)–(4.6)*, then, for each compact set $K \subset \mathbb{R}^d$,*

$$\sup_{v\in\Pi_M,\, x\in K} E^v_{x,i}[\tau] < \infty\,.$$

It is well known that harmonic functions play an important role in the study of recurrence and ergodicity of Markov processes [3]. Therefore, we now turn to the analysis of some properties of the harmonic functions of the process $\big(X(\cdot), S(\cdot)\big)$ under the Markov policy v. The function f is called *L^v-harmonic* in D if it is bounded on compact subsets of D, and for all $x \in D$, $i \in \mathcal{S}$,

$$f(x,i) = E^v_{x,i} f\big(X(\tau_{V,J}), S(\tau_{V,J})\big) \tag{4.7}$$

for every neighborhood V of x having compact closure $\overline{V}$ in D and every subset $J \subset \mathcal{S}$ containing i. It is clear that if f is L^v-harmonic then

$$f(x,i) = E^v_{x,i} f\big(X(\tau_V), S(\tau_V)\big). \tag{4.8}$$

On the other hand, if (4.8) holds, then by conditioning on $\mathfrak{F}_{\tau_{V,J}}$ we obtain

$$\begin{aligned} f(x,i) &= E^v_{x,i}\Big[E^v\big[f\big(X(\tau_V), S(\tau_V)\big) \mid \mathfrak{F}_{\tau_{V,J}}\big]\Big] \\ &= E^v_{x,i}\Big[E^v_{X_{\tau_{V,J}}, S_{\tau_{V,J}}}\big[f\big(X(\tau_V - \tau_{V,J}), S(\tau_V - \tau_{V,J})\big)\big]\Big] \\ &= E^v_{x,i}\big[f\big(X(\tau_{V,J}), S(\tau_{V,J})\big)\big]\,, \end{aligned}$$

concluding that (4.7) and (4.8) are actually equivalent.

LEMMA 4.3. *Let $D \subset \mathbb{R}^d$ be open. Then we have the following:*

(i) *Every L^v-harmonic function in D is continuous in D.*

(ii) *If $L^v f = 0$ in D and $f \in W^{2,p}(D \times \mathcal{S})$, then f is L^v-harmonic. Conversely, if f is L^v-harmonic and $f \in W^{2,p}_{\ell oc}(D \times \mathcal{S})$, then $L^v f = 0$ in D.*

(iii) *(Maximum principle) Let D be connected and $f \ge 0$ and L^v-harmonic in D. Then f is either strictly positive in $D \times \mathcal{S}$ or identically zero.*

Proof. The proof of (i) is standard [3], [12, Vol. II, Chap. 13], and (ii) can easily be proved using the generalized Ito formula [18]. Let $x_0 \in D$, $i_0 \in \mathcal{S}$, and $r > 0$ be such that $f(x_0, i_0) = 0$ and $\overline{B}(x_0, r) \subset D$, where $\overline{B}(x_0, r) = \big\{x \in \mathbb{R}^d : \|x - x_0\| \le r\big\}$. Then

$$0 = f(x_0, i_0) = \sum_{j \in \mathcal{S}} \int_{\partial B(x_0, r)} f(y, j) P^v_{x_0, i_0}\big(X(\tau_{B(x_0,r)}) \in dy, S(\tau_{B(x_0,r)}) = j\big).$$

Then, by Lemma 4.1, we can show using standard arguments [16, Chap. 6] that the support of the measure $P^v_{x_0,i_0}\big(X(\tau_{B(x_0,r)}) \in dy, S(\tau_{B(x_0,r)}) = j\big)$ is $\partial B(x_0, r) \times \mathcal{S}$. Hence,

$$f(y, j) = 0, \qquad \text{for all } y \in \partial B(x_0, r),\ j \in \mathcal{S}\,.$$

It follows that the set $\big\{y : f(y, j) = 0,\ j \in \mathcal{S}\big\}$ is open in D, and since D is connected, the result follows. □

We next state Harnack's inequality for L^v-harmonic functions, which extends a very important result in partial differential equations. This inequality plays a crucial role in proving the existence of a solution to the HJB equation via the vanishing discount method, as is done in section 6. As far as we know, this result is not known in the literature on partial differential equations. The detailed proof of Harnack's inequality is quite elaborate and can be found in the appendix. The proof follows the method introduced for diffusions by Krylov and Safonov [19] for deriving estimates for the oscillation of a harmonic function. For the system of coupled elliptic operators

characterizing switching diffusions, considerable complications arise in trying to follow the same methodology due to the vector-valued nature of the L^v-harmonic functions. A crucial step in the proof is "coupling" together the oscillations of the distinct components of the harmonic function. The irreducibility of the matrix $\widetilde{\Lambda}$ is essential in accomplishing this task.

THEOREM 4.1 (Harnack's inequality). *Let $\Omega \subset \mathbb{R}^d$ be a bounded domain and $K \subset \Omega$ a closed set. There exists a constant $C > 0$, depending only on Ω, K, the dimension d, N, the bounds m, m_0, and the Lipschitz constant γ_0 introduced in Assumption* 3.1*, such that for any nonnegative function $f \in W^{2,p}_{\ell oc}(\Omega \times \mathcal{S})$, $p \in [1, \infty)$, satisfying $L^v f = 0$ in $\Omega \times \mathcal{S}$, for some Markov policy v,*

$$f(x,i) \le C f(y,j) \qquad \forall\, x, y \in K \quad \forall\, i, j \in \mathcal{S}\,.$$

We now discuss the recurrence properties of switching diffusions. Our treatment closely follows [3], so we skip the details in several places. A point $(x,i) \in \mathbb{R}^d \times \mathcal{S}$ is said to be *recurrent* if, given any $\varepsilon > 0$,

$$P^v_{x,i}\big(X(t_n) \in B(x,\varepsilon),\ S(t_n) = i,\ \text{for a sequence } t_n \uparrow \infty\big) = 1. \tag{4.9}$$

A point (x,i) is *transient* if

$$P^v_{x,i}\big(\|X(t)\| \to \infty,\ \text{as } t \to \infty\big) = 1. \tag{4.10}$$

If all points of the switching diffusion are recurrent, then it is called recurrent. A transient switching diffusion is similarly defined. Note that the discrete component of the process has been ignored in the definition (4.10). The reason for doing so is that, in view of Assumption 3.1 (iii), we can show that, provided the continuous component visits a bounded set infinitely often with probability 1, then the discrete component is recurrent. More generally, a switching diffusion exhibits a dichotomy in that it is either recurrent or transient, as we will later show.

LEMMA 4.4. *The following statements are equivalent.*

(i) *The switching diffusion is recurrent;*

(ii) $P^v_{x,i}\big(X(t) \in D,\ S(t) = j,\ \text{for some } t \ge 0\big) = 1$*, for any open set $D \subset \mathbb{R}^d$ and any $j \in \mathcal{S}$.*

Proof. We prove (i) $\to$ (ii) (the converse is easier). We distinguish two cases.

Case 1. Let $x \in D$, $i \ne j$. Let $B = B(x,\varepsilon)$ and B_1 be bounded open sets such that $\overline{B} \subset B_1$ and $\overline{B}_1 \subset D$. Let

$$\eta_1 = \inf\big\{t \ge 0\ :\ X(t) \in \partial B_1\big\},$$

and inductively, for $n = 1, 2, \dots,$

$$\eta_{2n} = \inf\big\{t > \eta_{2n-1}\ :\ X(t) \in \partial B\big\},$$
$$\eta_{2n+1} = \inf\big\{t > \eta_{2n}\ :\ X(t) \in \partial B_1\big\}.$$

Then, by recurrence, $\eta_n < \infty$, $P^v_{x,i}$ a.s. Note that

$$y, \ell \mapsto P^v_{y,\ell}\big(\tau_{(\overline{B}\times\{j\})^c} < \tau_{B_1}\big)$$

is L^v-harmonic in $B_1 \times \mathcal{S}$ and not identically zero. Therefore, by Lemma 4.3,

$$\inf_{(y,\ell)\in\overline{B}\times\mathcal{S}} P^v_{y,\ell}\big(\tau_{(\overline{B}\times\{j\})^c} < \tau_{B_1}\big) > \delta_1 > 0 \tag{4.11}$$

for some $\delta_1 > 0$. Next we define

$$A_0 = \{S(t) = j \text{ for some } t \in [0, \eta_1)\},$$
$$A_n = \{S(t) = j \text{ for some } t \in [\eta_{2n}, \eta_{2n+1})\}.$$

By (4.11) and the strong Markov property,

$$P^v_{x,i}(A_0^c) \le (1 - \delta_1), \qquad P^v_{x,i}\Big(\bigcap_{k=0}^{n} A_k^c\Big) \le (1 - \delta_1)^{n+1}.$$

Now,

$$\begin{aligned} P^v_{x,i}\big(X(t) \in D,\ S(t) = j \text{ for no } t \ge 0\big) &\le P^v_{x,i}\big(X(t) \in \overline{B}_1,\ S(t) = j \text{ for no } t \ge 0\big) \\ &\le \lim_{n\to\infty} P^v_{x,i}\Big(\bigcap_{k=0}^{n} A_k^c\Big) = 0. \end{aligned}$$

Case 2. Suppose $x \notin D$ and let $B = B(x, \varepsilon)$, B_1, and D_1 be bounded open sets such that $B \bigcap D = \emptyset$, $\overline{B}_1 \subset D$, and $\overline{B} \bigcup \overline{B}_1 \subset D_1$. Let

$$\eta'_1 = \tau_{D_1},$$
$$\eta'_{2n} = \{t > \eta'_{2n-1} \,:\, X(t) \in \partial B\},$$
$$\eta'_{2n+1} = \{t > \eta'_{2n} \,:\, X(t) \in \partial D_1\}.$$

Let $\delta_2 > 0$ be such that

$$\inf_{(y,\ell)\in \partial D_1 \times \mathcal{S}} P^v_{y,\ell}\big(\tau_{(\overline{B}_1 \times \{j\})^c} < \tau_{(\overline{B}\times\{i\})^c}\big) > \delta_2 > 0.$$

Define

$$A'_n = \{X(t) \in \overline{B}_1, S(t) = j \text{ for some } t \in [\eta_{2n-1}, \eta_{2n})\}.$$

Then, as in the previous case,

$$P^v_{x,i}\big(X(t) \in D,\ S(t) = j \text{ for no } t \ge 0\big) = 0. \qquad \square$$

In view of Lemma 4.4, the following results can be proved the same way as in [3], [4].

LEMMA 4.5. *The following statements are equivalent.*

(i) *The switching diffusion is recurrent.*

(ii) $P^v_{x,i}\big(X(t) \in D \text{ for some } t \ge 0\big) = 1$ *for all* $x \in \mathbb{R}^d$, $i \in \mathcal{S}$, *and any nonempty open set* D.

(iii) *There exists a compact set* $K \subset \mathbb{R}^d$ *such that* $P^v_{x,i}\big(X(t) \in K \text{ for some } t \ge 0\big) = 1$ *for all* $(x, i) \in \mathbb{R}^d \times \mathcal{S}$.

(iv) $P^v_{x,i}\big(X(t_n) \in D, \text{ for a sequence } t_n \uparrow \infty\big) = 1$ *for all* $x \in \mathbb{R}^d$, $i \in \mathcal{S}$, *and any nonempty open set* D.

(v) *There exists a point* $z \in \mathbb{R}^d$, *a pair of numbers* r_0, r_1, $0 < r_0 < r_1$, *and a point* $y \in \partial B(z, r_1)$ *such that* $P^v_{y,i}\big(\tau_{\overline{B}(z,r_0)^c} < \infty\big) = 1$ *for any* $i \in \mathcal{S}$.

THEOREM 4.2. *For any Markov policy, the switching diffusion is either recurrent or transient.*

A recurrent switching diffusion admits a unique (up to a constant multiple) σ-finite invariant measure. The switching diffusion is called *positive recurrent* if it is recurrent and admits a finite invariant measure.

A Markov policy v is called *stable* if the corresponding process is positive recurrent; the corresponding invariant probability measure is denoted by η_v.

As is well known from the general theory of dynamical systems, even if L_i^v generates a positive recurrent diffusion, for each $i \in \mathcal{S}$, and the parametric Markov chain is ergodic, there is no reason to expect that the policy v is stable; i.e., the switching diffusion is positive recurrent. Indeed, as the following example shows, the hybrid process can be anything from transient to positive recurrent.

Example 4.1. We first consider a piecewise deterministic system with state dependent Markovian switching. Let $E_+, E_- \subset \mathbb{R}^2$ be defined as follows:

$$E_+ = \{(x_1, x_2) \,:\, x_1 > 0\} \bigcup \{x_2 \le 0, x_1 = 0\},$$
$$E_- = \{(x_1, x_2) \,:\, x_1 < 0\} \bigcup \{x_2 \ge 0, x_1 = 0\}.$$

Let

$$A_0 = \begin{bmatrix} 2 & 1 \\ -1 & 2 \end{bmatrix}, \qquad A_1 = \begin{bmatrix} -3 & 1 \\ -1 & -3 \end{bmatrix}.$$

Consider two stable dynamical systems $\mathcal{D}_0$ and $\mathcal{D}_1$ defined by

$$\mathcal{D}_0 : \quad \dot{x} = \begin{cases} A_0 x, & x \in E_+, \\ A_1 x, & x \in E_-, \end{cases}$$

and

$$\mathcal{D}_1 : \quad \dot{x} = \begin{cases} A_1 x, & x \in E_+, \\ A_0 x, & x \in E_-. \end{cases}$$

For $\delta > 0$, let Z be a (parameterized) Markov chain taking values in $\{0, 1\}$ with rate matrix

$$\begin{bmatrix} -\delta & \delta \\ \frac{1}{\delta} & -\frac{1}{\delta} \end{bmatrix} \text{ on } E_+ \quad \text{and} \quad \begin{bmatrix} -\frac{1}{\delta} & \frac{1}{\delta} \\ \delta & -\delta \end{bmatrix} \text{ on } E_-$$

and consider the dynamical system

$$\mathcal{D} := \mathcal{D}_Z.$$

If we define η by

$$\eta = \begin{cases} Z, & x \in E_+, \\ 1 - Z, & x \in E_-, \end{cases}$$

then η is Markovian with rate matrix

$$\begin{bmatrix} -\delta & \delta \\ \frac{1}{\delta} & -\frac{1}{\delta} \end{bmatrix},$$

and $\mathcal{D}$ can represented as

$$\dot{x} = A_\eta x.$$

Define

$$T_0(t) = \{\tau \le t \,:\, \eta(\tau) = 0\},$$
$$T_1(t) = \{\tau \le t \,:\, \eta(\tau) = 1\},$$

and $\lambda_0(t) = m(T_0(t))$, $\lambda_1(t) = m(T_1(t))$, where m is the Lebesgue measure on $\mathbb{R}_+$. Then, the solution to $\mathcal{D}$ can be expressed as

$$x(t) = \exp\bigl(2\lambda_0(t) - 3\lambda_1(t)\bigr) \begin{bmatrix} \cos t & -\sin t \\ \sin t & \cos t \end{bmatrix} x(0).$$

By the ergodic theory of Markov processes [25, Chap. 1], as $t \to \infty$,

$$\lambda_1(t) \sim \frac{\delta^2 t}{1+\delta^2}\,, \qquad \lambda_0(t) \sim \frac{t}{1+\delta^2}\,.$$

Thus,

$$2\lambda_0(t) - 3\lambda_1(t) \sim \frac{2-3\delta^2}{1+\delta^2}\,t\,.$$

Therefore, $\mathcal{D}$ is stable for $\delta < \sqrt{\frac{2}{3}}$ and unstable for $\delta \ge \sqrt{\frac{2}{3}}$. The matrices A_0, A_1 can be suitably altered to exhibit various other possibilities.

Now let $X(t)$ be defined as $dX(t) = A_{\eta(t)}X(t)dt + \sigma dW(t)$, where $W(\cdot)$ is a standard two-dimensional Wiener process and $\sigma\sigma'$ is a 2×2 positive definite matrix with constant entries. Then it is easily shown that the stability (instability) of $\mathcal{D}$ implies the positive recurrence (transience) of $X(t)$. Note that in this example the drift is unbounded. However, in the study of recurrence, boundedness of the drift can be replaced by local boundedness.

Remark 4.1. In view of the above example, it is clear that two positive recurrent processes with suitable switching may result in a transient process. Similarly, the random combination of two transient processes may give rise to a positive recurrent process. This phenomenon can be exploited in many practical situations such as fault-tolerant control systems, flexible manufacturing systems, etc. In a control system with multiple modes, we can trade off the stability of some (or all) nodes to gain a desired degree of flexibility. Addition of a few redundant nodes and/or the incorporation of a suitable switching mechanism among the nodes could result in global stability of the system, thereby gaining flexibility without sacrificing reliability.

A general criterion for positive recurrence of a switching diffusion is provided by the following theorem.

THEOREM 4.3. *Let z, r_0, r_1 be as in Lemma* 4.5(v). *Then the switching diffusion is positive recurrent if*

$$\sup_{y \in \partial B(z,r_1), i \in \mathcal{S}} E^v_{y,i}\bigl[\tau_{\overline{B}(z,r_0)^c}\bigr] < \infty\,. \tag{4.12}$$

The proof is standard [3]. Note that it may be very difficult to verify (4.12) for general b, σ, λ. One usually verifies (4.12) by constructing a Lyapunov function [3]. For switching diffusions such a construction seems difficult, since it involves solving a system of ordinary differential equations in closed form. However, we present some criteria for positive recurrence and discuss some implications.

(C1) There exists a $w \in C^2(\mathbb{R}^d \times \mathcal{S})$, $w \geq 0$, such that

(i) $w(x,i) \to \infty$, as $\|x\| \to \infty$.

(ii) For each $v \in \Pi_M$, $E^v_{x,i}\big[w\big(X(t),S(t)\big)\big]$ and $E^v_{x,i}\big|L^v w\big(X(t),S(t)\big)\big|$ are locally bounded.

(iii) There exists $p > 0$, $q > 0$ such that $L^u w(x,i) \leq p - qw(x,i)$, for each $u \in U$.

(C2) There exists a C^2 function $w : \mathbb{R}^d \times \mathcal{S} \to \mathbb{R}_+$ such that

(i) $\lim_{\|x\|\to\infty} w(x,i) = +\infty$.

(ii) There exists $a > 0$ and $\varepsilon > 0$ such that for $\|x\| > a$, $L^u w(x,i) < -\varepsilon$, for all $u \in U$, $i \in \mathcal{S}$, and $\|\nabla w(x,i)\|^2 \geq m^{-1}$, where m is the constant in Assumption 3.1 (ii).

(iii) $w(x,i)$ and $\|\nabla w(x,i)\|$ have polynomial growth.

THEOREM 4.4. *Under either* (C1) *or* (C2)*, the process* $\big(X(\cdot),S(\cdot)\big)$ *under any Markov policy* v *is positive recurrent. Thus, all Markov policies are stable.*

Proof. Under (C1), the result follows from [25, Theorem 25, p. 70]. Under (C2), the technique of the proof of [6, Lemma 6.2.2, p. 150] can be closely paralleled to draw the desired conclusion. □

Remark 4.2. If $\sigma \equiv I$ and $\bar{b}$ is such that $\langle \bar{b}(x,i,u), x\rangle < -(d+1)/2$ for all $i \in \mathcal{S}$ and $\|x\|$ sufficiently large, then $w(x) = \|x\|^2$ is a Lyapunov function for the system. We can construct several examples using this idea. Note that in this case all the diffusion generators L^u_i give rise to positive recurrent diffusions and have a common Lyapunov function (i.e., one which is independent of i). If all L^u_i have a common Lyapunov function, then switching does not destabilize the hybrid system. Of course, this is a very strong condition and is rarely met.

5. Existence of an optimal policy. In this section we establish the existence of a stable, nonrandomized Markov optimal policy under certain conditions. We follow the methodology developed in [6], [8], [9], [10] for controlled diffusions. For switching diffusions, similar techniques carry through with some extra effort. Therefore, we present the main ideas, skipping some of the technical details.

Let Π_{SM} and Π_{SMD} denote the set of stable Markov and stable nonrandomized Markov policies, respectively. Since we are searching for an optimal policy in Π_{SMD}, it is natural to assume that Π_{SM} is nonempty. Let $v \in \Pi_{SM}$. Then

$$\begin{aligned}\rho_v &:= \sum_{i\in\mathcal{S}} \int_{\mathbb{R}^d} c\big(x,i,v(x,i)\big)\eta_v(dx,i) \\ &= \lim_{T\to\infty} \frac{1}{T} \int_0^T c\big(X(s),S(s),v(X(s),S(s))\big)ds \quad \text{a.s.}\end{aligned} \tag{5.1}$$

Let

$$\rho^* := \inf_{v\in\Pi_{SM}} \big\{\rho_v\big\}. \tag{5.2}$$

We assume that $\rho^* < \infty$. We now state a condition on the cost function which penalizes unstable behavior.

(C3) Assume that for each $i \in \mathcal{S}$,

$$\liminf_{\|x\|\to\infty} \big\{ \inf_{u\in U} \bar{c}(x,i,u)\big\} > \rho^*. \tag{5.3}$$

Intuitively, (5.3) penalizes trajectories lying outside the set $\inf_{u\in U}\big\{\bar{c}(x,i,u)\big\} \leq \rho^*$, forcing an optimal process to spend a nonvanishing fraction of time in a bounded

neighborhood of this compact set. This behavior results in the stability of every optimal policy. If $\overline{c}(x,i,u) = K(\|x\|)$ for some increasing function $K : \mathbb{R}_+ \to \mathbb{R}_+$ then it can be easily seen that (5.3) holds. Such cost functions arise quite often in practice. Condition (C3) is referred to as the *near-monotonicity* condition [6, Chap. 6].

For $v \in \Pi_{SM}$ (or Π_{SMD}), we define the ergodic occupation measure $\mu[v] \in \boldsymbol{\mathcal{P}}(\mathbb{R}^d \times \mathcal{S} \times U)$ as

$$\mu[v](dx,i,du) = \eta_v(dx,i)v(x,i)(du). \tag{5.4}$$

Let

$$I_1 = \big\{\mu[v] \,:\, v \in \Pi_{SM}\big\},$$
$$I_2 = \big\{\mu[v] \,:\, v \in \Pi_{SMD}\big\}.$$

The following results can be proved as in [10], [14].

LEMMA 5.1. *The sets I_1, I_2 are closed, I_1 is convex, and the set of extreme points of I_1 lies in I_2.*

Let $v(\cdot)$ be an arbitrary admissible policy. Define the $\boldsymbol{\mathcal{P}}(\mathbb{R}^d \times \mathcal{S} \times U)$-valued empirical process $\mu_t(v)$ for $t > 0$ by

$$\mu_t(v)\big(A \times \{i\} \times B\big) = \frac{1}{t}\int_0^t I\big\{X(s) \in A, S(s) = i\big\}\, v(s)(B)\, ds\,, \tag{5.5}$$

with $A \in \mathfrak{B}(\mathbb{R}^d)$, $B \in \mathfrak{B}(U)$, and $i \in \mathcal{S}$. Let $\overline{\mathbb{R}}^d = \mathbb{R}^d \bigcup \{\infty\}$ be the one-point compactification of $\mathbb{R}^d$. We identify $\mu_t(v)$ with an element of $\boldsymbol{\mathcal{P}}(\overline{\mathbb{R}}^d \times \mathcal{S} \times U)$ by assigning zero mass at $\{\infty\} \times \mathcal{S} \times U$. Since $\boldsymbol{\mathcal{P}}(\overline{\mathbb{R}}^d \times \mathcal{S} \times U)$ is compact, $\{\mu_t(v)\}$, viewed as a $\boldsymbol{\mathcal{P}}(\overline{\mathbb{R}}^d \times \mathcal{S} \times U)$-valued process, converges to a sample path-dependent compact limit set in $\boldsymbol{\mathcal{P}}(\overline{\mathbb{R}}^d \times \mathcal{S} \times U)$. Note that any element $\mu \in \boldsymbol{\mathcal{P}}(\overline{\mathbb{R}}^d \times \mathcal{S} \times U)$ can be decomposed as

$$\mu(C) = \delta_\mu\, \mu'\big(C \bigcap (\mathbb{R}^d \times \mathcal{S} \times U)\big) + (1 - \delta_\mu)\mu''\big(C \bigcap (\{\infty\} \times \mathcal{S} \times U)\big)\,, \tag{5.6}$$

for $C \in \mathfrak{B}\big(\overline{\mathbb{R}}^d \times \mathcal{S} \times U\big)$. In this decomposition $\delta_\mu \in [0,1]$ is always uniquely defined, and $\mu' \in \boldsymbol{\mathcal{P}}(\mathbb{R}^d \times \mathcal{S} \times U)$ (respectively, $\mu'' \in \boldsymbol{\mathcal{P}}(\{\infty\} \times \mathcal{S} \times U)$) is also unique if $\delta_\mu > 0$ (respectively, $\delta_\mu < 1$). We may render μ', μ'' unique at all times by imposing an arbitrary fixed choice thereof when $\delta_\mu = 0$, respectively, 1.

Combining the results in [20] with the technique in [6, Lemma 6.1.1, p. 144], we establish the following lemma.

LEMMA 5.2. *If $\mu \in \boldsymbol{\mathcal{P}}(\mathbb{R}^d \times \mathcal{S})$ satisfies*

$$\sum_{i \in \mathcal{S}} \int_{\mathbb{R}^d} L^v f(x,i)\mu(dx,i) = 0 \qquad \forall\ f \in H \tag{5.7}$$

for some Markov policy v, where H is a dense subset of $C_0^2(\mathbb{R}^d \times \mathcal{S})$, then $\mu = \eta_v$.

Proof. Using the usual approximation procedure we can show that (5.7) is true for all $f \in C_b^2(\mathbb{R}^d \times \mathcal{S})$. Let $\big(X(\cdot), S(\cdot)\big)$ be the process corresponding to the policy v with initial law μ. The law μ_t of this process, for $t > 0$, satisfies the Kolmogorov forward equation

$$\sum_{i \in \mathcal{S}} \int_{\mathbb{R}^d} f(x,i)\mu_t(dx,i) = \sum_{i \in \mathcal{S}} \int_{\mathbb{R}^d} f(x,i)\mu(dx,i) + \sum_{i \in \mathcal{S}} \int_0^t \int_{\mathbb{R}^d} L^v f(x,i)\mu_s(dx,i)\, ds$$

for all $f \in C_b^2(\mathbb{R}^d \times \mathcal{S})$. The uniqueness of the solution to the above equation is established in [26]. Since $\mu_t \equiv \mu$ is a solution to (5.7), it follows that $\mu = \eta_v$. ∎

We disintegrate $\mu' \in \boldsymbol{\mathcal{P}}(\mathbb{R}^d \times \mathcal{S} \times U)$ as

$$\mu'(dx, i, du) = \mu^*(dx, i) v_\mu(x, i)(du)\,, \tag{5.8}$$

where μ^* is the marginal of μ' on $\mathbb{R}^d \times \mathcal{S}$ and v_μ is a version of the regular conditional law defined as μ^* a.s. We select an arbitrary version and keep it fixed henceforth. Using the martingale stability theorem, the following characterization of the limit points of $\{\mu_t(\cdot)\}$ can be established as in [6, Lemma 6.1.2].

LEMMA 5.3. *Outside a set of zero probability, each limit point μ of $\{\mu_t(\cdot)\}$ for which $\delta_\mu > 0$ satisfies $\mu^* = \eta_{v_\mu}$.*

We now establish the existence of an optimal policy under (C3). Since the proof closely follows the steps in [6, Theorem 6.1.1], we only present a brief sketch.

THEOREM 5.1. *Under* (C3)*, there exists a stable Markov policy which is a.s. optimal.*

Proof. Let $v_n \in \Pi_{SM}$ be such that

$$\int \overline{c}\, d\mu[v_n] \downarrow \rho^*.$$

We extend $\mu[v_n]$ to $\boldsymbol{\mathcal{P}}(\overline{\mathbb{R}}^d \times \mathcal{S} \times U)$ in the usual manner and denote it also by $\mu[v_n]$. Let μ_∞ be a limit point of $\{\mu[v_n]\}$ and denote $v_\infty = v_{\mu_\infty}$, where v_{μ_∞} is obtained from μ_∞ by the decomposition in (5.6) and (5.8). Then, for $f \in C_0^2(\mathbb{R}^d \times \mathcal{S})$,

$$\sum_{i \in \mathcal{S}} \int_{\mathbb{R}^d} L^{v_n} f(x, i) \eta_{v_n}(dx, i) = \sum_{i \in \mathcal{S}} \int_{\overline{\mathbb{R}}^d \times U} L^u f(x, i) \mu[v_n](dx, i, du) = 0\,.$$

Hence,

$$\sum_{i \in \mathcal{S}} \int_{\overline{\mathbb{R}}^d \times U} L^u f(x, i) \mu_\infty(dx, i, du) = 0\,.$$

Thus, by Lemmas 5.2 and 5.3, $\mu_\infty^* = \eta_{v_\infty}$, if $\delta_{\mu_\infty} > 0$. Using (C3), we can demonstrate as in [6, Lemma 6.1.3] that this is indeed the case. Therefore,

$$\min_{v \in \Pi_{SM}} \int \overline{c}\, d\mu[v] = \int \overline{c}\, d\mu[v_\infty] = \rho^*.$$

Finally, following the technique in [6, Lemma 6.1.3], we can now show that for an arbitrary policy u,

$$\liminf_{T \to \infty} \frac{1}{T} \int_0^T c\big(X(s), S(s), u(s)\big) ds \geq \rho^* \quad \text{a.s.},$$

which establishes the optimality of v_∞ in a much stronger sense. ∎

THEOREM 5.2. *Under* (C3) *there exists a $v^* \in \Pi_{SMD}$ which is a.s. optimal.*

Proof. We have already established the existence of $v_\infty \in \Pi_{SM}$ which is a.s. optimal. We argue as in [7, p. 58]. Embed I_1 in $\boldsymbol{\mathcal{P}}(\overline{\mathbb{R}}^d \times \mathcal{S} \times U)$ by assigning zero mass at $\{\infty\} \times \mathcal{S} \times U$. Let $\overline{I}_1$ denote the closure of I_1 in $\boldsymbol{\mathcal{P}}(\overline{\mathbb{R}}^d \times \mathcal{S} \times U)$. Then $\overline{I}_1$ is a compact convex set. By Choquet's theorem [24], each element μ of $\overline{I}_1$ is the

barycenter of a probability measure m supported on the set of extreme points of $\overline{I}_1$. Now, each extreme point of I_1 must be an extreme point of $\overline{I}_1$, since otherwise it would be assigning a strictly positive mass to $\{\infty\} \times \mathcal{S} \times U$. If m assigns a strictly positive mass to extreme points of $\overline{I}_1$, which are not extreme points of I_1, then μ must assign a strictly positive probability to $\{\infty\} \times \mathcal{S} \times U$, which is not true. Thus, m must be supported on the set I_1^e consisting of the extreme points of I_1. In particular,

$$\int \overline{c}\, d\mu[v_\infty] = \int_{I_1^e} \left(\int \overline{c}\, d\nu \right) m(d\nu)\,.$$

It follows that there exists a $v^* \in \Pi_{SMD}$ such that

$$\int \overline{c}\, d\mu[v_\infty] = \int \overline{c}\, d\mu[v^*],$$

and since $v_\infty \in \Pi_{SM}$ is optimal, the optimality of $v^* \in \Pi_{SMD}$ follows. □

We now investigate the existence of an optimal Markov policy under the blanket stability conditions in (C1)–(C2).

LEMMA 5.4. *Under either* (C1) *or* (C2)*, for any admissible policy $v \in \Pi$, the empirical process $\{\mu_t(v)\}$ defined in* (5.5) *is tight.*

The proof of Lemma 5.4 closely follows the arguments in the proof of [6, Theorem 6.2.2]. Topologize the space Π_M as in [6], [14]. We now state another result, the proof of which closely follows [14, Theorem 3.3, Lemma 4.4].

LEMMA 5.5. *Under either* (C1) *or* (C2)*, the sets I_1, I_2 are compact in total variation and the map $v \mapsto \mu[v]$ (as defined in* (5.4)*) is continuous.*

THEOREM 5.3. *Under either* (C1) *or* (C2)*, there exists a $v^* \in \Pi_{SMD}$ which is a.s. optimal.*

Proof. First note that under (C1) or (C2), $\Pi_{SM} = \Pi_M$ and $\Pi_{SMD} = \Pi_{SD}$. By Lemma 5.5, there exists a $\overline{v} \in \Pi_{SM}$ such that

$$\min_{v \in \Pi_{SM}} \int \overline{c}\, d\mu[v] = \int \overline{c}\, d\mu[\overline{v}].$$

In view of Lemma 5.4 and the decomposition and disintegration of the measure as defined in (5.6), (5.8), it suffices to confine our attention to Π_{SM} for optimality. Thus, the existence of an a.s. optimal $v^* \in \Pi_{SMD}$ then follows via Choquet's theorem as in Theorem 5.2. □

6. HJB Equations. In this section, we study the HJB equations and characterize the optimal policy in terms of their solution. We introduce the following condition:

(C4) The cost function $\overline{c}$ is bounded, continuous, and Lipschitz in its first argument uniformly with respect to the third.

We follow the vanishing discount approach; i.e., we derive the HJB equations for the ergodic criterion by taking the limit of the HJB equations for the discounted criterion as the discount factor approaches zero. The results and the broad outline of these proofs follow those of [9]. However, they differ in important technical details.

Let $V_\alpha(x,i)$ denote the discounted value function with discount factor $\alpha > 0$; i.e.,

$$V_\alpha(x,i) = \inf_{v \in \Pi} E^v_{x,i}\left[\int_0^\infty e^{-\alpha t} c\big(X(t), S(t), u(t)\big) dt \right], \qquad x \in \mathbb{R}^d, \quad i \in \mathcal{S}\,. \tag{6.1}$$

The following result is proved in [14].

THEOREM 6.1. *Under* (C4), V_α *is the unique solution in* $C^2\big(\mathbb{R}^d\times\mathcal{S}\big)\bigcap C_b\big(\mathbb{R}^d\times\mathcal{S}\big)$ *of*

$$\inf_{u\in U}\big\{L^uV_\alpha(x,i)+\overline{c}(x,i,u)\big\}=\alpha V_\alpha(x,i). \tag{6.2}$$

For $i\in\mathcal{S}$, define

$$G_i:=\big\{x\in\mathbb{R}^d\ :\ \inf_{u\in U}\overline{c}(x,i,u)\le\rho^*\big\},$$
$$G:=\bigcup_{i\in\mathcal{S}}G_i\,. \tag{6.3}$$

Observe that by (C3), G is compact.

The following result plays a very crucial role.

LEMMA 6.1. *Under* (C3) *and* (C4), *there exists* $\alpha_0\in(0,1)$ *such that if* $\alpha\in(0,\alpha_0]$, $\inf_{(x,i)\in\mathbb{R}^d\times\mathcal{S}}V_\alpha(x,i)$ *is attained on the set* G *as defined in* (6.3).

Proof. Let $v_\alpha\in\Pi_{MD}$ be an optimal policy for the discount factor α. By the results of [14], for $i\in\mathcal{S}$,

(6.4)
$$\sum_{k=1}^d\overline{b}_k(x,i,v_\alpha(x,i))\frac{\partial V_\alpha(x,i)}{\partial x_k}+\sum_{j\in\mathcal{S}}\overline{\lambda}_{ij}(x,v_\alpha(x,i))V_\alpha(x,j)+\overline{c}(x,i,v_\alpha(x,i))$$
$$=\inf_{u\in U}\left\{\sum_{k=1}^d\overline{b}_k(x,i,u)\frac{\partial V_\alpha(x,i)}{\partial x_k}+\sum_{j\in\mathcal{S}}\overline{\lambda}_{ij}(x,u)V_\alpha(x,j)+\overline{c}(x,i,u)\right\}\quad\text{a.e.}$$

We let $\|x_n\|\to\infty$ in $\mathbb{R}^d$ and fix $i\in\mathcal{S}$. For given α, let $\big(X^n(\cdot),S^n(\cdot)\big)$ be the process under the policy v_α with $X^n(0)=x_n$ and $S^n(0)=i$. We can show as in [21] that $\{X^n(\cdot)-x_n\}$ are tight as $C\big([0,\infty);\mathbb{R}^d\big)$-valued random variables. Dropping to a subsequence and using Skorohod's theorem [16, p. 9] we may assume that they are defined on a common probability space and converge a.s. in $C\big([0,\infty);\mathbb{R}^d\big)$ to some process $Y(\cdot)$. Hence, $\|X^n(t)\|\to\infty$ uniformly in $t\in[0,T]$ for each $T<\infty$, a.s. By (C3), there exist $\varepsilon>0$ and $M>0$, such that

$$\inf_{u\in U}\big\{\overline{c}(x,i,u)\big\}>\rho^*+2\varepsilon\qquad\text{if}\quad\|x\|>M\,,\quad\forall\ i\in\mathcal{S}\,.$$

We select a constant T_α such that

$$(\rho^*+2\varepsilon)\big(1-e^{-\alpha T_\alpha}\big)>\rho^*+\varepsilon;$$

i.e., $e^{-\alpha T_\alpha}<\frac{\varepsilon}{(\rho^*+2\varepsilon)}$. Since

$$V_\alpha(x_n,i)\ge E^{v_\alpha}_{x_n,i}\left[\int_0^{T_\alpha}e^{-\alpha t}c\big(X^n(t),S^n(t),v_\alpha(X^n(t),S^n(t))\big)dt\right],$$

it follows that

$$V_\alpha(x_n,i)>\frac{\rho^*+\varepsilon}{\alpha} \tag{6.5}$$

for n sufficiently large. On the other hand, by a standard Tauberian theorem,

$$\limsup_{\alpha\to0}\big\{\alpha V_\alpha(x,i)\big\}\le\rho^*\qquad\forall\ (x,i)\in\mathbb{R}^d\times\mathcal{S}\,. \tag{6.6}$$

Fix $x_0 \in \mathbb{R}^d$. By (6.6), there exists $\alpha_0 = \alpha_0(x_0)$ such that $V_\alpha(x_0,\cdot) \le (\rho^* + \frac{\varepsilon}{2})/\alpha$, for all $\alpha \le \alpha_0$. Hence, it follows from (6.5) that if $\alpha \le \alpha_0$, then $\inf_{x\in\mathbb{R}^d} V_\alpha(x,i)$ is attained in a set $\{x \in \mathbb{R}^d : \|x\| \le R(\alpha)\}$ for all $i \in \mathcal{S}$. Let

$$x_{\alpha,i} := \underset{x\in\mathbb{R}^d}{\to} \arg\min\{V_\alpha(x,i)\}, \qquad (x_\alpha, i_\alpha) := \underset{i\in\mathcal{S}}{\to} \arg\min\{V_\alpha(x_{\alpha,i},i)\}.$$

Using (6.2) and the fact that, at a minimum, the gradient of $V_\alpha(\cdot,i)$ vanishes and its Hessian is positive semidefinite, we have, for $\alpha \le \alpha_0$,

$$\inf_{u\in U}\left\{\overline{c}\big(x_{\alpha,i},i,u\big) + \sum_{j\in\mathcal{S}} \overline{\lambda}_{ij}\big(x_{\alpha,i},u\big)V_\alpha\big(x_{\alpha,i},j\big)\right\} \le \alpha V_\alpha(x_{\alpha,i},i). \tag{6.7}$$

In turn, from (6.7),

$$\inf_{u\in U}\big\{\overline{c}(x_\alpha,i_\alpha,u)\big\} \le \alpha V_\alpha(x_\alpha,i_\alpha) \qquad \forall\ \alpha \le \alpha_0\,. \tag{6.8}$$

We claim that $\alpha V_\alpha(x_\alpha,i_\alpha) \le \rho^*$ for all $\alpha > 0$. Indeed, for any $v \in \Pi_{SM}$,

$$V_\alpha(x,i) \le E^v_{x,i}\left[\int_0^\infty e^{-\alpha t} c\big(X(t),S(t),v(X(t),S(t))\big)dt\right] \qquad \forall\ (x,i) \in \mathbb{R}^d \times \mathcal{S}\,. \tag{6.9}$$

Integrating both sides of (6.9) with respect to $\eta_v(dx,i)$ and using Fubini's theorem, we obtain

$$\sum_{i\in\mathcal{S}} \int_{\mathbb{R}^d} V_\alpha(x,i)\eta_v(dx,i) \le \frac{\rho_v}{\alpha}\,.$$

Hence,

$$V_\alpha(x_\alpha,i_\alpha) \le \frac{\rho^*}{\alpha}\,. \tag{6.10}$$

From (6.8),

$$\inf_{u\in U}\big\{c(x_\alpha,i_\alpha,u)\big\} \le \rho^*,$$

concluding that $(x_\alpha,i_\alpha) \in G \times \mathcal{S}$. □

LEMMA 6.2. *Under* (C3) *and* (C4)*, the map* $(x,y,i,j) \mapsto |V_\alpha(x,i) - V_\alpha(y,j)|$ *is bounded on compact subsets, uniformly in* $\alpha \in (0,\alpha_0]$.

Proof. Let $\overline{V}_\alpha(\cdot,\cdot) := V_\alpha(\cdot,\cdot) - V_\alpha(x_\alpha,i_\alpha)$. In view of Lemma 6.1, it suffices to prove that $\overline{V}_\alpha$ is uniformly bounded on compacta. By (6.2) and (6.4),

$$L^{v_\alpha}V_\alpha(x,i) = \alpha V_\alpha(x,i) - \overline{c}\big(x,i,v_\alpha(x,i)\big) \quad \text{a.e.}$$

Let $R > 0$ be large enough so that $G \subset B(0,R)$. Let $\big(X(\cdot),S(\cdot)\big)$ be the process under the policy v_α and define $\tau = \inf\{t \ge 0 : X(t) \notin B(0,2R)\}$. Then for $x \in B(0,R)$, using the strong Markov property,

$$\begin{aligned} V_\alpha(x,i) &= E^{v_\alpha}_{x,i}\left[\int_0^\infty e^{-\alpha t}\overline{c}\big(X(t),S(t),v_\alpha(X(t),S(t))\big)dt\right] \\ &= E^{v_\alpha}_{x,i}\left[\int_0^\tau e^{-\alpha t}\Big\{\overline{c}\big(X(t),S(t),v_\alpha(X(t),S(t))\big) - \alpha V_\alpha\big(X(\tau),S(\tau)\big)\Big\}dt\right] \\ &\qquad + E^{v_\alpha}_{x,i}\Big[V_\alpha\big(X(\tau),S(\tau)\big)\Big]. \end{aligned}$$

Thus,

$$\begin{aligned}&\big|V_\alpha(x,i)-E^{v_\alpha}_{x,i}V_\alpha\big(X(\tau),S(\tau)\big)\big|\\&\qquad=\left|E^{v_\alpha}_{x,i}\int_0^\tau e^{-\alpha t}\Big\{\overline{c}\big(X(t),S(t),v_\alpha(X(t),S(t))\big)-\alpha V_\alpha\big(X(\tau),S(\tau)\big)\Big\}dt\right|.\end{aligned}$$

Using (C4) and Lemma 4.2, we deduce that there exists a constant C_1 (independent of α) such that

$$\big|V_\alpha(x,i)-E^{v_\alpha}_{x,i}V_\alpha\big(X(\tau),S(\tau)\big)\big|\le C_1\qquad\forall\ (x,i)\in B(0,R)\times\mathcal{S}\,. \tag{6.11}$$

We write

$$\begin{aligned}V_\alpha(x,i)-V_\alpha(x_\alpha,i_\alpha)&=\Big(V_\alpha(x,i)-E^{v_\alpha}_{x,i}V_\alpha\big(X(\tau),S(\tau)\big)\Big)\\&+\Big(E^{v_\alpha}_{x,i}V_\alpha\big(X(\tau),S(\tau)\big)-V_\alpha(x_\alpha,i_\alpha)\Big).\end{aligned} \tag{6.12}$$

Let

$$f(x,i)=E^{v_\alpha}_{x,i}V_\alpha\big(X(\tau),S(\tau)\big)-V_\alpha(x_\alpha,i_\alpha).$$

We observe that $f\ge 0$ and $L^{v_\alpha}f=0$ in $W^{2,p}\big(B(0,2R)\times\mathcal{S}\big)$, $2\le p<\infty$. Then, by Theorem 4.1, there exists a constant C_2 (independent of α) such that, in view of (6.11),

$$f(x,i)\le C_2 f(x_\alpha,i_\alpha)\le C_1C_2\qquad\forall\ (x,i)\in B(0,R)\times\mathcal{S}\,.$$

Hence,

$$V_\alpha(x,i)-V_\alpha(x_\alpha,i_\alpha)\le C_1(1+C_2)\qquad\forall\ (x,i)\in B(0,R)\times\mathcal{S}\,.\qquad\square$$

COROLLARY 6.1. *For any $\varepsilon>0$ and any compact $K\subset\mathbb{R}^d$, there exists $\alpha_\varepsilon\in(0,\alpha_0]$ such that for all $x\in K$, $i\in\mathcal{S}$, and $\alpha\in(0,\alpha_\varepsilon)$,*

$$\alpha V_\alpha(x,i)<\rho^*+\varepsilon. \tag{6.13}$$

Proof. The proof follows directly from Lemma 6.2 and (6.10). $\square$

THEOREM 6.2. *Under* (C3) *and* (C4)*, there exists a function $V\in C^2(\mathbb{R}^d\times\mathcal{S})$ and a scalar $\rho\in\mathbb{R}$ such that for some fixed $i_0\in\mathcal{S}$,*

$$\rho\le\rho^*,\qquad V(0,i_0)=0,\qquad\inf_{(x,i)\in\mathbb{R}^d\times\mathcal{S}}V(x,i)>-\infty \tag{6.14}$$

and the pair (V,ρ) satisfies the HJB equations given by

$$\inf_{u\in U}\big\{L^uV(x,i)+\overline{c}(x,i,u)\big\}=\rho. \tag{6.15}$$

Moreover, among all pairs $(\varphi,\rho)\in W^{2,p}_{\ell oc}(\mathbb{R}^d\times\mathcal{S})\times\mathbb{R}$, $2\le p<\infty$, satisfying (6.15)*, (V,ρ^*) is the unique one satisfying* (6.14).

Proof. Set $\overline{V}_\alpha(x,i)=V_\alpha(x,i)-V_\alpha(0,i_0)$. Then $\overline{V}(0,i_0)=0$, and by (6.2), (6.4),

$$L^{v_\alpha}\overline{V}_\alpha(x,i)=\alpha V_\alpha(x,i)-\overline{c}(x,i,v_\alpha(x,i)).$$

By Corollary 6.1, Lemma 6.2, and the interior estimates for solutions of uniformly elliptic systems [22, pp. 398–402], we can show using a standard bootstrap argument that for any $R > 0$, $2 \le p < \infty$,

$$\sup_{\alpha\in(0,\alpha_\varepsilon)} \left\| \overline{V}_\alpha(\cdot,\cdot) \right\|_{W^{2,p}(B(0,R)\times\mathcal{S})} \le C$$

for some constant C. Since $W^{2,p}_{\ell oc} \hookrightarrow W^{1,p}_{\ell oc}$ is compact for $p \ge 1$, $\{\overline{V}_\alpha(\cdot), \alpha \in (0,\alpha_\varepsilon)\}$ is sequentially compact in $W^{1,p}_{\ell oc}$. Let $\alpha_n \to 0$ in $(0,\alpha_\varepsilon)$. By dropping to a subsequence, if necessary, let $\overline{V}_{\alpha_n} \to V$ in $W^{1,p}_{\ell oc}$ for some V. By the Sobolev imbedding theorem, this convergence is also uniform on compact subsets of $\mathbb{R}^d$. Let ρ be a limit point of $\alpha_n V_{\alpha_n}(0,i_0)$ and hence of $\alpha_n V_{\alpha_n}(x,i)$ for any $(x,i) \in \mathbb{R}^d \times \mathcal{S}$, in view of Lemma 6.2. By (6.13), $\rho \le \rho^*$. It can be shown as in [2], [22, p. 420] that

$$\inf_{u\in U} \left\{ \sum_{k=1}^d \overline{b}_k(x,i,u) \frac{\partial \overline{V}_{\alpha_n}(x,i)}{\partial x_k} + \sum_{j\in\mathcal{S}} \overline{\lambda}_{ij}(x,u) \overline{V}_{\alpha_n}(x,j) + \overline{c}(x,i,u) \right\}$$
$$\underset{n\to\infty}{\longrightarrow} \inf_{u\in U} \left\{ \sum_{k=1}^d \overline{b}_k(x,i,u) \frac{\partial V(x,i)}{\partial x_k} + \sum_{j\in\mathcal{S}} \overline{\lambda}_{ij}(x,u) V(x,j) + \overline{c}(x,i,u) \right\}$$

in $L^p_{\ell oc}$ strongly. From the above discussion, it follows that $V \in W^{1,p}_{\ell oc}$, for any $2 \le p < \infty$, and V satisfies (6.15) in D' (i.e., in the sense of distributions). By elliptic regularity, $V \in W^{2,p}_{\ell oc}$, $2 \le p < \infty$. In turn, by the Sobolev imbedding theorem, $V \in C^{1,\gamma}(\mathbb{R}^d \times \mathcal{S})$ for $0 < \gamma < 1$, γ arbitrarily close to 1. Hence by (C4), it is easy to see that

$$\inf_{u\in U} \left\{ \sum_{k=1}^d \overline{b}_k(x,i,u) \frac{\partial V(x,i)}{\partial x_k} + \sum_{j\in\mathcal{S}} \overline{\lambda}_{ij}(x,u) V(x,j) + \overline{c}(x,i,u) \right\}$$

is in $C^{0,\gamma}(\mathbb{R}^d \times \mathcal{S})$. By elliptic regularity [15, p. 287] applied to (6.15), we conclude that $V \in C^{2,\gamma}(\mathbb{R}^d \times \mathcal{S})$. Clearly, $V(0,i_0) = 0$. It suffices to show that V is bounded below. For any $x \in \mathbb{R}^d$, $i \in \mathcal{S}$,

(6.16)
$$V(x,i) = \lim_{n\to\infty} \left[V_{\alpha_n}(x,i) - V_{\alpha_n}(0,i_0) \right]$$
$$\ge \lim_{n\to\infty} \left[V_{\alpha_n}(x_{\alpha_n},i) - V_{\alpha_n}(0,i_0) \right] + \lim_{n\to\infty} \left[V_{\alpha_n}(x_{\alpha_n},i_{\alpha_n}) - V_{\alpha_n}(x_{\alpha_n},i) \right].$$

Using Lemmas 6.1 and 6.2, it follows from (6.16) that for each $i \in \mathcal{S}$,

$$\inf_{(x,i)\in\mathbb{R}^d\times\mathcal{S}} V(x,i) > -\infty,$$

and the proof of the first part of the theorem is complete. The second assertion can be shown by following the methodology in [9]. □

Further, based on Lemmas 6.1 and 6.2 and Theorem 4.1, the following theorem can be proved using the techniques presented in [9]. We therefore skip the proof.

THEOREM 6.3. *Assume* (C3) *and* (C4). *Let* $v^* \in \Pi_{MD}$ *be such that for each* i

(6.17)
$$\inf_{u\in U}\left\{\sum_{k=1}^{d}\overline{b}_k(x,i,u)\frac{\partial V(x,i)}{\partial x_k}+\sum_{j\in\mathcal{S}}\overline{\lambda}_{ij}(x,u)V(x,j)+\overline{c}(x,i,u)\right\}$$
$$=\sum_{k=1}^{d}\overline{b}_k\big(x,i,v^*(x,i)\big)\frac{\partial V(x,i)}{\partial x_k}+\sum_{j\in\mathcal{S}}\overline{\lambda}_{ij}\big(x,v^*(x,i)\big)V(x,j)+\overline{c}\big(x,i,v^*(x,i)\big)\quad a.e.$$

Then $v^* \in \Pi_{SMD}$. *The scalar* ρ *in* (6.15) *equals* ρ^*, *and* v^* *is a.s. optimal. Moreover,* $v \in \Pi_{SMD}$ *is a.s. optimal if and only if it satisfies* (6.17).

Remark 6.1. The boundedness condition on the cost function $\overline{c}$ may be relaxed. For unbounded $\overline{c}$ we can use a suitable truncation procedure to approximate $\overline{c}$ by a sequence of bounded functions. Then the arguments in [9, p. 202] can be paralleled to establish the results in Theorems 6.2–6.3.

We now study the HJB equation under (C1) and (C4). Recall that under (C1), $\Pi_M = \Pi_{SM}$.

LEMMA 6.3. *Let* w *satisfy* (C1). *Then for any* $v \in \Pi_{SM}$,

(i) $\sum_{i\in\mathcal{S}}\int_{\mathbb{R}^d} w(x,i)\eta_v(dx,i) < \infty$,

(ii) $\lim_{t\to\infty}\frac{1}{t}E^v_{x,i}\big[w\big(X(t),S(t)\big)\big]=0$.

Proof. Let $R>0$ and τ_R be the exit time of $X(t)$ from $B(0,R)$. Then by Ito's formula

$$E^v_{x,i}\Big[w\big(X(t\wedge\tau_R),S(t\wedge\tau_R)\big)\Big]-w(x,i)=E^v_{x,i}\left[\int_0^{t\wedge\tau_R}L^v w\big(X(s),S(s)\big)ds\right].$$

Letting $R\to\infty$, we have

$$E^v_{x,i}\big[w\big(X(t),S(t)\big)\big]-w(x,i)=E^v_{x,i}\left[\int_0^{t}L^v w\big(X(s),S(s)\big)ds\right].$$

Therefore, by using (C1), we have

$$\frac{d}{dt}E^v_{x,i}\big[w\big(X(t),S(t)\big)\big]\le p-qE^v_{x,i}\big[w\big(X(t),S(t)\big)\big].$$

Then by Gronwall's inequality,

$$E^v_{x,i}\big[w\big(X(t),S(t)\big)\big]\le\frac{p}{q}+w(x,i)e^{-qt}. \tag{6.18}$$

Both (i) and (ii) follow directly from (6.18). □

LEMMA 6.4. *Assume* (C1) *holds. Let* $a>0$ *be such that*

$$L^u w(x,i)\le -1\qquad \textit{for all}\quad \|x\|>a\,,\ u\in U\,,\ i\in\mathcal{S}\,.$$

If

$$\tau_a:=\inf\big\{t\ge 0\,:\,\|X(t)\|\le a\big\}\,, \tag{6.19}$$

then, for all $v\in\Pi_M$, $\|x\|>a$, *and* $i\in\mathcal{S}$,

$$E^v_{x,i}\big[\tau_a\big]\le w(x,i)\,. \tag{6.20}$$

Proof. Let $v \in \Pi_M$. Choose $R > 0$ such that $a < \|x\| < R$. Let

$$\tau'_R = \inf\{t \geq 0 \,:\, X(t) \notin B(0,R) \setminus B(0,a)\}.$$

Then by Ito's formula

$$E^v_{x,i}\Big[w\big(X(t\wedge\tau'_R), S(t\wedge\tau'_R)\big)\Big] = w(x,i) + E^v_{x,i}\bigg[\int_0^{t\wedge\tau'_R} L^v w\big(X(s),S(s)\big)ds\bigg].$$

Therefore,

$$E^v_{x,i}\Big[w\big(X(t\wedge\tau'_R), S(t\wedge\tau'_R)\big)\Big] \leq w(x,i) - E^v_{x,i}\big[t\wedge\tau'_R\big].$$

Thus,

$$E^v_{x,i}\big[t\wedge\tau'_R\big] \leq w(x,i). \tag{6.21}$$

Letting first $t\to\infty$ and then $R\to\infty$, invoking Fatou's lemma at each step, we obtain (6.20). ∎

THEOREM 6.4. *Under* (C1) *and* (C4)*, the HJB equation* (6.15) *admits a unique solution* (V,ρ) *in the class* $C^2(\mathbb{R}^d\times\mathcal{S})\bigcap\mathcal{O}(w)$*, satisfying* $V(0,i_0)=0$ *for some fixed* $i_0\in\mathcal{S}$.

Proof. Let $v^*\in\Pi_{SMD}$ be a.s. optimal. The existence of such a v^* is guaranteed by Theorem 5.3. Let

$$K_1 = \sup_{x,i,u}\{\overline{c}(x,i,u)\},$$
$$K_2 = \sup_{v\in\Pi_{SMD}}\int \overline{c}\,d\mu[v].$$

We select an arbitrary sequence of smooth functions $\psi_n:\mathbb{R}^d\to[0,K_1+4K_2]$, $n\geq 1$, that are zero on $B(0,n)$ and equal to K_1+4K_2 on the complement of $B(0,n+1)$, and define

$$c_{1n}(x,i,u) = \frac{1}{2}\big[\overline{c}(x,i,u)+\psi_n(x)\big],$$
$$c_{2n}(x,i,u) = \frac{1}{2}\big[\psi_n(x)-\overline{c}(x,i,u)\big].$$

Then, for a sufficiently large n, c_{1n} and c_{2n} both satisfy the penalizing condition (C3). We select one such term of the sequence from now on and drop the subscript n for notational convenience. Let $\big(X(\cdot),S(\cdot)\big)$ be the process under the policy v^*. For $\alpha>0$, we define

$$V_{\alpha,1}(x,i) = E^{v^*}_{x,i}\bigg[\int_0^\infty e^{-\alpha t}c_1\big(X(t),S(t),v^*(X(t),S(t))\big)dt\bigg],$$
$$V_{\alpha,2}(x,i) = E^{v^*}_{x,i}\bigg[\int_0^\infty e^{-\alpha t}c_2\big(X(t),S(t),v^*(X(t),S(t))\big)dt\bigg],$$
$$V_{\alpha}(x,i) = E^{v^*}_{x,i}\bigg[\int_0^\infty e^{-\alpha t}\overline{c}\big(X(t),S(t),v^*(X(t),S(t))\big)dt\bigg].$$

Then we can modify the arguments in the proof of Lemma 6.2 to conclude that for a fixed $i_0 \in \mathcal{S}$, $\big(V_{\alpha,1}(x,i) - V_{\alpha,1}(0,i_0)\big)$ and $\big(V_{\alpha,2}(x,i) - V_{\alpha,2}(0,i_0)\big)$ are bounded on compacta uniformly in $\alpha \in (0,\alpha_0]$, for some $\alpha_0 > 0$. Hence,

$$\begin{aligned}\overline{V}_\alpha(x,i) &:= V_\alpha(x,i) - V_\alpha(0,i_0)\\ &= \big[V_{\alpha,1}(x,i) - V_{\alpha,1}(0,i_0)\big] - \big[V_{\alpha,2}(x,i) - V_{\alpha,2}(0,i_0)\big]\end{aligned}$$

is bounded on compact sets, uniformly in $\alpha \in (0,\alpha_0]$. Arguing as in the proof of Theorem 6.2 we conclude that $\overline{V}_\alpha(x,i) \to V(x,i)$, as $\alpha \to 0$, uniformly on compacta and in $W^{2,p}_{\ell oc}(\mathbb{R}^d \times \mathcal{S})$ for any $p \in [2,\infty)$, and that the limit V satisfies

$$L^{v^*} V(x,i) + \overline{c}(x,i,v^*(x,i)) = \rho^*,$$

with $V(0,i_0) = 0$. Using the strong Markov property, relative to the stopping time τ_a in (6.19), we obtain

$$\begin{aligned}\overline{V}_\alpha(x,i) = E^{v^*}_{x,i}\bigg[\int_0^{\tau_a} e^{-\alpha t}\Big\{\overline{c}\big(X(t),S(t),v^*(X(t),S(t))\big) - \alpha V_\alpha(0,i_0)\Big\}dt\bigg]\\ + E^{v^*}_{x,i}\Big[e^{-\alpha\tau_a}\overline{V}_\alpha\big(X(\tau_a),S(\tau_a)\big)\Big].\end{aligned}$$

Hence, by Lemma 6.4, for $\alpha \in (0,\alpha_0]$ and $\|x\| > a$,

$$\begin{aligned}\big|\overline{V}_\alpha(x,i)\big| &\le C_1 + C_2 E^{v^*}_{x,i}[\tau_a]\\ &\le C_1 + C_2 w(x,i),\end{aligned}$$

where C_1, C_2 are positive constants independent of α. Passing to the limit as $\alpha \to 0$, it follows that V is in the class $\mathcal{O}(w)$. Next we let $v \in \Pi_{SMD}$ be such that for each $i \in \mathcal{S}$,

$$\begin{aligned}&\sum_{k=1}^d \overline{b}_k(x,i,v(x,i))\frac{\partial V(x,i)}{\partial x_k} + \sum_{j\in\mathcal{S}} \overline{\lambda}_{ij}(x,v(x,i))V(x,j) + \overline{c}(x,i,v(x,i))\\ &\qquad = \inf_{u\in U}\left\{\sum_{k=1}^d \overline{b}_k(x,i,u)\frac{\partial V(x,i)}{\partial x_k} + \sum_{j\in\mathcal{S}} \overline{\lambda}_{ij}(x,u)V(x,j) + \overline{c}(x,i,u)\right\} \quad \text{a.e.}\end{aligned}$$

Suppose that for some $i' \in \mathcal{S}$, there exist $\delta > 0$ such that the set

$$D = \Big\{x \in \mathbb{R}^d \,:\, L^v V(x,i') \le \rho^* - \overline{c}\big(x,i',v(x,i')\big) - \delta\Big\}$$

has positive Lebesgue measure. By Ito's formula

$$E^v_{x,i}\big[V\big(X(t),S(t)\big)\big] - V(x,i) = E^v_{x,i}\bigg[\int_0^t L^v V\big(X(s),S(s)\big)ds\bigg].$$

This is justified because V is $\mathcal{O}(w)$. Therefore,

$$\begin{aligned}E^v_{x,i}\big[V\big(X(t),S(t)\big)\big] - V(x,i) \le E^v_{x,i}\bigg[\int_0^t \Big[\rho^* - \overline{c}\big(X(s),S(s),v(X(s),S(s))\big)\Big]ds\bigg]\\ - \delta E^v_{x,i}\bigg[\int_0^t I\big\{X(s)\in D, S(s) = i'\big\}ds\bigg].\end{aligned}$$

Dividing by t, letting $t \to \infty$, and using Lemma 6.3, we have

$$\rho_v \le \rho^* - \delta\eta_v\big(D \times \{i'\}\big).$$

Lemma 4.1 implies that η_v is mutually absolutely continuous with respect to the Lebesgue measure. Therefore, $\eta_v\big(D \times \{i'\}\big) > 0$. Hence, $\rho_v < \rho^*$, which contradicts the optimality of v^*. Thus, for each $i \in \mathcal{S}$,

$$\inf_{u \in U} \big\{ L^u V(x,i) + \overline{c}(x,i,u) \big\} = \rho^* \quad \text{a.e.} \tag{6.22}$$

Similar arguments as in the proof of Theorem 6.2 establish that $V \in C^{2,\gamma}(\mathbb{R}^d \times \mathcal{S})$, where $0 < \gamma < 1$, γ arbitrarily close to 1. We now proceed to show uniqueness. Let (V', ρ') be another solution of (6.15) in the desired class satisfying $V'(0, i_0) = 0$. Using Ito's formula and Lemma 6.3, it again follows that $\rho' = \rho^*$. Therefore,

$$L^{v^*}\big(V'(x,i) - V(x,i)\big) \ge 0.$$

Let $\big(X(t), S(t)\big)$ be the process governed by v^* and with initial law η_{v^*}. Then,

$$M(t) := V'\big(X(t), S(t)\big) - V\big(X(t), S(t)\big)$$

is a submartingale satisfying

$$\sup_{t \ge 0} E^{v^*} |M(t)| \le C_1' + C_2' \sum_{i \in \mathcal{S}} \int_{\mathbb{R}^d} w(x,i) \eta_{v^*}(dx, i) < \infty,$$

by Lemma 6.3, where C_1', C_2' are suitable constants. Here we are using the fact that both V and V' are of $\mathcal{O}(w)$. By the submartingale convergence theorem, $M(t)$ converges a.s. Since $\big(X(t), S(t)\big)$ is ergodic and irreducible under v^*, it follows that $V'(x,i) - V(x,i)$ must be constant a.s. This constant must be zero, since $V'(0, i_0) - V(0, i_0) = 0$. □

Remark 6.2. For the stable case we have carried out our analysis under the Lyapunov condition (C1). Analogous results can be derived under the condition (C2).

7. Conclusions. We have analyzed the optimal control of switching diffusions with a pathwise average cost criterion. Under certain conditions we have established the existence of a stable, nonrandomized Markov policy which is a.s. optimal in the class of all admissible policies. Also, we demonstrate the existence of a unique solution to the associated HJB equations in C^2, under varying conditions, and the optimal policy is characterized as a minimizing selector of the Hamiltonian. We have applied our results to a manufacturing model of Bielecki and Kumar and have shown that our methodology affords both greater generality and ease of solution. By studying the recurrence and ergodic properties of switching diffusions we have also obtained two new results in partial differential equations, viz. a strong maximum principle and Harnack's inequality for a weakly coupled elliptic system.

Appendix. This appendix is devoted to the proof of Theorem 4.1.

Given a domain $\Omega \subset \mathbb{R}^d$, a real function $\boldsymbol{u}$ defined on $\Omega \times \mathcal{S}$ is viewed as a vector-valued function $\boldsymbol{u} = (u_1, \dots, u_N)$, with each component u_i being a real function on Ω.

Consider a second order operator L defined by (note that L_k is different from the operator in (3.7))

$$(L\boldsymbol{u})_k(x) := L_k u_k(x) + \sum_{\substack{j\in\mathcal{S}\\ j\neq k}} c_{kj}(x)u_j(x)\,, \qquad k\in\mathcal{S},$$

$$L_k := \sum_{i,j=1}^{d} a_{ij}^k(x)\frac{\partial^2}{\partial x_i\partial x_j} + \sum_{i=1}^{d} b_i^k(x)\frac{\partial}{\partial x_i} + c_{kk}(x)\,. \tag{A.1}$$

Let $m, \overline{m}, \overline{\gamma}$, and ε_Ω be given positive constants, the last depending on the choice of a bounded domain Ω. We denote by $\mathfrak{L} = \mathfrak{L}(m, \overline{m}, \overline{\gamma}, \varepsilon_\Omega)$ the class of all such operators L, with coefficients $a_{ij}^k(\cdot) \in C^{0,1}(\mathbb{R}^d)$ and $b_i^k(\cdot), c_{kj}(\cdot) \in L^\infty(\mathbb{R}^d)$, satisfying

(A.2) $m\|\zeta\|^2 \le \sum_{i,j=1}^{d} a_{ij}^k(x)\zeta_i\zeta_j \le \overline{m}\|\zeta\|^2$ for all $x, \zeta \in \mathbb{R}^d$, $k \in \mathcal{S}$.

(A.3) $\|b_i^k\|_\infty \le \overline{m}$, $\|c_{k\ell}\|_\infty \le \overline{m}$ and $\|a_{ij}^k(x) - a_{ij}^k(y)\|_\infty \le \overline{\gamma}\|x-y\|$ for all $x, y \in \mathbb{R}^d$, $i, j \in \{1, \dots, d\}$, $k, \ell \in \mathcal{S}$.

(A.4) $\sum_{i\in\mathcal{S}} c_{ki}(\cdot) = 0$ and $c_{kj} \ge 0$, for $j \neq k$.

(A.5) The matrix $\boldsymbol{C}(x;\varepsilon_\Omega) := \big[c_{ij}(x) \,:\, c_{ij}(x) \ge \varepsilon_\Omega, i \neq j\big]$ is irreducible at each $x \in \Omega$.

We denote by $\mathfrak{U}_\Omega$ the class of all nonnegative functions $\boldsymbol{u} \in W^{2,d}_{\ell oc}(\Omega\times\mathcal{S}) \bigcap C^0(\overline{\Omega}\times\mathcal{S})$, satisfying $L\boldsymbol{u} = 0$ in Ω, for some $L \in \mathfrak{L}$. If $\xi \in \mathbb{R}$, then $\boldsymbol{u} \ge \xi$ is to be interpreted as $u_i \ge \xi$ for all $i \in \mathcal{S}$, and if $\boldsymbol{\xi} = (\xi_1, \dots, \xi_N) \in \mathbb{R}^N$, then $\boldsymbol{u} \ge \boldsymbol{\xi} \iff u_i \ge \xi_i$ for all $i \in \mathcal{S}$. For better clarity, we denote all $\mathbb{R}^N$-valued quantities by a bold letter. Also, operations such as "inf" on $\mathbb{R}^N$-valued functions are meant to be componentwise. If Γ is a closed subset of Ω, we define, for $x \in \Omega$ and $\boldsymbol{\xi} \in \mathbb{R}^N_+$,

$$\boldsymbol{\Psi}_x\big(\mathfrak{U}_\Omega, \Gamma; \boldsymbol{\xi}\big) := \inf_{\boldsymbol{u}\in\mathfrak{U}_\Omega} \big\{\boldsymbol{u}(x) \,:\, \boldsymbol{u} \ge \boldsymbol{\xi} \text{ on } \Gamma\big\}\,.$$

Deviating from the usual vector space notation, if D is a cube in $\mathbb{R}^d$ and $\delta > 0$, δD denotes the cube which is *concentric* to D and whose edges are δ times as long. For a measurable set $A \subset \mathbb{R}^d$, $|A|$ denotes the Lebesgue measure of A, while $\mathcal{B}(A)$ and $L^d(A)$ denote the sets of real-valued, measurable functions on A such that

$$\|f\|_{\mathcal{B}(A)} := \operatorname*{ess\,sup}_{x\in A} |f(x)| < \infty \qquad \forall\ f \in \mathcal{B}(A)$$

and

$$\|f\|_{d;A} := \left(\int_A |f(x)|^d dx\right)^{1/d} < \infty \qquad \forall\ f \in L^d(A)\,.$$

We use quite frequently the following comparison principle, which can be viewed as a weaker version of the maximum principle in that it holds even without condition (A.5): If $\boldsymbol{\varphi}, \boldsymbol{\psi} \in W^{2,d}_{\ell oc}(\Omega\times\mathcal{S}) \bigcap C^0(\overline{\Omega}\times\mathcal{S})$ satisfy $L\boldsymbol{\varphi} \le L\boldsymbol{\psi}$ in Ω and $\boldsymbol{\varphi} \ge \boldsymbol{\psi}$ on $\partial\Omega$, then $\boldsymbol{\varphi} \ge \boldsymbol{\psi}$ in $\overline{\Omega}$. The same comparison principle holds for $\varphi, \psi \in W^{2,d}_{\ell oc}(\Omega) \bigcap C^0(\overline{\Omega})$ relative to the set of operators $\{L_k\}_{k\in\mathcal{S}}$ as defined in (A.1).

We start with a measure-theoretic result, announced in [19].

LEMMA A.1. *Let $K \subset \mathbb{R}^d$ be a cube, $\Gamma \subset K$ be a closed subset, and $0 < \alpha < 1$. Define*

$$\boldsymbol{\mathcal{Q}} := \big\{Q \,:\, Q \text{ is a subcube of } K \text{ and } |Q \bigcap \Gamma| \geq \alpha |Q|\big\},$$
$$\widetilde{\Gamma} := \bigcup_{Q \in \boldsymbol{\mathcal{Q}}} (3Q \bigcap K)\,.$$

Then either $\widetilde{\Gamma} = K$ or $|\widetilde{\Gamma}| \geq \frac{1}{\alpha}|\Gamma|$.

Proof. If $|\Gamma| \geq \alpha |K|$, then $K \in \boldsymbol{\mathcal{Q}}$ and $\widetilde{\Gamma} = K$. So we assume $|\Gamma| < \alpha |K|$ or, equivalently, $K \notin \boldsymbol{\mathcal{Q}}$. We subdivide K into 2^d congruent subcubes with disjoint interiors. We select the ones in $\boldsymbol{\mathcal{Q}}$, while the remaining ones are similarly subdivided and the process is repeated indefinitely. Let $\boldsymbol{\mathcal{Q}}_0$ be the collection thus obtained, and with $\widehat{Q}$ denoting the ancestor of Q, we define

$$\widehat{\Gamma} := \bigcup_{Q \in \boldsymbol{\mathcal{Q}}_0} \widehat{Q}.$$

Clearly, $\widehat{Q} \subset 3Q \bigcap K$; hence, $\widetilde{\Gamma} \supset \widehat{\Gamma}$. Note that, discarding repetitions, $\widehat{\Gamma}$ can be represented as a disjoint union of cubes $\widehat{Q}$ which are not in $\boldsymbol{\mathcal{Q}}$. Therefore, each member $\widehat{Q}$ of this union satisfies $|\widehat{Q} \bigcap \Gamma| < \alpha |\widehat{Q}|$, and by σ-additivity, we obtain

$$|\widehat{\Gamma} \bigcap \Gamma| < \alpha |\widehat{\Gamma}| \leq \alpha |\widetilde{\Gamma}|\,.$$

By the regularity properties of the Lebesgue measure, $|\widehat{\Gamma} \bigcap \Gamma| = |\Gamma|$ and the proof is complete. ∎

Next we state without proof a ramification of the weak maximum principle of A. D. Aleksandroff.

LEMMA A.2. *There exist constants $C_1 > 0$ and $\kappa_0 \in (0, 1]$ such that if $D \subset \mathbb{R}^d$ is any cube of volume $|D| \leq \kappa_0$ and $\varphi \in W^{2,d}_{\ell oc}(D) \bigcap C^0(\overline{D})$, $f \in L^d(D)$ satisfy $L_k \varphi \geq f$ in D, and $\varphi = 0$ on ∂D for some $L \in \mathfrak{L}$, then*

$$\sup_{x \in D} \big\{\varphi(x)\big\} \leq C_1 |D|^{1/d} \|f\|_{d;D}\,.$$

For the remainder of this appendix, D will denote an open cube in $\mathbb{R}^d$ of volume not exceeding the constant κ_0 in Lemma A.2.

LEMMA A.3. *There exist constants $\beta_0 > 0$ and $\alpha_0 < 1$ such that, if Γ is a closed subset of some cube $D \subset \mathbb{R}^d$, satisfying $|\Gamma| \geq \alpha_0 |D|$, then*

$$\inf_{x \in \frac{1}{3} D} \boldsymbol{\Psi}_x\big(\mathfrak{U}_D, \Gamma; \boldsymbol{\xi}\big) \geq \beta_0 \boldsymbol{\xi} \qquad \forall\, \boldsymbol{\xi} \in \mathbb{R}^N_+\,.$$

Proof. Observe that if $\boldsymbol{u} \in \mathfrak{U}_D$, then each component u_k satisfies $L_k u_k \leq 0$ in D. Define $\varphi', \varphi'' \in W^{2,d}_{\ell oc}(D) \bigcap C^0(\overline{D})$ by

$$L_k \varphi'(x) = -I_\Gamma(x), \quad L_k \varphi''(x) = -I_{\Gamma^c}(x) \qquad \text{in } \ D$$
$$\text{and} \quad \varphi'(x) = \varphi''(x) = 0 \qquad \text{on } \ \partial D\,.$$

Then $\varphi := \varphi' + \varphi''$ satisfies $L_k \varphi = -1$ in D and $\varphi = 0$ on ∂D. Without loss of generality, suppose that D is centered at the origin and consider the function

$$\psi(x) := \prod_{i=1}^{d} \big(|D|^{2/d} - 4x_i^2\big).$$

Note that $\psi = 0$ on ∂D and $\psi > 0$ in D. In addition, there exists a positive constant C_2 such that

$$\inf_{x\in\frac{1}{3}D}\{\psi(x)\} \ge C_2|D|^{2/d}\|L_k\psi\|_{\mathcal{B}(D)} \qquad \forall\, L\in\mathfrak{L}\,.$$

Therefore, by the comparison principle,

$$\varphi(x) \ge \frac{\psi(x)}{\|L_k\psi\|_{\mathcal{B}(D)}} \ge C_2|D|^{2/d} \qquad \forall\, x\in\tfrac{1}{3}D\,. \tag{A.6}$$

Using Lemma A.2, we obtain

$$\begin{aligned} \varphi' &\le C_1|D|^{1/d}|\Gamma|^{1/d} = C_1|D|^{2/d}\Big(\tfrac{|\Gamma|}{|D|}\Big)^{1/d},\\ \varphi'' &\le C_1|D|^{1/d}|\Gamma^c|^{1/d} = C_1|D|^{2/d}\Big(1-\tfrac{|\Gamma|}{|D|}\Big)^{1/d}. \end{aligned} \tag{A.7}$$

By (A.6) and (A.7),

$$\varphi'(x) \ge C_2|D|^{2/d} - C_1|D|^{2/d}\Big(1-\tfrac{|\Gamma|}{|D|}\Big)^{1/d} \qquad \forall\, x\in\tfrac{1}{3}D\,.$$

On the other hand, since $L_k\varphi' = 0$ in $D\setminus\Gamma$ and $\varphi' = 0$ on ∂D, the comparison principle yields

$$\inf_{x\in\frac{1}{3}D}\{u_k(x)\} \ge \xi_k\frac{C_2 - C_1\Big(1-\frac{|\Gamma|}{|D|}\Big)^{1/d}}{C_1\Big(\frac{|\Gamma|}{|D|}\Big)^{1/d}}\,. \tag{A.8}$$

Selecting α_0 to satisfy

$$\alpha_0 \ge 1 - \Big(\frac{C_2}{2C_1}\Big)^d,$$

(A.8) yields

$$\inf_{x\in\frac{1}{3}D}\{u_k(x)\} \ge \frac{C_2\xi_k}{2C_1}.$$

Hence, the claim follows with $\beta_0 = \frac{C_2}{2C_1}$. □

LEMMA A.4. *For each $\delta > 0$, there exists a constant $k'_\delta > 0$ such that if $Q \subset (1-\delta)D$ is a subcube of an open cube $D\subset\mathbb{R}^d$, then*

$$\boldsymbol{\Psi}_x\big(\mathfrak{U}_D, \tfrac{1}{3}Q;\boldsymbol{\xi}\big) \ge k'_\delta\boldsymbol{\xi} \qquad \forall\, x\in 3Q\bigcap(1-\delta)D \quad \forall\,\boldsymbol{\xi}\in\mathbb{R}^N_+\,.$$

Proof. Let $B(r)\subset\mathbb{R}^d$ denote the ball of radius r centered at the origin. We claim that there exists a constant $m_0 > 0$ such that if $r\le 1$, then

$$\inf_{x\in B(\frac{3r}{4})}\boldsymbol{\Psi}_x\big(\mathfrak{U}_{B(r)}, B\big(\tfrac{r}{4}\big);\boldsymbol{\xi}\big) \ge m_0\,\boldsymbol{\xi} \qquad \forall\,\boldsymbol{\xi}\in\mathbb{R}^N_+\,. \tag{A.9}$$

In order to establish (A.9) we use the function

$$\varphi(x) := \exp\Big\{a\Big(1-\tfrac{\|x\|^2}{r^2}\Big)\Big\} - 1\,, \qquad a := \tfrac{\overline{m}}{m}(16d+2)\,, \qquad x\in B(r)\,,$$

which satisfies $L_k\varphi(x) \ge 0$ for all $L \in \mathfrak{L}$, provided $\|x\| \ge \frac{r}{4}$ and $r \le 1$. By the comparison principle, (A.9) holds with

$$m_0 = \frac{e^{\frac{7a}{16}} - 1}{e^{\frac{15a}{16}} - 1}\,.$$

It follows that if $B(r,y)$ is a ball of radius r centered at y, and x is an arbitrary point in D such that the distance between ∂D and the line segment joining x and y is at least r, then

$$\boldsymbol{\Psi}_x\big(\mathfrak{U}_D, B\big(\tfrac{r}{4}, y\big); \boldsymbol{\xi}\big) \ge (m_0)^{\ell}\, \boldsymbol{\xi}\,, \qquad \text{with} \quad \ell = \left\lceil \tfrac{4\|x-y\|-r}{2r} \right\rceil \qquad \forall\, \boldsymbol{\xi} \in \mathbb{R}^N_+\,. \tag{A.10}$$

Choosing $r = \min\big\{\frac{2}{3}, \frac{\delta}{2}\big\}|Q|^{1/d}$ and applying (A.10), an easy calculation shows that the result holds with

$$k'_\delta := m_0^{\ell(\delta)}\,, \qquad \ell(\delta) := \left\lceil \frac{6\sqrt{d}}{\min\{1,\delta\}} \right\rceil. \qquad \square$$

LEMMA A.5. *Suppose that there exist constants ε and θ such that if $\Gamma \subset (1-\delta)D$ is a closed subset of some cube D and $\boldsymbol{\xi} \in \mathbb{R}^N_+$, then*

$$\inf_{x \in \frac{1}{3}D} \boldsymbol{\Psi}_x\big(\mathfrak{U}_D, \Gamma; \boldsymbol{\xi}\big) \ge \varepsilon \boldsymbol{\xi} \quad \textit{whenever} \quad |\Gamma| \ge \theta |D|\,.$$

Then there exists a constant $k_\delta > 0$ such that

$$\inf_{x \in \frac{1}{3}D} \boldsymbol{\Psi}_x\big(\mathfrak{U}_D, \Gamma; \boldsymbol{\xi}\big) \ge \varepsilon k_\delta \boldsymbol{\xi} \quad \textit{whenever} \quad |\Gamma| \ge \alpha_0 \theta |D|\,,$$

where α_0 is the constant in Lemma A.3.

Proof. Suppose $|\Gamma| \ge \alpha_0\theta|D|$ and let $y \in \widetilde{\Gamma}$, with $\widetilde{\Gamma}$ as defined in Lemma A.1 corresponding to $\alpha = \alpha_0$ and $K = (1-\delta)D$. Then there exists a subcube $Q \subset K$ such that $|\Gamma \bigcap Q| \ge \alpha_0|Q|$ and $y \in 3Q \bigcap K$. We use the identities

$$\boldsymbol{\Psi}_x\big(\mathfrak{U}_D, \Gamma; \boldsymbol{\xi}\big) \ge \boldsymbol{\Psi}_x\Big(\mathfrak{U}_D, \widetilde{\Gamma};\ \inf_{y \in \widetilde{\Gamma}} \boldsymbol{\Psi}_y\big(\mathfrak{U}_D, \Gamma; \boldsymbol{\xi}\big)\Big) \tag{A.11}$$

and

$$\begin{aligned} \boldsymbol{\Psi}_y\big(\mathfrak{U}_D, \Gamma; \boldsymbol{\xi}\big) &\ge \boldsymbol{\Psi}_y\Big(\mathfrak{U}_D, \tfrac{1}{3}Q;\ \inf_{z \in \frac{1}{3}Q} \boldsymbol{\Psi}_z\big(\mathfrak{U}_D, \Gamma; \boldsymbol{\xi}\big)\Big) \\ &\ge \boldsymbol{\Psi}_y\Big(\mathfrak{U}_D, \tfrac{1}{3}Q;\ \inf_{z \in \frac{1}{3}Q} \boldsymbol{\Psi}_z\big(\mathfrak{U}_Q, \Gamma \bigcap Q; \boldsymbol{\xi}\big)\Big)\,. \end{aligned} \tag{A.12}$$

From Lemma A.3, we have

$$\inf_{z \in \frac{1}{3}Q} \boldsymbol{\Psi}_z\big(\mathfrak{U}_Q, \Gamma \bigcap Q; \boldsymbol{\xi}\big) \ge \beta_0 \boldsymbol{\xi}\,. \tag{A.13}$$

From Lemma A.4, we obtain $\boldsymbol{\Psi}_y\big(\mathfrak{U}_D, \frac{1}{3}Q; \beta_0\boldsymbol{\xi}\big) \ge \beta_0 k'_\delta \boldsymbol{\xi}$, for all $y \in 3Q \bigcap K$. Hence, combining (A.12) and (A.13) yields

$$\inf_{y \in \widetilde{\Gamma}} \boldsymbol{\Psi}_y\big(\mathfrak{U}_D, \Gamma; \boldsymbol{\xi}\big) \ge k_\delta \boldsymbol{\xi}\,, \qquad \text{with} \quad k_\delta := \beta_0 k'_\delta\,. \tag{A.14}$$

From Lemma A.1, $|\widetilde{\Gamma}| \ge \frac{1}{\alpha_0}|\Gamma| \ge \theta|D|$. Therefore, by hypothesis,

$$\inf_{x\in\frac{1}{3}D} \boldsymbol{\Psi}_x\big(\mathfrak{U}_D, \widetilde{\Gamma}; k_\delta\boldsymbol{\xi}\big) \ge \varepsilon k_\delta\boldsymbol{\xi}\,,$$

which along with (A.11) and (A.14) yield the desired result. ∎

THEOREM A.1. *The following estimates hold.*

(i) *Let D be a cube and $\Gamma \subset (1-\delta)D$ a closed subset. Then for all $\boldsymbol{\xi} \in \mathbb{R}^N_+$,*

$$\inf_{x\in\frac{1}{3}D} \boldsymbol{\Psi}_x\big(\mathfrak{U}_D, \Gamma; \boldsymbol{\xi}\big) \ge \beta_0 \left(\tfrac{|\Gamma|}{|D|}\right)^{\rho(\delta)} \boldsymbol{\xi}\,, \qquad \rho(\delta) := \frac{\log k_\delta}{\log \alpha_0}, \tag{A.15}$$

where the constants α_0, β_0, and k_δ are as in Lemmas A.3 *and* A.5.

(ii) *There exists a real function F defined in $[0,1]$, with $F(\theta) > 0$ if $\theta > 0$, such that if $\Gamma \subset D$ is a closed subset of a cube D, then*

$$\inf_{x\in\frac{1}{3}D} \boldsymbol{\Psi}_x\big(\mathfrak{U}_D, \Gamma; \boldsymbol{\xi}\big) \ge F\left(\tfrac{|\Gamma|}{|D|}\right) \boldsymbol{\xi} \qquad \forall\ \boldsymbol{\xi} \in \mathbb{R}^N_+\,. \tag{A.16}$$

Proof. Part (i) is a direct consequence of Lemmas A.3 and A.5. For part (ii), choose $\delta = \frac{|\Gamma|}{4d|D|}$. Then,

$$\frac{|\Gamma \bigcap (1-\delta)D|}{|D|} \ge \frac{|\Gamma|}{|D|} - \big(1-(1-\delta)^d\big) \ge \frac{|\Gamma|}{|D|} - d\delta \ge \frac{3|\Gamma|}{4|D|}\,. \tag{A.17}$$

Since

$$\boldsymbol{\Psi}_x\big(\mathfrak{U}_D, \Gamma; \boldsymbol{\xi}\big) \ge \boldsymbol{\Psi}_x\big(\mathfrak{U}_D, \Gamma \bigcap (1-\delta)D; \boldsymbol{\xi}\big)\,,$$

the bound in (A.16) follows from (A.15) and (A.17), with

$$F(\theta) := \beta_0 \left(\tfrac{3\theta}{4}\right)^{\rho(\frac{\theta}{4d})}\,. \qquad \blacksquare$$

Definition A.1. If $A \subset \Omega$ we define the oscillation of a function $\boldsymbol{u} \in C^0(\overline{\Omega} \times \mathcal{S})$ over A by

$$\operatorname{osc}(\boldsymbol{u}; A) = \max_{k\in\mathcal{S}} \sup_{x\in A}\big\{u_k(x)\big\} - \min_{k\in\mathcal{S}} \inf_{x\in A}\big\{u_k(x)\big\}\,.$$

The oscillation of a function in $C^0(\overline{\Omega})$ is defined in the usual manner.

THEOREM A.2. *If D is a cube, $\boldsymbol{u} \in \mathfrak{U}_D$ and $q = F\big(\frac{1}{2}\big)$, with $F(\cdot)$ as defined in Theorem* A.1 (ii)*, then*

$$\operatorname{osc}(u_k; \tfrac{1}{3}D) \le \big(1 - \tfrac{q}{2}\big) \operatorname{osc}(\boldsymbol{u}; D) \qquad \forall\ k \in \mathcal{S}\,.$$

Proof. Let

$$M^a_k := \sup_{x\in\frac{1}{3}D}\big\{u_k(x)\big\}\,, \qquad M^a := \max_{k\in\mathcal{S}} M^a_k,$$
$$m^a_k := \inf_{x\in\frac{1}{3}D}\big\{u_k(x)\big\}\,, \qquad m^a := \min_{k\in\mathcal{S}} m^a_k,$$

and M^b, m^b be the corresponding quantities relative to D. Consider the sets

$$\Gamma_1^{(k)} := \Big\{ x \in D \,:\, u_k(x) \le \tfrac{M^b+m^b}{2} \Big\},$$
$$\Gamma_2^{(k)} := \Big\{ x \in D \,:\, u_k(x) \ge \tfrac{M^b+m^b}{2} \Big\}.$$

Suppose $|\Gamma_2^{(k)}| \ge \frac{1}{2}|D|$. Since $\boldsymbol{u} - m^b$ is nonnegative and $u_k - m^b \ge \frac{M^b-m^b}{2}$ in $\Gamma_2^{(k)}$, applying Theorem A.1 (ii) yields

$$u_k(x) - m^b \ge q\tfrac{M^b-m^b}{2} \qquad \forall\, x \in \tfrac{1}{3}D\,.$$

Consequently, $m_k^a \ge m^b + q\frac{M^b-m^b}{2}$, and since $M^a \le M^b$, we obtain

$$M^a - m_k^a \le M^b - m^b - q\tfrac{M^b-m^b}{2} \le \big(1-\tfrac{q}{2}\big)(M^b-m^b)\,. \tag{A.18}$$

On the other hand, if $|\Gamma_1^{(k)}| \ge \frac{1}{2}|D|$, then using the nonnegative function $M^b - \boldsymbol{u}$, we similarly obtain

$$M_k^a - m^a \le \big(1-\tfrac{q}{2}\big)(M^b-m^b)\,, \tag{A.19}$$

and the result follows by (A.18)–(A.19). □

THEOREM A.3. *There exists a constant $M_1 > 0$ such that, for any $\boldsymbol{u} \in \mathfrak{U}_D$,*

$$\sup_{x\in\frac{1}{9}D} \big\{u_i(x)\big\} \le M_1 \max_{k\in\mathcal{S}} \inf_{x\in\frac{1}{9}D} \big\{u_k(x)\big\} \qquad \forall\, i \in \mathcal{S}\,.$$

Proof. Let β_0 be as given in Lemma A.3, and with $\rho(\cdot)$ and q as in (A.15) and Theorem A.2, respectively, define

$$\rho := \frac{1}{d\rho(\frac{2}{3})} \qquad \text{and} \qquad q_0 := \frac{(1-\frac{q}{4})}{(1-\frac{q}{2})}\,. \tag{A.20}$$

We claim that the value of the constant M_1 may be chosen as

$$M_1 := \frac{4q_0}{q\beta_0}\left[\frac{27N^{1/d}}{2\big(q_0^\rho-1\big)}\right]^{1/\rho}. \tag{A.21}$$

We argue by contradiction. Suppose $\boldsymbol{u} \in \mathfrak{U}_D$ violates this bound. Let $\big\{x^{(1)},\dots,x^{(N)}\big\}$ denote the points in $\frac{1}{9}\overline{D}$ where the minima of $\boldsymbol{u}$ are attained; i.e.,

$$\inf_{x\in\frac{1}{9}D} \big\{u_k(x)\big\} = u_k(x^{(k)})\,, \qquad k \in \mathcal{S}\,.$$

Without loss of generality, suppose that $\max_{k\in\mathcal{S}}\big\{u_k(x^{(k)})\big\} = 1$ ($\boldsymbol{u}$ can always be scaled to satisfy this) and that for some $y_0 \in \frac{1}{9}D$ and $k_0 \in \mathcal{S}$, $u_{k_0}(y_0) = M > aM_1$ with $a > 1$. Using the estimate for the growth of the oscillation of $\boldsymbol{u}$ in Theorem A.2, we will show that $\boldsymbol{u}$ has to be unbounded in $\frac{1}{3}D$. By hypothesis, $\frac{M}{a}$ exceeds M_1 in (A.21), and in order to facilitate the construction that follows, we choose to express this as

$$\tfrac{1}{9} + 3N^{1/d}\big(\tfrac{4a}{q\beta_0 M}\big)^\rho \sum_{n=0}^{\infty} \big(\tfrac{1}{q_0}\big)^{n\rho} < \tfrac{1}{3}\,. \tag{A.22}$$

For $\xi > 0$, define

$$\mathcal{D}_k^{(\xi)} := \big\{ z \in \tfrac{1}{3}\overline{D} \,:\, u_k(z) \ge \xi \big\}\,, \qquad \mathcal{D}^{(\xi)} := \bigcup_{k\in\mathcal{S}} \mathcal{D}_k^{(\xi)}\,.$$

If $\mathbf{1}_k \in \mathbb{R}^N_+$ stands for the vector whose kth component is equal to 1 and the others 0, then

$$\boldsymbol{u}(x^{(k)}) \ge \boldsymbol{\Psi}_{x^{(k)}}\big(\mathfrak{U}_D, \mathcal{D}_k^{(\xi)}; \xi\mathbf{1}_k\big) \qquad \forall\, k \in \mathcal{S}\,, \tag{A.23}$$

while, on the other hand, Theorem A.1 yields,

$$\boldsymbol{\Psi}_{x^{(k)}}\big(\mathfrak{U}_D, \mathcal{D}_k^{(\xi)}; \xi\mathbf{1}_k\big) \ge \beta_0 \left(\frac{|\mathcal{D}_k^{(\xi)}|}{|D|}\right)^{\rho(\frac{2}{3})} \xi\mathbf{1}_k \qquad \forall\, k \in \mathcal{S}\,. \tag{A.24}$$

By (A.23)–(A.24) and using (A.20), we obtain the estimate

$$|\mathcal{D}^{(\xi)}| \le \sum_{k\in\mathcal{S}} |\mathcal{D}_k^{(\xi)}| \le \sum_{k\in\mathcal{S}} \left(\frac{u_k(x^{(k)})}{\xi\beta_0}\right)^{\rho d} |D| \le N\left(\frac{1}{\xi\beta_0}\right)^{\rho d} |D| \qquad \forall\, \xi > 0\,. \tag{A.25}$$

Choosing $\xi = \frac{qM}{4}$, we have by (A.25)

$$\left| \Big\{ x \in \tfrac{1}{3}D \,:\, \max_{k\in\mathcal{S}} \big\{u_k(x)\big\} \ge \tfrac{qM}{4} \Big\} \right| \le N(\tfrac{4}{q\beta_0 M})^{\rho d} |D|\,.$$

Hence, if Q_0 is a cube of volume $|Q_0| = N\big(\frac{4a}{q\beta_0 M}\big)^{\rho d}|D|$ centered at y_0, then

$$\operatorname{osc}(u_{k_0}; Q_0) \ge \big(1 - \tfrac{q}{4}\big) M\,. \tag{A.26}$$

By Theorem A.2, we obtain from (A.26)

$$\operatorname{osc}(\boldsymbol{u}; 3Q_0) \ge \frac{(1-\frac{q}{4})}{(1-\frac{q}{2})} M = q_0 M\,. \tag{A.27}$$

Since $\boldsymbol{u}$ is nonnegative, (A.27) implies that there exists $y^{(1)} \in 3Q_0$ and $k_1 \in \mathcal{S}$ such that

$$u_{k_1}(y^{(1)}) \ge q_0 M\,.$$

Note that (A.22) implies that $3Q_0 \subset \frac{1}{3}D$. Therefore, we can repeat the argument, now choosing $\xi = q_0 \frac{qM}{4}$ in (A.25) and a cube Q_1 of volume $N\big(\frac{4a}{q_0 q\beta_0 M}\big)^{\rho d}|D|$ centered at $y^{(1)}$, to conclude that there exists $y^{(2)} \in 3Q_1$ and $k_2 \in \mathcal{S}$ such that $u_{k_2}(y^{(2)}) \ge q_0^2 M$. Inductively, we can construct a sequence $\big\{y^{(n)}, k_n, Q_n\big\}_{n=0}^{\infty}$ satisfying, for all $n = 0, 1, \ldots,$

$$\begin{aligned} y^{(0)} = y_0 \in \tfrac{1}{9}\overline{D} \bigcap Q_0\,, \qquad y^{(n)} \in Q_n \bigcap 3Q_{n-1}\,, \\ |Q_n|^{1/d} = N^{1/d}\big(\tfrac{1}{q_0}\big)^{n\rho}\big(\tfrac{4a}{q\beta_0 M}\big)^{\rho}|D|^{1/d}\,, \\ u_{k_n}(y^{(n)}) \ge q_0^n M\,. \end{aligned} \tag{A.28}$$

The inequality in (A.22) guarantees that $y^{(n)} \in \frac{1}{3}D$ for all n. But (A.28) implies that $\boldsymbol{u}$ is unbounded in $\frac{1}{3}D$, which is a contradiction. ∎

Remark A.1. By the comparison principle, Lemmas A.3–A.5 and Theorem A.1 clearly hold unmodified for the class of L_k-superharmonic, nonnegative functions, i.e., functions $u \in W^{2,d}_{\ell oc}(D) \bigcap C^0(\overline{D})$, satisfying $L_k u \le 0$ in D, for some $k \in \mathcal{S}$ and $L \in \mathfrak{L}$. This fact will be used in the next result.

LEMMA A.6. *Let $L \in \mathfrak{L}$, $k \in \mathcal{S}$ and suppose φ is a solution to the Dirichlet problem $L_k\varphi = -f$ in a cube $D \subset \mathbb{R}^d$, with $\varphi = 0$ on ∂D, with f satisfying*

$$0 \le f(x) \le M \quad \forall\, x \in D \qquad \text{and} \qquad \|f\|_{d;D} \ge \varepsilon > 0$$

for some constants M and ε. Then there exists a constant $C' = C'(M, \varepsilon, m, \overline{m}, \overline{\gamma})$ such that

$$\inf_{x \in \frac{1}{3}D} \{\varphi(x)\} \ge C'.$$

Proof. First note that the Dirichlet problem as defined has a unique strong solution $\varphi \in W^{2,p}_{\ell oc}(D) \bigcap C^0(\overline{D})$ for all $p \in [d, \infty)$. We argue by contradiction. Suppose there exists a sequence of operators $\{L^{(n)}\}_{n=1}^{\infty} \subset \mathfrak{L}$ and a sequence of functions $\{f^{(n)}\}_{n=1}^{\infty}$, in accord with the hypotheses of the lemma, such that the corresponding solutions $\{\varphi^{(n)}\}_{n=1}^{\infty}$ of $L_k^{(n)}\varphi^{(n)} = -f^{(n)}$ satisfy

$$\inf_{x \in \frac{1}{3}D} \{\varphi^{(n)}(x)\} < \frac{1}{n^2}\,, \qquad n = 1, 2, \dots\,.$$

Thus, by Theorem A.1,

$$\left|\{x \in D \,:\, \varphi^{(n)}(x) \ge \tfrac{1}{n}\}\right| \le \left(\frac{1}{\beta_0 n}\right)^{\rho d} |D|\,,$$

with ρ as defined in (A.20). Since the sequence $\varphi^{(n)}$ is bounded in $L^\infty(D)$ (by Lemma A.2), it follows that $\varphi^{(n)} \to 0$ in $L^p(D)$, as $n \to \infty$, for all $p \in [1, \infty)$. Let $D' = \delta D$, with $\delta < 1$, be a subcube of D, and let $\|\cdot\|_{2,p;D'}$ denote the standard norm of $W^{2,p}(D')$. We use the well-known estimate

$$\|\varphi^{(n)}\|_{2,p;D'} \le C''\big(\|\varphi^{(n)}\|_{p;D} + \|f^{(n)}\|_{p;D}\big)\,,$$

for some constant $C'' = C''(|D|, p, \delta, d, m, \overline{m}, \overline{\gamma})$, to conclude that the first and second derivatives of $\varphi^{(n)}$ converge weakly to 0 in $L^p(D')$, for all $p \in [1, \infty)$. In turn, since $W^{2,p}_0(D') \hookrightarrow W^{1,p}_0(D')$ is compact for $p > d$, using the standard approximation argument we deduce that $\frac{\partial \varphi^{(n)}}{\partial x_i}$ converges in $L^p(D')$ strongly for all $i = 1, \dots, d$. Also, since the second order coefficients of $L_k^{(n)}$ are uniformly Lipschitz, we can extract a subsequence, along which they converge uniformly. Combining all the previous arguments, we deduce that the sequence $\{L_k^{(n)}\varphi^{(n)}\}$ converges weakly to 0 in $L^p(D')$, $p \in [1, \infty)$. On the other hand, if we choose $\delta \ge (1 - \frac{\varepsilon}{2M|D|})^{1/d}$, an easy calculation yields

$$\int_{D'} f^{(n)}(x)\, dx \ge \frac{\varepsilon}{2}\,, \qquad n = 1, 2, \dots\,,$$

resulting in a contradiction. ∎

We pause to note that (A.5) has not been utilized in any of the results obtained thus far. It will be used in the next result to provide the necessary "coupling" between distinct components of the harmonic function.

LEMMA A.7. *For each cube $D \subset \mathbb{R}^d$ there exists a constant $M_2 > 0$ such that, for any $\boldsymbol{u} \in \mathfrak{U}_D$,*

$$\inf_{x \in \frac{1}{9}D} \{u_i(x)\} \le M_2 \inf_{x \in \frac{1}{9}D} \{u_j(x)\} \qquad \forall\, i, j \in \mathcal{S}\,.$$

Proof. Let ε_D be the constant in hypothesis (A.5). Define a collection of functions $\{\varphi_{ij}(x),\ \ i, j \in \mathcal{S}\} \subset W^{2,d}_{\ell oc}(\frac{1}{3}D) \bigcap C^0(\frac{1}{3}\overline{D})$, relative to some $L \in \mathfrak{L}$, by

$$\begin{aligned} &L_i\varphi_{ij}(x) = -c_{ij}(x) \quad \text{in } \tfrac{1}{3}D \qquad \text{and} \quad \varphi_{ij}(x) = 0 \quad \text{on } \partial\big(\tfrac{1}{3}D\big) \qquad \text{if} \quad i \neq j, \\ &\varphi_{ij}(x) = 0 \qquad \text{if} \quad i = j\,, \end{aligned} \tag{A.29}$$

and let $\boldsymbol{\Phi}(x)$, $\boldsymbol{C}(x)$ denote the matrices with elements $\{\varphi_{ij}(x)\}$ and $\{c_{ij}(x)\}_{i \neq j}$, respectively. By (A.4), there exists a constant irreducible matrix $\boldsymbol{C}_D \subset \mathbb{R}^{N \times N}$, with elements equal to 0 or 1 such that

$$\Big|\big\{x \in \tfrac{1}{3}D \,:\, \boldsymbol{C}(x) \ge \varepsilon_D \boldsymbol{C}_D\big\}\Big| \ge \tfrac{1}{N^2 3^d}|D|\,. \tag{A.30}$$

It follows by (A.29), (A.30), and Lemma A.6 that there exists a constant $\varepsilon'_D > 0$ such that

$$\boldsymbol{\Phi}(x) \ge \varepsilon'_D \boldsymbol{C}_D \qquad \forall\, x \in \tfrac{1}{9}D\,, \tag{A.31}$$

and (A.31) holds relative to any $L \in \mathfrak{L}$ used to generate φ_{ij}. Therefore, if $\boldsymbol{u} \in \mathfrak{U}_D$ and we define $\underline{\boldsymbol{u}} := \inf_{x \in \frac{1}{9}D} \boldsymbol{u}(x)$ and $\underline{\boldsymbol{u}}' := \inf_{x \in \frac{1}{3}D} \boldsymbol{u}(x)$, it is a direct consequence of the comparison principle that

$$\boldsymbol{u}(x) \ge \boldsymbol{\Phi}(x)\underline{\boldsymbol{u}}' \qquad \forall\, x \in \tfrac{1}{3}D\,. \tag{A.32}$$

On the other hand, by Theorem A.1,

$$\underline{\boldsymbol{u}}' \ge F\big(\tfrac{1}{9^d}\big)\underline{\boldsymbol{u}}\,. \tag{A.33}$$

By (A.31)–(A.33),

$$\boldsymbol{u}(x) \ge \varepsilon'_D F\big(\tfrac{1}{9^d}\big)\boldsymbol{C}_D\underline{\boldsymbol{u}} \qquad \forall\, x \in \tfrac{1}{9}D\,,$$

which yields $\underline{\boldsymbol{u}} \ge \varepsilon'_D F\big(\frac{1}{9^d}\big)\boldsymbol{C}_D\underline{\boldsymbol{u}}$. In turn, the irreducibility of $\boldsymbol{C}_D$ implies that

$$\underline{\boldsymbol{u}}_i \ge \Big(\varepsilon'_D F\big(\tfrac{1}{9^d}\big)\Big)^{N-1} \underline{\boldsymbol{u}}_j \qquad \forall\, i, j \in \mathcal{S}\,. \qquad \square$$

Combining Theorem A.3 and Lemma A.7 and letting $M := M_1 M_2$, we have the following theorem.

THEOREM A.4. *For each cube $D \subset \mathbb{R}^d$, $|D| \le \kappa_0$, there exists a constant $M > 0$ such that, for any $\boldsymbol{u} \in \mathfrak{U}_D$,*

$$u_i(y) \le M u_j(x) \qquad \forall\, x, y \in \tfrac{1}{9}D \quad \forall\, i, j \in \mathcal{S}\,.$$

Theorem 4.1 easily follows from Theorem A.4 by covering the domain Ω with a collection of congruent cubes D of suitable size. For an elegant exposition of this technique, see [11, p. 153]. The existence of a constant $\varepsilon_\Omega > 0$ satisfying (A.5) is guaranteed by the continuity and irreducibility conditions in Assumption 3.1 (i) and (iii), along with the compactness of U. Concerning (A.3), (A.4), and the upper bound in (A.2), observe that for each bounded domain Ω, Assumption 3.1 (i) implies the existence of constants $\overline{m}$ and $\overline{\gamma}$ satisfying all these conditions in Ω. This suffices for our purposes.

Acknowledgment. The authors wish to thank Prof. S. R. S. Varadhan for explaining to us the work of Krylov and Safonov. The appendix in this paper is both inspired by and based in part on his notes on the proof of Harnack's inequality for a uniformly elliptic operator.

REFERENCES

[1] R. Akella and P. R. Kumar, *Optimal control of production rate in a failure prone manufacturing system*, IEEE Trans. Automat. Control, AC-31 (1986), pp. 116–126.
[2] A. Bensoussan and P. L. Lions, *Optimal control of random evolutions*, Stochastics, 5 (1981), pp. 169–199.
[3] R. N. Bhattacharya, *Criteria for recurrence and existence of invariant measures for multidimensional diffusions*, Ann. Probab., 6 (1978), pp. 541–553.
[4] R. N. Bhattacharya and S. Ramasubramanian, *Recurrence and ergodicity of diffusions*, J. Multivariate Anal., 12 (1982), pp. 95–112.
[5] T. Bielecki and P. R. Kumar, *Optimality of zero-inventory policies for unreliable manufacturing systems*, Oper. Res., 36 (1988), pp. 532–541.
[6] V. S. Borkar, *Optimal Control of Diffusion Processes*, Pitman Research Notes in Math. 203, Longman, Harlow, UK, 1989.
[7] V. S. Borkar, *Topics in Controlled Markov Chains*, Pitman Research Notes in Math. 240, Longman, Harlow, UK, 1991.
[8] V. S. Borkar and M. K. Ghosh, *Ergodic control of multidimensional diffusions,* I*: The existence results*, SIAM J. Control Optim., 26 (1988), pp. 112–126.
[9] V. S. Borkar and M. K. Ghosh, *Ergodic control of multidimensional diffusions,* II*: Adaptive control*, Appl. Math. Optim., 21 (1990), pp. 191–220.
[10] V. S. Borkar and M. K. Ghosh, *Controlled diffusions with constraints*, II, J. Math. Anal. Appl., 176 (1993), pp. 310–321.
[11] C. Caratheodory, *Theory of Functions of a Complex Variable* I, 2nd ed., Chelsea, New York, 1964.
[12] E. B. Dynkin, *Markov Processes Vols.* I *and* II, Springer-Verlag, New York, 1965.
[13] M. K. Ghosh, A. Arapostathis, and S. I. Marcus, *An optimal control problem arising in flexible manufacturing systems*, in Proc. 30th IEEE Conf. on Decision and Control, Brighton, England, 1991, pp. 1844–1849.
[14] M. K. Ghosh, A. Arapostathis, and S. I. Marcus, *Optimal control of switching diffusions with application to flexible manufacturing systems*, SIAM J. Control Optim., 31 (1993), pp. 1183–1204.
[15] P. Grisvard, *Elliptic Problems in Non-Smooth Domains*, Pitman, Boston, 1965.
[16] N. Ikeda and S. Watanabe, *Stochastic Differential Equations and Diffusion Processes*, North-Holland, Amsterdam, 1981.
[17] J. Jacod and A. N. Shiryayev, *Limit Theorems for Stochastic Processes*, Springer-Verlag, New York, 1980.
[18] N. V. Krylov, *Controlled Diffusion Processes*, Applications of Mathematics 14, Springer-Verlag, New York, 1980.
[19] N. V. Krylov and M. V. Safonov, *An estimate of the probability that a diffusion process hits a set of positive measure*, Soviet Math. Dokl., 20 (1979), pp. 253–255.
[20] T. G. Kurtz and D. L. Ocone, *Unique characterization of conditional distributions in nonlinear filtering*, Ann. Probab., 16 (1988), pp. 80–107.
[21] H. J. Kushner, *Existence results for optimal stochastic control*, J. Optim. Theory Appl., 15 (1975), pp. 347–359.
[22] O. A. Ladyzhenskaya and N. N. Ural'ceva, *Linear and Quasilinear Elliptic Equations*, Academic Press, New York, 1968.
[23] M. Mariton, *Jump Linear Systems in Automatic Control*, Marcel Dekker, New York, 1990.
[24] R. Phelps, *Lectures on Choquet's Theorem*, Van Nostrand, New York, 1966.
[25] A. V. Skorohod, *Asymptotic Methods in the Theory of Stochastic Differential Equations*, AMS, Providence, 1989.
[26] R. H. Stockbridge, *Time-average control of a martingale problem: Existence of a stationary solution*, Ann. Probab., 18 (1990), pp. 88–108.

SIAM J. CONTROL OPTIM.
Vol. 35, No. 6, pp. 1989–2020, November 1997

007

AVERAGING THEOREMS FOR HIGHLY OSCILLATORY DIFFERENTIAL EQUATIONS AND ITERATED LIE BRACKETS*

WENSHENG LIU†

Abstract. Using averaging techniques and developing proper algebraic formalisms, we study the limiting process of ordinary differential equations with highly oscillatory right-hand sides. We give sufficient conditions, generalizing earlier work by Kurzweil and Jarnik, for a sequence $\{u^j = (u_1^j, \dots, u_m^j)\} \subseteq L^1([0,T], \mathbb{R}^m)$ to be such that, for every choice of smooth vector fields $f_k, k = 1, \dots, m$, on a smooth manifold, the trajectories of $\dot{x} = \sum_{k=1}^m u_k^j(t) f_k(x)$ converge to the trajectories of an "extended system" $\dot{x} = \sum_{k=1}^r v_k(t) f_k(x)$, where the new directions $f_{m+1}, \dots, f_r$ are Lie brackets of $f_1, \dots, f_m$.

Key words. control affine systems, extended inputs, averaging, Lie brackets, continuous dependence

AMS subject classifications. 34E10, 34C29, 34A12, 93B29, 93C15

PII. S0363012994268667

1. Introduction.

Consider control-affine systems of the form

$$\dot{x} = \sum_{k=1}^{m} u_k(t) f_k(x), \tag{1}$$

where $f_1, \dots, f_m$ are smooth vector fields on a smooth manifold M and the inputs $u = (u_1, \dots, u_m)$ are functions belonging to $L^1([0,T], \mathbb{R}^m)$. It is well known in the control theory literature that trajectories of (1) generated by a sequence of highly oscillatory inputs may converge to a function that is no longer a trajectory of (1). In many cases it happens that the limiting function is a solution of a differential equation whose right-hand side involves not only the vector fields $f_1, \dots, f_m$ but also various Lie brackets of them; cf. the example below. The purpose of this paper is to give a systematic study of the Lie brackets that can occur in the limit and to clarify the underlying algebraic structures. By developing proper algebraic formalisms, we can study the limiting processes of (1) corresponding to sequences $\{u^j\} \subseteq L^1([0,T], \mathbb{R}^m)$ in pure algebraic levels, and it turns out that the limiting processes are closely related to the limiting behavior of the *Chen–Fliess series* determined by the u^j. (The Chen–Fliess series techniques have been widely used in control theory, e.g., by Fliess, Sussmann, etc.; cf. [4], [9], [11].) We then show that under very general conditions the trajectories of (1) generated by a sequence $\{u^j\} \subseteq L^1([0,T], \mathbb{R}^m)$ converge to trajectories of a system of the form

$$\dot{x} = \sum_{k=1}^{r} v_k(t) f_k(x), \tag{2}$$

where the first m vector fields $f_1, \dots, f_m$ are the same as in (1), and $f_{m+1}, \dots, f_r$ are Lie brackets of the f_k, $k \in \{1, \dots, m\}$, so the limiting equation involves resonance terms of Lie brackets of $f_1, \dots, f_m$.

*Received by the editors June 1, 1994; accepted for publication (in revised form) August 21, 1996. http://www.siam.org/journals/sicon/35-6/26866.html

†Department of Mathematics, Rutgers University, New Brunswick, NJ 08903 (wliu@hilbert.rutgers.edu).

For a simple illustration of the phenomenon that Lie brackets can occur in limiting equations, consider the system

$$\dot{x}_1 = u_1,$$
$$\dot{x}_2 = u_2,$$
$$\dot{x}_3 = u_2 x_1,$$

with initial condition $(0,0,0)$. If $u^j(t) = j^{\frac{1}{2}}(\cos jt, \sin jt)$, then the u^j "converge to 0" in the sense that their indefinite integrals $U^j(t) = \int_0^t u^j(s)\,ds$ converge to 0 uniformly. So, if we let $u^\infty = (0,0)$, we might think that the solutions x^j converge uniformly to $(0,0,0)$. However, a simple calculation shows that this is not so, and in fact the x^j converge uniformly to $x^\infty = (0,0,\frac{t}{2})$. On the other hand, the vector fields f_1, f_2 have components $(1,0,0)$ and $(0,1,x_1)$, respectively. So the Lie bracket $f_3 = [f_1, f_2]$ has components $(0,0,1)$, and the limiting function x^∞ turns out to satisfy $\dot{x} = \frac{1}{2} f_3(x)$.

Using integration by parts, it is easy to show that the sequence $\{u^j\}$ in the above example has the stronger convergence property that for *every* initial condition $x(0) = p \in \mathbb{R}^n$, *every* choice of C^2 vector fields f_1, f_2 on $\mathbb{R}^n$, the solutions of $\dot{x} = u_1^j(t) f_1(x) + u_2^j(t) f_2(x), x(0) = p$ are defined on $[0,T]$ for j large enough and converge uniformly to the solution x^∞ of $\dot{x} = \frac{1}{2}[f_1, f_2](x)$, $x(0) = p$, provided only that the latter exists on $[0,T]$. In this paper we will explore these "universal convergence" properties for sequences $\{u^j\} \subseteq L^1([0,T], \mathbb{R}^m)$. We will give general convergence results which in particular include the above example as a special case.

The main technical difficulty of studying the limiting process of (1) corresponding to sequences $\{u^j\}$ is not to prove the convergence of trajectories but to show that the limiting equations are of the form (2) and to have simple formulas to compute the v_k. In fact, it is fairly easy to establish the convergence of trajectories and to show that the right-hand sides of the limiting equations are equal to finite linear combinations of functions that involve Jacobian matrices of $f_1, \ldots, f_m$, but it is not obvious a priori that these linear combinations give rise to (time-varying) vector fields. In [5], [6], [7], Kurzweil and Jarnik have studied the problem of finding the proper forms of limiting equations for some special input sequences. For those special cases, using lengthy combinatorial proofs, they have shown that the limiting equations are of the form (2). Here we provide a different approach. Using proper algebraic formalisms, that the limiting equations are of the form (2) follows naturally, and we can give simple algebraic formulas to compute the v_k.

The key point for our algebraic formalism is to reformulate the problem of convergence of trajectories in terms of *convergence of inputs*. For instance, we can reformulate the universal convergence property of the sequence $\{u^j\}$ in the above example in the language of "input convergence" as follows. First we introduce formal noncommuting *indeterminates* $X_1, X_2, \ldots, X_m$ and define an *extended input value* (EIV) to be a linear combination of $X_1, \ldots, X_m$ and various Lie brackets such as $[X_1, X_2]$, $[X_1, [X_1, X_2]]$, etc. Define an *extended input* to be a Lebesgue integrable function on an interval $[0,T]$ whose values are EIVs. So, for example (if $m = 2$), an expression such as

$$(3)\quad v(t) = v_1(t)X_1 + v_2(t)X_2 + v_3(t)[X_1, X_2] + v_4(t)[X_1, [X_1, X_2]] + v_5(t)[X_2, [X_1, X_2]],$$

with integrable coefficient functions v_k on $[0,T]$, is an extended input. It is clear that an ordinary input can then be regarded in a natural way as a special kind of extended input, namely, one that contains only terms with $X_1, \ldots, X_m$ but no higher-order Lie brackets. For any m-tuple of vector fields $(f_1, \ldots, f_m)$ on M, extended

inputs can be plugged into a control system such as (1) exactly as ordinary inputs can, by substituting the vector fields f_k for the indeterminates X_k. The results are (nonautonomous) ordinary differential equations. For example, the extended input above gives rise to the following differential equation:

$$\dot{x} = v_1(t)f_1(x)+v_2(t)f_2(x)+v_3(t)[f_1,f_2](x)+v_4(t)[f_1,[f_1,f_2]](x)+v_5(t)[f_2,[f_1,f_2]](x).$$

Therefore we can talk about trajectories corresponding to an extended input once the vector fields $f_1,\dots,f_m$ are known. (This requires only that the f_k be smooth enough for the appropriate Lie brackets to exist.) We can then say that a sequence $\{u^j\}$ of ordinary inputs *converges* to an extended input u^∞ if for *every* initial condition $x(0)=p$, *every* choice of sufficiently smooth vector fields f_k, the solutions x^j generated by the u^j converge uniformly to the solution x^∞ generated by u^∞.[1] With this terminology, our example above simply says that the ordinary inputs $u^j(t) = j^{\frac{1}{2}}\cos(jt)X_1 + j^{\frac{1}{2}}\sin(jt)X_2$ converge to the extended input $u^\infty(t) = \frac{1}{2}[X_1,X_2]$. In this paper, we will give various sufficient conditions for a sequence of ordinary inputs to converge to an extended input.

The convergence theorems in this paper are *high-order* generalizations of the convergence results for control affine systems discussed in [13]. In there it is given a convergence result that roughly says the following. A sequence $\{u^j = u_1^jX_1+\cdots+u_m^jX_m\}$ of ordinary inputs converges to an ordinary input $u^\infty = u_1^\infty X_1+\cdots+u_m^\infty X_m$ if and only if

(c1) the indefinite integrals $U^j(t) = \int_0^t u^j(s)\,ds$ converge uniformly on $[0,T]$ to $U^\infty(t) = \int_0^t u^\infty(s)\,ds$; i.e., u^∞ is the averaged limit of the u^j.

(c2) the functions u^j are uniformly bounded in L^1; i.e., there exists a finite constant C such that $\int_0^T \|u^j(t)\|dt \le C$ for all j.

This is a first-order convergence result since the limiting extended input u^∞ does not contain any high-order Lie brackets. The fact that u^∞ is still an ordinary input is due to the boundedness assumption (c2). The above example shows that if this condition fails, it is indeed possible for the u^j to converge to an extended input that is not an ordinary input.

Our generalization of this first-order convergence result to high-order extended inputs is as follows. Instead of (c1) we will assume that all the iterated integrals

$$U_I^j(t) \stackrel{\text{def}}{=} \int_0^t\int_0^{t_k}\int_0^{t_{k-1}}\cdots\int_0^{t_2} u_{i_k}^j(t_k)u_{i_{k-1}}^j(t_{k-1})\dots u_{i_1}^j(t_1)dt_1\dots dt_k \tag{4}$$

converge uniformly, as $j\to\infty$, to the indefinite integrals H_I of certain L^1 functions h_I, for all multiindices $I=(i_1,\dots,i_k)\in\{1,\dots,m\}^k$, $k=1,2,\dots$, such that $|I| \stackrel{\text{def}}{=} k \le r$. (In that case, we will say that the u^j *converge to* $H=\{H_I\}_{|I|\le r}$ *in the* r*th-order iterated integral sense* or, for short, that they *ii*(r)*-converge* to H.) As will be explained in section 2, the *ii*(r)-convergence is equivalent to the convergence of the trajectories for one special system, the "rth-order truncated formal system."

Instead of (c2), we will assume an r*th-order boundedness condition* (c2(r)), which states that certain sequences $\{\widetilde{uv}_I^j\}$ of functions associated with the multiindices I such that $|I|=r$ are bounded in L^1 norm. The conclusion is that the ordinary inputs

[1] This is not made precise yet, since the solutions may have explosions. The precise definition is given in section 3.

u^j converge as $j \to \infty$ to an extended input u^∞ given by

$$u^\infty(t) = \sum_{|I| \le r} u_I^\infty(t)[X_I],$$

where, for any index $I = (i_1, \dots, i_k)$, we define $[X_I]$ by

$$[X_I] \stackrel{\text{def}}{=} [X_{i_1}, [X_{i_2}, [\dots, [X_{i_{k-1}}, X_{i_k}] \cdots],$$

and the u_I^∞ are certain integrable real-valued functions on $[0, T]$. Unfortunately, the statement of the convergence result is somewhat less transparent than might be desired, because the limiting extended input involves the u_I^∞ rather than the h_I, and the boundedness condition (c2(r)) involves another family $\{\widetilde{uv}_I^j\}$ of sequences of functions. (For more applicability, the $\widetilde{uv}_I^j$ are allowed to depend on another collection $\{v_I^j\}$ of sequences of functions; cf. section 2.) It turns out, however, that the u_I^∞ and the $\widetilde{uv}_I^j$ can be computed from the h_I, u^j, and v_I^j by very simple algebraic formulas, provided that everything is reformulated in an appropriate algebraic context; cf. (9) and (18). See also (46) and (47).

One particular situation where the definition of the u_I^∞ is extremely simple is the case when the only terms that occur in the limit are the rth-order ones, i.e., when the h_I for $|I| < r$ vanish. In that case the u_I^∞ vanish for $|I| < r$ and are equal to $\frac{h_I}{r}$ for $|I| = r$. (This special case was studied, under stronger assumptions, in [5], [6], [7].)

In the formulation of the convergence theorems, it is convenient to let $U_k(t) = \int_0^t u_k(s)\, ds$ and to rewrite equation (1) in the following form:

$$dx = \sum_{k=1}^{m} f_k(x)\, dU_k. \tag{5}$$

However, this is not simply a rewriting of (1) since for this system we can use more general inputs, e.g., continuous inputs with bounded variations, or Hölder continuous inputs with Hölder exponents $> \frac{1}{2}$. Moreover, one can consider stochastic differential equations, where $U = (U_1, \dots, U_m)$ is a standard m-dimensional Brownian motion. When the U_k are absolutely continuous, system (5) has no difference from (1).

In this paper we will restrict ourselves to study the limiting process of (5) corresponding to a sequence $\{U^j\}$ of functions in $BVC([0, T], \mathbb{R}^m)$. (Here $BVC([0, T], \mathbb{R}^m)$ denotes the set of $\mathbb{R}^m$-valued functions on $[0, T]$ that are continuous and of bounded variations.) It turns out that the limiting theorems presented in this paper are also true if we use Hölder continuous inputs, and can be generalized to stochastic differential equations too. These will be given in forthcoming papers.

The organization of the paper is as follows. In section 2 we develop the needed algebraic formalisms. We begin by reviewing some basic terminologies on free associative algebras and free Lie algebras. We then introduce the concepts of *extended inputs, formal trajectories, generalized differences*, etc., and study systematically the relations between them in pure algebraic level. Using the algebraic formalisms, in section 3, we present the main convergence theorems. In section 4 we discuss briefly the necessity of the conditions of our main convergence theorems. We conclude the paper with an appendix that contains the proof of a lemma and an approximation result.

2. Algebraic preliminaries and formalisms. In this section we develop the necessary algebraic formalisms for our convergence theorems. We follow the notation and definitions of [11].

2.1. Review of basic terminology on free associative algebras and free Lie algebras. As in [11], let $\mathbf{X} = \{X_1, \ldots, X_m\}$ be a finite sequence of objects, which will be called *indeterminates*. We let $A(\mathbf{X})$ denote the free associative algebra generated over $\mathbb{R}$ by $\mathbf{X}$. For any multiindex $I = (i_1, \ldots, i_k)$ with $i_1, \ldots, i_k \in \{1, \ldots, m\}$, we let $X_I = X_{i_1} \cdots X_{i_k}$. (There is a special multiindex $I = \emptyset$. It is understood that $X_\emptyset = 1$.) Then $A(\mathbf{X})$ is the set of all sums $\sum_I a_I X_I$, where the coefficients a_I are real numbers, the summation runs over all possible multiindices I, and all but finitely many a_I vanish. Therefore the *monomials* X_I form a basis of $A(\mathbf{X})$ and every element of $A(\mathbf{X})$ is a finite linear combination of the X_I.

We also consider the algebra $\hat{A}(\mathbf{X})$ of all formal power series in $\mathbf{X}$. The elements of $\hat{A}(\mathbf{X})$ are the formal sums $\sum_I a_I X_I$, where I ranges over all multiindices. This is the sum as above except that the a_I are no longer required to vanish for all but finitely many I. In both $A(\mathbf{X})$ and $\hat{A}(\mathbf{X})$, addition is done componentwise, and multiplication is carried out using the formula $X_I X_J = X_{IJ}$, where IJ is the concatenation of I and J, namely, the multiindex obtained by writing, in order, first the components of I and then those of J.

For any integer $r \geq 0$, let us use $A^r(\mathbf{X})$ to denote the *free nilpotent associative algebra of step* $r+1$ in the indeterminates $\mathbf{X}$. Therefore $A^r(\mathbf{X})$ is defined like $\hat{A}(\mathbf{X})$, except that now all the monomials X_I with $|I| > r$ are set equal to zero. (Here $|I|$ is the length of I; i.e., $|I| = k$ if $I = (i_1, \ldots, i_k)$.) Then $A^r(\mathbf{X})$ can be thought of as the quotient of $A(\mathbf{X})$ or $\hat{A}(\mathbf{X})$ modulo the two-side ideal of all sums of monomials of degree strictly larger than r. (The degree of a monomial X_I is $|I|$.) The canonical projection $\mathbf{Tr}(r)$ from $\hat{A}(\mathbf{X})$ to $A^r(\mathbf{X})$ is the operator that assigns to each series $S \in \hat{A}(\mathbf{X})$ the finite series $\mathbf{Tr}(r)(S)$ obtained from S by deleting all the terms of degree $> r$. (The symbol $\mathbf{Tr}$ comes from the word "truncation." It is used to indicate that the map $\mathbf{Tr}(r) : \hat{A}(\mathbf{X}) \to A^r(\mathbf{X})$ is in essence a truncation map; i.e., for any $S \in \hat{A}(\mathbf{X})$, $\mathbf{Tr}(r)(S)$ is the truncation of S "up to order r.") The kernel of $\mathbf{Tr}(r)$ is denoted by $\hat{A}_r(\mathbf{X})$. In particular, $\hat{A}_0(\mathbf{X})$ is the set of all formal power series $\sum_I a_I X_I$ for which $a_\emptyset = 0$. The exponential map is a well-defined bijection

$$\exp : \ \hat{A}_0(\mathbf{X}) \to 1 + \hat{A}_0(\mathbf{X}),$$

whose inverse is a map from $1 + \hat{A}_0(\mathbf{X})$ to $\hat{A}_0(\mathbf{X})$ denoted by "log." (Here $1 + \hat{A}_0(\mathbf{X})$ is the subset of $\hat{A}(\mathbf{X})$ that contains all the elements S such that $S - 1 \in \hat{A}_0(\mathbf{X})$.) If $S \in \hat{A}_0(\mathbf{X})$, then $\exp(S)$ and $\log(1+S)$ are given by the usual series

$$\exp(S) = \sum_{n=0}^{\infty} \frac{S^n}{n!},$$

$$\log(1+S) = \sum_{n=1}^{\infty} \frac{(-1)^{(n-1)} S^n}{n}.$$

One can also define $A^r_\tau(\mathbf{X})$ to be the set of all elements of $A^r(\mathbf{X})$ that are linear combinations of monomials of degree $> \tau$. Then $A^r_\tau(\mathbf{X}) = \mathbf{Tr}(r)(\hat{A}_\tau(\mathbf{X}))$. The exponential map

$$\exp_r : \ A^r_0(\mathbf{X}) \to 1 + A^r_0(\mathbf{X})$$

and its inverse $\log_r$ are given in the case by power series that are actually finite sums due to the nilpotency of $A^r(\mathbf{X})$.

The algebras $A(\mathbf{X}), \hat{A}(\mathbf{X}), A^r(\mathbf{X})$ are Lie algebras in the usual way. (If A with a product $(a, b) \to ab$ is an associative algebra, then with the bracket product $[a, b] \stackrel{\text{def}}{=}$

$ab-ba$, A is a Lie algebra.) We let $L(\mathbf{X})$ denote the Lie subalgebra of $A(\mathbf{X})$ generated by the indeterminates $X_1, \ldots, X_m$. An element S of $A(\mathbf{X})$ will be said to be a *Lie element* if $S \in L(\mathbf{X})$. It is clear that an $S \in A(\mathbf{X})$ is a Lie element iff all the *homogeneous* components of S are Lie elements. (An $S \in \hat{A}(\mathbf{X})$ is homogeneous if it is a linear combination of monomials with equal degree.)

We can also define $\hat{L}(\mathbf{X})$ to be the set of all those elements of $\hat{A}(\mathbf{X})$ whose components are Lie brackets in $X_1, \ldots, X_m$. Therefore $\hat{L}(\mathbf{X})$ contains those $S \in \hat{A}(\mathbf{X})$ whose homogeneous components are Lie elements. The elements of $\hat{L}(\mathbf{X})$ are called *Lie series* in $X_1, \ldots, X_m$. For each multiple index $I = (i_1, \ldots, i_k)$, let $[X_I] \stackrel{\text{def}}{=} [X_{i_1}, [X_{i_2}, [\ldots, [X_{i_{k-1}}, X_{i_k}] \cdots]$. Then $L(\mathbf{X})$ and $\hat{L}(\mathbf{X})$ are spanned by the $[X_I]$. Naturally the $[X_I]$ are not linearly independent. There are several systematic procedures of figuring out a basis of $L(\mathbf{X})$. But we will not need this here.

Let $\hat{G}(\mathbf{X}) = \{\exp(Z), Z \in \hat{L}(\mathbf{X})\}$, the set of exponentials of the elements of $\hat{L}(\mathbf{X})$. The Campbell–Hausdorff formula implies (cf., e.g., [3]) that $\hat{G}(\mathbf{X})$ is in fact a group under the operation of multiplication in $\hat{A}(\mathbf{X})$. The elements of $\hat{G}(\mathbf{X})$ are called *exponential Lie series*. We let $L^r(\mathbf{X})$ be the Lie subalgebra of $A^r(\mathbf{X})$ generated by $\mathbf{X}$ and define $G^r(\mathbf{X})$ to be the subset of $A^r(\mathbf{X})$ consisting of all the exponentials of elements of $L^r(\mathbf{X})$. Now $L^r(\mathbf{X})$ is a finite-dimensional Lie algebra and $G^r(\mathbf{X})$ is its corresponding simply connected Lie group.

2.2. Polynomial inputs and formal trajectories. Let $BVC[0,T]$ be the set of real-valued functions U on $[0,T]$ which are continuous and of bounded variations. We will say that a function is *B-continuous* if it is in $BVC[0,T]$. If $f \in BVC[0,T]$, we will use $TV[f;0,T]$ to denote the total variation of f on $[0,T]$.

Let V be a function on $[0,T]$ with values in $\hat{A}(\mathbf{X})$. We say that V is *B-continuous* if we write $V(t) = \sum_I V_I(t) X_I$; then all the functions V_I are in $BVC[0,T]$.

We define two bilinear products on $BVC([0,T], A)$,

$$(F, G) \to F \stackrel{\rightarrow}{*} G,$$
$$(F, G) \to F \stackrel{\leftarrow}{*} G,$$

where

$$(F \stackrel{\rightarrow}{*} G)(t) \stackrel{\text{def}}{=} \int_0^t F(s)\, dG(s),$$
$$(F \stackrel{\leftarrow}{*} G)(t) \stackrel{\text{def}}{=} \int_0^t dF(s)\, G(s),$$

and A could be $A(\mathbf{X})$, $\hat{A}(\mathbf{X})$, or $A^r(\mathbf{X})$ for some integer $r \geq 0$. (Here $BVC([0,T], A)$ denotes the set of B-continuous A-valued functions.) If $F = \sum_I a_I X_I$ and $G = \sum_I b_I X_I$, then by definition

$$(F \stackrel{\rightarrow}{*} G)(t) = \sum_I \left(\sum_{J_1 J_2 = I} (a_{J_1} \stackrel{\rightarrow}{*} b_{J_2})(t) \right) X_I = \sum_I \left(\sum_{J_1 J_2 = I} \int_0^t a_{J_1}(s) db_{J_2}(s) \right) X_I,$$

$$(F \stackrel{\leftarrow}{*} G)(t) = \sum_I \left(\sum_{J_1 J_2 = I} (a_{J_1} \stackrel{\leftarrow}{*} b_{J_2})(t) \right) X_I = \sum_I \left(\sum_{J_1 J_2 = I} \int_0^t b_{J_2}(s) da_{J_1}(s) \right) X_I,$$

where the inner summation above runs over all ways of expressing the multiindex I as a concatenation $J_1 J_2$ of 2 indices, and the integral is the usual Riemann–Stieltjes

integral. If $F_1, \dots, F_k \in BVC([0,T], A)$, we use the conventions

$$(F_1 \overrightarrow{*} F_2 \overrightarrow{*} \cdots \overrightarrow{*} F_k)(t) = (\cdots (F_1 \overrightarrow{*} F_2) \overrightarrow{*} F_3) \overrightarrow{*} \cdots) \overrightarrow{*} F_k)(t),$$
$$(F_1 \overleftarrow{*} F_2 \overleftarrow{*} \cdots \overleftarrow{*} F_k)(t) = (F_1 \overleftarrow{*} (F_2 \overleftarrow{*} (\cdots \overleftarrow{*} (F_{k-1} \overleftarrow{*} F_k) \cdots)(t).$$

Note that when $r = 0$, $A^r(\mathbf{X}) = A^0(\mathbf{X})$ can be identified with $\mathbb{R}$, so by definition $BVC([0,T], A^0(\mathbf{X})) = BVC[0,T]$. In this case $\overrightarrow{*}, \overleftarrow{*}$ reduce to maps from $BVC[0,T] \times BVC[0,T] \to BVC[0,T]$. Let f and g be two functions in $BVC[0,T]$; then by definition

$$(f \overrightarrow{*} g)(t) = \int_0^t f(s)\, dg(s),$$
$$(f \overleftarrow{*} g)(t) = \int_0^t g(s)\, df(s).$$

The functions $f \overrightarrow{*} g, f \overleftarrow{*} g$ are clearly in $BVC[0,T]$.

Let us say that two B-continuous $\hat{A}_0(\mathbf{X})$-valued functions V_1, V_2 on $[0,T]$ are *equivalent* if $V_1 - V_2$ is an element of $\hat{A}_0(\mathbf{X})$; i.e., $V_1 - V_2$ does not depend on t. Then the set of all B-continuous $\hat{A}_0(\mathbf{X})$-valued functions is divided into equivalence classes. Let us use $\mathcal{P}$ to denote the set of all equivalence classes. For each B-continuous $\hat{A}_0(\mathbf{X})$-valued function V we use $\mathbf{V}$ to denote the equivalence class determined by V, and V will be called a *representative* of $\mathbf{V}$.

DEFINITION 1. *Any element $\mathbf{V}$ of $\mathcal{P}$ will be called a polynomial input.*

Therefore a polynomial input is an equivalence class $\mathbf{V}$, whose representatives V are B-continuous functions on $[0,T]$ with values in $\hat{A}_0(\mathbf{X})$. In general we will use $\mathbf{V}$ to denote a polynomial input and V to denote a representative of $\mathbf{V}$. But, if there is no confusion, we sometimes write $\mathbf{V} = V$ to denote the equivalence class determined by V.

DEFINITION 2. *A polynomial input $\mathbf{V}$ is an extended input if $\mathbf{V}$ has a representative that is $\hat{L}(\mathbf{X})$-valued.*

An ordinary input $U = (U_1, \dots, U_m) \in BVC([0,T], \mathbb{R}^m)$ can be regarded as a polynomial input by identifying it with the equivalence class $\mathbf{U}$ determined by $U = U_1X_1 + \cdots + U_mX_m$. It is an extended input in fact by the above definition. In most cases we will make no difference between $U = (U_1, \dots, U_m)$ and $U = U_1X_1 + \cdots + U_mX_m$. We will call the equivalence class $\mathbf{U} = U$ an ordinary input too. Therefore an ordinary input $\mathbf{U}$ is an equivalence class that has a representative U whose values are linear combinations of the $X_k, k = 1, \dots, m$.

Let $V = \sum_I V_I X_I$ be an $\hat{A}(\mathbf{X})$-valued function on $[0,T]$. We say that V is B-continuous, absolutely continuous, differentiable, etc., on $[0,T]$ if all the V_I are B-continuous, absolutely continuous, differentiable, etc. We say that a sequence $\{V^j = \sum_I V_I^j X_I\}$, $j \in \{1, 2, \dots, \} \cup \{\infty\}$, of $\hat{A}(\mathbf{X})$-valued functions on $[0,T]$ *converges* to V^∞ uniformly if, for each I, the V_I^j converge to V_I^∞ uniformly on $[0,T]$ as $j \to \infty$. We say that a polynomial input $\mathbf{V}$ is *absolutely continuous* or *differentiable* if there is a representative of $\mathbf{V}$ which is absolutely continuous or differentiable. Therefore if $\mathbf{V}$ is absolutely continuous (differentiable) all the representatives of $\mathbf{V}$ are absolutely continuous (differentiable). And if $\mathbf{V}$ is an absolutely continuous polynomial input, then $\mathbf{v} = \dot{\mathbf{V}}$ is well defined. The result is an $\hat{A}_0(\mathbf{X})$-valued integrable function on $[0,T]$.

DEFINITION 3. *Let $\mathbf{V}$ be a polynomial input. The Chen–Fliess series determined by $\mathbf{V}$ is the $\hat{A}(\mathbf{X})$-valued function $S_{\mathbf{V}}$ on $[0,T]$ that satisfies the initial value problem*

$$dS = SdV, S(t) \in \hat{A}(\mathbf{X}), \tag{6}$$

$$S(0) = 1, \tag{7}$$

where V is any representative of $\mathbf{V}$.

A solution of (6) and (7) is an $\hat{A}(\mathbf{X})$-valued function S on $[0,T]$, which is B-continuous, satisfying

$$S = 1 + S \overrightarrow{*} V.$$

Such a solution clearly exists, is unique, and depends only on $\mathbf{V}$. Moreover, if $S_{\mathbf{V}}$ is the solution of (6) and (7), it is clear that

$$S_{\mathbf{V}} = 1 + \sum_{k=1}^{\infty} 1 \overrightarrow{*} \overbrace{V \overrightarrow{*} V \overrightarrow{*} \cdots \overrightarrow{*} V}^{k}.$$

Equation (6) will be called the *formal equation* determined by $\mathbf{V}$. The Chen–Fliess series $S_{\mathbf{V}}$ will be called the *formal trajectory* of $\mathbf{V}$.

By definition, the function $t \to S_{\mathbf{V}}(t)$ is B-continuous as an $\hat{A}(\mathbf{X})$-valued function. It is in fact $1 + \hat{A}_0(\mathbf{X})$-valued. Conversely, define a *formal trajectory* to be a B-continuous $1 + \hat{A}_0(\mathbf{X})$-valued function S on $[0,T]$. Then every formal trajectory S is the formal trajectory of a polynomial input $\mathbf{V}$ given by $\mathbf{V} = (S^{-1} \overrightarrow{*} S)$. Therefore the map $\mathbf{V} \to S_{\mathbf{V}}$ is a one-to-one correspondence between the set of polynomial inputs and that of formal trajectories, whose inverse is given by $\mathbf{V} = (S_{\mathbf{V}}^{-1} \overrightarrow{*} S_{\mathbf{V}})$. (Note that if $S = \sum_I a_I X_I$ is an element of $\hat{A}(\mathbf{X})$ with $a_\emptyset \neq 0$, then S^{-1} exists and is given by (for simplicity assume $a_\emptyset = 1$)

$$\begin{aligned} S^{-1} &= \left(1 + \sum_{|I|>0} a_I X_I\right)^{-1} = \sum_{k=0}^{\infty} (-1)^k \left(\sum_{|I|>0} a_I X_I\right)^k \\ &= 1 + \sum_{|I|>0} \left(\sum_{k=1}^{\infty} (-1)^k \sum_{J_1 \ldots J_k = I} a_{J_1} \ldots a_{J_k}\right) X_I, \end{aligned} \tag{8}$$

where the inner summation in the last equality above is over all ways of expressing the multiindex I as a concatenation $J_1 \cdots J_k$ of k indices.)

Remark 1. It follows from the results of [10] that $S_{\mathbf{V}}$ is $\hat{G}(\mathbf{X})$-valued if $\mathbf{V}$ is an extended input. The converse of this is also true; i.e., if S is a $\hat{G}(\mathbf{X})$-valued formal trajectory, then $\mathbf{V} = (S^{-1} \overrightarrow{*} S)$ is an extended input. To see this, it is enough to prove that in this case the function $(S^{-1} \overrightarrow{*} S)$ is $\hat{L}(\mathbf{X})$-valued. This is true when S is C^∞, since in that case $(S^{-1} \overrightarrow{*} S)(t) = \int_0^t S^{-1}(\tau)\dot{S}(\tau)\,d\tau$, and $S^{-1}(t)\dot{S}(t) = \lim_{h\to 0} \frac{1}{h} S^{-1}(t)(S(t+h) - S(t)) = \lim_{h\to 0} \frac{1}{h}(S^{-1}(t)S(t+h) - 1)$. Using the Campbell–Hausdorff formula we conclude that $S^{-1}(t)S(t+h) = \exp(\Lambda(t,h))$, where $\Lambda(t,h)$ is a Lie series that goes to zero as $h \to 0$. So $S^{-1}(t)\dot{S}(t) = \lim_{h\to 0} \frac{1}{h}(\Lambda(t,h) + \frac{1}{2}\Lambda(t,h)^2 + \cdots) = \lim_{h\to 0} \frac{\Lambda(t,h)}{h}$. So $S^{-1}\dot{S}$ is Lie series valued, which implies that $\mathbf{V} = (S^{-1} \overrightarrow{*} S)$ is an extended input. In the general case, let $S(t) = \exp(Z(t))$, where $Z(t) = \sum_{|I|>0} Z_I(t)[X_I]$, $Z_I \in BVC[0,T]$, $Z(0) = 0$. From Proposition 4 in the appendix, for each I, we can take a sequence $\{Z_I^j\} \subseteq C^\infty[0,T]$, $Z_I^j(0) = 0$, such that the Z_I^j converge to Z_I uniformly as $j \to \infty$ and the $\|\dot{Z}_I^j\|_{L^1}$ are uniformly bounded in j. Let $Z^j = \sum_{|I|>0} Z_I^j[X_I]$. Then we know that the Z^j converge to Z uniformly in $\hat{A}_0(\mathbf{X})$. Let $S^j(t) = \exp(Z^j(t))$. We have $S^j \to \exp(Z) = S$ uniformly. Since $(S^j)^{-1} = \exp(-Z^j)$, the $(S^j)^{-1}$ converge to $\exp(-Z) = S^{-1}$ uniformly as

$j \to \infty$. Since the $\dot{Z}_I^j$ are uniformly bounded in $L^1[0,T]$ for each I, if we write $S^j = 1 + \sum_{|I|>0} H_I^j X_I$, then the L^1 norms of the $\dot{H}_I^j$ are also uniformly bounded in j for each I. The following lemma implies that the $((S^j)^{-1} \overrightarrow{*} S^j)$ converge to $(S^{-1} \overrightarrow{*} S)$ uniformly. Since the $((S^j)^{-1} \overrightarrow{*} S^j)$ are Lie series valued functions, $(S^{-1} \overrightarrow{*} S)$ is a Lie series. Therefore $\mathbf{V} = (S^{-1} \overrightarrow{*} S)$ is an extended input.

LEMMA 1. *Let $\{f^j\}$ be a sequence of continuous functions on $[0,T]$ that converges to f uniformly on $[0,T]$ as $j \to \infty$. Let $\{g^j\}$ be a sequence of functions belonging to $BVC[0,T]$. Assume that the functions g^j converge to g uniformly on $[0,T]$ as $j \to \infty$ and the total variations $TV[g^j;0,T]$ of the g^j on $[0,T]$ are uniformly bounded in j. Then the integrals $\int_0^t f^j(s)\,dg^j(s)$ converge to $\int_0^t f(s)\,dg(s)$ uniformly as $j \to \infty$.*

The proof of this lemma is given in the appendix.

The polynomial input $\mathbf{V}$ can be computed from $S_{\mathbf{V}}$ using standard algebraic tools. If $S_{\mathbf{V}} = 1 + \sum_{|I|>0} H_I X_I$, from (8) we know that

$$S_{\mathbf{V}}^{-1} = 1 + \sum_{|I|>0} \left(\sum_{k=1}^{\infty} (-1)^k \sum_{J_1 \dots J_k = I} H_{J_1} \dots H_{J_k} \right) X_I.$$

We have

$$(S_{\mathbf{V}}^{-1} \overrightarrow{*} S_{\mathbf{V}})(t) = \left(\left(1 + \sum_{|I|>0} \left(\sum_{k=1}^{\infty} (-1)^k \sum_{J_1 \dots J_k = I} H_{J_1} \dots H_{J_k} \right) X_I \right) \overrightarrow{*} \left(1 + \sum_{|I|>0} H_I X_I \right) \right)(t)$$

$$= \sum_{|I|>0} \left(H_I + \sum_{k=1}^{\infty} (-1)^k \sum_{J_1 \dots J_k J_{k+1} = I} (H_{J_1} \dots H_{J_k}) \overrightarrow{*} H_{J_{k+1}} \right) X_I,$$

where the inner summation above runs over all ways of expressing the multiindex I as a concatenation $J_1 \cdots J_k J_{k+1}$ of $k+1$ indices. So, if we let $V(t) = (S_{\mathbf{V}}^{-1} \overrightarrow{*} S_{\mathbf{V}})(t) = \sum_{|I|>0} V_I(t) X_I$, then the V_I are given by

$$V_I = H_I + \sum_{k=1}^{\infty} (-1)^k \sum_{J_1 \dots J_k J_{k+1} = I} (H_{J_1} \dots H_{J_k}) \overrightarrow{*} H_{J_{k+1}}. \tag{9}$$

Let π be the linear map of $A(\mathbf{X})$ onto $L(\mathbf{X})$ defined by $\pi(X_I) = \frac{1}{|I|}[X_I]$. It is well known (cf., e.g., [3]) that the restriction of π to $L(\mathbf{X})$ is the identity map; i.e., π is a projector of $A(\mathbf{X})$ onto $L(\mathbf{X})$. Let $\hat{\pi}$ be the linear projection map from $\hat{A}(\mathbf{X})$ to $\hat{L}(\mathbf{X})$ that extends π. From Remark 1 we know that if $S = 1 + \sum_{|I|>0} H_I X_I$ is a $\hat{G}(\mathbf{X})$-valued formal trajectory, $V = (S^{-1} \overrightarrow{*} S)$ is a Lie series valued function. In that case we have $\hat{\pi}(V) = V$, so

$$V(t) = \sum_{|I|>0} \frac{1}{|I|} V_I(t)[X_I], \tag{10}$$

where the V_I are given by (9).

In the particular case when $\mathbf{V}$ is an ordinary input $\mathbf{U} = U_1X_1 + \cdots + U_mX_m$, the Chen–Fliess series $S_{\mathbf{U}}$ is given by the formula

$$S_{\mathbf{U}}(t) = 1 + \sum_{|I|>0} \hat{U}_I(t)X_I,$$

where, if $I = (i_1, \ldots, i_k)$, then $\hat{U}_I$ is the iterated integral defined by

$$\hat{U}_I \stackrel{\text{def}}{=} 1 \overrightarrow{*} U_{i_1} \overrightarrow{*} U_{i_2} \overrightarrow{*} \cdots \overrightarrow{*} U_{i_k}. \tag{11}$$

DEFINITION 4. *A sequence $\{\mathbf{V}^j\}$ of polynomial inputs FT-converges (converges in the formal trajectory sense) to a polynomial input $\mathbf{V}^\infty$ if the corresponding formal trajectories $S_{\mathbf{V}^j}$ converge uniformly to $S_{\mathbf{V}^\infty}$ as $j \to \infty$.*

Hence, if $\{\mathbf{U}^j\}$ is a sequence of ordinary inputs and $\mathbf{V}$ is a polynomial input such that

$$S_{\mathbf{V}}(t) = 1 + \sum_{|I|>0} H_I(t)X_I, \tag{12}$$

then $\{\mathbf{U}^j\}$ FT-converges to $\mathbf{V}$ if and only if the sequence $\{\hat{U}_I^j\}$ of functions converges uniformly to H_I for each I. Using formula (9), we can explicitly compute the limiting polynomial input $\mathbf{V}$ in terms of the limits H_I of the iterated integrals $\hat{U}_I^j$.

DEFINITION 5. *Given two polynomial inputs $\mathbf{V}_1 = V_1$ and $\mathbf{V}_2 = V_2$, we define a generalized difference (GD for short) of $\mathbf{V}_1$ and $\mathbf{V}_2$ to be a B-continuous $\hat{A}_0(\mathbf{X})$-valued function W that satisfies*

$$dW = -dV_1W + dV_2 - dV_1, \quad W(t) \in \hat{A}_0(\mathbf{X}). \tag{13}$$

Clearly a solution W of (13) is uniquely determined by its initial condition $W(0)$. We will use $GD\{\mathbf{V}_1, \mathbf{V}_2\}$ to denote the set of all generalized differences of $\mathbf{V}_1$ and $\mathbf{V}_2$.

Remark 2. If $W(0) = 0$, then it is clear that W has some of the properties of a "difference of $\mathbf{V}_1$ and $\mathbf{V}_2$." (For instance, $W \equiv 0$ iff $\mathbf{V}_1 = \mathbf{V}_2$.) In general, if $W(0)$ is "small" in some sense, then it is reasonable to expect that $W(t)$ is "small" for all t if and only if $\mathbf{V}_1$ and $\mathbf{V}_2$ are "close." The convergence theorems will make this precise. Let W be a solution of (13) with initial condition $W(0) = W_0 \in \hat{A}_0(\mathbf{X})$. Then W satisfies the integral equation

$$W = W_0 - V_2(0) + V_1(0) - V_1 \overleftarrow{*} W + V_2 - V_1$$

and can be calculated by the formula

$$W = \sum_{k=0}^{\infty} (-1)^k \overbrace{V_1 \overleftarrow{*} V_1 \overleftarrow{*} \cdots \overleftarrow{*} V_1}^{k} \overleftarrow{*} (V_2 - V_1 + \tilde{W}), \tag{14}$$

where $\tilde{W} = W_0 - V_2(0) + V_1(0)$. Let $\mathbf{U}$ be an ordinary input and $\mathbf{V}$ be a polynomial input. Let $U = U_1X_1 + \cdots + U_mX_m$ and $V = \sum_{|I|>0} V_IX_I$ be a representative of $\mathbf{U}$ and $\mathbf{V}$, respectively. Let

$$(\mathbf{U} \stackrel{\text{g.d.}}{-} \mathbf{V})(t) = \sum_{|I|>0} \widetilde{UV}_I(t)X_I$$

be a generalized difference of $\mathbf{U}$ and $\mathbf{V}$ with initial value $(\mathbf{U} \overset{\text{g.d.}}{-} \mathbf{V})(0) = \sum_{|I|>0} \hat{W}_I X_I$. Let $\tilde{W} = \sum_{|I|>0} \tilde{W}_I X_I = \sum_{|I|>0} \hat{W}_I X_I - V(0) + U(0)$. From (14) we have

$$\widetilde{UV}_{i_1,\dots,i_k} = \sum_{\ell=0}^{k} (-1)^\ell U_{i_1} \overset{\leftarrow}{*} U_{i_2} \overset{\leftarrow}{*} \cdots \overset{\leftarrow}{*} U_{i_\ell} \overset{\leftarrow}{*} (V_{i_{\ell+1},\dots,i_k} + \tilde{W}_{i_{\ell+1},\dots,i_k}). \tag{15}$$

Formula (15) implies the recursive formula (this also follows directly from (13))

$$\widetilde{UV}_i = V_i + \tilde{W}_i - U_i,$$

$$\widetilde{UV}_{i_1,\dots,i_k} = V_{i_1,\dots,i_k} + \tilde{W}_{i_1,\dots,i_k} - U_{i_1} \overset{\leftarrow}{*} \widetilde{UV}_{i_2,\dots,i_k} = V_{i_1,\dots,i_k} + \tilde{W}_{i_1,\dots,i_k} - \widetilde{UV}_{i_2,\dots,i_k} \overset{\rightarrow}{*} U_{i_1}.$$

Remark 3. If $\tilde{W} = \sum_{|I|>0} \tilde{W}_I X_I = \sum_{|I|>0} \hat{W}_I X_I - V(0) + U(0) = 0$, then the above recursive formula becomes

$$\widetilde{UV}_{i_1,\dots,i_k} = V_{i_1,\dots,i_k} - U_{i_1} \overset{\leftarrow}{*} \widetilde{UV}_{i_2,\dots,i_k}. \tag{16}$$

For example, if we take a representative V of $\mathbf{V}$ with $V(0) = \hat{W} + U(0)$, then $\tilde{W} = 0$.

Remark 4. If $\mathbf{V}_1$ and $\mathbf{V}_2$ are two polynomial inputs, we use $(\mathbf{V}_1 \overset{\text{g.d.}}{-} \mathbf{V}_2)(t)$ to denote a generalized difference of $\mathbf{V}_1$ and $\mathbf{V}_2$. If $\mathbf{U}$ is an ordinary input and $\mathbf{V}$ is a polynomial input, we will use $\widetilde{UV}_I$ to denote the coefficients of X_I in $(\mathbf{U} \overset{\text{g.d.}}{-} \mathbf{V})(t)$ with the understanding that the $\widetilde{UV}_I$ are given by (15). The notations $(\mathbf{V}_1 \overset{\text{g.d.}}{-} \mathbf{V}_2)(t)$, $(\mathbf{U} \overset{\text{g.d.}}{-} \mathbf{V})(t)$, and $\widetilde{UV}_I(t)$ are slightly ambiguous because they depend also on the initial values $(\mathbf{V}_1 \overset{\text{g.d.}}{-} \mathbf{V}_2)(0)$, $(\mathbf{U} \overset{\text{g.d.}}{-} \mathbf{V})(0)$, and $\widetilde{UV}_I(0)$, but we will use them anyhow.

DEFINITION 6. *Let $\{\mathbf{V}_1^j\}$ and $\{\mathbf{V}_2^j\}$ be two sequences of polynomial inputs. We say that $\{\mathbf{V}_1^j, \mathbf{V}_2^j\}$ GD-converges to zero if there exist generalized differences $(\mathbf{V}_1^j \overset{\text{g.d.}}{-} \mathbf{V}_2^j)(t)$ of $\mathbf{V}_1^j$ and $\mathbf{V}_2^j$ that converge to zero uniformly as $j \to \infty$. We say that a sequence $\{\mathbf{V}^j\}$ of polynomial inputs GD-converges to a polynomial input $\mathbf{V}$ if $\{\mathbf{V}^j, \mathbf{V}\}$ GD-converges to* 0.

Remark 5. As the following simple examples shows, the positions of $\mathbf{V}_1^j$ and $\mathbf{V}_2^j$ are not symmetric in the definition of GD-convergence. Namely, the fact that $\{\mathbf{V}_1^j, \mathbf{V}_2^j\}$ GD-converges to 0 does not imply that $\{\mathbf{V}_2^j, \mathbf{V}_1^j\}$ GD-converges to 0. Similarly the condition that a sequence $\{\mathbf{V}^j\}$ GD-converges to $\mathbf{V}$ is not equivalent to the condition that $\{\mathbf{V}^j - \mathbf{V}\}$ GD-converges to 0.

Example. Let $m = 2$, $\mathbf{U}_1^j = j^{-\frac{1}{3}} \sin jt X_1 + j^{-\frac{1}{3}}(1 - \cos jt) X_2$, and $\mathbf{U}_2^j \equiv 0$ be two sequences of ordinary inputs. Then it is easy to see that $\{\mathbf{U}_2^j, \mathbf{U}_1^j\}$ GD-converges to 0. Indeed the differential equation (13) corresponding to $\mathbf{U}_2^j, \mathbf{U}_1^j$ is given by $dW = d\mathbf{U}_1^j$. On the other hand, the differential equation (13) corresponding to $\mathbf{U}_1^j, \mathbf{U}_2^j$ is given by

$$dW = -d\mathbf{U}_1^j W - d\mathbf{U}_1^j. \tag{17}$$

The solution $W^j(t) = \sum_{|I|>0} W_I^j(t) X_I$ of (17) with initial condition $W^j(0) = \hat{W}^j = \sum_{|I|>0} \hat{W}_I^j X_I$ can be computed explicitly and

$$W_{2,1}^j(t) = \hat{W}_{2,1}^j + j^{-\frac{1}{3}} \hat{W}_1^j (\cos jt - 1) + \frac{t}{2} j^{\frac{1}{3}} + \frac{1}{4} j^{-\frac{1}{3}} \sin 2jt,$$

which does not converge to 0 no matter what $\hat{W}$ is.

Remark 6. If $\{\mathbf{U}^j\}$ is a sequence of ordinary inputs and $\{\mathbf{V}^j\}$ a sequence of polynomial inputs, we will use the abused notation $\widetilde{UV}_I^j$ to denote the coefficients of a generalized difference of $\mathbf{U}^j$ and $\mathbf{V}^j$. (For simplicity we will use $\widetilde{UV}_I^j$ rather than $\widetilde{U^jV^j}_I$.) Therefore if $U^j = U_1^jX_1 + \cdots + U_m^jX_m$ and $V^j = \sum_{|I|>0} V_I^jX_I$ are representatives of $\mathbf{U}^j$ and $\mathbf{V}^j$, respectively, and $(\mathbf{U}^j \overset{\text{g.d.}}{-} \mathbf{V}^j)(t)$ is a generalized difference of $\mathbf{U}^j$ and $\mathbf{V}^j$ with $(\mathbf{U}^j \overset{\text{g.d.}}{-} \mathbf{V}^j)(0) = \sum_{|I|>0} \hat{W}_I^jX_I$, then the $\widetilde{UV}_I^j$ are given by

$$\widetilde{UV}_{i_1,\dots,i_k}^j = \sum_{\ell=0}^{k}(-1)^\ell U_{i_1}^j \overset{\leftarrow}{*} U_{i_2}^j \overset{\leftarrow}{*} \cdots \overset{\leftarrow}{*} U_{i_\ell}^j \overset{\leftarrow}{*} (V_{i_{\ell+1},\dots,i_k}^j + \tilde{W}_{i_{\ell+1},\dots,i_k}^j). \tag{18}$$

So it is clear that $\{\mathbf{U}^j, \mathbf{V}^j\}$ GD-converges to 0 if and only if there exist sequences of constants $\hat{W}_I^j$ such that the functions $\widetilde{UV}_I^j$ determined by the $\hat{W}_I^j$ converge to 0 uniformly for all multiindices $|I| > 0$.

DEFINITION 7. *Let $V = \sum_I V_IX_I \in BVC([0,T], \hat{A}(\mathbf{X}))$ be a B-continuous $\hat{A}(\mathbf{X})$-valued function and $\{V^j = \sum_I V_I^jX_I\}$ be a sequence in $BVC([0,T], \hat{A}(\mathbf{X}))$. We say that $\{V^j\}$ converges to V strongly if the functions V_I^j converge to V_I uniformly for all I and the total variations $TV[V_I^j; 0, T]$ of the V_I^j are uniformly bounded in j for each I.*

DEFINITION 8. *Let $\mathbf{V}$ be a polynomial input and $\{\mathbf{V}^j\}$ be a sequence of polynomial inputs. Then we say that $\{\mathbf{V}^j\}$ converges to $\mathbf{V}$ strongly if there are representatives V^j and V of $\mathbf{V}^j$ and $\mathbf{V}$, respectively, such that the V^j converge to V strongly.*

We will use $STCON(\mathbf{V})$ to denote the set of all sequences of polynomial inputs that converge to $\mathbf{V}$ strongly, so

$$STCON(\mathbf{V}) = \Big\{ \{\mathbf{V}^j\} \;\Big|\; \{\mathbf{V}^j\} \to \mathbf{V} \text{ strongly as } j \to \infty \Big\}.$$

Let $\{\mathbf{V}^j\}$ be an element of $STCON(\mathbf{V})$. Let us take a representative $V = \sum_{|I|>0} V_IX_I$ and $V^j = \sum_{|I|>0} V_I^jX_I$ of $\mathbf{V}$ and $\mathbf{V}^j$, respectively, such that the V^j converge to V strongly. Let $S_{\mathbf{V}^j} = 1 + \sum_{|I|>0} H_I^jX_I$ and $S_{\mathbf{V}} = 1 + \sum_{|I|>0} H_IX_I$ be the formal trajectories determined by $\mathbf{V}^j$ and $\mathbf{V}$, respectively. Then it is easy to see that the total variations $TV[H_I^j; 0, T]$ of the functions H_I^j are uniformly bounded in j for each I. From

$$\begin{aligned} S_{\mathbf{V}^j}(t) - S_{\mathbf{V}}(t) &= (S_{\mathbf{V}^j} \overset{\rightarrow}{*} V^j)(t) - (S_{\mathbf{V}} \overset{\rightarrow}{*} V)(t) \\ &= ((S_{\mathbf{V}^j} - S_{\mathbf{V}}) \overset{\rightarrow}{*} V^j)(t) + (S_{\mathbf{V}} \overset{\rightarrow}{*} (V^j - V))(t) \end{aligned}$$

we see that

$$\begin{aligned} H_\ell^j(t) - H_\ell(t) &= V_\ell^j(t) - V_\ell^j(0) - V_\ell(t) + V_\ell(0), \\ H_I^j(t) - H_I(t) &= \sum_{J_1J_2=I} \Big((H_{J_1}^j - H_{J_1}) \overset{\rightarrow}{*} V_{J_2}^j\Big)(t) + \sum_{J_1J_2=I} \Big(H_{J_1} \overset{\rightarrow}{*} (V_{J_2}^j - V_{J_2})\Big)(t) \\ &\quad + V_I^j(t) - V_I(t) - V_I^j(0) + V_I(0). \end{aligned}$$

Using this and induction we can show easily that the H_I^j converge to H_I for each I; cf. Lemma 1. Therefore $\{\mathbf{V}^j\} \in STCON(\mathbf{V})$ implies that $\{\mathbf{V}^j\}$ FT-converges to $\mathbf{V}$.

Let $\mathbf{U} = U_1X_1 + \cdots + U_mX_m$ be an ordinary input and $\mathbf{V} = \sum_{|I|>0} V_I X_I$ be a polynomial input. Assume that $U(0) = V(0) = 0$. Let $(\mathbf{U} \overset{\text{g.d.}}{-} \mathbf{V}) = \sum_{|I|>0} \widetilde{UV}_I X_I$ be a generalized difference of $\mathbf{U}$ and $\mathbf{V}$. Let $S_{\mathbf{U}}$ be the formal trajectory of $\mathbf{U}$. Now we establish a formula that relates $S_{\mathbf{U}}$, $\mathbf{V}$ and $(\mathbf{U} \overset{\text{g.d.}}{-} \mathbf{V})$. We know that the functions U_ℓ, V_I and $\widetilde{UV}_I$ are related by (15). Therefore we have

$$\begin{aligned} S_{\mathbf{U}}(t) &= 1 + \sum_{i_1=1}^{m} \int_0^t S_{\mathbf{U}}(\tau)\, dU_{i_1}(\tau)\, X_{i_1} \\ &= 1 + \sum_{i_1=1}^{m} \int_0^t S_{\mathbf{U}}(\tau)\, dV_{i_1}(\tau)\, X_{i_1} - \sum_{i_1=1}^{m} \int_0^t S_{\mathbf{U}}(\tau) d\widetilde{UV}_{i_1}(\tau)\, X_{i_1}. \end{aligned}$$

Applying integration by parts and noticing that $\widetilde{UV}_{i_1,i_2}(t) = \widetilde{UV}_{i_1,i_2}(0) + V_{i_1,i_2}(t) - (U_{i_1} \overset{\leftarrow}{*} \widetilde{UV}_{i_2})(t)$, we get

$$\begin{aligned} S_{\mathbf{U}}(t) &= 1 + \sum_{i_1=1}^{m} \int_0^t S_{\mathbf{U}}(\tau)\, dV_{i_1}(\tau)\, X_{i_1} + \sum_{i_1,i_2=1}^{m} \int_0^t S_{\mathbf{U}}(\tau)\, dV_{i_1,i_2}(\tau)\, X_{i_1}X_{i_2} \\ &- \sum_{i_1=1}^{m} S_{\mathbf{U}}(t)\widetilde{UV}_{i_1}(t)\, X_{i_1} + \sum_{i_1=1}^{m} \widetilde{UV}_{i_1}(0)\, X_{i_1} - \sum_{i_1,i_2=1}^{m} \int_0^t S_{\mathbf{U}}(\tau) d\widetilde{UV}_{i_1,i_2}(\tau)\, X_{i_1}X_{i_2}. \end{aligned}$$

Continuing the integration by parts in this way up to order k, we get

$$\begin{aligned} S_{\mathbf{U}}(t) &= 1 + \left(S_{\mathbf{U}} \overset{\rightarrow}{*} \left(\sum_{0<|I|\le k} V_I\, X_I \right) \right)(t) - \left(S_{\mathbf{U}} \overset{\rightarrow}{*} \left(\sum_{|I|=k} \widetilde{UV}_I X_I \right) \right)(t) \\ &- S_{\mathbf{U}}(t) \left(\sum_{0<|I|<k} \widetilde{UV}_I(t)\, X_I \right) + \sum_{0<|I|<k} \widetilde{UV}_I(0)\, X_I. \end{aligned} \tag{19}$$

PROPOSITION 1. *Let $\{\mathbf{U}^j\}$ be a sequence of ordinary inputs and $\mathbf{V}$ be a polynomial input. Then*

(a) *the following conditions are equivalent:*
   (i) *the $\mathbf{U}^j$ FT-converge to $\mathbf{V}$;*
   (ii) *the $\mathbf{U}^j$ GD-converge to $\mathbf{V}$;*
   (iii) *there is one element $\{\mathbf{V}^j\}$ of $STCON(\mathbf{V})$ such that $\{\mathbf{U}^j, \mathbf{V}^j\}$ GD-converges to 0;*
   (iv) *for any element $\{\mathbf{V}^j\}$ of $STCON(\mathbf{V})$ and any $(\mathbf{U}^j \overset{\text{g.d.}}{-} \mathbf{V}^j) \in GD\{\mathbf{U}^j, \mathbf{V}^j\}$, if the initial values $(\mathbf{U}^j \overset{\text{g.d.}}{-} \mathbf{V}^j)(0) \to 0$, then the sequence $\{(\mathbf{U}^j \overset{\text{g.d.}}{-} \mathbf{V}^j)\}$ converges to 0 uniformly as $j \to \infty$.*

(b) *if the $\mathbf{U}^j$ FT-converge to $\mathbf{V}$, then $\mathbf{V}$ is an extended input.*

Proof. First notice that conclusion (b) follows from Remark 1. (We know that the $S_{\mathbf{U}^j}$ are $\hat{G}(\mathbf{X})$ valued, so $S_{\mathbf{U}^j} = \exp(Z^j)$ for some $Z^j \in BVC([0,T], \hat{L}(\mathbf{X}))$. Now the fact that $\{S_{\mathbf{U}^j}\}$ converges to $S_{\mathbf{V}}$ uniformly implies that $\{Z^j\}$ converges to $\log(S_{\mathbf{V}})$ uniformly. Therefore $\log(S_{\mathbf{V}})$ is $\hat{L}(\mathbf{X})$-valued, which implies that $S_{\mathbf{V}}$ is a $\hat{G}(\mathbf{X})$-valued formal trajectory.)

Now we prove (a). It is clear that (iv) $\Rightarrow$ (ii) $\Rightarrow$ (iii). Therefore we need only to show that (i) implies (iv) and (iii) implies (i).

First we show that (i) implies (iv). Assume that $\{\mathbf{U}^j\}$ FT-converges to $\mathbf{V}$. Take any element $\{\mathbf{V}^j\}$ of $STCON(\mathbf{V})$ and generalized differences $(\mathbf{U}^j \overset{\text{g.d.}}{-} \mathbf{V}^j) = \sum_{|I|>0} \widetilde{UV}_I^j X_I$ of $\mathbf{U}^j$ and $\mathbf{V}^j$ with $(\mathbf{U}^j \overset{\text{g.d.}}{-} \mathbf{V}^j)(0) \to 0$. Let $U^j = U_1^j X_1 + \cdots + U_m^j X_m$ and $V = \sum_{|I|>0} V_I X_I$ be representatives of $\mathbf{U}^j$ and $\mathbf{V}$. Let $V^j = \sum_{|I|>0} V_I^j X_I$ be representatives of $\mathbf{V}^j$ that converge to V strongly. Without loss of generality we may assume that $U^j(0) = V^j(0) = V(0) = 0$. We use induction to show that the $\widetilde{UV}_I^j$ converge to 0 uniformly. Since the $\mathbf{U}^j$ FT-converge to $\mathbf{V}$, we know that $U_\ell^j \to V_\ell$ uniformly. (Notice that U_ℓ^j and V_ℓ are the coefficients of X_ℓ in $S_{\mathbf{U}^j}$ and $S_{\mathbf{V}}$, respectively.) It then follows that the $\widetilde{UV}_\ell^j = V_\ell^j - U_\ell^j + \widetilde{UV}_\ell^j(0)$ converge to 0 uniformly for $\ell = 1, \ldots, m$. Assume that $\widetilde{UV}_I^j \to 0$ uniformly for all I with $0 < |I| < k$. Now from (19) we get

$$S_{\mathbf{U}^j}(t) = 1 + \left(S_{\mathbf{U}^j} \overset{\rightarrow}{*} \left(\sum_{0<|I|\le k} V_I^j X_I\right)\right)(t) - \left(S_{\mathbf{U}^j} \overset{\rightarrow}{*} \left(\sum_{|I|=k} \widetilde{UV}_I^j X_I\right)\right)(t)$$

$$- S_{\mathbf{U}^j}(t) \left(\sum_{0<|I|<k} \widetilde{UV}_I^j(t)\, X_I\right) + \sum_{0<|I|<k} \widetilde{UV}_I^j(0)\, X_I. \tag{20}$$

Since $V^j(0) = 0$, we know that $V^j(t) = (S_{\mathbf{V}^j}^{-1} \overset{\rightarrow}{*} S_{\mathbf{V}^j})(t)$. Therefore

$$\sum_{0<|I|\le k} V_I^j(t)\, X_I = (S_{\mathbf{V}^j}^{-1} \overset{\rightarrow}{*} S_{\mathbf{V}^j})(t) - \sum_{|I|>k} V_I^j(t)\, X_I.$$

Notice that $\{S_{\mathbf{V}^j}\} \to S_{\mathbf{V}}$ uniformly implies that $\{S_{\mathbf{V}^j}^{-1}\}$ converges $S_{\mathbf{V}}^{-1}$ uniformly. Then we have

$$S_{\mathbf{U}^j} S_{\mathbf{V}^j}^{-1} \to S_{\mathbf{V}} S_{\mathbf{V}}^{-1} = 1 \tag{21}$$

uniformly as $j \to \infty$. Let $S_{\mathbf{V}^j} = 1 + \sum_{|I|>0} H_I^j X_I$ be the formal trajectories determined by $\mathbf{V}^j$. Then from the uniform boundedness of the $TV[V_I^j; 0, T]$ for each I we know that the $TV[H_I^j; 0, T]$ are also uniformly bounded in j for each I. Combining this with (21) we get that the $\hat{A}(\mathbf{X})$-valued functions

$$S_{\mathbf{U}^j}(t) - 1 - (S_{\mathbf{U}^j} \overset{\rightarrow}{*} (S_{\mathbf{V}^j}^{-1} \overset{\rightarrow}{*} S_{\mathbf{V}^j}))(t) = S_{\mathbf{U}^j}(t) - 1 - (S_{\mathbf{U}^j} S_{\mathbf{V}^j}^{-1} \overset{\rightarrow}{*} S_{\mathbf{V}^j})(t)$$

converge to zero uniformly. By induction, the $S_{\mathbf{U}^j}(\sum_{0<|I|<k} \widetilde{UV}_I^j X_I)$ converge to 0 uniformly. Therefore we have

$$S_{\mathbf{U}^j} \overset{\rightarrow}{*} \left(\sum_{|I|=k} \widetilde{UV}_I^j X_I\right) + S_{\mathbf{U}^j} \overset{\rightarrow}{*} \left(\sum_{|I|>k} V_I^j X_I\right) \to 0$$

uniformly. If we just take the degree k parts we get that the $\widetilde{UV}_I^j - \widetilde{UV}_I^j(0)$ converge to 0 uniformly for $|I| = k$, which implies that the $\widetilde{UV}_I^j$ converge to zero uniformly.

(iii) $\Longrightarrow$ (i). Let $\{\mathbf{V}^j\}$ be an element of $STCON(\mathbf{V})$ such that $\{\mathbf{U}^j, \mathbf{V}^j\}$ GD-converges to 0. Let $(\mathbf{U}^j \overset{\text{g.d.}}{-} \mathbf{V}^j) = \sum_{|I|>0} \widetilde{UV}_I^j X_I$ be generalized differences of $\mathbf{U}^j$

and $\mathbf{V}^j$ that converge to 0 uniformly. Let U^j, V, V^j be representatives of $\mathbf{U}^j, \mathbf{V}$, and $\mathbf{V}^j$ as above with $U^j(0) = V^j(0) = V(0) = 0$. Let $S_{\mathbf{U}^j} = 1 + \sum_{|I|>0} \hat{U}_I^j X_I$, $S_{\mathbf{V}} = 1 + \sum_{|I|>0} H_I X_I$, and $S_{\mathbf{V}^j} = 1 + \sum_{|I|>0} H_I^j X_I$ be the formal trajectories determined by $\mathbf{U}^j$, $\mathbf{V}$, and $\mathbf{V}^j$, respectively. We use induction to show that the $\hat{U}_I^j$ converge to H_I uniformly as $j \to \infty$ for each I. We know that $\{H_I^j\} \to H_I$ uniformly as $j \to \infty$. From (18) we have $\hat{U}_\ell^j = V_\ell^j - \widetilde{UV}_\ell^j + \widetilde{UV}_\ell^j(0) \to H_\ell$ uniformly for $\ell = 1, \dots, m$. Assume that the $\hat{U}_I^j$ converge to H_I for all $|I| < k$ with $k \geq 2$. Let I be such that $|I| = k$. Then from (20) we have

$$\hat{U}_I^j(t) = \sum_{J_1 J_2 = I} (\hat{U}_{J_1}^j \overrightarrow{*} V_{J_2}^j)(t) + V_I^j(t) - \widetilde{UV}_I^j(t) + \widetilde{UV}_I^j(0) - \sum_{J_1 J_2 = I} \hat{U}_{J_1}^j(t) \widetilde{UV}_{J_2}^j(t).$$

From this we see easily that the $\hat{U}_I^j$ converge to $V_I + \sum_{J_1 J_2 = I} H_{J_1} \overrightarrow{*} V_{J_2}$ uniformly. But since $S_{\mathbf{V}} = 1 + S_{\mathbf{V}} \overrightarrow{*} V$, we know that $H_I = V_I + \sum_{J_1 J_2 = I} H_{J_1} \overrightarrow{*} V_{J_2}$. Therefore we get that the $\hat{U}_I^j$ converge to H_I uniformly. □

2.3. Polynomial inputs of finite order and truncated formal trajectories. The concepts of polynomial inputs, extended inputs, formal trajectories, generalized differences, etc., have truncated analogues. Let us say that a polynomial input $\mathbf{V}$ has *order* $\leq r$ if there exists a representative of $\mathbf{V}$ whose values are linear combinations of monomials of degree $\leq r$. The smallest such an r is called the *order* of $\mathbf{V}$. We say that $\mathbf{V}$ is of *finite order* if it has order r for some integer $r > 0$.

If $\mathbf{V}$ is a polynomial input of order $\leq r$, we can regard $\mathbf{V}$ as an equivalence class of $A_0^r(\mathbf{X})$-valued rather than $\hat{A}_0(\mathbf{X})$-valued functions. The *rth-order truncated formal trajectory* determined by $\mathbf{V}$ is the solution of the initial value problem

$$dS = SdV, \quad S(t) \in A^r(\mathbf{X}), \tag{22}$$

$$S(0) = 1, \tag{23}$$

where V is any representative of $\mathbf{V}$. We will use $S_{\mathbf{V}}^r$ to denote the rth-order truncated formal trajectory determined by $\mathbf{V}$ in $A^r(\mathbf{X})$. Equation (22) is the *rth-order truncated formal equation* determined by $\mathbf{V}$.

We will say that a sequence $\{\mathbf{U}^j = U_1^j X_1 + \cdots + U_m^j X_m\}$ of ordinary inputs *FT(r)-converges* to a polynomial input $\mathbf{V}$ of order $\leq r$, with Chen–Fliess series $S_{\mathbf{V}} = 1 + \sum_{|I|>0} H_I X_I$, if the iterated integrals $\hat{U}_I^j$ converge uniformly to the functions H_I for all multiindices I such that $0 < |I| \leq r$. Equivalently, $\{\mathbf{U}^j\}$ FT(r)-converges to $\mathbf{V}$ if the rth-order truncated formal trajectories $S_{\mathbf{U}^j}^r$ converge uniformly to $S_{\mathbf{V}}^r$. In that case, the components V_I of a representative V of $\mathbf{V}$ are still given by (9). (In (9), we can use either the H_I of the series $S_{\mathbf{V}}$ or those of the truncated version, in which case we would set H_I to be 0 for $|I| > r$. Both yield the same result, because $\mathbf{V}$ is of order $\leq r$.)

If $\mathbf{V}_1$ and $\mathbf{V}_2$ are polynomial inputs of order $\leq r$, we can also define an *rth-order truncated generalized difference* of $\mathbf{V}_1$ and $\mathbf{V}_2$, which is a B-continuous $A_0^r(\mathbf{X})$-valued function that satisfies

$$dW = -dV_1 W + dV_2 - dV_1, \quad W(t) \in A_0^r(\mathbf{X}), \tag{24}$$

where V_1 and V_2 are representatives of $\mathbf{V}_1$ and $\mathbf{V}_2$, respectively. We will use the abused notation $(\mathbf{V}_1 \overset{\text{g.d.(r)}}{-} \mathbf{V}_2)$ to denote an rth-order truncated generalized difference of $\mathbf{V}_1$

and $\mathbf{V}_2$. It is uniquely determined by $(\mathbf{V}_1 \overset{\text{g.d.(r)}}{-} \mathbf{V}_2)(0)$. Let $U = U_1X_1 + \cdots + U_mX_m$ be an ordinary input and $\mathbf{V} = \sum_{0<|I|\le r} V_I X_I$ be a polynomial input of order $\le r$. Let $(\mathbf{U} \overset{\text{g.d.(r)}}{-} \mathbf{V})(t) = \sum_{0<|I|\le r} \widetilde{UV}_I(t) X_I$ be an rth-order truncated generalized difference of $\mathbf{U}$ and $\mathbf{V}$. Then the $\widetilde{UV}_I, 0 < |I| \le r$, can still be calculated by (15) once the initial values $\widetilde{UV}_I(0)$ are known. Similarly if $\{\mathbf{V}_1^j\}$ and $\{\mathbf{V}_2^j\}$ are two sequences of polynomial inputs of order $\le r$, then $\{\mathbf{V}_1^j, \mathbf{V}_2^j\}$ *GD(r)-converges* to 0 if there exist $(\mathbf{V}_1^j \overset{\text{g.d.(r)}}{-} \mathbf{V}_2^j)$ of $\mathbf{V}_1^j$ and $\mathbf{V}_2^j$ that converge to 0 uniformly as $j \to \infty$. Let $\{\mathbf{V}^j\}$ be a sequence of polynomial inputs of order $\le r$ and $\mathbf{V}$ be a polynomial input of order $\le r$. We say that $\mathbf{V}^j$ *GD(r)-converges* to $\mathbf{V}$ if $\{\mathbf{V}^j, \mathbf{V}\}$ GD(r)-converges to 0. If $\mathbf{V}$ is a polynomial input of order $\le r$, the set $STCON_r(\mathbf{V})$ contains those sequences $\{\mathbf{V}^j\}$ of polynomial inputs of order $\le r$ that there exist representatives $V^j = \sum_{0<|I|\le r} V_I^j X_I$ and $V = \sum_{0<|I|\le r} V_I X_I$ of $\mathbf{V}^j$ and $\mathbf{V}$, respectively, such that the V^j converge to V strongly. It is still true that a sequence $\{\mathbf{U}^j\}$ of ordinary inputs FT(r)-converges to a polynomial input $\mathbf{V}$ of order $\le r$ iff $\{\mathbf{U}^j, \mathbf{V}^j\}$ GD(r)-converge to zero for all elements $\{\mathbf{V}^j\}$ of $STCON_r(\mathbf{V})$. In particular $\{\mathbf{U}^j\}$ FT(r)-converges to $\mathbf{V}$ iff $\{\mathbf{U}^j\}$ GD(r)-converges to $\mathbf{V}$.

We conclude this section with a remark.

Remark 7. If we restrict ourselves to absolutely continuous polynomial inputs instead of considering B-continuous inputs, using the derivatives $\mathbf{v}$ of $\mathbf{V}$, everything can be reformulated in terms of $\mathbf{v}$. For example, we can define a polynomial input to be an $\hat{A}_0(\mathbf{X})$-valued integrable function $\mathbf{v} = \sum_{|I|>0} v_I X_I$ on $[0,T]$. An ordinary input $u = (u_1, \ldots, u_m) \in L^1([0,T], \mathbb{R}^m)$ can be identified with a polynomial input $\mathbf{u} = u_1X_1 + \cdots + u_mX_m$. In this case all the differential equations such as (6) and (13) become ordinary equations. We need not consider equivalence classes any more. Almost every concept can be defined in terms of $\mathbf{u}$ and $\mathbf{v}$. For example, the formal trajectory $S_{\mathbf{v}}$ determined by a polynomial input $\mathbf{v}$ is the solution of the following initial value problem:

$$\dot{S} = S\mathbf{v}, \quad S(t) \in \hat{A}(\mathbf{X}), \tag{25}$$

$$S(0) = 1. \tag{26}$$

Let $\mathbf{v} = \sum_{|I|>0} v_I X_I$ be a polynomial input and $\{\mathbf{v}^j = \sum_{|I|>0} v_I^j X_I\}$ be a sequence of polynomial inputs. Then $\{\mathbf{v}^j\}$ converges to $\mathbf{v}$ strongly if the indefinite integrals $\int_0^t v_I^j(s)ds$ converge to $\int_0^t v_I(s)ds$ uniformly and the L^1 norms of the v_I^j are uniformly bounded in j for each I. The $STCON(\mathbf{v})$ is the set that contains all the sequences $\{\mathbf{v}^j\}$ of polynomial inputs that converge to $\mathbf{v}$ strongly. If $\mathbf{v}_1$ and $\mathbf{v}_2$ are two polynomial inputs. A generalized difference of $\mathbf{v}_1$ and $\mathbf{v}_2$ is an absolutely continuous $\hat{A}_0(\mathbf{X})$-valued function W that satisfies

$$\dot{W} = -\mathbf{v}_1 W + \mathbf{v}_2 - \mathbf{v}_1.$$

Again if $\mathbf{u} = u_1X_1 + \cdots + u_mX_m$ is an ordinary input and $\mathbf{v} = \sum_{|I|>0} v_I X_I$ is a polynomial input, if we let $(\mathbf{u} \overset{\text{g.d.}}{-} \mathbf{v}) = \sum_{|I|>0} \widetilde{UV}_I X_I$ be a generalized difference of $\mathbf{u}$ and $\mathbf{v}$, then the functions $U_k(t) = \int_0^t u_k(s)ds, V_I(t) = \int_0^t v_I(s)ds$, and $\widetilde{UV}_I(t)$ are related by (15). The coefficients $\widetilde{UV}_I$ in this case are absolutely continuous functions. Similarly we can consider polynomial inputs of finite order. If r is a positive integer,

we can define rth-order truncated formal trajectories, rth-order truncated generalized differences, the set of $STCON_r(\mathbf{v})$, etc.

3. The convergence theorems. In this section we study the limiting behavior of trajectories of systems

$$dx = \sum_{k=1}^{m} f_k(x) dU_k \tag{27}$$

generated by a sequence $\{U^j = (U_1^j, \ldots, U_m^j)\} \subseteq BVC([0,T], \mathbb{R}^m)$ of inputs. Let U be a function in $BVC([0,T], \mathbb{R}^m)$ and $\bar{x}$ be a point in $\mathbb{R}^n$. First we review some existence and uniqueness results of solutions of the following initial value problem:

$$dx = \sum_{k=1}^{m} f_k(x)\, dU_k, \tag{28}$$

$$x(0) = \bar{x}, \tag{29}$$

where $f_k, k = 1, \ldots, m$, are continuous vector fields on $\mathbb{R}^n$. If the functions U_k in (28) are absolutely continuous, letting $u \in L^1([0,T], \mathbb{R}^m)$ be the derivative of U, then (28) and (29) becomes an ordinary initial value problem

$$\dot{x} = \sum_{k=1}^{m} u_k(t) f_k(x), \tag{30}$$

$$x(0) = \bar{x}. \tag{31}$$

For this initial value problem, the well-known Carathéodory theorem on existence and uniqueness of solutions says that

(1) if the vector fields f_k are continuous on $\mathbb{R}^n$, u_k are in $L^1[0,T]$, then the initial value problem (30), (31) has solutions;[2]
(2) if, moreover, the f_k are locally Lipschitz continuous, then uniqueness of solutions is guaranteed.

Let f_k, $k = 1, \ldots, m$, be continuous vector fields on $\mathbb{R}^n$. Let U be a function in $BVC([0,T], \mathbb{R}^m)$. By a *solution* of (28) and (29) we mean a continuous function x, defined on some subinterval $J \subseteq [0,T]$ that contains 0 and has positive length, that satisfies

$$x(t) = \bar{x} + \sum_{k=1}^{m} \int_0^t f_k(x(s))\, dU_k(s).$$

(Here the integral is the usual Riemann–Stieltjes integral.) The interval J is called the *domain* of the solution x. Clearly if x is a solution with domain J, then x is B-continuous on J. (We recall that a function x defined on an interval J (not necessarily closed) is of bounded variation if for any $a < b \in J$, x is of bounded variation on $[a,b]$ and $\sup_{a<b\in J}\{TV[x;a,b]\} < \infty$. A function x is *B-continuous* on an interval J if it is both continuous and of bounded variation on J.) Every solution x of (28) and (29) can be extended to a *maximal solution.*[3] We say that the initial value problem (28)

[2]Recall that a solution of (30) and (31) is an absolutely continuous function x on a subinterval $J \subseteq [0,T]$ that contains 0 and has positive length such that $x(0) = \bar{x}$ and (30) holds almost everywhere on J.

[3]A solution x of (28) and (29) with domain J is called a maximal solution if whenever $\hat{x}$ is another solution with domain $\hat{J}$, then $J \subseteq \hat{J}$, $x = \hat{x}$ on J imply that $J = \hat{J}$.

and (29) has a *unique* solution if it has one and only one maximal solution. For the initial value problem (28), (29), the Carathéodory kind of conditions on existence and uniqueness of solutions as 1 and 2 above for (30) and (31) is still true. More precisely we have the following proposition.

PROPOSITION 2. *Let U be a function in $BVC([0,T],\mathbb{R}^m)$ and $\bar{x}$ be a point in $\mathbb{R}^n$. Then the initial value problem* (28), (29) *has solutions if the vector fields f_k are continuous on $\mathbb{R}^n$. It has a unique solution if the f_k are locally Lipschitz continuous.*

The proof of this proposition is similar to the usual case when all the U_k are absolutely continuous.

As we said in the introduction, an extended input $\mathbf{V}$ of finite order can be thought of as an object that contains "slots" (the indeterminates X_k) where vector fields can be plugged in, giving rise to a differential equation. (The only requirement for this is that the vector fields should be smooth enough such that the various brackets in $\mathbf{V}$ exist.) For example, the extended input $\mathbf{V} = V_1(t)X_1 + V_2(t)X_2 + V_3(t)[X_1, X_2]$ gives rise, for each choice of C^1 vector fields $f_1, \ldots, f_m$ on $\mathbb{R}^n$, to the differential equation

$$dx = f_1(x)dV_1(t) + f_2(x)dV_2(t) + [f_1, f_2](x)dV_3(t).$$

More precisely, let $\mathbf{f} = (f_1, \ldots, f_m)$ be an m-tuple of vector fields on $\mathbb{R}^n$. We say that $\mathbf{f}$ is of *class* C^κ if all the f_ℓ are of class C^κ. Let $\mathbf{V} = \sum_{0<|I|\le r} V_I[X_I]$ be an extended input of order r on $[0,T]$. Let $\mathbf{f} = (f_1, \ldots, f_m)$ be a system of vector fields of class C^{r-1} on $\mathbb{R}^n$ and $\bar{x}$ be a point in $\mathbb{R}^n$. Then we can consider the following initial value problem:

$$dx = \sum_{0<|I|\le r} [f_I](x)dV_I(t), \tag{32}$$

$$x(0) = \bar{x}. \tag{33}$$

(Here we write

$$[f_I] \stackrel{\text{def}}{=} [f_{i_1}, [f_{i_2}, [\cdots, [f_{i_{k-1}}, f_{i_k}]\cdots]$$

for $I = (i_1, \ldots, i_k)$.) We will call (32), (33) the initial value problem determined by the triple $(\mathbf{f}, \mathbf{V}, \bar{x})$ and use $IVP(\mathbf{f}, \mathbf{V}, \bar{x})$ to denote it. We say that an $IVP(\mathbf{f}, \mathbf{V}, \bar{x})$ has the *uniqueness property* (UP for short), if it has a unique solution.

DEFINITION 9. *Let ξ be an $\mathbb{R}^n$-valued function whose domain is a subinterval I of $\mathbb{R}$. Let $\{\xi^j\}$ be a sequence of $\mathbb{R}^n$-valued functions. Let I^j be the domain of ξ^j, and assume that each I^j is a subinterval of $\mathbb{R}$. We say that $\{\xi^j\}$ converges to ξ on compact sets if for every compact subset $\tilde{I} \subseteq I$ there exists a J such that $\tilde{I} \subseteq I^j$ for $j \ge J$ and $\{\xi^j\}$ converges to ξ uniformly on $\tilde{I}$.*

DEFINITION 10. *Let $\mathbf{V}$ be an extended input of order $\le r$. A sequence $\{\mathbf{U}^j\}$ of ordinary inputs EI(r)-converges to $\mathbf{V}$ if for every integer $n > 0$, every point $\bar{x} \in \mathbb{R}^n$, every sequence $\{\bar{x}^j\} \subseteq \mathbb{R}^n$ that converges to $\bar{x}$, and every $\mathbf{f} = (f_1, \ldots, f_m)$ of class C^{r-1} on $\mathbb{R}^n$, the following holds. For each j let x^j be a maximal solution of the $IVP(\mathbf{f}, \mathbf{U}^j, \bar{x}^j)$. For every subsequence $\{x^{j(k)}\}$ of $\{x^j\}$, there exist a maximal solution x of the $IVP(\mathbf{f}, \mathbf{V}, \bar{x})$ and a further subsequence $\{x^{j(k(\ell))}\}$ of $\{x^{j(k)}\}$ such that the $x^{j(k(\ell))}$ converge to x on compact sets.*

The following proposition gives a stronger convergence result for the case when an $IVP(\mathbf{f}, \mathbf{V}, \bar{x})$ has UP.

PROPOSITION 3. *Assume that a sequence $\{\mathbf{U}^j\}$ of ordinary inputs EI(r)-converges to an extended input $\mathbf{V}$ of order $\le r$. Let $\bar{x} \in \mathbb{R}^n$ be a point and $\mathbf{f}$ be a system of*

class C^{r-1} on $\mathbb{R}^n$ such that the IVP$(\mathbf{f}, \mathbf{V}, \bar{x})$ has the UP. Let x be the unique maximal solution of the IVP$(\mathbf{f}, \mathbf{V}, \bar{x})$. Let $\{\bar{x}^j\}$ be a sequence of points in $\mathbb{R}^n$ that converges to $\bar{x}$. Let x^j be a maximal solution of the IVP$(\mathbf{f}, \mathbf{U}^j, \bar{x}^j)$. Then the x^j converge to x on compact sets.

The proof of the proposition follows directly from the definition of EI(r)-convergence.

With these preliminaries, we are ready to state the main convergence theorem of this section.

THEOREM 1. *Let $\{\mathbf{U}^j = U_1^j X_1 + \cdots + U_m^j X_m\}$ be a sequence of ordinary inputs. Let r be a positive integer, and $\mathbf{V} = \sum_{0<|I|\le r} V_I X_I$ be a polynomial input of order $\le r$. Let $\{\mathbf{V}^j\}$ be a sequence of polynomial inputs of order $\le r$ that converges to $\mathbf{V}$ strongly. Assume that there exist rth-order truncated generalized differences $(\mathbf{U}^j \overset{\text{g.d.(r)}}{-} \mathbf{V}^j) = \sum_{0<|I|\le r} \widetilde{UV}_I^j X_I$ of $\mathbf{U}^j$ and $\mathbf{V}^j$ such that*

(c1(r)) *the $(\mathbf{U}^j \overset{\text{g.d.(r)}}{-} \mathbf{V}^j)$ converge to 0 uniformly as $j \to \infty$;*

(c2(r)) *the total variations $TV[\widetilde{UV}_I^j; 0, T]$ of the $\widetilde{UV}_I^j$ are uniformly bounded for $|I| = r$.*

Then

(C1) $\mathbf{V}$ *is an extended input of order $\le r$, i.e.,*

$$\mathbf{V} = \sum_{0<|I|\le r} \frac{V_I}{|I|}[X_I];$$

(C2) *the $\mathbf{U}^j$ EI(r)-converge to $\mathbf{V}$.*

Remark 8. Condition (c1(r)) says that $\{\mathbf{U}^j, \mathbf{V}^j\}$ GD(r)-converges to 0, which is equivalent to the FT(r)-convergence of $\{\mathbf{U}^j\}$ to $\mathbf{V}$, i.e., the convergence of the trajectories of $\mathbf{U}^j$ to that of $\mathbf{V}$ for one particular initial value problem $\mathbf{U}(r)$, namely, the rth-order truncated formal equation (22) with initial condition (23). Therefore this initial value problem plays the "universal" role for the rth-order convergence. Theorem 1 simply says that, if the trajectories of $\mathbf{U}^j$ converge to that of $\mathbf{V}$ for the initial value problem $\mathbf{U}(r)$, then they converge for every problem with sufficiently smooth vector fields f_k, provided that a boundedness condition (c2(r)) holds to prevent the occurrence of brackets of higher order.

Proof of Theorem 1. From Proposition 1 we know that, under condition (c1(r)), $\mathbf{V}$ is an extended input of order $\le r$. We show (C2). Let $n > 0$ be an integer. Let $\bar{x}$ be a point in $\mathbb{R}^n$ and $\{\bar{x}^j\} \subseteq \mathbb{R}^n$ be a sequence of points that converges to $\bar{x}$ as $j \to \infty$. Let $\mathbf{f} = (f_1, \ldots, f_m)$ be a system of vector fields of class C^{r-1} on $\mathbb{R}^n$. For each j, let x^j be a maximal solution of the IVP$(\mathbf{f}, \mathbf{U}^j, \bar{x}^j)$ with domain I^j.

Let $\{K^\ell\}_{\ell=1}^\infty$ be a sequence of compact subsets of $\mathbb{R}^n$ such that (1) $\bar{x} \in \mathrm{Int}(K^1)$, (2) $K^j \subseteq \mathrm{Int}(K^{j+1})$, and (3) $\cup_j K^j = \mathbb{R}^n$. (Here $\mathrm{Int}(K)$ denotes the interior of K.) Since the $\bar{x}^j \to \bar{x}$, we may assume that all the $\bar{x}^j$ are contained in the interior of K^1. Now for each integer $\ell > 0$, let $a_\ell^j = \sup\{a : x^j(t) \in K^\ell \text{ for } t \in [0, a]\}$. Then for each j, ℓ, either $a_\ell^j = T$ or $x^j(a_\ell^j) \in \partial K^\ell$, the boundary of K^ℓ. Let x_ℓ^j denote the restriction $x^j|_{[0,a_\ell^j]}$ of x^j to $[0, a_\ell^j]$. Our next step is to show that for each $\ell > 0$, the sequence $\{x_\ell^j\}_{j=1}^\infty$ of functions is equicontinuous. Let's fix an $\ell > 0$. Let θ be a smooth real-valued function on $\mathbb{R}^n$ with compact support such that $\theta(x) = 1$ on K^ℓ. Let $\tilde{\mathbf{f}} = (\theta f_1, \ldots, \theta f_m)$. Then x_ℓ^j is a solution of the IVP$(\tilde{\mathbf{f}}, \mathbf{U}^j, \bar{x}^j)$. Let $\tilde{x}^j$ be a maximal solution of the IVP$(\tilde{\mathbf{f}}, \mathbf{U}^j, \bar{x}^j)$ that extends x_ℓ^j. Then $\tilde{x}^j$ is defined on

$[0,T]$. If we can show that $\{\tilde{x}^j\}_{j=1}^{\infty}$ is equicontinuous, then the $\{x_\ell^j\}_{j=1}^{\infty}$ would be also equicontinuous. So in order to show that the $\{x_\ell^j\}_{j=1}^{\infty}$ is equicontinuous, we may without loss of generality assume that the f_i are compactly supported and the x^j are defined on $[0,T]$, and show that $\{x^j\}$ is equicontinuous.

By definition, each x^j satisfies

$$dx = \sum_{k=1}^{m} f_k(x)\, dU_k^j, \tag{34}$$

$$x(0) = \bar{x}^j. \tag{35}$$

Let $V^j = \sum_{0<|I|\le r} V_I^j X_I$ be representatives of $\mathbf{V}^j$ that converge to $V = \sum_{0<|I|\le r} V_I X_I$ uniformly and the $TV[V_I^j;0,T]$ are uniformly bounded. Then the functions U_ℓ^j, V_I^j and $\widetilde{UV}_I^j$ are related by (18). Let φ be a smooth function from $\mathbb{R}^n$ to $\mathbb{R}$. From (34) and (35), by repeated integration by parts (cf. the proof of Proposition 1), we have

$$\varphi(x^j(t)) = \varphi(\bar{x}^j) + \sum_{0<|I|\le r} \int_0^t (f_I\varphi)(x^j(s))\, dV_I^j(s) - \sum_{|I|=r} \int_0^t (f_I\varphi)(x^j(s))\, d\widetilde{UV}_I^j(s)$$

$$- \sum_{0<|I|<r} \Big\{ \widetilde{UV}_I^j(t)\,(f_I\varphi)(x^j(t)) - \widetilde{UV}_I^j(0)(f_I\varphi)(\bar{x}^j) \Big\}, \tag{36}$$

where $(f_I\varphi)$ denotes the function $(f_{i_1}\cdots f_{i_k}\varphi)$ for $I=(i_1,\dots,i_k)$. (Each f_ℓ can be viewed as a differential operator on smooth real-valued functions on $\mathbb{R}^n$, so here $(f_\ell\varphi)$ is the derivative of φ in the direction of f_ℓ.) By our assumption we know that the $\widetilde{UV}_I^j$ converge to 0 uniformly for all $0<|I|\le r$ as j goes to ∞. Therefore we have

$$\varphi(x^j(t)) = \varphi(\bar{x}^j) + \sum_{0<|I|\le r} \int_0^t (f_I\varphi)(x^j(s))\, dV_I^j(s)$$

$$- \sum_{|I|=r} \int_0^t (f_I\varphi)(x^j(s))\, d\widetilde{UV}_I^j(s) \; + o(1), \tag{37}$$

where $o(1)$ denotes the term $-\sum_{0<|I|<r}\{\widetilde{UV}_I^j(t)(f_I\varphi)(x^j(t)) - \widetilde{UV}_I^j(0)(f_I\varphi)(\bar{x}^j)\}$, which converges to 0 uniformly. Equality (37) clearly holds for vector functions too. So we can apply it to the identity map from $\mathbb{R}^n$ to $\mathbb{R}^n$. Now let us still use φ to denote the identity map from $\mathbb{R}^n$ to $\mathbb{R}^n$. We get

$$x^j(t) = \bar{x}^j + \sum_{0<|I|\le r} \int_0^t (f_I\varphi)(x^j(s))\, dV_I^j(s) - \sum_{|I|=r} \int_0^t (f_I\varphi)(x^j(s))\, d\widetilde{UV}_I^j(s) + o(1). \tag{38}$$

To show that $\{x^j\}$ is equicontinuous, we need a lemma.

LEMMA 2. *Let $g_\alpha : \mathbb{R}^n \to \mathbb{R}^n$ be continuous and bounded functions, where $\alpha \in A$ and A is some finite index set. Let $\{V_\alpha^j\}$ be sequences of functions in $BVC[0,T]$ with the $TV[V_\alpha^j;0,T]$ being uniformly bounded in j for all $\alpha \in A$. Assume that $\{V_\alpha^j\}$ converges to $H_\alpha \in BVC[0,T]$ uniformly for each $\alpha \in A$. Let $\{\xi^j\} : [0,T] \to \mathbb{R}^n$ be*

a sequence of vector-valued functions which is equicontinuous and uniformly bounded. Let x^j be continuous functions on $[0,T]$ that satisfy

$$x^j(t) = \sum_{\alpha\in A} \int_0^t g_\alpha(x^j(s))\, dV_\alpha^j(s) + \xi^j(t).$$

Then $\{x^j\}$ is uniformly bounded and equicontinuous.

Proof. The sequence $\{x^j\}$ is clearly uniformly bounded. Without loss of generality we may assume that the g_α are compactly supported in $\mathbb{R}^n$. (Otherwise, let K be a compact set in $\mathbb{R}^n$ that contains all the x^j and ξ^j in its interior. Then multiply each g_α by a smooth compactly supported function on $\mathbb{R}^n$ which is equal to 1 on K.) In order to show that $\{x^j\}$ is equicontinuous, we need only show that every subsequence of $\{x^j\}$ is equicontinuous. Therefore we may further assume that $\{\xi^j\}$ converges to some continuous function ξ uniformly.

Let $K = \sup_j\{\sum_{\alpha\in A} TV[V_\alpha^j;0,T]\}$. For any given $\varepsilon > 0$, since the g_α are uniformly continuous on $\mathbb{R}^n$, there exists a $\delta > 0$ such that, for all $\alpha \in A$, $\|g_\alpha(x_2) - g_\alpha(x_1)\| < \frac{\varepsilon}{4K}$ if $\|x_2 - x_1\| < \delta$. We may assume $\delta < \varepsilon$. Choose a smooth vector-valued function η on $[0,T]$ such that $\|\eta(t) - \xi(t)\| < \delta/8$ for all $t \in [0,T]$. Let

$$y^j(t) = \sum_{\alpha\in A} \int_0^t g_\alpha(x^j(s))\, dV_\alpha^j(s) + \eta(t).$$

Then

$$\|y^j(t) - x^j(t)\| \le \|\xi^j(t) - \xi(t)\| + \frac{\delta}{8}.$$

Since the ξ^j converge to ξ uniformly, take J large enough such that $\|\xi^j - \xi\| < \frac{\delta}{8}$ if $j \ge J$. So, when $j \ge J$, $\|x^j(t) - y^j(t)\| < \frac{\delta}{4} < \frac{\varepsilon}{4}$. In order to show that $\{x^j\}$ is equicontinuous, we need only show that the sequence $\{y^j\}$ is equicontinuous. Let

$$\zeta^j(t) = \sum_{\alpha\in A} \int_0^t g_\alpha(y^j(s))\, dV_\alpha^j(s) + \eta(t).$$

Then

$$\|\zeta^j(t) - y^j(t)\| \le \sum_{\alpha\in A} \Big\| \int_0^t (g_\alpha(y^j(s)) - g_\alpha(x^j(s)))dV_\alpha^j(s)\Big\| < \frac{\varepsilon}{4}$$

for $j \ge J$. Therefore we need only to show that $\{\zeta^j\}$ is equicontinuous.

For each $\alpha \in A$, let $G_\alpha : \mathbb{R}^n \to \mathbb{R}^n$, which is of class C^1 and has compact support, be such that

$$\|G_\alpha - g_\alpha\| \stackrel{\text{def}}{=} \sup_{x\in\mathbb{R}^n} \|G_\alpha(x) - g_\alpha(x)\| < \frac{\varepsilon}{8K}.$$

Then

$$\zeta^j(t) = \sum_{\alpha\in A} \int_0^t G_\alpha(y^j(s))dV_\alpha^j(s) + \sum_{\alpha\in A} \int_0^t (g_\alpha(y^j(s)) - G_\alpha(y^j(s)))dV_\alpha^j(s) + \eta(t).$$

The second summation in the right-hand side above is clearly less than $\frac{\varepsilon}{8}$, so

$$\sum_{\alpha\in A} \Big\| \int_{t_1}^{t_2} (g_\alpha(y^j(s)) - G_\alpha(y^j(s)))dV_\alpha^j(s)\Big\| < \frac{\varepsilon}{4}$$

for any $t_1, t_2 \in [0,T]$. Now η is a fixed continuous function. So, in order to show that $\{\zeta^j\}$ is equicontinuous, all we need is to show that the sequence $\{\int_0^t G_\alpha(y^j(s))dV_\alpha^j(s)\}$ is equicontinuous for each $\alpha \in A$. We can write

$$\int_0^t G_\alpha(y^j(s))dV_\alpha^j(s) = \int_0^t G_\alpha(y^j(s))d(V_\alpha^j(s) - V_\alpha(s)) + \int_0^t G_\alpha(y^j(s))dV_\alpha(s).$$

Since $\{G_\alpha(y^j(t))\}$ is uniformly bounded and V_α is a fixed function in $BVC[0,T]$, the sequence $\{\int_0^t G_\alpha(y^j(s))dV_\alpha(s)\}$ is equicontinuous. By integration by parts we can rewrite the first integral $\int_0^t G_\alpha(y^j(s))d(V_\alpha^j(s) - V_\alpha(s))$ as

$$\begin{aligned}\int_0^t G_\alpha(y^j(s))\, d\,(V_\alpha^j(s) - V_\alpha(s)) &= G_\alpha(y^j(t))(V_\alpha^j(t) - V_\alpha(t)) - G_\alpha(y^j(0))(V_\alpha^j(0) - V_\alpha(0)) \\ &\quad - \sum_{\beta\in A}\int_0^t (V_\alpha^j(s) - V_\alpha(s))DG_\alpha(y^j(s))g_\beta(x^j(s))dV_\beta^j(s) \\ &\quad - \int_0^t (V_\alpha^j(s) - V_\alpha(s))DG_\alpha(y^j(s))\dot\eta(s)ds.\end{aligned}$$

(Here DG_α denotes the Jacobian matrix of G_α.) Since the $V_\alpha^j - V_\alpha$ converge to 0 uniformly, the $\|g_\beta\|_{\sup}$, $\|G_\alpha\|_{\sup}$, $\|DG_\alpha\|_{\sup}$, and $\|\dot\eta\|_{\sup}$ are fixed constants and the total variations $TV[V_\beta^j; 0, T]$ are uniformly bounded, we see that all the four terms above go to zero uniformly as $j \to \infty$. Therefore $\{\int_0^t G_\alpha(y^j(s))d(V_\alpha^j(s) - V_\alpha(s))\} \to 0$ uniformly, which implies that $\{\int_0^t G_\alpha(y^j(s))dV_\alpha^j(s)\}$ is equicontinuous. We then conclude that the sequence $\{\zeta^j\}$ is equicontinuous. This completes the proof of the lemma. □

Now we go back to the proof of Theorem 1. Lemma 2 implies that for each ℓ, the $\{x_\ell^j\}_{j=1}^\infty$ is equicontinuous.

Now let $\ell = 1$. Since the sequence $\{x_1^j\}_{j=1}^\infty$ is equicontinuous, there exists a subsequence $\{x_1^{j_1(k)}\}$ of $\{x_1^j\}_{j=1}^\infty$ that converges to a function x_1^∞ on some interval $[0, a_1^\infty]$ uniformly. (We say that a sequence $\{\xi^j\}_{j=1}^\infty$ of $\mathbb{R}^n$-valued functions with domains $[a^j, b^j]$ converges to a function ξ with domain $[a,b]$ uniformly if $\{a^j\}_{j=1}^\infty$ converges to a, $\{b^j\}_{j=1}^\infty$ converges to b, and, for any $t^j \in [a^j, b^j]$ such that the t^j converge to t, $\xi^j(t^j) \to \xi(t)$.) So $\{x^{j_1(k)}\}$ is a subsequence of $\{x^j\}$. Now let $\ell = 2$. Since the $\{x_2^{j_1(k)}\}$ is also equicontinuous, there exists a further subsequence $\{x^{j_2(k)}\}$ of $\{x^{j_1(k)}\}$ such that the $x_2^{j_2(k)}$ converge to a function x_2^∞ with domain $[0, a_2^\infty]$ uniformly. Continuing this way we can construct a sequence $\{x_\ell^\infty\}_{\ell=1}^\infty$ of functions and a collection $\{x^{j_\ell(k)}\}, \ell = 1, 2, \ldots$, of subsequences of $\{x^j\}$ such that

(1) the domain of x_ℓ^∞ is $[0, a_\ell^\infty]$, $a_\ell^\infty \le a_{\ell+1}^\infty$, and $x_{\ell+1}^\infty|_{[0,a_\ell^\infty]} = x_\ell^\infty$;
(2) each $\{x^{j_\ell(k)}\}$ is a subsequence of $\{x^{j_{\ell-1}(k)}\}$, and $\{x^{j_1(k)}\}$ is a subsequence of $\{x^j\}$;
(3) for each fixed ℓ, the $x_\ell^{j_\ell(k)}$ converge to x_ℓ^∞ uniformly.

Let $I = \cup_\ell [0, a_\ell^\infty]$. Then there is a well-defined function x^∞ on I given by $x^\infty(t) = x_\ell^\infty(t)$ if $t \in [0, a_\ell^\infty]$. Moreover, for each fixed ℓ, the sequence $\{x_\ell^{j_\ell(k)}\}$ converges to $x^\infty|_{[0,a_\ell^\infty]}$ uniformly. A standard diagonal argument implies that there exists a subsequence $\{x^{j(k)}\}$ of $\{x^j\}$ such that $\{x^{j(k)}\}$ is a subsequence of $\{x^{j_\ell(k)}\}$ for every ℓ. This implies that the $x^{j(k)}$ converge to x^∞ on compact sets. Now we show that x^∞ on I is a solution of the $IVP(\mathbf{f}, \mathbf{V}, \bar{x})$.

For any $t \in I$, then there exists a J large enough such that t is in the domain of $x^{j(k)}$ if $j(k) \geq J$. From (38) we have

$$\begin{aligned} x^{j(k)}(t) = \bar{x}^{j(k)} &+ \sum_{0<|I|\leq r} \int_0^t (f_I\varphi)(x^{j(k)}(s))\, dV_I^{j(k)}(s) \\ &- \sum_{|I|=r} \int_0^t (f_I\varphi)(x^{j(k)})(s))\, d\widetilde{UV}_I^{j(k)}(s) + o(1). \end{aligned} \tag{39}$$

From Lemma 1, by letting $j(k) \to \infty$ in (39), we get

$$x^\infty(t) = \bar{x} + \sum_{0<|I|\leq r} \int_0^t (f_I\varphi)(x^\infty(s))\, dV_I(s);$$

i.e., x^∞ satisfies

$$\begin{aligned} dx &= \sum_{0<|I|\leq r} (f_I\varphi)(x)\, dV_I, \\ x(0) &= \bar{x}. \end{aligned}$$

Now we need to establish that for any continuous function $\xi : [0,T] \to \mathbb{R}^n$,

$$\sum_{0<|I|\leq r} \int_0^t (f_I\varphi)(\xi(s))\, dV_I(s) = \sum_{0<|I|\leq r} \int_0^t \frac{1}{|I|}[f_I](\xi(s))\, dV_I(s). \tag{40}$$

From (C1) we know that $\mathbf{V} = \sum_{0<|I|\leq r} V_I\, X_I$ is an extended input. Therefore

$$\sum_{0<|I|\leq r} V_I\, X_I = \sum_{0<|I|\leq r} \frac{1}{|I|} V_I\, [X_I]. \tag{41}$$

Let $C^\kappa(\mathbb{R}^n)$, $\kappa \in \{0,1,2,\dots,\} \cup \{\infty\}$, be the set of real-valued C^κ functions on $\mathbb{R}^n$. Now each member f_k of $\mathbf{f} = (f_1, \dots, f_m)$ can be viewed as a differential operator from $C^\infty(\mathbb{R}^n)$ to $C^{r-1}(\mathbb{R}^n)$. For each $I = (i_1, \dots, i_k)$, $k \leq r$, the

$$f_I \stackrel{\text{def}}{=} f_{i_1} f_{i_2} \cdots f_{i_k}$$

can be viewed as a high-order differential operator that maps each $\theta \in C^\infty(\mathbb{R}^n)$ to the function $(f_I\theta) \in C^{r-k}(\mathbb{R}^n)$. Let $A_0^r(\mathbf{f})$ be the vector space of differential operators that map $C^\infty(\mathbb{R}^n)$ to $C^0(\mathbb{R}^n)$ spanned by $\{f_I, 0 < |I| \leq r\}$ over $\mathbb{R}$, i.e., $A_0^r(\mathbf{f}) = \{\sum_{0<|I|\leq r} a_I f_I : a_I \in \mathbb{R}\}$. Let $L^r(\mathbf{f})$ be the vector space of vector fields on $\mathbb{R}^n$ spanned by $\{[f_I], 0 < |I| \leq r\}$ over $\mathbb{R}$, so $L^r(\mathbf{f}) = \{\sum_{0<|I|\leq r} a_I\, [f_I] : a_I \in \mathbb{R}\}$. Then each element of $L^r(\mathbf{f})$ also gives rise to a differential operator from $C^\infty(\mathbb{R}^n)$ to $C^0(\mathbb{R}^n)$ and $L^r(\mathbf{f})$ can be viewed as a subset of $A_0^r(\mathbf{f})$. There is a well-defined *evaluation* map

$$\mathrm{Ev}(\mathbf{f}) : A_0^r(\mathbf{X}) \to A_0^r(\mathbf{f})$$

obtained by "plugging in the f_k for the X_k" so that

$$\mathrm{Ev}(\mathbf{f})(\textstyle\sum_I a_I X_I) = \sum_I a_I f_I.$$

The evaluation map $\mathrm{Ev}(\mathbf{f})$ can be restricted to $L^r(\mathbf{X})$. Let us still denote by $\mathrm{Ev}(\mathbf{f})$ the restriction of $\mathrm{Ev}(\mathbf{f})$ to $L^r(\mathbf{X})$. It is obvious that $\mathrm{Ev}(\mathbf{f})$ maps $L^r(\mathbf{X})$ onto $L^r(\mathbf{f})$. Letting $\mathrm{Ev}(\mathbf{f})$ act on (41) we get

$$\sum_{0<|I|\le r} f_I\, V_I(t) = \sum_{0<|I|\le r} \frac{1}{|I|}[f_I]\, V_I(t),$$

i.e., for any function $\theta \in C^\infty(\mathbb{R}^n)$, $x \in \mathbb{R}^n$,

$$\sum_{0<|I|\le r} (f_I\theta)(x)\, V_I(t) = \sum_{0<|I|\le r} \frac{1}{|I|}([f_I]\theta)(x)\, V_I(t). \tag{42}$$

Equality (42) also holds for vector functions too. So we can apply it to the identity map φ from $\mathbb{R}^n$ to $\mathbb{R}^n$. It is clear that $[f_I](x) = ([f_I]\varphi)(x)$. We then get

$$\sum_{0<|I|\le r} (f_I\varphi)(x)\, V_I(t) = \sum_{0<|I|\le r} \frac{1}{|I|}[f_I](x)\, V_I(t) \tag{43}$$

for any $x \in \mathbb{R}^n$ and $t \in [0,T]$. Take a sequence of partitions $\Gamma_\rho = \{0 = t_0 < t_1 < \cdots < t_\rho = t\}$ of $[0,t]$ with $|\Gamma_\rho| = \max_i\{t_{i+1} - t_i\} \to 0$ as $\rho \to \infty$. It follows by the definition of Riemann–Stieltjes integrals that

$$\begin{aligned}
\sum_{0<|I|\le r} \int_0^t (f_I\varphi)(\xi(s))dV_I(s) &= \sum_{0<|I|\le r} \lim_{\rho\to\infty} \sum_{i=0}^{\rho-1} (f_I\varphi)(\xi(t_i))(V_I(t_{i+1}) - V_I(t_i)) \\
&= \sum_{0<|I|\le r} \lim_{\rho\to\infty} \sum_{i=0}^{\rho-1} \frac{[f_I](\xi(t_i))}{|I|}(V_I(t_{i+1}) - V_I(t_i)) \\
&= \sum_{0<|I|\le r} \int_0^t \frac{[f_I](\xi(s))}{|I|} dV_I(s),
\end{aligned}$$

which establishes (40). Therefore x^∞ satisfies

$$dx = \sum_{0<|I|\le r} \frac{1}{|I|}[f_I](x)\, dV_I, \tag{44}$$

$$x(0) = \bar{x}; \tag{45}$$

i.e., x^∞ is a solution of the $IVP(\mathbf{f}, \mathbf{V}, \bar{x})$.

It remains to show that x^∞ is a maximal solution. Assume that there was an extension $\tilde{x}^\infty : \tilde{I} \to \mathbb{R}^n$ of x^∞ to a strictly large interval $\tilde{I} \subseteq [0,T]$. Let τ be a point in $\tilde{I}$ that is not contained in I. Then $\tau > a_\ell^\infty$ for all ℓ. The set $\{\tilde{x}(t) : t \in [0,\tau]\}$ is a compact subset of $\mathbb{R}^n$, so it is contained in some K^i. This implies that $x^\infty(a_\ell^\infty) \in K^i$ for all ℓ. Now for every ℓ, we know that $\{x^{j(k)}\}$ is a subsequence of $\{x^{j_\ell(k)}\}$. So the $a_\ell^{j(k)}$ converges to a_ℓ^∞ as $j(k) \to \infty$ for every ℓ. In particular $a_{i+1}^{j(k)} < \tau \le T$ for $j(k)$ large enough. By the definition of the a_ℓ^j, this implies that $x^{j(k)}(a_{i+1}^{j(k)}) \in \partial K^{i+1}$ for $j(k)$ large enough. On the other hand, the $x^{j(k)}(a_{i+1}^{j(k)})$ converge to $x^\infty(a_{i+1}^\infty) \in K^i$, so we have reached a contradiction since K^i is contained in the interior of K^{i+1}. The contradiction shows that x^∞ is a maximal solution of the $IVP(\mathbf{f}, \mathbf{V}, \bar{x})$ and the proof of Theorem 1 is therefore complete. □

Remark 9. In applications, the typical situation is that we just have the sequence $\{\mathbf{U}^j\}$. We do not know a priori if it is convergent or what its limit $\mathbf{V}$ is even if it converges. We have to find an r, sequences $\{\mathbf{V}^j\}$ and $(\mathbf{U}^j \overset{\text{g.d.(r)}}{-} \mathbf{V}^j)$ so that the conditions of Theorem 1 are satisfied. If we can find an integer r and B-continuous functions V_I^j for $0 < |I| \le r$ such that

(c1′(r)) the V_I^j converge to B-continuous functions V_I uniformly for all $0 < |I| \le r$,

(c2′(r)) the $\widetilde{UV}_I^j$ determined recursively by

$$\widetilde{UV}_i^j(t) = V_i^j(t) - U_i^j(t), \tag{46}$$

$$\widetilde{UV}_{i_1,\dots,i_k}^j(t) = V_{i_1,\dots,i_k}^j(t) - (U_{i_1}^j \overset{\leftarrow}{*} \widetilde{UV}_{i_2,\dots,i_k}^j)(t) \tag{47}$$

converge to 0 uniformly for $0 < |I| \le r$, and

(c3′(r)) the $TV[\widetilde{UV}_I^j; 0, T]$ for $|I| = r$ and the $TV[V_I^j; 0, T]$ for $0 < |I| \le r$ are uniformly bounded, then the conclusions (C1) and (C2) of Theorem 1 hold, and the EI(r)-limit of the $\mathbf{U}^j$ is equal to $\mathbf{V} = \sum_{0<|I|\le r} V_I(t)X_I$. To show how Theorem 1 can be applied, we consider the following example. We consider the case when there are two inputs, the $\mathbf{U}^j = U_1^j X_1 + U_2^j X_2$ are absolutely continuous, and $\mathbf{u}^j = \dot{\mathbf{U}}^j = u_1^j X_1 + u_2^j X_2$ are given (which is typical in applications).

Example. Consider the ordinary input sequence $\mathbf{u}^j = u_1^j X_1 + u_2^j X_2$ with

$$u_1^j(t) = \eta_1(t) + j^{\frac{2}{3}} \cos jt,$$
$$u_2^j(t) = \eta_2(t) + j^{\frac{2}{3}} \eta_3(t) \cos 2jt,$$

where the η_i are functions of class C^1 on $[0, T]$. Using Theorem 1 we can show that $\{u^j\}$ converges to some $\mathbf{V}$, and we can find $\mathbf{V}$ explicitly.

As the first step, we let

$$U_1^j(t) = \int_0^t u_1^j(s)ds = \int_0^t \eta_1(s)\, ds + j^{-\frac{1}{3}} \sin jt,$$
$$U_2^j(t) = \int_0^t u_2^j(s)ds = \int_0^t \eta_2(s)\, ds + \frac{j^{-\frac{1}{3}}}{2}\eta_3(t) \sin 2jt - \frac{j^{-\frac{1}{3}}}{2} \int_0^t \eta_3'(s) \sin 2js\, ds.$$

Letting

$$V_1^j(t) = \int_0^t \eta_1(s)\, ds,$$
$$V_2^j(t) = \int_0^t \eta_2(s)\, ds - \frac{j^{-\frac{1}{3}}}{2} \int_0^t \eta_3'(s) \sin 2js\, ds$$

and using (46) and (47) we have

$$\widetilde{UV}_1^j(t) = -j^{-\frac{1}{3}} \sin jt, \quad \widetilde{UV}_2^j(t) = -\frac{j^{-\frac{1}{3}}}{2}\eta_3(t) \sin 2jt.$$

Now it is easily computed that

$$\int_0^t u_1^j(s)\widetilde{UV}_1^j(s)\, ds = \int_0^t \eta_1(s)\widetilde{UV}_1^j(s)ds + \frac{j^{-\frac{2}{3}}}{4}[\cos 2jt - 1],$$

$$\int_0^t u_1^j(s)\widetilde{UV}_2^j(s)\,ds = \int_0^t \eta_1(s)\widetilde{UV}_2^j(s)ds + \frac{j^{-\frac{2}{3}}\eta_3(t)}{4}\left[\frac{1}{3}\cos 3jt + \cos jt\right]$$
$$-\frac{1}{3}j^{-\frac{2}{3}}\eta_3(0) - \frac{j^{-\frac{2}{3}}}{4}\int_0^t \eta_3'(s)\Big[\frac{1}{3}\cos 3js + \cos js\Big]ds,$$
$$\int_0^t u_2^j(s)\widetilde{UV}_1^j(s)\,ds = \int_0^t \eta_2(s)\widetilde{UV}_1^j(s)ds + \frac{j^{-\frac{2}{3}}\eta_3(t)}{2}\Big[\frac{1}{3}\cos 3jt - \cos jt\Big]$$
$$+\frac{1}{3}j^{-\frac{2}{3}}\eta_3(0) - \frac{j^{-\frac{2}{3}}}{2}\int_0^t \eta_3'(s)\Big[\frac{1}{3}\cos 3js - \cos js\Big]ds,$$

$$\int_0^t u_2^j(s)\widetilde{UV}_2^j(s)\,ds = \int_0^t \eta_2(s)\widetilde{UV}_2^j(s)ds + \frac{j^{-\frac{2}{3}}}{16}\Big[\eta_3^2(t)\cos 4jt - \eta_3^2(0)\Big]$$
$$-\frac{j^{-\frac{2}{3}}}{8}\int_0^t \eta_3(s)\eta_3'(s)\cos 4js\,ds.$$

So if we let

$$V_{1,1}^j(t) = \int_0^t \eta_1(s)\widetilde{UV}_1^j(s)ds - \frac{j^{-\frac{2}{3}}}{4},$$
$$V_{1,2}^j(t) = \int_0^t \eta_1(s)\widetilde{UV}_2^j(s)ds - \frac{1}{3}j^{-\frac{2}{3}}\eta_3(0) - \frac{j^{-\frac{2}{3}}}{4}\int_0^t \eta_3'(s)\Big[\frac{1}{3}\cos 3js + \cos js\Big]ds,$$
$$V_{2,1}^j(t) = \int_0^t \eta_2(s)\widetilde{UV}_1^j(s)ds + \frac{1}{3}j^{-\frac{2}{3}}\eta_3(0) - \frac{j^{-\frac{2}{3}}}{2}\int_0^t \eta_3'(s)\Big[\frac{1}{3}\cos 3js - \cos js\Big]ds,$$
$$V_{2,2}^j(t) = \int_0^t \eta_2(s)\widetilde{UV}_2^j(s)ds - \frac{j^{-\frac{2}{3}}\eta_3^2(0)}{16} - \frac{j^{-\frac{2}{3}}}{8}\int_0^t \eta_3(s)\eta_3'(s)\cos 4js\,ds,$$

then we get

$$\widetilde{UV}_{1,1}^j(t) = -\frac{j^{-\frac{2}{3}}}{4}\cos 2jt,$$
$$\widetilde{UV}_{1,2}^j(t) = -\frac{j^{-\frac{2}{3}}\eta_3(t)}{4}\Big[\frac{1}{3}\cos 3jt + \cos jt\Big],$$
$$\widetilde{UV}_{2,1}^j(t) = -\frac{j^{-\frac{2}{3}}\eta_3(t)}{2}\Big[\frac{1}{3}\cos 3jt - \cos jt\Big],$$
$$\widetilde{UV}_{2,2}^j(t) = -\frac{j^{-\frac{2}{3}}\eta_3^2(t)}{16}\cos 4jt.$$

There are eight indices of degree 3, namely, $(1,1,1), (1,1,2), (1,2,1), (1,2,2), (2,1,1), (2,1,2), (2,2,1)$, and $(2,2,2)$. We simply let $V_{i_1,i_2,i_3}^j(t) = \int_0^t u_{i_1}^j(s)\widetilde{UV}_{i_2,i_3}^j(s)ds$ so that $\widetilde{UV}_{i_1,i_2,i_3}^j(t) \equiv 0$. The V_{i_1,i_2,i_3}^j can be computed explicitly. We just compute $V_{1,1,2}^j$. By definition we have

$$V_{1,1,2}^j(t) = \int_0^t u_1^j(s)\widetilde{UV}_{1,2}^j(s)$$
$$= \int_0^t \eta_1(s)\widetilde{UV}_{1,2}^j(s)\,ds - \frac{1}{4}\int_0^t \eta_3(s)\Big[\frac{1}{3}\cos js\cos 3js + \cos^2 js\Big]ds$$
$$= -\frac{1}{8}\int_0^t \eta_3(s)\,ds + o(1),$$

where $o(1)$ denotes the terms that converge to 0 uniformly as $j \to \infty$. Similarly one can compute all the other $V^j_{i_1,i_2,i_3}$.

From the definition of the V^j_I we can get that

$$\lim_{j\to\infty} V^j_1(t) = \int_0^t \eta_1(s)\,ds,\ \lim_{j\to\infty} V^j_2(t) = \int_0^t \eta_2(s)\,ds,\ \lim_{j\to\infty} V^j_{1,1,2}(t) = -\frac{1}{8}\int_0^t \eta_3(s)ds,$$

$$\lim_{j\to\infty} V^j_{1,2,1}(t) = \frac{1}{4}\int_0^t \eta_3(s)ds,\ \lim_{j\to\infty} V^j_{2,1,1}(t) = -\frac{1}{8}\int_0^t \eta_3(s)ds,$$

and all the other V^j_I for $0 < |I| \le 3$ converge to 0 uniformly. It is also clear that the $\widetilde{UV}^j_I$ and V^j_I satisfy conditions (c2′(r)) and (c3′(r)). Therefore if we let

$$\begin{aligned}\mathbf{V} &= \int_0^t \eta_1(s)ds\,X_1 + \int_0^t \eta_2(s)ds\,X_2 - \frac{1}{8}\int_0^t \eta_3(s)ds\ [X_1X_1X_2 - 2X_1X_2X_1 + X_2X_1X_1]\\ &= \int_0^t \eta_1(s)ds\,X_1 + \int_0^t \eta_2(s)ds\,X_2 - \frac{1}{8}\int_0^t \eta_3(s)\,ds\,[X_1,[X_1,X_2]],\end{aligned}$$

from Theorem 1 we can conclude that the $\mathbf{u}^j$ EI(3)-converge to $\mathbf{V}$.

In control theory, one often considers systems with a *drift* term, i.e., systems of the form

$$\dot{x} = f_0(x) + \sum_{k=1}^m u_k(t) f_k(x).$$

Since such systems arise frequently in applications, we state explicitly the result generalizing Theorem 1 to systems with a drift. First we give a definition similar to EI(r)-convergence.

DEFINITION 11. *We say that a sequence $\{\mathbf{U}^j = \sum_{i=1}^m U^j_i X_i\}$ of ordinary inputs EI′(r)-converges to an extended input $\mathbf{V} = \sum_{0<|I|\le r} V_I[X_I]$ of order $\le r$ if the following holds. For every integer $n > 0$, any vector field f_0 of class C^0, and vector fields f_k, $k = 1,\dots,m$, of class C^{r-1} on $\mathbb{R}^n$, if*

(1) *$\{\mathbf{U}^{j(k)}\}$ is a subsequence of $\{\mathbf{U}^j\}$,*

(2) *$\{\bar{x}^{j(k)}\}$ is a sequence of points in $\mathbb{R}^n$ that converges to a limit $\bar{x} \in \mathbb{R}^n$,*

(3) *$x^{j(k)}$ is a maximal solution of*

$$\begin{aligned} dx &= f_0(x)dt + \sum_{i=1}^m f_i(x)dU_i^{j(k)}(t),\\ x(0) &= \bar{x}^{j(k)},\end{aligned}$$

then there exist a maximal solution x^∞ of

$$dx = f_0(x)dt + \sum_{0<|I|\le r} [f_I](x)dV_I(t), \tag{48}$$

$$x(0) = \bar{x}, \tag{49}$$

and a subsequence $\{x^{j(k(\ell))}\}$ of $\{x^{j(k)}\}$ that converges to x^∞ on compact sets.

Then similar to Theorem 1, we have the following theorem.

THEOREM 2. *Let $\{\mathbf{U}^j\}$, r, and $\mathbf{V}$ be as in Theorem 1. Then under the same conditions of Theorem 1 the $\{\mathbf{U}^j\}$ EI′(r)-converge to $\mathbf{V}$.*

One can also consider time-varying systems

$$dx = f_0(x,t)dt + \sum_{k=1}^{m} f_k(x,t)\, dU_k \tag{50}$$

with a drift, where $f_k : \mathbb{R}^n \times [0,T] \to \mathbb{R}^n$, $k = 0, 1, \dots m$, are time-varying vector fields. Let $\mathbf{f} = (f_1, \dots, f_m)$ be an m-tuple of time-varying vector fields on $\mathbb{R}^n$. We say that $\mathbf{f} \in C_t^{r-1}(\mathbb{R}^n)$ if for any smooth function ψ from $\mathbb{R}^n$ to $\mathbb{R}$, and for any $I = (i_1, \dots, i_k)$ with $|I| < r$, $(f_{i_1} \cdots f_{i_k}\psi)(x,t)$ is of class C^1 on $\mathbb{R}^n \times [0,T]$. Then one can similarly prove the following.

THEOREM 3. *Let $\{\mathbf{U}^j\}$, r, and $\mathbf{V}$ be as in Theorem* 1. *Let $f_k : \mathbb{R}^n \times [0,T] \to \mathbb{R}^n, k = 0, 1, \dots, m$, be time-varying vector fields on $\mathbb{R}^n$ and let $\mathbf{f} = (f_1, \dots, f_m)$. Assume that $\mathbf{f} \in C_t^{r-1}(\mathbb{R}^n)$ and f_0 is continuous. Let $\bar{x} \in \mathbb{R}^n$ be a point and $\{\bar{x}^j\}$ be a sequence of points in $\mathbb{R}^n$ converging to $\bar{x}$. Then under the same conditions of Theorem* 1 *the following holds. Let x^j be a maximal solution of* (50) *corresponding to inputs $\mathbf{U}^j$ with initial values $x(0) = \bar{x}^j$. Then for any subsequence $\{x^{j(k)}\}$ of $\{x^j\}$, there exist a maximal solution x^∞ of*

$$dx = f_0(x,t)\, dt + \sum_{0<|I|\le r} \frac{1}{|I|}[f_I](x,t)\, dV_I, \qquad x(0) = \bar{x}, \tag{51}$$

and a further subsequence $\{x^{j(k(\ell))}\}$ of $\{x^{j(k)}\}$ such that the $x^{j(k(\ell))}$ converge to x^∞ on compact sets.

It is easy to see that the main theorems of [5], [7] are consequences of Theorem 3.

Remark 10. We point out that one can weaken the regularity assumptions on the vector fields $f_1, \dots, f_m$ in the definition of EI(r)-convergence in many ways. To see how this can be done, cf. [16].

4. The necessity of the conditions of Theorem 2. When $r = 1$, Theorem 2 reduces to the following theorem discussed in [13].

THEOREM 4. *Let $\{\mathbf{U}^j\}$, $j \in \{1, 2, \dots\} \cup \{\infty\}$, be a sequence of ordinary inputs. Assume that there exist representatives $U^j = U_1^j X_1 + \dots + U_m^j X_m$ of the $\mathbf{U}^j$ such that*

(c1) *the U^j converge to U^∞ uniformly on $[0,T]$,*

(c2) *the total variations $TV[U_i^j; 0, T]$ are uniformly bounded; i.e., there exists a finite constant C such that $TV[U_i^j; 0, T] \le C$ for $i = 1, \dots, m$ and all j.*

Then $\{\mathbf{U}^j\}$ EI$'$(1)-converges to $\mathbf{U}^\infty$.

As point out in [13], conditions (c1), (c2) are also necessary for $\{\mathbf{U}^j\}$ to EI$'$(1)-converge to $\mathbf{U}^\infty$; namely, the following theorem holds.

THEOREM 5. *Let $\{\mathbf{U}^j\}$, $j \in \{1, 2, \dots\} \cup \{\infty\}$, be a sequence of ordinary inputs. Then the $\mathbf{U}^j$ EI$'$(1)-converge to $\mathbf{U}^\infty$ iff there are representatives $U^j = U_1^j X_1 + \cdots + U_m^j X_m$ of the $\mathbf{U}^j$ that satisfy conditions* (c1) *and* (c2) *in Theorem* 4.

This theorem is proven in [13] for the case when all the U_i^j, U_i^∞ are absolutely continuous. For completeness we include the proof here since it is very simple anyhow.

Proof of Theorem 5. Assume that the $\mathbf{U}^j$ EI$'$(1)-converge to $\mathbf{U}^\infty$. For each j let $U^j = U_1^j X_1 + \cdots + U_m^j X_m$ be a representative of $\mathbf{U}^j$ with $U_i^j(0) = 0$ for $i = 1, \dots, m$. Condition (c1) follows from applying the EI$'$(1)-convergence of the $\mathbf{U}^j$ to $\mathbf{U}^\infty$ with $n = m$, $\bar{x}^j = \bar{x} = (0, \dots, 0)$, $f_0(x) = (0, \dots, 0)$, $f_1 = (1, 0, \dots, 0), \dots, f_m = (0, \dots, 0, 1)$.

To show (c2), let φ be a function in $C[0,T]$, the set of all real-valued continuous functions on $[0,T]$. Let $n = 2$, and fix $1 \le \ell \le m$. For any $x = (x_1, x_2) \in \mathbb{R}^2$, let

$f_0(x) = (1, 0), f_i(x) = (0, 0)$ except for $i = \ell$, and $f_\ell(x) = (0, \varphi(x_1))$. For $1 \le j \le \infty$, let x^j be the unique maximal solution of

$$dx = f_0(x)dt + \sum_{i=1}^{m} f_i(x)dU_i^j(t), \quad x(0) = (0, 0).$$

Then we have $x_1^j(t) = t$, and $x_2^j(t) = \int_0^t \varphi(s)dU_\ell^j(s)$. Since the $\mathbf{U}^j$ $\mathrm{EI}'(1)$-converge to $\mathbf{U}^\infty$, we know that the functions $x_2^j(t) = \int_0^t \varphi(s)dU_\ell^j(s)$ converge to $x_2^\infty(t) = \int_0^t \varphi(s)dU_\ell^\infty(s)$ uniformly. In particular the integrals $\int_0^T \varphi(s)dU_\ell^j(s)$ converge to $\int_0^T \varphi(s)dU_\ell^j(s)$. Now each U_ℓ^j can be viewed as a linear functional on the Banach space $C[0, T]$ given by $\langle U_\ell^j, \varphi\rangle = \int_0^T \varphi(s)dU_\ell^j(s)$. It is well known that if U is any function in $BVC([0, T], \mathbb{R})$, the norm of the linear functional $\varphi \longrightarrow \int_0^T \varphi(s)dU(s)$ is equal to the total variation of U on $[0, T]$, i.e., $TV[U; 0, T]$. Now the above implies that for each φ, the set $\{\langle U_\ell^j, \varphi\rangle : j = 1, 2, \ldots\}$ is bounded. By the Banach–Steinhaus theorem we know that the set $\{TV[U_\ell^j; 0, T] : j = 1, 2, \ldots\}$ is also bounded. Since ℓ is arbitrary between 1 and m, we know that (c2) holds. ∎

It is easy to see that condition (c1(r)) is still necessary for a sequence $\{\mathbf{U}^j\}$ to $\mathrm{EI}'(r)$-converge to an extended input $\mathbf{V}$ of order r; namely, if a sequence $\{\mathbf{U}^j\}$ of ordinary inputs $\mathrm{EI}'(r)$-converges to an extended $\mathbf{V}$ of order r, then $\{\mathbf{U}^j\}$ $\mathrm{FT}(r)$-converges to $\mathbf{V}$ (cf. Remark 8). To see this, let $F_k^r, k = 0, \ldots, m$, be the linear vector fields on $A^r(\mathbf{X})$ given by $F_0^r(S) = 0, F_k^r(S) = SX_k$ for $k = 1, \ldots, m$. Since the $\mathbf{U}^j$ $\mathrm{EI}'(r)$-converge to $\mathbf{V}$, in particular the solutions of the initial value problems

$$dS = F_0^r(S)dt + \sum_{k=1}^{m} F_k^r(S)dU_k^j,$$
$$S(0) = 1$$

converge to the function S that satisfies $dS = Sd\mathbf{V}, S(0) = 1$ on $[0, T]$ in $A^r(\mathbf{X})$, which is equivalent to the $\mathrm{FT}(r)$-convergence of the $\mathbf{U}^j$ to $\mathbf{V}$. However, condition (c2(r)) may not be necessary anymore. An interesting problem is of course to get simple characterizations for sequences of ordinary inputs to $\mathrm{EI}'(r)$-converge to extended inputs of order r for $r \ge 2$.

Appendix: Proof of Lemma 1 and an auxiliary result. We now complete the proof of Lemma 1 and give an approximation result of B-continuous functions by smooth functions (in the weak convergence sense), which was used in section 2. This result is well known, but we provide a proof for completeness.

Proof of Lemma 1. Let $\tilde{g}^j = g^j - g$ and $K = \sup_j\{TV[g^j; 0, T]\}$. Our assumptions imply that f is continuous and g is B-continuous with $TV[g; 0, T] \le K$. We have

$$\Big|\int_0^t f^j(s)\,dg^j(s)\Big| \le \Big|\int_0^t (f^j(s) - f(s))\,dg^j(s)\Big| + \Big|\int_0^t f(s)\,dg^j(s)\Big|$$
$$\le K|f^j - f|_\infty + \Big|\int_0^t f(s)\,dg^j(s)\Big|.$$

(Here $|f^j - f|_\infty$ denotes the sup-norm of the function $f^j - f$ on $[0, T]$.) Since the f^j converge to f uniformly, the $K|f^j - f|_\infty$ converge to 0. The integrals $\int_0^t f(s)\,dg^j(s)$ can be rewritten as $\int_0^t f(s)d\tilde{g}^j(s) + \int_0^t f(s)dg(s)$. So all we need is to show that the $\int_0^t f(s)d\tilde{g}^j(s)$ converge to 0 uniformly as $j \to \infty$. For any given $\varepsilon > 0$, let $\tilde{f}$ be a

smooth function on $[0,T]$ such that $|f-\tilde{f}|_\infty < \frac{\varepsilon}{4K}$. Then

$$\Big|\int_0^t f(s)d\tilde{g}^j(s)\Big| \le \Big|\int_0^t (f(s)-\tilde{f}(s))d\tilde{g}^j(s)\Big| + \Big|\int_0^t \tilde{f}(s)d\tilde{g}^j(s)\Big|.$$

The first term in the right-hand side is less than $\frac{\varepsilon}{2}$. The second integral can be rewritten via integration by parts as

$$\int_0^t \tilde{f}(s)d\tilde{g}^j(s) = \tilde{f}(t)\tilde{g}^j(t) - \tilde{f}(0)\tilde{g}^j(0) - \int_0^t \dot{\tilde{f}}(s)\tilde{g}^j(s)ds.$$

Since the $\tilde{g}^j$ converge to 0 uniformly, we see that the $\int_0^t \tilde{f}(s)d\tilde{g}^j(s)$ converge to 0 uniformly as $j\to\infty$. This completes the proof of the lemma. ∎

PROPOSITION 4. *Let V be a function in $BVC[0,T]$. Then there exists a sequence $\{v_n\}$ of smooth functions on $[0,T]$ with the $\|v_n\|_{L^1}$ being uniformly bounded such that, for any $g\in C[0,T]$,*

$$\lim_{n\to\infty}\int_0^t g(s)v_n(s)\,ds = \int_0^t g(s)\,dV(s)$$

uniformly in t.

Proof. Without lost of generality we may assume that $V\in BVC[0,T]$ is increasing and $V(0)=0$. Divide $[0,T]$ into 2^n equal parts. Let $\Gamma_n = \{0=t_0^n < t_1^n < \cdots < t_{2^n}^n = T\}$ be the partition. Let $\tilde{v}_n$ be functions on $[0,T]$ defined $\tilde{v}_n(t) = \frac{V(t_l^n)-V(t_{l-1}^n)}{t_l^n - t_{l-1}^n}$ on $[t_{l-1}^n, t_l^n)$ and $\tilde{v}_n(T) = V(T)$. Clearly $\int_0^T |\tilde{v}_n(s)|\,ds = V(T)$. Let $g\in C[0,T]$ and $M = \max_{0\le t\le T}|g(t)|$. We first show that

$$\lim_{n\to\infty}\int_0^t g(s)\tilde{v}_n(s)\,ds = \int_0^t g(s)\,dV(s)$$

uniformly. Given any $\varepsilon>0$, since V is continuous, there exists a $\delta>0$ such that $|V(t_2)-V(t_1)| < \frac{\varepsilon}{3M}$ if $|t_2-t_1|<\delta$. Let m_l^n, M_l^n be the minimum and the maximum values of g on $[t_{l-1}^n, t_l^n]$, respectively. Let

$$s_{\Gamma_n} = \sum_{l=1}^{2^n} m_l^n(V(t_l^n)-V(t_{l-1}^n)),$$
$$S_{\Gamma_n} = \sum_{l=1}^{2^n} M_l^n(V(t_l^n)-V(t_{l-1}^n)).$$

Then we know that

$$0 \le S_{\Gamma_n} - s_{\Gamma_n} \to 0 \quad \text{as } n\to\infty.$$

For the given $\varepsilon>0$, take N_1 large enough such that $0\le S_{\Gamma_n}-s_{\Gamma_n} < \frac{\varepsilon}{3}$ when $n>N_1$. Let N be such that $\frac{T}{2^N} < \delta$ and $N>N_1$. For any $t\in[0,T]$, let $t(n)$ be the integer such that $t\in[t_{t(n)}^n, t_{t(n)+1}^n)$. Then, via the mean value theorem, we have

$$\int_0^t g(s)\tilde{v}^n(s)\,ds = \sum_{l=1}^{t(n)}\int_{t_{l-1}^n}^{t_l^n} g(s)\tilde{v}_n(s)\,ds + \int_{t_{t(n)}^n}^t g(s)\tilde{v}_n(s)\,ds$$
$$= \sum_{l=1}^{t(n)} g(\xi_l^n)(V(t_l^n)-V(t_{l-1}^n)) + \frac{V(t_{t(n)+1}^n)-V(t_{t(n)}^n)}{t_{t(n)+1}^n - t_{t(n)}^n} g(\xi_{t(n)+1}^n)(t-t_{t(n)}^n),$$

where $\xi_l^n \in [t_{l-1}^n, t_l^n]$. We know that when $n > N$

$$\left| \frac{V(t_{t(n)+1}^n) - V(t_{t(n)}^n)}{t_{t(n)+1}^n - t_{t(n)}^n} g(\xi_{t(n)+1}^n)(t - t_{t(n)}^n) \right| < \frac{\varepsilon}{3}.$$

So we have

$$-\frac{\varepsilon}{3} + \sum_{l=1}^{t(n)} m_l^n (V(t_l^n) - V(t_{l-1}^n)) \le \int_0^t g(s)\tilde{v}_n(s)\,ds \le \frac{\varepsilon}{3} + \sum_{l=1}^{t(n)} M_l^n (V(t_l^n) - V(t_{l-1}^n)).$$

We also have

$$\sum_{l=1}^{t(n)} m_l^n (V(t_l^n) - V(t_{l-1}^n)) + m_{t(n)+1}^n (V(t) - V(t_{t(n)}^n)) \le \int_0^t g(s)\,dV(s)$$

$$\le \sum_{l=1}^{t(n)} M_l^n (V(t_l^n) - V(t_{l-1}^n)) + M_{t(n)+1}^n (V(t) - V(t_{t(n)}^n)).$$

So when $n > N$, we have

$$\int_0^t g(s)\,dV(s) - \int_0^t g(s)\tilde{v}_n(s)\,ds \le \sum_{l=1}^{t(n)} (M_l^n - m_l^n)(V(t_l^n) - V(t_{l-1}^n)) + M_{t(n)+1}^n (V(t) - V(t_{t(n)}^n)) + \frac{\varepsilon}{3}$$

$$\le S_{\Gamma_n} - s_{\Gamma_n} + \frac{\varepsilon}{3} + M_{t(n)+1}^n (V(t) - V(t_{t(n)}^n)) \le \varepsilon$$

and

$$\int_0^t g(s)\,dV(s) - \int_0^t g(s)\tilde{v}_n(s)\,ds \ge \sum_{l=1}^{t(n)} (m_l^n - M_l^n)(V(t_l^n) - V(t_{l-1}^n)) + m_{t(n)+1}^n (V(t) - V(t_{t(n)}^n)) - \frac{\varepsilon}{3}$$

$$\ge s_{\Gamma_n} - S_{\Gamma_n} - \frac{\varepsilon}{3} + m_{t(n)+1}^n (V(t) - V(t_{t(n)}^n)) \ge -\varepsilon,$$

i.e.,

$$\left| \int_0^t g(s)\,dV(s) - \int_0^t g(s)\tilde{v}_n(s)\,ds \right| \le \varepsilon.$$

Now taking $\{v_n\} \in C^\infty[0,T]$ such that $\|\tilde{v}_n - v_n\|_{L^1} \to 0$ as $n \to \infty$, we get

$$\lim_{n\to\infty} \left| \int_0^t g(s)\,dV(s) - \int_0^t g(s)v_n(s)\,ds \right| \le \lim_{n\to\infty} \left\{ \left| \int_0^t g(s)\,dV(s) - \int_0^t g(s)\tilde{v}_n(s)\,ds \right| + \int_0^t |g(s)|\,|\tilde{v}_n(s) - v_n(s)|\,ds \right\} = 0$$

uniformly on t. Clearly the $\|v_n\|_{L^1}$ are uniformly bounded. □

Acknowledgment. The author is very grateful to Professor H. J. Sussmann for proposing to him the topic, for his help, and for many fruitful discussions.

REFERENCES

[1] Z. Artstein, *The limiting equations of nonautonomous ordinary differential equations*, J. Differential Equations, 25 (1977), pp. 184–202.
[2] Z. Artstein, *Continuous dependence of solutions of Volterra integral equations*, SIAM J. Math. Anal., 6 (1975), pp. 446–456.
[3] N. Bourbaki, *Elements of Mathematics: Lie Groups and Lie Algebras*, Hermann, Paris, 1975.
[4] M. Fliess, *Réalisation locale des systémes non linéaires algébres de Lie filtrées transitives et séries génératices non commutatives*, Invent. Math., 71 (1983), pp. 521–535.
[5] J. Kurzweil and J. Jarnik, *Limit process in ordinary differential equations*, Journal of Applied Mathematics and Physics, 38 (1987), pp. 241–256.
[6] J. Kurzweil and J. Jarnik, *Iterated Lie brackets in limit processes in ordinary differential equations*, Results Math., 14 (1988), pp. 125–137.
[7] J. Kurzweil and J. Jarnik, *A convergence effect in ordinary differential equations*, in Asymptotic Methods of the Mathematical Physics, 301, Naukova Dumka, Kiev, 1989.
[8] E. B. Lee and L. Markus, *Foundations of Optimal Control Theory*, John Wiley & Sons, New York, 1968.
[9] H. J. Sussmann, *Lie brackets and local controllability, a sufficiency condition for scalar-input systems*, SIAM J. Control Optim., 21 (1983), pp. 686–713.
[10] H. J. Sussmann, *A Product Expansion for the Chen Series, Theory and Applications of Nonlinear Control Systems*, C. Byrnes and A. Lindquist, eds., North-Holland, Amsterdam, 1986, pp. 323–335.
[11] H. J. Sussmann, *A general theorem on local controllability*, SIAM J. Control Optim., 25 (1987), pp. 158–194.
[12] H. J. Sussmann and W. Liu, *Limiting Behavior of Trajectories for Highly Oscillating Controls*, Technical Report SYCON–91-02.
[13] H. J. Sussmann and W. Liu, *Continuous dependence with respect to the input of trajectories of control-affine systems*, SIAM J. Control Optim., to appear.
[14] H. J. Sussmann and W. Liu, *Averaged Convergence of Vector Fields and Continuous Dependence of Solutions of Ordinary Differential Equations*, preprint.
[15] H. J. Sussmann and W. Liu, *Limits of highly oscillatory controls and approximation of general paths by admissible trajectories*, in Proc. 30th IEEE Control and Design Conference, Brighton, UK, Dec. 1991.
[16] W. Liu, *Averaging Theorems for Highly Oscillatory Ordinary Differential Equations and the Approximation of General Paths by Admissible Trajectories for Nonholonomic Systems*, Ph.D. thesis, Rutgers University, 1992.

SIAM J. CONTROL OPTIM.
Vol. 35, No. 6, pp. 2021–2049, November 1997

008

RISK-SENSITIVE AND ROBUST ESCAPE CRITERIA*

PAUL DUPUIS† AND WILLIAM M. MCENEANEY‡

Abstract. The problem of controlling a noisy process so as to prevent it from leaving a prescribed set has a number of interesting applications. In this paper, new approaches to this problem are considered. First, a risk-sensitive criterion for a stochastic diffusion process model is examined, and it is shown that the value is a classical solution of a related PDE. The qualitative properties of this criterion are favorably contrasted with those of existing criteria in the risk-averse limit. It is proved that in the risk-averse limit the value of the risk-sensitive criterion converges to a viscosity solution of a first-order PDE. It is then demonstrated that the value function of a deterministic differential game is also a viscosity solution to the PDE. This game gives a robust control formulation of the escape time problem and is analogous to H^∞ control. In particular, the opposing player attempts to push the process out of the prescribed set and suffers an L^2 cost for his efforts. Lower bounds on the escape time as a function of this cost are obtained.

Key words. exit time control, risk-sensitive control, large deviations, robust control

AMS subject classifications. Primary, 35B37, 49L25, 60F10, 90D25, 93B36, 93C10, 93E20; Secondary, 35F99, 93C90

PII. S0363012995281626

1. Introduction.

In problems that cover a range of interesting applications, there is a stochastic process model for a system, and the goal is to keep the process in a given fixed open set G. In this paper we will consider such problems in the context of optimization and control. Here, as with many other problems to which control theory is applied, the selection of a cost is subjective. Typically, one chooses a cost with the purpose of inducing desirable qualitative behavior. For example, in many control problems the goal is to keep the controlled process near a certain operating point. In this case one would choose a criterion so that the corresponding optimal controls will "stabilize" the process about this operating point. For our problem the situation is different, in that the goal is not so much to keep the process near a particular point as it is to keep it away from the "bad" set G^c, where the c denotes complement. For a problem to fit well into such a framework, it must be the case that entry into the set G^c is in a certain sense catastrophic, and avoiding such an event must be a high priority. Examples are the failure of a machine, loss of data in a communication network, loss of "lock" in an adaptive tracking device, and entry of bistable adaptive control algorithms, such as those using ALOHA-type protocols, into the "bad" region.

There are two criteria that are often associated with the problem described above. The first criterion is the probability of escape over some interval $[0, T]$:

$$P_x \{X_t \notin G \text{ for some } t \in [0, T]\},$$

where P_x denotes probability conditioned on $X_0 = x$. If the process is controlled, then obviously one would like to choose the control to minimize this probability. The

*Received by the editors February 15, 1995; accepted for publication (in revised form) August 22, 1996.

http://www.siam.org/journals/sicon/35-6/28162.html

†Division of Applied Mathematics, Brown University, Providence, RI 02912 (dupuis@cfm.brown.edu). The research of this author was supported in part by NSF grant DMS-9403820.

‡Department of Mathematics, North Carolina State University, Raleigh, NC 27695-8205 (wmm@eos.ncsu.edu). The research of this author was supported in part by the ARO and the NSF through the Center for Nonlinear Analysis, and by AFOSR grant F49620-95-1-0296.

second criterion is the mean escape time, $E_x\tau$, where τ is the first time the process X_t escapes from G. Here we maximize when a control is available. As we discuss in detail in section 3, both these cost criteria have some theoretical and computational shortcomings. One of the contributions of the present paper is the introduction of new cost criteria for this problem. To have a concrete model for the purposes of discussion, consider the case of a Markov process described by the stochastic differential equation

$$dX_t^\varepsilon = f(X_t^\varepsilon)dt + \varepsilon^{1/2}\sigma(X_t^\varepsilon)dB_t, \qquad X_0^\varepsilon = x,$$

where the dimensions of the Wiener process B_t, f, and σ are compatible. The quantity $\varepsilon > 0$ is a parameter whose role will be explained shortly.

One of the criteria we consider takes the form

$$E_x \exp\left[-\theta\tau^\varepsilon/\varepsilon\right], \tag{1.1}$$

where θ is a positive design parameter and τ^ε is the first time the process X_t^ε exits G. If the process is controlled, we choose the control to minimize this quantity. We will consider the qualitative properties of this cost from two different perspectives. The first perspective is that of control of "small noise" systems. In this setting, the term small noise essentially means that escape from G is a relatively rare event. For the particular model under consideration, and under appropriate conditions on f and σ, this corresponds to $\varepsilon > 0$ being small. Many problems of interest, such as the control of communication devices, fall into this "small noise" category, since design tolerances are quite strict and a system in which escape was common would not even be worth considering. As we discuss in section 3, the quantity (1.1) has desirable properties when compared to standard criteria for problems of this type.

One can also view (1.1) as a "risk-sensitive" analogue of a more standard criterion. We consider the same limit as in the small noise perspective. Based on the now well-known connection between limits of risk-sensitive control problems and nonlinear robust control, one might expect the limit $\varepsilon \to 0$ to lead to a robust control problem for escape time problems that is analogous to H^∞ control. The second cost criterion we study is the one associated with this robust control problem. In this setting, we do not assume that the true system is "small noise."

Hence in both cases we are interested in the limit problem obtained when $\varepsilon \to 0$ in the cost (1.1). In order to obtain a well-defined limit, it is necessary to work with the scaled quantity

$$-\varepsilon \log E_x \exp\left[-\theta\tau^\varepsilon/\varepsilon\right]. \tag{1.2}$$

Because of the additional minus sign, the objective is to maximize this quantity. Depending on the means by which the designer may influence the system, we distinguish three classes of progressively more difficult optimization problems. The first case is that of "performance analysis." Here the designer cannot really influence the system at all. Instead, the designer would be interested in approximating (1.2) or its limit as $\varepsilon \to 0$ so that it could be used to compare two or more system designs. The second case is one we call "parametric optimization." Here the designer has control over a collection of parameters that determine the dynamics of the system. In this case the designer could use the limit of (1.2) as a convenient criterion when optimizing with respect to these parameters. In both of these cases, the limit of (1.2) as $\varepsilon \to 0$ can be characterized as the solution to a minimal cost deterministic optimal control problem. The last case is that in which the designer can choose an active, state-dependent control. In this case the limit as $\varepsilon \to 0$ of the supremum of (1.2) over admissible controls

can be characterized as the solution to a deterministic differential game. The control that seeks to maximize the limit of (1.2) will be opposed by a minimizing control. In order to distinguish the prelimit control problem for this case from the limit control problem obtained in the previous two, we refer to it as the case of "maximizing control." To simplify the exposition, in several places the case without maximizing control is treated separately and before the more general case.

A summary of the paper is as follows. In order to illustrate some of the main issues and suggest some of the applications, we describe several motivational examples in section 2. In section 3 we turn our attention to possible design criteria. We describe shortcomings of the standard criteria from the small noise perspective, and show how these shortcomings are avoided by considering the limit of (1.2) as the design criterion. The "risk-sensitive" interpretation of (1.2) is given in this section, and we also discuss generalizations of (1.1). Since the qualitative comparisons do not rely heavily on the use of a specific model, the discussion in sections 2 and 3 is in a general setting. However, for the asymptotic analysis and interpretation of the limit problem it is convenient to fix a model, and accordingly a diffusion model and the associated model for the robust control problem are assumed throughout sections 4–6. In section 4 we characterize the limit of (1.2) for the various cases when $\varepsilon \to 0$, and state the convergence theorem. As noted previously, the established connections between risk-sensitive control, risk-averse limits, and robust control suggest that the limit problem for the case of maximizing control would define a control that has an interpretation as a robust control. This is the case, and the precise interpretation is stated in section 5. To ease the presentation, a number of proofs are postponed to section 6.

2. Examples. In this section we describe examples that fit into the framework of the last section. The examples are intended only to be illustrative. They provide the basis for the qualitative comparisons of the next section and motivate the criteria that will be considered later. Some of the examples fall outside the class of diffusion processes, and hence indicate interesting extensions that are not covered by the theory developed in this paper. However, since the qualitative properties discussed in section 3 require only a certain type of large deviation behavior, the remarks of this section apply to all the examples.

Problems for which escape time criteria are appropriate fit into one of two categories. In the first category, exit from the region of desired operation essentially causes the system to shut down, and the system is more or less "off-line" until the state of the system can be steered back into the good operating region. For example, when an ALOHA-type system exits its stable operating region the entire system is shut down and then restarted. In Example 1 below, exit from the operating region means that the pair of communicating satellites must suspend data transmission and initiate the procedure to regain "lock." In the second category exit from the domain is not fatal, but still an event to be avoided because of a drastic decline in performance when outside the good region. An example in this category is the queueing example given below, in which escape from G corresponds to a nonnegligible fraction of incoming customers being turned away. For additional examples the reader can consult the references in [28].

Example 1 (satellite laser communication problem). In space-based laser communication an essential role is played by the tracking and pointing subsystems of the satellites that are involved. In particular, data may be transmitted between two satellites by a laser communication crosslink. In order that the communication links not be broken, the pointing system of each satellite must keep the laser focused on the

detector of the other. Since the communication beams are very narrow, the pointing requirements are rather stringent [26]. In this example the set G is defined in terms of an angular cone of allowable orientations of the gimbal-mounted optics used to control the beam direction. At any given time one of the satellites is allowed to transmit data at a high rate while the other transmits a "beacon" signal at a low rate, the purpose of which is to help maintain the "lock." Owing to "noises" (such as internal vibrations) the operating state of the system can be driven from G, at which time there will be a communication interruption, and lock between the transmitter and receiver will have to be re-established via a separate acquisition algorithm. It is desirable to have the time between losses of lock be quite large (e.g., several months).

For such a system linear methods automatically make little sense, since the effective state space in which the system can operate is not itself a linear space. This is true even if the system dynamics are linear. The criterion described in the introduction and the robust criterion we will define in section 5 are more natural for this problem, since they focus on the event that is actually of interest. In this problem one may be interested in parametric optimization or feedback control.

Example 2 (tracking problem). Many synchronization systems in advanced communication systems are digital. One might model such a system as

$$X_{i+1}^n = X_i^n + \frac{1}{n} b(X_i^n, \xi_i).$$

In this equation ξ_i is a random sequence composed of noise and the inputs to the system. The discrete time parameter i is used because time is "slotted," and data are communicated only at discrete times. The factor $1/n$ reflects the fact that the state X_i^n of the system changes slowly as a function of each new piece of data ξ^i, although the data rates are very high. In order that the system operate properly, it is crucial that the transmitter and receiver both be on the same "clock," so that it is clear when a discrete time interval begins and ends. In a synchronization system, one of the components of X_i^n (say $(X_i^n)_1$) will be the difference between an estimate of a phase timing indicator and the true value. For accurate communication or tracking, one needs a very good estimate. Here the set G will be of the form $\{x : -a_1 \le x_1 \le a_2\}$. As part of the design procedure "stabilizing" dynamics are always built into the system in order to keep X_i^n in the acceptable region G. However, owing to the presence of noise, the difference between the estimate of the phase timing indicator and the true value is eventually driven from G. A risk-sensitive escape criterion is natural in this context. For these problems, one is typically interested only in performance analysis (for purposes of comparing competing designs), or at most in parametric optimization.

This example is illustrative of a large class of similar problems from statistics and adaptive stochastic algorithms [4]. The large deviation analysis, numerical computations, and simulations for a related analog device known as a phase-locked loop appear in [7, 9].

Example 3 (queueing problems). Numerous problems involving the design and optimization of queues can be cast in terms of the risk-sensitive criterion. Escape criteria will be suitable whenever the main purpose of the design is to reduce the possibility of very large buffers. A number of examples and references to additional examples are given in the book [30]. Although there is at the present time no theory of robust control for queueing systems, one would expect by analogy with the case for diffusions that controls designed on the basis of a risk-sensitive criterion would enjoy the features one would desire of such a robust control.

Escape criteria are suitable whenever buffer overflow constitutes a critical event. This is often true for problems associated with the analysis and design of high-speed data networks. These networks carry many types of data in digitized form, e.g., voice, computer, and video data. Each of these classes has its own requirements and characteristics. In addition, for some classes of data there may be contractually agreed upon requirements regarding network performance. These requirements will be stringent (e.g., probabilities of data loss in the 10^{-9} range). Data loss occurs when buffer capacities are exceeded, and thus corresponds to an "escape." There are numerous difficult and interesting design and control problems that are associated with such networks.

Example 4 (power system stability). In this example a diffusion model is believed to be appropriate for describing the time evolution of the system (cf. [6] and the references therein). The state of the system is a vector whose components consist of various generator frequencies, voltage phase angles, and voltage magnitudes. The set G is defined to be a region in which the "security" of the system is acceptable. Exit from the region may require substantial intervention to return the state of the system to the stable region, and hence is an event to be avoided. Coefficient matrices in the system of equations that describe operation provide the system parameters over which optimization can be performed. In addition, feedback control in various forms can also be possible.

3. Comparison and interpretation of the cost criterion. In this section we study the optimization criterion

$$E_x \exp\left[-\theta\tau^\varepsilon/\varepsilon\right] \tag{3.1}$$

from a qualitative point of view. In section 3.1 a comparison is made between this cost and more standard cost criteria in the small noise setting. In section 3.2, (3.1) is interpreted from the risk-sensitive point of view. Generalizations of the cost are briefly discussed in section 3.3.

3.1. The small noise setting. In many of the problems for which a criterion based on escape times is appropriate, it is useful to assume that the stochastic process model is in some sense a "small noise" model. Indeed, if this were not the case, then (essentially regardless of system design) escapes from the acceptable operating region would be common, and for many problems the models might not be worth considering.

As remarked in the introduction, there are two criteria that are often associated with such problems, namely, the probability of escape over some interval $[0, T]$,

$$P_x\left\{X_t^\varepsilon \notin G \text{ for some } t \in [0,T]\right\}$$

and the mean escape time

$$E_x\tau^\varepsilon, \text{ where } \tau^\varepsilon = \inf\{t : X_t^\varepsilon \notin G\}.$$

The parameter $\varepsilon > 0$ indicates the "strength" of the noise, with zero noise in the limit $\varepsilon \to 0$.

Unfortunately, both of these criteria have shortcomings. We first consider the escape probability. While this criterion might be acceptable for problems for which the duration T is fixed and known, it is probably not appropriate otherwise. In many problems one has a rough idea of the interval of interest, but not much more than that. The escape probability criterion can be inconvenient to work with even when

the interval of interest is known. For example, in the control setting it typically results in controls that are not stationary. A second difficulty is of a computational nature. In cases where the control space is large, there can be numerical difficulties due to the fact that the controls become somewhat singular near the final time T. A final problem is greater sensitivity (relative to more well-behaved functionals) with regard to the parameter that measures the strength of the noise. Hence escape probabilities can be more difficult to approximate than smoother functionals of a process (such as an expectation of a smooth function of the process). Because of this, the behavior of systems designed on such a criterion is less reliably predicted by the asymptotic theory.

While the mean escape time criterion has the advantage of yielding time-independent feedback controls, it too can be problematic, especially in the desired situation where the probability of escape over $O(1)$ time intervals is rare. In this setting, computation and approximation of the mean escape time, even for the case of a fixed control, can be difficult. In order to make this statement precise, we return to the small noise diffusion model

$$dX_t^\varepsilon = f(X_t^\varepsilon)dt + \varepsilon^{1/2}\sigma(X_t^\varepsilon)dB_t, \qquad X_0^\varepsilon = x. \tag{3.2}$$

It is well known (under some assumptions) that $W^\varepsilon(x) \doteq E_x\tau^\varepsilon$ satisfies a second-order PDE

$$\mathcal{L}^\varepsilon W^\varepsilon(x) = -1, \quad x \in G^0, \;\; W^\varepsilon(x) = 0, \quad x \in \partial G, \tag{3.3}$$

where

$$\mathcal{L}^\varepsilon g(x) \doteq \frac{\varepsilon}{2}\mathrm{tr}\left[g_{xx}(x)a(x)\right] + \langle f(x), g_x(x)\rangle, \qquad a(x) = \sigma(x)\sigma^T(x),$$

and $\mathrm{tr}B$ denotes the trace of the square matrix B. The case with control involves obvious modifications of this equation. The smallness of the noise translates into the small ε coefficient in front of the second derivative term. This produces a solution that is essentially flat over much of G^0 but with steep gradients close to ∂G. Because of these properties, it is very difficult to accurately approximate the solution numerically. This is rather unsatisfactory, since the very conditions that are required for escapes to be a rare event lead to difficult computational problems.

A problem that may be even more important than the one just discussed is that in the small noise setting the mean escape time (and by extension any controls that are designed using it as the performance criterion) may focus on events that are not of any practical interest, since the dominant contribution to this criterion comes from paths that take a very long time to escape. Under certain technical conditions on f, σ, and their relation to ∂G, one can show [20] there exists a constant $C^* > 0$ such that the limit

$$\lim_{\varepsilon\to 0} \varepsilon \log E_x\tau^\varepsilon = C^*$$

holds uniformly for x in compact subsets of G. For any $\delta \in (0, C^*)$, the scaling in ε of this quantity implies that when $\varepsilon > 0$ is sufficiently small the dominant contribution to $E_x\tau^\varepsilon$ is due to sample paths that take at least $\exp\left([C^* - \delta]/\varepsilon\right)$ units of time to escape. While this may be a moot point if one is absolutely certain that the mean escape time is the criterion of interest, it is an important point when this is not the case. Consider, for example, a telecommunication problem involving routing of data through a network. For obvious reasons, the true process representing loading

of the network is nonstationary, with cyclical variations (e.g., periods of one day). For this problem the mean escape criterion, even if it were computable, would be an inappropriate basis on which to design controls if the dominant contribution is due to sample paths that take longer than one day to escape. Such a criterion is clearly not desirable, since controls designed on it may allow a large number of escapes over a short time interval, as long as this is balanced by relatively few sample paths which take the very long time $\exp([C^* - \delta]/\varepsilon)$ to escape. Moreover, these paths would not even contribute to the overall performance of the true system, owing to its nonstationarity.

As we will see, the escape time criterion (3.1) and controls based on its asymptotic behavior avoid these difficulties. The controls will automatically be independent of time, and by appropriately choosing the design parameter θ, one can to some degree "tune" the system to focus on escapes over different $O(1)$ time intervals. By working with the logarithmic transform of (3.1), we obtain a Hamilton–Jacobi equation that is much more well behaved than (3.3), especially with regard to computational approximations. We also note that in comparison with the escape probability, the relative smoothness of the functional mapping $X^\varepsilon \to \tau^\varepsilon$ under the distribution of X^ε suggests that (3.1) (or rather the logarithmic transform of (3.1)) should be more reliably predicted by the asymptotic theory when $\varepsilon \to 0$.

3.2. The risk-sensitive interpretation. The theory of risk-sensitive control investigates the manner in which modifications of cost structures affect the associated optimal policies. For example, one might be interested in the effect of so-called "risk-sensitizing transformations." At an intuitive level, the goal of such a transformation is to amplify the effect that certain outcomes have in determining the overall cost, and thereby force the optimal control (or any nearly optimal control) to "pay more attention" to these more heavily weighted outcomes. In particular, in the risk-averse case, "bad" events are weighed more heavily, and the control becomes more conservative with regard to allowing these events to occur.

From this perspective, for each fixed $\varepsilon > 0$, one may view the criterion (3.1) as a risk-sensitive version of the mean escape time criterion $E_x\tau^\varepsilon$. The effect of the nonlinear mapping $\tau \to \exp -\theta\tau/\varepsilon$ is to shift the attention toward paths that escape in a relatively short time, at the expense of optimizing the mean escape time. As noted in the previous section, this makes sense for many problems, especially those that may involve some type of nonstationarity. The design parameter $\theta > 0$ controls the degree to which these short time escapes are emphasized, with larger values of θ focusing attention more heavily on such escapes. On the other hand, if we fix $\varepsilon > 0$ and take the limit $\theta \to 0$, one can show (under suitable uniform integrability conditions) that the design criterion (3.1) becomes equivalent in this limit to the mean escape time.

For various reasons, including computational simplicity, model simplification, and because of the connection with robust control design, one may also be interested in taking limits with respect to a family of risk-sensitizing transformations. This is in fact the second motivation for the asymptotic analysis to be carried out in the next few sections. The interpretation of the limit in terms of robust control will be discussed further in section 5. As the reader can easily check, unless the dynamics of the process model are modified in an ε-dependent way such as that of (3.2), then it will be difficult to normalize the quantity (3.1) so as to obtain a well-defined limit as $\varepsilon \to 0$. The limit problem will be interpreted as a robust control problem in section 5. The link between the two allows one to connect these two rather different modeling perspectives. It is important to observe that the robust interpretation of the limit control problem is independent of the asymptotic analysis.

The relationship between risk-sensitive control and robust control in the linear setting was discussed at length in [33]. Connections between the two have also been examined in several nonlinear contexts. It was shown in [14, 15, 23] that the values of certain risk-sensitive infinite time horizon control problems converge to the values of H^∞ disturbance attenuation problems. In [16, 27] the connection between finite time horizon risk-sensitive control and robust finite time horizon control was made. Finally, in [29] and [12] the connection was examined in the Markov chain and hidden Markov cases, respectively.

3.3. Generalizations of the cost. Depending on the problem, one might consider various generalizations of the cost criterion

$$E_x \exp\left[-\theta\tau^\varepsilon/\varepsilon\right].$$

For example, one can measure "time until escape" in a state-dependent way by considering a cost of the form

$$E_x \exp\left[-\int_0^{\tau^\varepsilon} \theta(X_s^\varepsilon)ds/\varepsilon\right],$$

where $\theta(\cdot)$ is a continuous function on $\bar{G}$. As long as $\inf_{x\in\bar{G}}\theta(x) > 0$, the analysis for this case is essentially the same as for the case $\theta(\cdot) \doteq \theta$ that is considered in sections 4–6.

For maximizing control problems where the control space is potentially unbounded, one might consider a cost of the form

$$E_x \exp\left[-\int_0^{\tau^\varepsilon} \left[\theta - \langle u_s, Au_s\rangle\right] ds/\varepsilon\right],$$

where A is a positive definite matrix of appropriate dimensions. Because of the lack of compactness in the control variable, the analysis in this case is more complicated than that given in sections 4–6.

4. Asymptotic analysis. For the remainder of this paper attention is restricted to controlled and uncontrolled diffusion processes, and it is further assumed that the control only affects the drift. Hence the dynamics of the controlled process are given by

$$dX_t^\varepsilon = f(X_t^\varepsilon, u_t)\,dt + \varepsilon^{1/2}\,\sigma(X_t^\varepsilon)\,dB_t, \tag{4.1}$$

with the obvious modification for the uncontrolled case. B is a Brownian motion with sample space Ω, filtration F_t, and measure P. The control process u is F_t-progressively measurable (see [18]) and takes values in a compact set U.

In this section, the results for the small noise/risk-sensitive escape problem are developed. Since the proofs for the case with a maximizing control are not significantly more complex than those for the case without control, all the results in this section will be stated only for the case with control. For the case without control, the modifications are obvious.

The following conditions will be assumed for the remainder of the paper; no additional assumptions will appear. Recall that $G \subseteq \Re^n$ denotes the set one wishes to keep the process in, with G open and $\overline{G}$ compact. Let ∂G denote the boundary

of G and for $y \in \Re^n$ let $B_a(y) \doteq \{x : |x-y| \le a\}$, where $|\cdot|$ denotes the Euclidean norm.

CONDITION 4.1. *$\overline{G}$ satisfies a uniform exterior sphere condition. Thus there exists $r > 0$ such that for any $x \in \partial G$ there is $y \in (\overline{G})^c$ with $|x-y| = r$ and*

$$B_r(y) \cap \overline{G} = \{x\}.$$

CONDITION 4.2. *$f \in C(\overline{G}, U)$, and furthermore f is uniformly Lipschitz in x in the sense that there exists $K_f < \infty$ such that*

$$|f(x,u) - f(y,u)| \le K_f |x-y| \qquad \text{for all } (x,y,u) \in \overline{G} \times \overline{G} \times U.$$

CONDITION 4.3. *$\sigma \in C^1(\overline{G})$, and there exists $\mu > 0$ such that*

$$\xi^T \sigma(x) \sigma^T(x) \xi \ge \mu |\xi|^2 \qquad \text{for all } (x,\xi) \in \overline{G} \times \Re^n.$$

Note that the last assumption implies that σ is Lipschitz on $\overline{G}$ with some constant $K_\sigma < \infty$.

Let $\mathcal{U}_\nu$ be the set of F_t-progressively measurable control processes with values in U with respect to the reference probability system $\nu = (\Omega, \{F_t\}, P, B)$ (see [18]). Although X^ε and τ^ε depend on $u \in \mathcal{U}_\nu$, we omit this dependence from the notation. Define

$$\Phi^\varepsilon(x) \doteq \inf_{u \in \mathcal{U}_\nu} E_x \exp\left[-\theta \tau^\varepsilon / \varepsilon\right],$$

where θ is a positive constant, $\tau^\varepsilon \doteq \inf\{t : X_t^\varepsilon \notin G\}$, and E_x indicates expectation conditioned on an initial state x. As is well known (and easy to prove using a large deviation calculation), for each fixed control the quantity

$$E_x \exp\left[-\theta \tau^\varepsilon / \varepsilon\right]$$

scales exponentially in ε as $\varepsilon \to 0$, in the sense that for each $x \in G$ there exists $c(x) \in \mathbb{R}$ such that

$$-\varepsilon \log E_x \exp\left[-\theta \tau^\varepsilon / \varepsilon\right] \to c(x)$$

as $\varepsilon \to 0$.

Thus it is natural to apply a logarithmic transform (see, e.g., [13, 18]) and consider a criterion of the form

$$V^\varepsilon(x) \doteq -\varepsilon \log \Phi^\varepsilon(x) = \sup_{u \in \mathcal{U}_\nu} -\varepsilon \log E_x \exp\left[-\theta \tau^\varepsilon / \varepsilon\right]. \tag{4.2}$$

(As noted in section 3, the cost in (4.2) is similar to $E_x\{\theta\tau\}$ but with the insertion of a risk-sensitizing transformation.) Starting with the quasi-linear PDE satisfied by Φ^ε and taking the log transform, one formally obtains the PDE

$$\begin{aligned} &\theta + \frac{\varepsilon}{2} \operatorname{tr}\left[a(x) V_{xx}(x)\right] + H(x, V_x(x)) = 0, \quad x \in G, \\ &V(x) = 0, \qquad\qquad\qquad\qquad\qquad\qquad\quad x \in \partial G, \end{aligned} \tag{4.3}$$

with

$$H(x,p) \doteq \max_{v \in U} \langle f(x,v), p \rangle - \frac{1}{2} \langle p, a(x) p \rangle, \tag{4.4}$$

and $a(x) \doteq \sigma(x) \sigma^T(x)$.

Criterion (4.2) is analogous to the risk-sensitive criteria applied in many stochastic control problems (see [14, 15, 16, 23, 27, 29], among others). The following results support this interpretation. The first lemma is standard (see, e.g., [18, Theorem 15.18]).

LEMMA 4.1. *There exists a solution* $\widetilde{V}^\varepsilon \in C^2(G) \cap C^0(\overline{G})$ *to* (4.3).

Define the cost for a fixed control $u \in \mathcal{U}_\nu$ by

$$J^\varepsilon(x,u) \doteq -\varepsilon \log E_x \exp\left[-\theta\tau^\varepsilon/\varepsilon\right].$$

THEOREM 4.2. *Given any* $x \in \overline{G}$ *and* $u \in \mathcal{U}_\nu$ $\widetilde{V}^\varepsilon(x) \geq J^\varepsilon(x,u)$. *Let* $\bar{u}$ *be a Borel measurable function such that*

$$\bar{u}(x) \in \operatorname{argmax}_{v\in U}\langle f(x,v), \widetilde{V}^\varepsilon_x(x)\rangle,$$

and let $\bar{X}^\varepsilon$ *be a solution of* (4.1) *with* $u_t \doteq \bar{u}(\bar{X}^\varepsilon_t)$ [32]. *Then* u *is in* $\mathcal{U}_\nu$, *and* $\widetilde{V}^\varepsilon(x) = J^\varepsilon(x,u)$. *Consequently,* $V^\varepsilon(x) = \widetilde{V}^\varepsilon(x)$ *for all* $x \in \overline{G}$.

Although the proof of this assertion is delayed until section 6, we note here that Theorem 4.2 follows from Girsanov's Theorem and an application of Ito's rule.

For reasons discussed in section 3, we would like to determine the limit as $\varepsilon \downarrow 0$ for this problem. Two common techniques for proving convergence are based on large deviations ideas (see [8], among others) and viscosity solutions (see [19], among others). The first is a probabilistic method, while the second is a PDE approach. The viscosity solution approach is used here. Since it is assumed that the prelimit ($\varepsilon > 0$) problem is uniformly nondegenerate, the viscosity solution approach is relatively easy to apply. Before applying it we must obtain bounds on the behavior of V^ε which are uniform in $\varepsilon > 0$. These are supplied by the following two lemmas. Their proofs, which are given in section 6, are standard.

LEMMA 4.3. *There exists* $M_1 < \infty$ *such that* $0 \leq V^\varepsilon(x) \leq M_1$ *for all* $x \in \overline{G}$ *and all* $\varepsilon > 0$.

LEMMA 4.4. *Given any* $\varepsilon_0 < \infty$, *there exists* $M_2 < \infty$ *such that* $|V^\varepsilon(x) - V^\varepsilon(y)| \leq M_2|x-y|$ *for all* $x \in \partial G$, *all* $y \in \overline{G}$, *and all* $\varepsilon \in (0, \varepsilon_0)$.

As $\varepsilon \downarrow 0$, one formally obtains from (4.3) the limit PDE problem

$$\begin{aligned} -\theta - H(x, V_x(x)) &= 0, \quad x \in G, \\ V(x) &= 0, \qquad\qquad x \in \partial G, \end{aligned} \tag{4.5}$$

where H is given by (4.4). (The minus sign in (4.5) is a consequence of the definition of viscosity solutions used in section 5.) In section 5 we will prove the existence of a unique continuous viscosity solution to (4.5) which satisfies the boundary condition pointwise. Let this solution be denoted by W. In that section we will also apply the method of Barles and Perthame [2, 18] to show that

$$\lim_{\varepsilon\downarrow 0} V^\varepsilon(x) = W(x) \qquad \forall\, x \in \overline{G}. \tag{4.6}$$

The cited method does not require gradient bounds which are uniform in $\varepsilon > 0$ and $x \in \overline{G}$, and in fact such bounds are only needed on the boundary (Lemma 4.4). Define

$$\begin{aligned} V^*(x) &\doteq \limsup\{V^\varepsilon(y) : \ y \to x, \ \varepsilon \downarrow 0, \ y \in \overline{G}\}, \\ V_*(x) &\doteq \liminf\{V^\varepsilon(y) : \ y \to x, \ \varepsilon \downarrow 0, \ y \in \overline{G}\}. \end{aligned} \tag{4.7}$$

It follows from Lemma 4.4 that

$$V^*(x) = 0 = V_*(x) \qquad \forall\, x \in \partial G.$$

Because $W(x) = 0$ for all $x \in \partial G$, one also has

$$V^*(x) = V_*(x) = W(x) \qquad \forall\, x \in \partial G. \tag{4.8}$$

Since a consequence of the definitions given in (4.7) is

$$V^*(x) \geq V_*(x) \qquad \forall\, x \in G,$$

to prove (4.6) it suffices to show

$$V^*(x) \leq W(x) \leq V_*(x) \qquad \forall\, x \in \overline{G}.$$

The proof of this last statement is delayed until section 6. However, it immediately implies the following.

THEOREM 4.5. *$V^\varepsilon(x) \to W(x)$ uniformly on $\overline{G}$.*

Thus one has the characterization of the limit of the value function as a continuous viscosity solution to (4.5). Furthermore, the comparison result at the heart of the proof of Theorem 4.5 also implies uniqueness of the solution of (4.5) among the class of continuous viscosity solutions. In section 5 it will be shown that W is the value of the deterministic game that corresponds to the associated robust control problem. (In the case where there is no control in the risk-sensitive problem, W is simply the value of a deterministic minimizing control problem rather than a game.) This game serves two roles. In the small noise problem it provides a convenient starting point for the analysis and construction of controls for the prelimit problem. It will also serve as the starting point for the interpretation of the maximizing control in the game as a robust control. In this paper we will only use W in the latter role. For an example of how it can be used in the first role, we refer to [8].

5. The robust limit problem. In this section we consider the robust problem corresponding to the limit $\varepsilon \to 0$. The term robust is used here to denote a system where the effect on the system output due to a given disturbance is bounded by a function of the power of that disturbance. This notion takes different forms in different contexts. The most well-known example is H^∞ control. In the state-space formulation, a system is said to satisfy an H^∞ bound if there exists a bound on the L^2 norm of an output in the form of a product of disturbance attenuation constant and the L^2 norm of the disturbance. This H^∞ disturbance attenuation control problem may be formulated as a deterministic differential game. The robust (maximizing) control escape time problem will also be formulated as a game. In particular, the maximizing player in the game will correspond to the original maximizing control, and an opposing player will be introduced who will try to minimize the same payoff. In the robust problem, the "noise" is a process chosen by this new minimizing player, rather than a stochastic process. The player will select the noise so as to drive the process from the set G, but must pay a quadratic cost. Since there is only one player for the case without maximizing control (i.e., the performance evaluation or parameter optimization problems discussed in section 1), the robust formulation in this case is a control problem rather than a game.

In the analysis of the limit problem the case with no control is significantly easier than the case with a maximizing controller. To simplify the presentation, we consider these two cases separately.

5.1. The case without control. Consider the deterministic dynamics

$$\frac{dY_t}{dt} = f(Y_t) + \sigma(Y_t)w_t, \qquad Y_0 = x \tag{5.1}$$

for $t \in [0, \tau]$, where $\tau \doteq \inf\{t : Y_t \notin G\}$. The functions f and σ are those introduced in section 4. Note that the disturbance w is now a deterministic process. The function w is assumed to be in

$$\mathcal{W}^0 \doteq \left\{ w : [0,\infty) \to \Re^m : w \text{ is measurable and } \int_0^T |w_t|^2 dt < \infty \text{ for all } T < \infty \right\}.$$

The cost criterion takes the form

$$J(x,w) \doteq \int_0^\tau \left[\theta + \frac{1}{2}|w_t|^2\right] dt, \tag{5.2}$$

where again $\theta > 0$ is a constant. The control problem which will yield the robust value, $\widetilde{W}$, is

$$\widetilde{W}(x) \doteq \inf_{w \in \mathcal{W}^0} J(x,w). \tag{5.3}$$

The Hamilton–Jacobi–Bellman (HJB) equation corresponding to control problem (5.1)–(5.3) is

$$-\theta - H_{\rm nc}(x, W_x(x)) = 0, \quad x \in G, \qquad W(x) = 0, \quad x \in \partial G\ , \tag{5.4}$$

where

$$H_{\rm nc}(x,p) \doteq \langle f(x), p\rangle + \min_{w \in \Re^m} \left[\langle \sigma(x)w, p\rangle + \frac{1}{2}|w|^2\right].$$

Evaluation of the minimum yields

$$H_{\rm nc}(x,p) = \langle f(x), p\rangle - \frac{1}{2}\langle p, a(x)p\rangle.$$

Hence (5.4) is the same as (4.5) for the case with no (maximizing) control. For easy reference, we recall the definition of viscosity solution that is appropriate for our problem. $W \in C(\overline{G})$ is a continuous viscosity subsolution of (5.4) if

$$-\theta - H_{\rm nc}(x_0, g_x(x_0)) \le 0$$

whenever $g \in C^1(G)$ and $W - g$ attains a maximum at $x_0 \in G$. $W \in C(\overline{G})$ is a continuous viscosity supersolution of (5.4) if

$$-\theta - H_{\rm nc}(x_0, g_x(x_0)) \ge 0$$

whenever $g \in C^1(G)$ and $W - g$ attains a minimum at $x_0 \in G$. If W is both a subsolution and a supersolution, then it is a solution. Due to the nondegeneracy (Condition 4.3), it is not necessary to formulate the boundary conditions in the viscosity framework; only solutions satisfying the boundary conditions pointwise will be considered. In particular, all assertions of uniqueness are within the class of continuous viscosity solutions satisfying the boundary conditions pointwise. (For modifications relevant

to extending the proof of Theorem 4.5 to viscosity solutions satisfying the weaker viscosity form of the boundary conditions, see [18, section 7.8].)

THEOREM 5.1. *$\widetilde{W}$ is the unique continuous viscosity solution of* (5.4).

Proof. Theorem 5.1 is just a variant of what are by now standard results. With this in mind, only an outline of the proof is provided.

First note that there exists $M_3 < \infty$ such that

$$0 \leq \inf_{w \in \mathcal{W}^0} J(x,w) \leq M_3$$

for all $x \in \overline{G}$. This follows from the fact that one can choose a constant control w which guarantees exit prior to some fixed time T, where T is independent of x. The nonnegativity of θ then implies the existence of $C < \infty$ such that $\int_0^\tau |w_s|^2 ds \leq C$ for all $x \in \overline{G}$ and all w such that $J(x,w) \leq \widetilde{W}(x) + 1$. Thus,

$$\widetilde{W}(x) = \inf_{w \in \mathcal{W}_b^0} J(x,w)$$

where

$$\mathcal{W}_b^0 \doteq \left\{ w \in \mathcal{W}^0 : \ \|w\|_{L^2(0,\infty)} \leq C \right\}.$$

The bounds on f and σ imply that given $\delta > 0$ there is an $\eta > 0$ such that

$$|Y_t - x| \leq \delta \qquad \forall\, t \in [0,\eta], \ \ \forall\, w \in \mathcal{W}_b^0 \tag{5.5}$$

and also such that for all $w \in \mathcal{W}_b^0$, $\tau = \tau(x,w)$ is bounded from below by a function $h(x)$ that is strictly positive on G.

Dynamic programming principles for $\widetilde{W}$ are easily obtained, and in fact are contained as special cases of the corresponding results in the next subsection. In particular, one has

$$\widetilde{W}(x) = \inf_{w \in \mathcal{W}_b^0} \left[\int_0^{T \wedge \tau} \left(\theta + \frac{1}{2}|w_t|^2 \right) dt + \widetilde{W}(Y_{T \wedge \tau}) \right]$$

for all $x \in G$ and $T \in [0,\infty)$, and

$$\widetilde{W}(x) \leq \inf_{w \in \widehat{\mathcal{W}}_m^0} \left[\int_0^{T \wedge \tau} \left(\theta + \frac{1}{2}|w_t|^2 \right) dt + \widetilde{W}(Y_{T \wedge \tau}) \right]$$

for all $x \in G$, $T \in [0,\infty)$, and $m \in (0,\infty)$, where

$$\widehat{\mathcal{W}}_m^0 \doteq \left\{ w \in \mathcal{W}^0 : |w_t| \leq m \ \forall\, t \right\}.$$

The last two paragraphs provide all that is needed for the proof of Theorem 5.1. The continuity of $\widetilde{W}$ follows from the nondegeneracy of σ and the boundedness of f. In particular, these properties imply that given $\delta > 0$ there exists $\varepsilon > 0$ such that for any point y within ε of x, one can construct a control that moves Y_t from x to y with cost less than δ and in time less than δ. This in turn implies the continuity. Uniqueness (among the class of continuous viscosity solutions satisfying the boundary conditions pointwise) follows from the comparison principle used in the proof of Theorem 4.5 (cf.

section 6.1.4). The proof of the present theorem is completed by showing that $\widetilde{W}$ is both a viscosity subsolution and supersolution of (5.4).

We first prove that $\widetilde{W}$ is a viscosity supersolution to (5.4). Suppose that $g \in C^1(\overline{G})$ and that $\widetilde{W} - g$ has a local minimum at $x_0 \in G$. To prove that $\widetilde{W}$ is a viscosity supersolution, it must be shown that

$$-\theta - H_{\mathrm{nc}}(x_0, g_x(x_0)) \geq 0.$$

If this inequality is not valid, then there exists $\alpha > 0$ such that

$$\theta + H_{\mathrm{nc}}(x_0, g_x(x_0)) > \alpha.$$

By using (5.5), the definition of H_{nc}, and the uniform lower bound $\tau = \tau(x, w) \geq h(x) > 0$, it follows that there is $\eta_0 \in (0, \tau)$ such that for all $\eta \in (0, \eta_0)$ and $w \in \mathcal{W}_b^0$

$$\int_0^\eta \left\{ \theta + \frac{1}{2}|w_t|^2 + \langle f(Y_t) + \sigma(Y_t) w_t, g_x(Y_t) \rangle \right\} dt \geq \frac{\eta \alpha}{2} > 0, \tag{5.6}$$

where Y_t is given by (5.1) with initial condition x_0. However, because $\widetilde{W} - g$ has a local minimum at x_0, for sufficiently small $\eta \in (0, \eta_0)$,

$$\widetilde{W}(x_0) - g(x_0) \leq \widetilde{W}(Y_t) - g(Y_t) \qquad \forall\, t \in [0, \eta].$$

By the first dynamic programming principle above, this implies

$$\inf_{w \in \mathcal{W}_b^0} \left\{ \int_0^\eta \left[\theta + \frac{1}{2}|w_t|^2 \right] dt + g(Y_\eta) - g(x_0) \right\} \leq 0. \tag{5.7}$$

But (5.6) and (5.7) form a contradiction, and consequently $\widetilde{W}$ is a viscosity supersolution. The proof that $\widetilde{W}$ is a viscosity subsolution is analogous and employs the second dynamic programming principle above. Therefore $\widetilde{W}$ is a continuous viscosity solution.

5.2. Case with control. In this subsection we consider the following deterministic differential game. For $t \in [0, \tau]$ the dynamics are given by

$$\frac{dY_t}{dt} = f(Y_t, u_t) + \sigma(Y_t) w_t, \qquad Y_0 = x, \tag{5.8}$$

where τ is the time of first escape from G. The function u is the (deterministic) measurable control for the maximizing player, which takes values in U. Let this set of controls be denoted by $\mathcal{U}^0$. The function w is the deterministic control for the minimizing player, and we assume

$$w \in \mathcal{W}^0 \doteq \left\{ w : [0, \infty) \to \Re^m : w \text{ is measurable and } \int_0^T |w_t|^2 dt < \infty \text{ for all } T < \infty \right\}.$$

The Elliott–Kalton [10] definition of the game will be used, and consequently the set of strategies for each player must be defined. A strategy for the maximizing player is a mapping $\phi : \mathcal{W}^0 \to \mathcal{U}^0$ which is nonanticipating in the following sense. Let any $t \in [0, \infty)$ be given. If $w, \tilde{w} \in \mathcal{W}^0$ satisfy $w_r = \tilde{w}_r$ for a.e. $r \in [0, t]$, then we require $\phi[w]_r = \phi[\tilde{w}]_r$ for a.e. $r \in [0, t]$. We use Φ to denote the set of strategies for the

maximizing player. A strategy for the minimizing player is a mapping $\lambda : \mathcal{U}^0 \to \mathcal{W}^0$ which is nonanticipating in the analogous sense. Let Λ denote the set of strategies for the minimizing player. The payoff for the game is

$$J(x,u,w) \doteq \int_0^\tau \left[\theta + \frac{1}{2}|w_t|^2\right] dt.$$

The upper and lower values in the Elliott–Kalton sense are given by

$$\widetilde{W}(x) \doteq \sup_{\phi\in\Phi} \inf_{w\in\mathcal{W}^0} J(x,\phi[w],w) \text{ and } \widehat{W}(x) \doteq \inf_{\lambda\in\Lambda} \sup_{u\in\mathcal{U}^0} J(x,u,\lambda[u]).$$

If $\widetilde{W} = \widehat{W}$, then the game is said to have value.

The Isaacs equation corresponding to this game is given by

$$-\theta - H(x, W_x(x)) = 0, \quad x \in G, \qquad W(x) = 0, \quad x \in \partial G, \tag{5.9}$$

where

$$H(x,p) \doteq \max_{v\in U}\langle f(x,v),p\rangle + \min_{w\in\Re^m}\left[\langle\sigma(x)w,p\rangle + \frac{1}{2}|w|^2\right].$$

Evaluation of the minimum gives

$$H(x,p) = \max_{v\in U}\langle f(x,v),p\rangle - \frac{1}{2}\langle p, a(x)p\rangle.$$

We next show that the upper value, $\widetilde{W}$, is a continuous viscosity solution of (5.9). The same result holds for the lower value. For a fixed finite time horizon problem under stronger assumptions, Evans–Souganidis [11] showed that a class of deterministic differential games had value and that this common value function was a viscosity solution of the corresponding Isaacs equation. The method of proof was to first obtain the dynamic programming principle and then combine this relation with some arguments regarding the continuity of the state trajectories to prove that the value was the viscosity solution. In adapting this approach, some technical difficulties arise due to the unbounded controls for the minimizing player, the weaker assumptions on the dynamics and payoff, and the fact that this is not a fixed finite time horizon problem. Thus, some preliminary lemmas are necessary, as are some variations on the method of proof. In the statement and proofs of these lemmas, Y_t will denote the solution to (5.6) for the given controls and τ will be $\inf\{t : Y_t \notin G\}$.

LEMMA 5.2. *There exists $M_3 < \infty$ such that*

$$0 \le \inf_{w\in\mathcal{W}^0} J(x,\phi[w],w) \le M_3 \qquad \forall\,\phi\in\Phi,\ \forall\, x\in\overline{G}.$$

The proof of Lemma 5.2 will be delayed until section 6. The method is standard and involves the construction of a piecewise constant control $w.$ which guarantees exit prior to some fixed time T that is independent of x and ϕ. Since the control so constructed will also be bounded, the result will follow. We note that a similar approach is used in [5]. The following result is an immediate consequence of Lemma 5.2 and the form of the cost.

LEMMA 5.3. *Let $\varepsilon_0 > 0$ be fixed, and let M_3 satisfy the conclusion of Lemma* 5.2. *Then for all $x \in \overline{G}$, $\phi \in \Phi$, and any $\tilde{w}$ that satisfies*

$$J(x,\phi[\tilde{w}],\tilde{w}) \le \inf_{w\in\mathcal{W}^0} J(x,\phi[w],w) + \epsilon_0,$$

we have the bound

$$\int_0^\tau |\tilde{w}_t|^2\,dt \le 2(M_3+\epsilon_0).$$

Now let

$$\mathcal{W}_b^0 \dodoteq \left\{w \in \mathcal{W}^0 :\ \|w\|_{L^2(0,\infty)} \le 2(M_3+\epsilon_0)\right\},$$

and note that

$$\widetilde{W}(x) = \sup_{\phi\in\Phi}\inf_{w\in\mathcal{W}_b^0} J(x,\phi[w],w).$$

LEMMA 5.4. *Let $w \in \mathcal{W}_b^0$ and $u \in \mathcal{U}^0$ be given. Then there exist $B_1, B_2 < \infty$ such that*

$$|Y_t - x| \le B_1 t + B_2\sqrt{t} \qquad \forall\, t \in [0,\tau].$$

Proof. By (5.8)

$$|Y_t - x| \le \int_0^t |f(Y_r,u_r)|\,dr + \int_0^t |\sigma(Y_r)||w_r|\,dr.$$

Under Conditions 4.1 and 4.2, there exist $C_f, C_\sigma < \infty$ such that the right-hand side of this inequality is less than or equal to

$$C_f t + C_\sigma \int_0^t |w_r|\,dr.$$

According to Lemma 5.3 this can be bounded above by

$$C_f t + C_\sigma[2(M_3+\varepsilon_0)]^{1/2}\sqrt{t}. \qquad \square$$

From Lemma 5.4, one immediately obtains the following result.

LEMMA 5.5. *Let $x \in G$, $w \in \mathcal{W}_b^0$, and $u \in \mathcal{U}^0$ be given. Then*

$$B_1\tau + B_2\sqrt{\tau} \ge d(x, G^c).$$

Lemmas 5.2–5.5 serve to bound the controls for the minimizing player and demonstrate continuity of the state with respect to time. Next, dynamic programming principles will be obtained.

THEOREM 5.6.

$$\widetilde{W}(x) = \sup_{\phi\in\Phi}\inf_{w\in\mathcal{W}_b^0}\left[\int_0^{T\wedge\tau}\left(\theta + \frac{1}{2}|w_t|^2\right)\,dt + \widetilde{W}(Y_{T\wedge\tau})\right]$$

for all $x \in G$ and $T \in [0,\infty)$.

The proof of this theorem is standard and by now a little tedious; it will be delayed until section 6. The proof of the following variation on this theorem involves only simple modifications of the proof of Theorem 5.6.

THEOREM 5.7.

$$\widetilde{W}(x) \le \sup_{\phi\in\Phi}\inf_{w\in\widehat{\mathcal{W}}_m^0}\left[\int_0^{T\wedge\tau}\left(\theta + \frac{1}{2}|w_t|^2\right)\,dt + \widetilde{W}(Y_{T\wedge\tau})\right]$$

for all $x \in G$, $T \in [0,\infty)$ *and* $m > 0$, *where*

$$\widehat{\mathcal{W}}_m^0 \doteq \left\{ w \in \mathcal{W}^0 : |w_t| \le m \ \forall\, t \right\}.$$

Given the validity of these dynamic programming principles, one can prove that the upper value $\widetilde{W}$ is a continuous viscosity solution of the Isaacs equation (5.9). The next two lemmas imply certain semicontinuity properties of $\widetilde{W}(Y.)$ that are needed.

LEMMA 5.8. *Let* $g \in C^1(\overline{G})$ *satisfy*

$$0 > -\alpha \ge -\theta - H(x_0, g_x(x_0))$$

for some $x_0 \in G$ *and* $\alpha > 0$. *Then there exists* $\phi \in \Phi$ *and* $\eta_0 > 0$ *such that for all* $w \in \mathcal{W}_b^0$ *and all* $\eta \in (0, \eta_0)$,

$$\int_0^{\eta \wedge \tau} \left\{ \theta + \frac{1}{2}|w_t|^2 + \langle f(Y_t, \phi[w]_t) + \sigma(Y_t) w_t, g_x(Y_t) \rangle \right\} dt \ge \frac{\eta \alpha}{2},$$

where Y_t *is given by* (5.8) *with initial condition* x_0.

Proof. Define

$$F(x,u,w) \doteq \theta + \frac{1}{2}|w|^2 + \langle f(x,u) + \sigma(x)w, g_x(x) \rangle, \tag{5.10}$$

and let $u_0 \in \operatorname{argmax} F(x_0, u, w)$. Note that

$$\theta + H(x, g_x(x)) = \max_{u \in U} \min_{w \in \mathbb{R}^m} F(x,u,w)$$

and that u_0 is independent of w since the Isaacs condition is satisfied. The assumption on g and x_0 implies

$$F(x_0, u_0, w) \ge \theta + H(x_0, g_x(x_0)) \ge \alpha \qquad \forall\, w \in \Re^m.$$

Let $w \in \mathcal{W}_b^0$ and $\phi[w]_t \doteq u_0$. Then by Lemma 5.4 there exists $\eta > 0$ such that

$$F(Y_t, \phi[w]_t, w_t) \ge \frac{\alpha}{2}$$

for all $t \in [0, \eta]$ and $w \in \mathcal{W}_b^0$. Integrating and using Lemma 5.5 to assert that $\eta \wedge \tau = \eta$ for η sufficiently small yields the result. □

LEMMA 5.9. *Let* $g \in C^1(\overline{G})$ *satisfy*

$$0 < \alpha \le -\theta - H(x_0, g_x(x_0))$$

for some $x_0 \in G$ *and* $\alpha > 0$. *Then there exists* $\eta > 0$ *and a bounded* $w \in \mathcal{W}^0$ *such that for all* $\phi \in \Phi$

$$\int_0^{\eta \wedge \tau} \left\{ \theta + \frac{1}{2}|w_t|^2 + \langle f(Y_t, \phi[w]_t) + \sigma(Y_t) w_t, g_x(Y_t) \rangle \right\} dt \le -\frac{\eta \alpha}{2},$$

where Y_t *is given by* (5.8) *with initial condition* x_0.

The proof of Lemma 5.9 is very similar to that of Lemma 5.8. In particular, here one lets $w_t \doteq w^*$, where $w^* \in \operatorname{argmin} F(x_0, u, w)$, with F as given by (5.10) (note that w^* is independent of u).

The way is now clear for the main theorem of this subsection. The proof is delayed until section 6.

THEOREM 5.10. $\widetilde{W}$ *is the unique continuous viscosity solution of* (5.9).

5.3. Robust interpretation. After reviewing the nonlinear H^∞ disturbance attenuation problem, the analogous robust interpretation for the limit problem considered here will be presented. For the nonlinear H^∞ disturbance attenuation problem, one typically considers games with payoffs of the form

$$\int_0^T \left[L(X_t, u_t) - \gamma^2 |w_t|^2\right] \, dt,$$

where X is the state, u is the (true) control for the minimizing player, and w is the disturbance which becomes the control for the maximizing player. L is the running cost, and it is often taken to be a quadratic function. Further, one assumes an initial value for the state, $X_0 = x_0$, such that $L(x_0, 0) = 0$. If there exists a strategy for the minimizing player (ideally a feedback control leading to well-defined dynamics) such that the value of the game is zero for all T, then

$$\left[\int_0^T L(X_t, u_t)\, dt\right]^{1/2} \leq \gamma \|w\|_{L^2(0,T)} \qquad \forall\, T < \infty.$$

Thus there is a bound for the cost in the form of a product of the disturbance attenuation constant γ and the L^2 norm of the disturbance. This bound is valid for all time horizons and disturbances. See, for instance, [1, 3, 22, 24, 31].

For the escape time problem, large time averages do not make any sense, since it is the transient behavior that is of primary importance. Because we cannot eliminate escape entirely, the best one should hope for in the robust problem is a bound on the escape time in terms of the energy of the player representing the disturbance. Clearly, this bound will depend on the initial position of the controlled process. For the case without maximizing control, let

$$\overline{W}(x) \doteq \widetilde{W}(x) = \inf_{w \in \mathcal{W}^0} J(x, w) \qquad \forall\, x \in G.$$

(This new notation is being introduced so that both the controlled and uncontrolled cases can be treated together.) For the case with maximizing control, choose some (optimal or suboptimal) strategy ϕ_0 given in a feedback form such that the dynamics are well defined. Let $\overline{W}$ be the value of the game with this control, i.e.,

$$\overline{W}(x) \doteq \inf_{w \in \mathcal{W}^0} J(x, \phi_0[w], w) \qquad \forall\, x \in G.$$

Then by (5.1) and (5.2) in the case without control (or by their analogues in the case with maximizing control),

$$\overline{W}(x) \leq \left[\theta + \frac{1}{2\tau} \int_0^\tau |w_t|^2 \, dt\right] \tau \qquad \forall\, w \in L^2, \ \forall\, x \in G,$$

or

$$\tau \geq \frac{\overline{W}(x)}{\theta + \frac{1}{2\tau} \int_0^\tau |w_t|^2 \, dt} \qquad \forall\, w \in L^2, \ \forall\, x \in G. \tag{5.11}$$

Let

$$\mathcal{W}^P \doteq \left\{ w : [0, \infty) \to \mathbb{R}^m : w \text{ is measurable and } \frac{1}{T} \int_0^T |w_t|^2 \, dt \leq P \quad \forall\, T \in [0, \infty) \right\}.$$

Then (5.11) has the interpretation

$$\tau \geq \frac{\overline{W}(x)}{\theta + \frac{1}{2}P} \qquad \forall\, w \in \mathcal{W}^P,\ \forall\, x \in G.$$

This is a lower bound on the escape time as a function of the power of the input noise. It is analogous to the attenuation bound of H^∞ control.

6. Proofs. This section contains proofs for results that appeared in sections 4 and 5.

6.1. Proofs for section 4.

6.1.1. Proof of Theorem 4.2. Let $u \in \mathcal{U}_\nu$ and $x \in G$. By (4.1), for any bounded, F_t-progressively measurable process w, one has

$$\begin{aligned} X_t^\varepsilon &= x + \int_0^t f(X_r^\varepsilon, u_r)\, dr + \sqrt{\varepsilon} \int_0^t \sigma(X_r^\varepsilon)\, dB_r \\ &= x + \int_0^t [f(X_r^\varepsilon, u_r) + \sigma(X_r^\varepsilon) w_r]\, dr + \sqrt{\varepsilon} \int_0^t \sigma(X_r^\varepsilon)\, dB_r - \int_0^t \sigma(X_r^\varepsilon) w_r\, dr \\ &= x + \int_0^t [f(X_r^\varepsilon, u_r) + \sigma(X_r^\varepsilon) w_r]\, dr + \sqrt{\varepsilon} \int_0^t \sigma(X_r^\varepsilon)\, dB_r^0, \end{aligned} \tag{6.1}$$

where the last equality defines B^0.

Fix any $T < \infty$. We would like a probability measure P^0 under which B^0 is a Brownian motion. By Girsanov's theorem [25, p. 191], such a measure exists on (Ω, F_T). For any set A in F_T, this measure satisfies

$$P^0(A) = \int_A \exp\left[\sqrt{\frac{1}{\varepsilon}} \int_0^T \langle w_r,\, dB_r^0 \rangle + \frac{1}{2\varepsilon} \int_0^T |w_r|^2\, dr \right] P(d\omega).$$

In terms of this measure, we can write

$$E_x \exp\left[-\frac{\theta(\tau^\varepsilon \wedge T)}{\varepsilon} \right] = E_x^0 \exp \frac{1}{\varepsilon} \left[-\int_0^{\tau^\varepsilon \wedge T} \left(\theta + \frac{1}{2}|w_r|^2 \right) dr - \varepsilon^{1/2} \int_0^{\tau^\varepsilon \wedge T} \langle w_r,\, dB_r^0 \rangle \right]. \tag{6.2}$$

Since $\widetilde{V}^\varepsilon$ is a solution of the PDE (4.3), an application of Ito's rule with the new dynamics (6.1) gives

$$\begin{aligned} \widetilde{V}^\varepsilon(X_t^\varepsilon) - \widetilde{V}^\varepsilon(x) &= \int_0^t \left[\frac{\varepsilon}{2} \mathrm{tr} \left[a(X_r^\varepsilon) \widetilde{V}_{xx}^\varepsilon(X_r^\varepsilon) \right] + \langle f(X_r^\varepsilon, u_r), \widetilde{V}_x^\varepsilon(X_r^\varepsilon) \rangle \right. \\ &\quad \left. + \langle \sigma(X_r^\varepsilon) w_r, \widetilde{V}_x^\varepsilon(X_r^\varepsilon) \rangle \right] dr \\ &\quad + \varepsilon^{1/2} \int_0^t \langle \widetilde{V}_x^\varepsilon(X_r^\varepsilon), \sigma(X_r^\varepsilon) dB_r^0 \rangle \\ &\leq \int_0^t [-\theta + \frac{1}{2} \langle \widetilde{V}_x^\varepsilon(X_r^\varepsilon), a(X_r^\varepsilon) \widetilde{V}_x^\varepsilon(X_r^\varepsilon) \rangle + \langle \sigma(X_r^\varepsilon) w_r, \widetilde{V}_x^\varepsilon(X_r^\varepsilon) \rangle] \, dr \\ &\quad + \varepsilon^{1/2} \int_0^t \langle \widetilde{V}_x^\varepsilon(X_r^\varepsilon), \sigma(X_r^\varepsilon) dB_r^0 \rangle. \end{aligned} \tag{6.3}$$

If the disturbance process is taken to be $w_r = -\sigma^T(X_r^\varepsilon)\widetilde{V}_x^\varepsilon(X_r^\varepsilon)$, then (6.3) implies

$$\begin{aligned}\widetilde{V}^\varepsilon(X_t^\varepsilon) - \widetilde{V}^\varepsilon(x) &\le -\int_0^t \left(\theta + \frac{1}{2}|w_r|^2\right) dr + \varepsilon^{1/2}\int_0^t \langle \widetilde{V}_x^\varepsilon(X_r^\varepsilon), \sigma(X_r^\varepsilon) dB_r^0\rangle \\ &= -\int_0^t \left(\theta + \frac{1}{2}|w_r|^2\right) dr - \varepsilon^{1/2}\int_0^t \langle w_r,\, dB_r^0\rangle.\end{aligned} \tag{6.4}$$

Combining (6.2) and (6.4), one obtains

$$\begin{aligned}E_x \exp\left[-\frac{\theta(\tau^\varepsilon \wedge T)}{\varepsilon}\right] &\ge E_x^0 \exp\frac{1}{\varepsilon}[\widetilde{V}^\varepsilon(X^\varepsilon_{\tau^\varepsilon\wedge T}) - \widetilde{V}^\varepsilon(x)] \\ &= \exp\left[-\frac{1}{\varepsilon}\widetilde{V}^\varepsilon(x)\right] E_x^0 \exp\frac{1}{\varepsilon}\widetilde{V}^\varepsilon(X^\varepsilon_{\tau^\varepsilon\wedge T}).\end{aligned}$$

Note that $\widetilde{V}^\varepsilon(\cdot)$ is bounded on $\overline{G}$ and that $\widetilde{V}^\varepsilon(X^\varepsilon_{\tau^\varepsilon\wedge T}) \to 0$ in distribution under E_x^0 as $T \to \infty$. Sending $T \to \infty$ and applying the dominated convergence theorem to the left side of the last display gives

$$E_x \exp\left[-\frac{\theta\tau^\varepsilon}{\varepsilon}\right] \ge \exp\left[-\frac{1}{\varepsilon}\widetilde{V}^\varepsilon(x)\right],$$

which implies the first assertion of the theorem.

To obtain the second assertion, simply note that such a $\bar{u}$ exists (see, for example, [17]) and use $u_t = \bar{u}(\bar{X}_t^\varepsilon)$ to obtain equality in the argument above. □

6.1.2. Proof of Lemma 4.3. The bounds are obtained from standard applications of the comparison principle. In particular, for the lower bound, one compares with $Z(x) \doteq 0$. For the upper bound, one compares V^ε with $Z(x) \doteq A + q \cdot x$, where A and q are constants independent of ε. Using the uniform ellipticity, it is easy to show that for $|q|$ sufficiently large there exists $\delta > 0$ such that

$$0 > -\delta \ge \theta + \frac{\varepsilon}{2}\mathrm{tr}\,[a(x)Z_{xx}(x)] + H(x, Z_x(x)) \qquad \forall\, x \in G,\ \ \forall\, \varepsilon > 0.$$

Further, for A sufficiently large, $Z(x) \ge 0 = V^\varepsilon(x)$ for all $x \in \partial G$. The upper bound then follows from the comparison principle and the compactness of $\overline{G}$. □

6.1.3. Proof of Lemma 4.4. The barrier method (see, for instance, [21]) will be used. Fix $x_0 \in \partial G$ and note that by Lemma 4.3

$$V^\varepsilon(x) \ge 0 = V^\varepsilon(x_0) \qquad \forall\, x \in \overline{G}. \tag{6.5}$$

Recall Condition 4.1, which states that the uniform exterior sphere condition holds. This implies there exists r independent of x_0 and $y \in \overline{G}^c$ such that $|x_0 - y| = r$ and

$$B_r(y) \cap \overline{G} = \{x_0\}.$$

Fix this value of y, and define

$$Z(x) \doteq \alpha(|x - y| - r) \qquad \forall\, x \in \overline{G}, \tag{6.6}$$

where α is a positive constant that will be chosen independent of $\varepsilon \in (0, \varepsilon_0)$ and $x_0 \in \partial G$. By (6.5) and (6.6), it suffices to prove that

$$V^\varepsilon(x) \le Z(x) \qquad \forall\, x \in G,\ \ \forall\, \varepsilon \in (0, \varepsilon_0). \tag{6.7}$$

On the boundary, of course,

$$V^\varepsilon(x) = 0 \le Z(x). \tag{6.8}$$

In terms of $v_x \doteq \frac{x-y}{|x-y|} = \frac{1}{\alpha} Z_x(x)$,

$$\begin{aligned}
&\theta + \frac{\varepsilon}{2}\mathrm{tr}\,[a(x)Z_{xx}(x)] + H(x, Z_x(x)) \\
&\quad = \theta + \frac{\varepsilon}{2}\frac{\alpha}{|x-y|}\left[\mathrm{tr}(a(x)) - \langle v_x, a(x)v_x\rangle\right] + \alpha \max_{v\in U}\langle f(x,v), v_x\rangle - \frac{\alpha^2}{2}\langle v_x, a(x)v_x\rangle \\
&\quad \le \theta + \alpha\left\{\frac{\varepsilon}{2|x-y|}\left[C_a - \langle v_x, a(x)v_x\rangle\right] + C_f - \frac{\alpha}{2}\langle v_x, a(x)v_x\rangle\right\},
\end{aligned}$$

where

$$C_a \doteq \max_{x\in\overline{G}} |\mathrm{tr}(a(x))| \text{ and } C_f \doteq \max_{(x,v)\in\overline{G}\times U} |f(x,v)|.$$

Using the uniform ellipticity of a, for α sufficiently large we have

$$\theta + \frac{\varepsilon}{2}\mathrm{tr}\,[a(x)Z_{xx}(x)] + H(x, Z_x(x)) \le \theta + \alpha\left\{\frac{C_a\varepsilon}{2r} + C_f - \frac{\alpha}{2}\mu\right\} < 0 \tag{6.9}$$

for all $\varepsilon \in (0, \varepsilon_0)$ and $x_0 \in \partial G$. From (6.8), (6.9), and the comparison principle, one obtains (6.7). □

6.1.4. Proof of Theorem 4.5. We recall the definitions of V^* and V_* given in equation (4.7). As noted in section 4, it is sufficient to prove that

$$V^*(x) \le W(x) \le V_*(x) \qquad \forall\, x \in \overline{G}.$$

By a simple modification of the proof of [18, Prop. 7.6.1], one can show that V^* is a subsolution of (4.5) and V_* is a supersolution. In addition, Lemma 4.4 implies that the boundary conditions are achieved pointwise. We now show that $W \le V_*$ on G; the proof that $V^* \le W$ is similar and is thus omitted.

Let $\delta > 0$ and suppose

$$\tilde{a} \doteq \min_{x\in\overline{G}}\left[(1+\delta)V_*(x) - W(x)\right] < 0. \tag{6.10}$$

By (4.8) and the respective semicontinuity and continuity properties of V_* and $\widetilde{W}$, the minimum in (6.10) occurs at some point $\tilde{x} \in G$. Let

$$\phi^\eta(x,y) \doteq (1+\delta)V_*(x) - W(y) + \frac{1}{2\eta}|x-y|^2 \tag{6.11}$$

and

$$(x^\eta, y^\eta) \in \mathrm{argmin}\,\phi^\eta(x,y).$$

It is easy to see that

$$|x^\eta - y^\eta| \to 0 \ \text{ as } \ \eta \downarrow 0. \tag{6.12}$$

Further, since (x^η, y^η) minimizes ϕ^η,

$$\phi^\eta(x^\eta, y^\eta) \le \phi^\eta(x^\eta, x^\eta).$$

When combined with (6.11) this implies

$$\frac{|x^\eta - y^\eta|^2}{\eta} \le 2m_W(|x^\eta - y^\eta|),$$

where $m_W(\cdot)$ is the modulus of continuity of W over $\overline{G}$. Thus by (6.12)

$$\frac{|x^\eta - y^\eta|^2}{\eta} \to 0 \quad \text{as} \quad \eta \downarrow 0. \tag{6.13}$$

Since $\overline{G}$ is compact, there exists a sequence $\eta_k \downarrow 0$ and an $x^0 \in \overline{G}$ such that

$$x^{\eta_k} \to x^0 \text{ and } y^{\eta_k} \to x^0$$

as $k \to \infty$. To simplify the notation, we retain η as the index of this convergent sequence. By the choice of (x^η, y^η), (6.10), and (6.11),

$$\phi^\eta(x^\eta, y^\eta) \le \tilde{a} < 0 \qquad \forall\, \eta > 0.$$

Then by (6.13), the lower semicontinuity of V_*, and the continuity of W,

$$(1+\delta)V_*(x^0) - W(x^0) < 0. \tag{6.14}$$

If x^0 were in ∂G, then (6.14) would contradict (4.7) and (4.8). Therefore $x^0 \in G$, which implies that for η sufficiently small (in our reindexed subsequence)

$$x^\eta, y^\eta \in G.$$

Now let

$$\psi(x) \doteq \frac{1}{1+\delta}\left[W(y^\eta) - \frac{1}{2\eta}|x - y^\eta|^2\right]$$

and note that $V_* - \psi$ has a minimum at x^η. Since V_* is a supersolution, one has

$$-\theta - H(x^\eta, \psi_x(x^\eta)) \ge 0,$$

which implies

$$-(1+\delta)\theta \ge \frac{1}{\eta}\max_{v\in U}\langle -f(x^\eta, v), (x^\eta - y^\eta)\rangle - \frac{1}{2\eta^2(1+\delta)}\langle (x^\eta - y^\eta), a(x^\eta)(x^\eta - y^\eta)\rangle.$$

(6.15)

Also, let

$$\widetilde{\psi}(y) \doteq (1+\delta)V_*(x^\eta) + \frac{1}{2\eta}|x^\eta - y|^2,$$

which implies that $W - \widetilde{\psi}$ has a maximum at y^η. Since W is a subsolution (as well as a supersolution), one has

$$-\theta - H(y^\eta, \widetilde{\psi}_x(y^\eta)) \le 0,$$

which implies

$$\theta \geq -\frac{1}{\eta}\max_{v\in U}\langle -f(y^\eta, v), (x^\eta - y^\eta)\rangle + \frac{1}{2\eta^2}\langle (x^\eta - y^\eta), a(y^\eta)(x^\eta - y^\eta)\rangle. \tag{6.16}$$

Adding (6.15) and (6.16) yields

$$\begin{aligned} -\delta\theta \geq{}& \frac{1}{\eta}\max_{v\in U}\langle -f(x^\eta, v), (x^\eta - y^\eta)\rangle - \frac{1}{\eta}\max_{v\in U}\langle -f(y^\eta, v), (x^\eta - y^\eta)\rangle \\ &+ \frac{1}{2\eta^2}\left\langle (x^\eta - y^\eta), \left[a(y^\eta) - \frac{1}{1+\delta}a(x^\eta)\right](x^\eta - y^\eta)\right\rangle. \end{aligned} \tag{6.17}$$

Recall that by Conditions 4.2 and 4.3, f and a are Lipschitz continuous with constants K_f and K_a on $\overline{G}$, respectively. Consequently, (6.17) implies

$$\begin{aligned} -\delta\theta \geq{}& \frac{|x^\eta - y^\eta|^2}{2\eta^2}\left[-2\eta K_f - K_a|x^\eta - y^\eta|\right] \\ &+ \frac{1}{2\eta^2}\left\langle (x^\eta - y^\eta), \left(1 - \frac{1}{1+\delta}\right)a(x^\eta)(x^\eta - y^\eta)\right\rangle, \end{aligned}$$

and since a is uniformly elliptic with constant μ,

$$-\delta\theta \geq \frac{|x^\eta - y^\eta|^2}{2\eta^2}\left[-2\eta K_f - K_a|x^\eta - y^\eta| + \left(1 - \frac{1}{1+\delta}\right)\mu\right]. \tag{6.18}$$

But by (6.12), for η sufficiently small,

$$-2\eta K_f - K_a|x^\eta - y^\eta| + \left(1 - \frac{1}{1+\delta}\right)\mu > 0.$$

This implies that for η sufficiently small the right-hand side of (6.18) is nonnegative, which is a contradiction. Therefore, (6.10) is false, and consequently

$$\min_{x\in\overline{G}}(1+\delta)V_*(x) - W(x) \geq 0.$$

Since this is true for all $\delta > 0$, $W \leq V_*$ for all $x \in \overline{G}$. □

6.2. Proofs for section 5.

6.2.1. Proof of Lemma 5.2. Since the lower bound is obvious, only the upper bound will be considered. It is sufficient to prove that for each $\phi \in \Phi$ there exists $w \in \mathcal{W}^0$ such that $J(x, \phi[w], w) \leq M_3$. The control w will be constructed in a feedback fashion; the existence of an open loop w with the same values will be clear.

Let $b \in \Re^n$ be any vector with $|b| = 1$. Let $t_n = n\Delta$ for all nonnegative integers n where the value of Δ is yet to be specified. The control w will be constant over each interval $[t_n, t_{n+1})$. Let C_f be a bound for f over $\overline{G}$ and fix $\phi \in \Phi$. Define

$$w_t \doteq w^0 \qquad \forall\, t \in [t_0, t_1),$$

where $w^0 = 2C_f\sigma^{-1}(x)b$ and x is the initial state. Let the dynamics over $[t_0, t_1)$ be given by (5.8) with controls w and $\phi[w]$. Then for $t \in [t_0, t_1]$,

$$\begin{aligned} \langle Y_t - x, b\rangle &\geq -C_f t + 2C_f t + 2C_f\int_0^t \langle b, \left[\sigma(Y_r)\sigma^{-1}(x) - I\right]b\rangle\, dr \\ &\geq C_f t + 2C_f\int_0^t \langle b, \left[\sigma(Y_r) - \sigma(x)\right]\sigma^{-1}(x)b\rangle\, dr. \end{aligned} \tag{6.19}$$

But by Condition 4.3, there exists $\overline{m}_\sigma < \infty$ such that $\|\sigma^{-1}(x)\| \le \overline{m}_\sigma$ for all $x \in \overline{G}$. Further, since $\sigma \in C^1(\overline{G})$, it is Lipschitz on $\overline{G}$ with constant K_σ. Employing these bounds in (6.19) yields

$$\langle Y_t - x, b\rangle \ge C_f t - 2C_f \overline{m}_\sigma K_\sigma \int_0^t |Y_r - x|\, dr.$$

Since there exists $\overline{C}$ (depending on the bounds on f, σ, and σ^{-1}) such that $|Y_t - x| \le \overline{C}t$, one has

$$\langle Y_t - x, b\rangle \ge C_f t - C_f \overline{m}_\sigma K_\sigma \overline{C} t^2,$$

which implies that there is a $\Delta > 0$ (independent of x) such that

$$\langle Y_t - x, b\rangle \ge \frac{C_f}{2} t \qquad \forall\, t \in [0, \Delta]. \tag{6.20}$$

This is the desired choice for Δ.

Turning now to the second segment, let

$$w_t = w^1 \doteq 2C_f \sigma^{-1}(Y_{t_1}) b \qquad \forall\, t \in [t_1, t_2).$$

Proceeding as for the first segment, one finds

$$\langle Y_t - Y_{t_1}, b\rangle \ge \frac{C_f}{2}(t - t_1) \qquad \forall\, t \in [t_1, t_2]. \tag{6.21}$$

Combining (6.20) and (6.21) yields

$$\langle Y_{t_2} - x, b\rangle \ge 2\frac{C_f}{2}\Delta.$$

Continuing this process, one has

$$w_t = 2C_f \sigma^{-1}(Y_{t_n}) b \qquad \forall\, t \in [t_n, t_{n+1}]$$

and

$$\langle Y_{t_{n+1}} - x, b\rangle \ge (n+1)\frac{C_f}{2}\Delta \qquad \forall\, n.$$

Therefore,

$$\tau \le \frac{2\mathrm{diam}(\overline{G})}{C_f},$$

and consequently

$$M_3 \doteq \frac{2\mathrm{diam}(\overline{G})}{C_f}\left[\theta + \frac{1}{2}[2C_f\overline{m}_\sigma]^2\right] \ge J(x, \phi[w], w). \qquad \square$$

6.2.2. Proof of Theorem 5.6. The proof follows the standard form. The equality in the dynamic programming principle is obtained by proving inequalities in both directions. Let τ_x indicate the time to escape given the initial state x.

Define

$$R(x) \doteq \sup_{\phi\in\Phi}\inf_{w\in\mathcal{W}_b^0}\left[\int_0^{T\wedge\tau_x}\left(\theta+\frac{1}{2}|w_t|^2\right)\,dt+\widetilde{W}(Y_{T\wedge\tau_x})\right], \tag{6.22}$$

and let $\varepsilon\in(0,1]$. Then there exists $\widetilde{\phi}\in\Phi$ such that

$$R(x)\le\inf_{w\in\mathcal{W}_b^0}\left[\int_0^{T\wedge\tau_x}\left(\theta+\frac{1}{2}|w_t|^2\right)\,dt+\widetilde{W}(Y_{T\wedge\tau_x})\right]+\varepsilon \tag{6.23}$$

when $\widetilde{\phi}[w]$ is used to define Y in (5.8). For any $y\in G$,

$$\widetilde{W}(y)=\sup_{\phi\in\Phi}\inf_{w\in\mathcal{W}_b^0}\left[\int_0^{\tau_y}\left(\theta+\frac{1}{2}|w_t|^2\right)\,dt\right],$$

which implies there exists $\widetilde{\phi}'_y\in\Phi$ such that

$$\widetilde{W}(y)\le\inf_{w\in\mathcal{W}_b^0}\left[\int_0^{\tau_y}\left(\theta+\frac{1}{2}|w_t|^2\right)\,dt\right]+\varepsilon \tag{6.24}$$

when $\widetilde{\phi}'_y[w]$ is used in the dynamics (5.8).

Define a strategy $\widehat{\phi}$ as follows. For each $w\in\mathcal{W}_b^0$, let

$$\widehat{\phi}[w]_t\doteq\begin{cases}\widetilde{\phi}[w]_t & \text{if } t\le T,\\ \widetilde{\phi}'_{Y_T}[w_{\cdot-T}]_{t-T} & \text{if } t>T.\end{cases}$$

Note that $\widehat{\phi}\in\Phi$. (6.23) and (6.24) imply

$$R(x)\le\inf_{w^1\in\mathcal{W}_b^0}\inf_{w^2\in\mathcal{W}_b^0}\left[\int_0^{T\wedge\tau_x}\left(\theta+\frac{1}{2}|w_t^1|^2\right)\,dt\;+I(T<\tau_x)\int_T^{\hat{\tau}_{Y_T}}\left(\theta+\frac{1}{2}|w_t^2|^2\right)\,dt\right]+2\varepsilon,$$

where

$$Y_t=x+\int_0^t\left[f(Y_r,\widehat{\phi}[\widehat{w}]_r)+\sigma(Y_r)\widehat{w}_r\right]\,dr,$$

$$\widehat{w}_t\doteq\begin{cases}w_t^1 & \text{if } t\le T,\\ w_t^2 & \text{if } t>T,\end{cases}$$

and where $\hat{\tau}_y$ is the first escape time of Y_t after T given $Y_T=y$. This implies

$$R(x)\le\inf_{w\in\mathcal{W}_b^0}\left[\int_0^{\tau_x}\left(\theta+\frac{1}{2}|w_t|^2\right)\,dt\right]+2\varepsilon\le\sup_{\phi\in\Phi}\inf_{w\in\mathcal{W}_b^0}J(x,\phi[w],w)+2\varepsilon,$$

which, since $\varepsilon\in(0,1]$ was arbitrary, implies

$$R(x)\le\widetilde{W}(x).$$

Now the reverse inequality is proved. Given $\varepsilon > 0$, there exists $\widetilde{\phi} \in \Phi$ such that

$$\widetilde{W}(x) \le \inf_{w \in \mathcal{W}_b^0} \left[\int_0^{\tau_x} \left(\theta + \frac{1}{2}|w_t|^2 \right) dt \right] + \varepsilon \tag{6.25}$$

when $\widetilde{\phi}[w]$ is used in the dynamics. On the other hand, by (6.22),

$$R(x) \ge \inf_{w \in \mathcal{W}_b^0} \left[\int_0^{T \wedge \tau_x} \left(\theta + \frac{1}{2}|w_t|^2 \right) dt + \widetilde{W}(Y_{T \wedge \tau_x}) \right],$$

where again $\widetilde{\phi}[w]$ is being used in the dynamics. This implies there exists $\widetilde{w} \in \mathcal{W}_b^0$ such that

$$R(x) \ge \int_0^{T \wedge \tau_x} \left(\theta + \frac{1}{2}|\widetilde{w}_t|^2 \right) dt + \widetilde{W}(Y_{T \wedge \tau_x}) - \varepsilon. \tag{6.26}$$

If $T < \tau_x$ then we can continue to use the strategy $\tilde{\phi}$ for $t \ge T$. In this case, the definition of $\widetilde{W}$ implies the existence of $\hat{w}'_t$ defined on $[T, \infty)$ such that $\hat{w}'_{\cdot - T} \in \mathcal{W}_b^0$ and

$$\widetilde{W}(Y_{T \wedge \tau_x}) \ge \int_T^{\hat{\tau}_{Y_T}} \left(\theta + \frac{1}{2}|\hat{w}'_t|^2 \right) dt - \varepsilon. \tag{6.27}$$

Now let

$$\hat{w}_t = \begin{cases} \widetilde{w}_t & \text{if } t \le T, \\ \hat{w}'_t & \text{if } t > T. \end{cases}$$

Then (6.26) and (6.27) imply

$$R(x) \ge \int_0^{\tau_x} \left(\theta + \frac{1}{2}|\hat{w}_t|^2 \right) dt - 2\varepsilon$$

when the strategy $\tilde{\phi}$ is used. Equation (6.25) then implies $R(x) \ge \widetilde{W}(x) - 3\varepsilon$. Since $\varepsilon > 0$ was arbitrary, the proof is complete. □

6.2.3. Proof of Theorem 5.10. First note that continuity of $\widetilde{W}$ follows from Lemma 5.4 and constructions similar to those used in the proof of Lemma 5.2. Once it has been proved that $\widetilde{W}$ is a viscosity solution, the uniqueness will then follow from the comparison principle used in the proof of Theorem 4.5. Thus, it is sufficient to prove that $\widetilde{W}$ is a viscosity solution. This is done by proving that it is both a supersolution and a subsolution.

Suppose we are given $g \in C^1(\overline{G})$ such that $\widetilde{W} - g$ has a local minimum at $x_0 \in G$. To prove that $\widetilde{W}$ is a viscosity supersolution, it must be shown that

$$-\theta - H(x_0, g_x(x_0)) \ge 0.$$

If the last inequality is not valid, then there exists $\alpha > 0$ such that

$$-\alpha \ge -\theta - H(x_0, g_x(x_0)).$$

Then, by Lemmas 5.5 and 5.8, there exist $\phi \in \Phi$ and $\eta > 0$ such that for all $w \in \mathcal{W}_b^0$

$$\int_0^{\eta} \left\{ \theta + \frac{1}{2}|w_t|^2 + \langle [f(Y_t, \phi[w]_t) + \sigma(Y_t) w_t], g_x(Y_t) \rangle \right\} dt \ge \frac{\eta \alpha}{2},$$

where Y_t is given by (5.8) with initial condition x_0 and controls w and $\phi[w]$. This implies

$$\sup_{\phi\in\Phi}\inf_{w\in\mathcal{W}_b^0}\left[\int_0^\eta\left\{\theta+\frac{1}{2}|w_t|^2+\langle[f(Y_t,\phi[w]_t)+\sigma(Y_t)w_t],g_x(Y_t)\rangle\right\}dt\right]\geq\frac{\eta\alpha}{2}. \tag{6.28}$$

On the other hand, since $\widetilde{W}-g$ has a local minimum at x_0, Lemma 5.4 implies

$$\widetilde{W}(x_0)-g(x_0)\leq\widetilde{W}(Y_t)-g(Y_t)\qquad\forall\, t\leq\eta,\ \forall\,\phi\in\Phi,\ \forall\, w\in\mathcal{W}_b^0, \tag{6.29}$$

where $\eta>0$ may have to be reduced in size. Also, since η has been chosen so that $\tau\geq\eta$, Theorem 5.6 implies

$$\widetilde{W}(x_0)=\sup_{\phi\in\Phi}\inf_{w\in\mathcal{W}_b^0}\left[\int_0^\eta\left(\theta+\frac{1}{2}|w_t|^2\right)dt+\widetilde{W}(Y_\eta)\right]. \tag{6.30}$$

Substituting (6.29) into (6.30) yields

$$\begin{aligned}0&\geq\sup_{\phi\in\Phi}\inf_{w\in\mathcal{W}_b^0}\left[\int_0^\eta\left(\theta+\frac{1}{2}|w_t|^2\right)dt+g(Y_\eta)-g(x_0)\right]\\&=\sup_{\phi\in\Phi}\inf_{w\in\mathcal{W}_b^0}\left[\int_0^\eta\left\{\left(\theta+\frac{1}{2}|w_t|^2\right)+\langle[f(Y_t,\phi[w]_t)+\sigma(Y_t)w_t],g_x(Y_t)\rangle\right\}dt\right].\end{aligned} \tag{6.31}$$

But (6.28) and (6.31) form a contradiction. Therefore $\widetilde{W}$ is a supersolution.

The analogous proof that $\widetilde{W}$ is a subsolution is as follows. Consider $g\in C^1(\overline{G})$ such that $\widetilde{W}-g$ has a local maximum at $x_0\in G$. To prove that $\widetilde{W}$ is a subsolution, it must be shown that

$$-\theta-H(x_0,g_x(x_0))\leq 0.$$

If this inequality is not true, then there exists $\alpha>0$ such that

$$\alpha\leq-\theta-H(x_0,g_x(x_0)).$$

Then, by Lemmas 5.5 and 5.9, there exist a bounded $w\in\mathcal{W}^0$ and $\eta>0$ such that for all $\phi\in\Phi$

$$\int_0^\eta\left\{\theta+\frac{1}{2}|w_t|^2+\langle[f(Y_t,\phi[w]_t)+\sigma(Y_t)w_t],g_x(Y_t)\rangle\right\}dt\leq-\frac{\eta\alpha}{2},$$

where Y_t is given by (5.8) with initial condition x_0 and controls w and $\phi[w]$. This implies

$$\sup_{\phi\in\Phi}\inf_{w\in\mathcal{W}_*^0}\left[\int_0^\eta\left\{\theta+\frac{1}{2}|w_t|^2+\langle[f(Y_t,\phi[w]_t)+\sigma(Y_t)w_t],g_x(Y_t)\rangle\right\}dt\right]\leq-\frac{\eta\alpha}{2}, \tag{6.32}$$

where $\mathcal{W}_*^0=\{w\in\mathcal{W}^0:\,|w_t|\leq M_*\ \forall\, t\}$ and M_* is the bound on the function w whose existence is asserted in Lemma 5.9.

On the other hand, since $\widetilde{W}-g$ has a local maximum at x_0, Lemma 5.4 implies

$$\widetilde{W}(x_0)-g(x_0)\geq\widetilde{W}(Y_t)-g(Y_t)\qquad\forall\, t\leq\eta,\ \forall\,\phi\in\Phi,\ \forall\, w\in\mathcal{W}_*^0, \tag{6.33}$$

where $\eta > 0$ may have to be reduced in size. Also, from Theorem 5.7,

$$\widetilde{W}(x_0) \le \sup_{\phi\in\Phi}\inf_{w\in\mathcal{W}^0_*}\left[\int_0^\eta \left(\theta + \frac{1}{2}|w_t|^2\right)\,dt + \widetilde{W}(Y_\eta)\right]. \tag{6.34}$$

Combining (6.33) and (6.34) yields

$$\begin{aligned} 0 &\le \sup_{\phi\in\Phi}\inf_{w\in\mathcal{W}^0_*}\left[\int_0^\eta \left(\theta + \frac{1}{2}|w_t|^2\right)\,dt + g(Y_\eta) - g(x_0)\right] \\ &= \sup_{\phi\in\Phi}\inf_{w\in\mathcal{W}^0_*}\left[\int_0^\eta \left\{\left(\theta + \frac{1}{2}|w_t|^2\right) + \langle [f(Y_t,\phi[w]_t) + \sigma(Y_t)w_t], g_x(Y_t)\rangle\right\}\,dt\right]. \end{aligned} \tag{6.35}$$

But (6.32) and (6.35) form a contradiction. Therefore $\widetilde{W}$ is a subsolution. ∎

Acknowledgment. The second author would like to thank H. Mete Soner for many helpful discussions.

REFERENCES

[1] J. A. Ball, J. W. Helton, and M. L. Walker, *H^∞ control for nonlinear systems with output feedback*, IEEE Trans. Automat. Control, 38 (1993), pp. 117–164.

[2] G. Barles and B. Perthame, *Exit time problems in optimal control and the vanishing viscosity method*, SIAM J. Control Optim., 26 (1988), pp. 1133–1148.

[3] T. Basar and P. Bernhard, *H^∞ Optimal Control and Related Minimax Design Problems*, Birkhauser, Boston, 1991.

[4] A. Benveniste, M. Metivier, and P. Priouret, *Adaptive Algorithms and Stochastic Approximation*, Springer-Verlag, Berlin, 1990.

[5] P. Cannarsa and C. Sinestrari, *Convexity Properties of the Minimum Time Function*, Tech. Rep., Dept. of Mathematics, Universita di Roma "Tor Vergata," 1994.

[6] C. L. DeMarco, *Random perturbations in power system dynamics: A large deviations approach to security assessment*, in Proceedings of the 22nd Annual Conference on Information Sciences and Systems, Dept. of Electrical Engineering, Princeton University, Princeton, NJ, March 1988.

[7] P. Dupuis and H. J. Kushner, *Large deviations estimates for systems with small noise effects, and applications to stochastic systems theory*, SIAM J. Control Optim., 24 (1986), pp. 979–1008.

[8] P. Dupuis and H. J. Kushner, *Minimizing exit probabilities; a large deviations approach*, SIAM J. Control Optim., 27 (1989), pp. 432–445.

[9] P. Dupuis and H. J. Kushner, *Stochastic systems with small noise, analysis and simulation; a phase locked loop example*, SIAM J. Appl. Math., 47 (1987), pp. 643–661.

[10] R. J. Elliott and N. J. Kalton, *The Existence of Value in Differential Games*, Mem. Amer. Math. Soc. 126, AMS, Providence, RI, 1972.

[11] L. C. Evans and P. E. Souganidis, *Differential games and representation formulas for solutions of Hamilton–Jacobi–Isaacs equations*, Indiana Univ. Math. J., 33 (1984), pp. 773–797.

[12] E. Fernandez-Gaucherand and S. I. Marcus, *Risk-sensitive optimal control of hidden Markov models: A case study*, in Proc. 33rd IEEE Conf. on Decision and Control, 1994, IEEE, Piscataway, NJ, pp. 1657–1662.

[13] W. H. Fleming, *Exit probabilities and optimal stochastic control*, Applied Math. Optim., 4 (1978), pp. 329–346.

[14] W. H. Fleming and W. M. McEneaney, *Risk sensitive control on an infinite time horizon*, SIAM J. Control Optim., 33 (1995), pp. 1881–1915.

[15] W. H. Fleming and W. M. McEneaney, *Risk sensitive control with ergodic cost criteria*, in Proc. 31st IEEE Conf. on Decision and Control, IEEE, Piscataway, NJ, 1992.

[16] W. H. Fleming and W. M. McEneaney, *Risk Sensitive Optimal Control and Differential Games*, Lecture Notes in Control and Inform. Sci. 184, Spinger-Verlag, Berlin, 1992, pp. 185–197.

[17] W. H. Fleming and R. Rishel, *Deterministic and Stochastic Optimal Control*, Springer-Verlag, Berlin, 1975.

[18] W. H. Fleming and H. M. Soner, *Controlled Markov Processes and Viscosity Solutions*, Springer-Verlag, New York, 1992.
[19] W. H. Fleming and P. E. Souganidis, *On the existence of value functions of two-player, zero-sum stochastic differential games*, Indiana Univ. Math. J., 38 (1989), pp. 293–314.
[20] M. I. Freidlin and A. D. Wentzell, *Random Perturbations of Dynamical Systems*, Springer-Verlag, New York, 1984.
[21] D. Gilbarg and N. S. Trudinger, *Elliptic Partial Differential Equations of Second Order*, 2nd ed., Springer-Verlag, Berlin, 1983.
[22] A. Isidori, *H^∞ control via measurement feedback for affine nonlinear systems*, in Proc. 31st IEEE Conf. on Decision and Control, IEEE, Piscataway, NJ, 1992.
[23] M. R. James, *Asymptotic analysis of nonlinear stochastic risk-sensitive control and differential games*, Math. Control, Signals Systems, 5 (1992), pp. 401–417.
[24] M. R. James, *A partial differential inequality for dissipative nonlinear systems*, Systems Control Lett., 21 (1993), pp. 315–320.
[25] I. Karatzas and S. E. Shreve, *Brownian Motion and Stochastic Calculus*, Springer-Verlag, New York, 1988.
[26] M. Katzman, Ed., *Laser Satellite Communications*, Prentice-Hall, Englewood Cliffs, NJ, 1987.
[27] W. M. McEneaney, *Uniqueness for viscosity solutions of nonstationary HJB equations under some a priori conditions (with applications)*, SIAM J. Control Optim., 33 (1995), pp. 156–1576.
[28] S.M. Meerkov and T. Runolfsson, *Residence time control*, IEEE Trans. Automat. Control, AC-33 (1988), pp. 323–332.
[29] T. Runolfsson, *Risk-sensitive control of Markov chains and differential games*, in Proc. 32nd IEEE Conf. on Decision and Control, IEEE, Piscataway, NJ, 1993.
[30] A. Shwartz and A. Weiss, *Large Deviations for Performance Analysis: Queues, Communication and Computing*, Chapman and Hall, New York, 1995.
[31] A. J. van der Schaft, *Nonlinear state space H^∞ control theory*, in Perspectives in Control, Birkhauser, Boston, 1993.
[32] A. J. Veretennikov, *On strong solutions and explicit formulas for solutions of stochastic integral equations*, Math. USSR-Sb., 39 (1981), pp. 387–403.
[33] P. Whittle, *Risk-Sensitive Optimal Control*, Wiley, New York, 1990.

SIAM J. CONTROL OPTIM.
Vol. 35, No. 6, pp. 2050–2069, November 1997

009

BOLZA PROBLEMS WITH GENERAL TIME CONSTRAINTS*

P. D. LOEWEN† AND R. T. ROCKAFELLAR‡

Abstract. This work provides necessary conditions for optimality in problems of optimal control expressed as instances of the generalized problem of Bolza, with the added feature that the fundamental planning interval is allowed to vary. A central product of the analysis is a generalization of the conservation-of-Hamiltonian condition for problems on either fixed or variable intervals. The results, which allow for unprecedented generality in the problem data, are derived from known properties of fixed-interval problems under the hypothesis that the time-dependence of the objective integrand has the same modest level of regularity as the state-dependence.

Key words. optimal control, calculus of variations, Bolza problem, free time, minimum-time problem, Erdmann condition, Euler–Lagrange condition, Hamiltonian condition, transversality condition, nonsmooth analysis

AMS subject classifications. 49K05, 49K24, 49K15

PII. S0363012996298801

1. Introduction.

This paper provides necessary conditions for local optimality in a general optimal control problem where the endpoints of the underlying time interval are choice variables. The problem is to choose a nondegenerate interval $[a, b]$ and an absolutely continuous function (or *arc*) $x\colon [a, b] \to \mathbb{R}^n$ in such a way as to

$$\text{(P)} \qquad \text{minimize} \quad l(a, x(a), b, x(b)) + \int_a^b L(t, x(t), \dot{x}(t))\, dt.$$

The usefulness of this simple-looking model is directly correlated to the mildness of the hypotheses under which conclusive results can be obtained. In the current work, both the endpoint cost l and the integrand L are allowed to be nondifferentiable and even to take the value $+\infty$. These features allow for enormous flexibility in modeling applied problems—in particular, a wide range of differential and endpoint constraints can be introduced implicitly by encoding them in l and L.

The introduction of extended-real-valued functionals as practical modeling tools in dynamic optimization can be traced to the early work of Rockafellar [17] in the convex case. Many authors have since addressed the technical challenges raised in this context: we mention in particular Rockafellar [18, 19] and Clarke [3, 4]. The hypotheses in this paper are in some respects weaker than those of Clarke [4] and moreover allow for a variable time interval. We handle this additional complication by using the classical Erdmann transform to reduce (P) to a fixed-time problem in which the time plays the role of an additional state variable. This forces us to assume a degree of regularity in the problem's time-dependence that matches what we require on the state-dependence. Clarke, Loewen, and Vinter [7], using other methods, have

*Received by the editors February 14, 1996; accepted for publication (in revised form) August 30, 1996.

http://www.siam.org/journals/sicon/35-6/29880.html

†Department of Mathematics, University of British Columbia, Vancouver, BC, Canada (loew@math.ubc.ca). The research of this author was supported by Canada's Natural Science and Engineering Research Council.

‡Department of Mathematics, University of Washington, Seattle, WA 98195 (rtr@math.washington.edu). The research of this author was supported by the National Science Foundation grant DMS-9200303.

discussed optimality conditions for related problems in which the time-dependence is merely measurable.

Our most general assertions about minimizers in problem (P) appear in Theorem 1.1 below. The remainder of section 1 is devoted to clarifying the scope and content of this result. In section 2 we review the implications of these developments for a special case of (P) arising frequently in practice, while in section 3 we explain how the hypotheses of Theorem 1.1 can be weakened in the presence of additional structure on the problem data. The technical details of most proofs are assembled in section 4. Finally, in section 5 we sketch the modifications required when unilateral state constraints are imposed on the basic problem.

1.1. Subgradients and normals. We will treat instances of (P) whose data lie well beyond the scope of classical first-order approximations. Thus the terminology and methods of nonsmooth analysis will be needed throughout. In this work, the symbols $\partial f(x)$ and $\partial^\infty f(x)$ stand for the sets of *limiting proximal subgradients* and *singular limiting proximal subgradients* associated with an extended-valued function $f\colon \mathbb{R}^m \to \mathbb{R}\cup\{\infty\}$ at a point x such that $f(x)$ is finite and the epigraph of f is locally closed near $(x, f(x))$. Given a closed set S, we will write Ψ_S for the corresponding *indicator function,* which equals 0 at points in S and $+\infty$ elsewhere. The set of *limiting proximal normals* associated with S at a point $s \in S$ will be denoted by $N_S(s)$; note $N_S(s) = \partial\Psi_S(s) = \partial^\infty\Psi_S(s)$. For simplicity, we refer to these limiting objects simply as "subgradients," "singular subgradients," and "normals"; further discussion of their constructions and relationships to other fundamental objects is now widely available (e.g., [5], [10], [15]). However, we will often refer to the basic relationships

$$\begin{aligned} \partial f(x) &= \{\xi \,:\, (\xi,-1) \in N_{\operatorname{epi} f}(x, f(x))\}, \\ \partial^\infty f(x) &= \{\xi \,:\, (\xi,0) \in N_{\operatorname{epi} f}(x, f(x))\}. \end{aligned} \tag{1.1}$$

We will write $\mathbb{B}$ for the *closed* unit ball centered at the origin in various Euclidean spaces distinguished by the context.

1.2. Hypotheses. An arc $\overline{x}$ is given, together with its associated interval of definition $[\overline{a}, \overline{b}]$; one has $\overline{b} - \overline{a} > 0$. For some $\rho > 0$, the open set

$$\Omega = \big\{(t,x) \,:\, |(t,x) - (r, \overline{x}(r))| < \rho \text{ for some } r \in [\overline{a}, \overline{b}]\big\}, \tag{1.2}$$

with sections $\Omega_t = \{x \,:\, (t,x) \in \Omega\}$, is one in which the data of problem (P) satisfy hypotheses (H1)–(H4) below. These conditions refer to the *Hamiltonian* $H\colon \Omega \times \mathbb{R}^n \to \mathbb{R}\cup\{\infty\}$, defined by

$$H(t,x,p) := \sup\{\langle p\,,\, v\rangle - L(t,x,v) \,:\, v \in \mathbb{R}^n\}. \tag{1.3}$$

In (H4) and throughout the paper, we use the shorthand $\overline{L}(t) := L\big(t, \overline{x}(t), \dot{\overline{x}}(t)\big)$.

(H1) The endpoint cost $l\colon \mathbb{R}\times\mathbb{R}^n\times\mathbb{R}\times\mathbb{R}^n \to \mathbb{R}\cup\{\infty\}$ is lower semicontinuous on the set $\big\{(a,x,b,y) \,:\, |(a,x) - (\overline{a}, \overline{x}(\overline{a}))| < \rho,\ \big|(b,y) - (\overline{b}, \overline{x}(\overline{b}))\big| < \rho\big\}$.

(H2) For each fixed (t,x) in Ω, the function $v \mapsto L(t,x,v)$ is convex on $\mathbb{R}^n$.

(H3) The function L is lower semicontinuous on $\Omega\times\mathbb{R}^n$ and epicontinuous in (t,x): that is, for any point (t,x,v) in $\Omega\times\mathbb{R}^n$ where $L(t,x,v)$ is finite, and any sequence $(t_k, x_k) \to (t,x)$, there exists a sequence $v_k \to v$ along which $L(t_k, x_k, v_k) \to L(t,x,v)$.

(H4) There are positive constants δ and κ such that the following statement is true for almost every t in $[\overline{a}, \overline{b}]$: for every point (r,x,v) in $\Omega\times\mathbb{R}^n$ satisfying the three conditions

(i) $|(r,x)-(t,\overline{x}(t))\,|<\rho$,
(ii) $\big|v-\dot{\overline{x}}(t)\big|<\delta\big[1+\big|\dot{\overline{x}}(t)\big|\big]$,
(iii) $\big|L(r,x,v)-L(t,\overline{x}(t),\dot{\overline{x}}(t))\big|<\delta\big[1+\big|L(t,\overline{x}(t),\dot{\overline{x}}(t))\big|\big]$,
one has the subgradient inequality

$$|(u,w)|\le\kappa[1+|p|+|H(r,x,p)|]\qquad\forall(u,w,p)\in\partial L(r,x,v).$$

To understand (H2)–(H4), consider this multifunction $E\colon\Omega\rightrightarrows\mathbb{R}^n\times\mathbb{R}$:

$$E(t,x):=\operatorname{epi}L(t,x,\,\cdot\,)=\{(v,\gamma)\,:\,\gamma\ge L(t,x,v)\}\,.$$

Note that a scalar arc y obeys $y(t)\ge y(a)+\int_a^t L(r,x(r),\dot{x}(r))\,dr$ if and only if $(\dot{x}(t),\dot{y}(t))\in E(t,x(t))$ a.e. This observation furnishes a link between the objective values in (P) and the endpoint values of trajectories for the differential inclusion based on E: this reformulation, detailed in [13], provides for a geometric perspective on the hypotheses.

Hypothesis (H2), together with the lower semicontinuity part of (H3), ensures that the set $E(t,x)$ is closed and convex for every (t,x)—an important condition in the existence theory for problem (P).

Hypothesis (H3) is equivalent to the requirement that the multifunction E be continuous on Ω, in the sense that for every (t,x) in Ω

$$\liminf_{(t',x')\to(t,x)}E(t',x')\supseteq E(t,x)\supseteq\limsup_{(t',x')\to(t,x)}E(t',x')\,. \tag{1.4}$$

As a consequence of Wijsman's theorem, which asserts that the correspondence between Lagrangian and Hamiltonian under the Legendre–Fenchel transform preserves epicontinuity, we can express (H3) in three equivalent ways (cf. [22, Prop. 2.1]):

(i) The (closed) set $\operatorname{epi}L(t,x,\,\cdot\,)$ depends continuously on (t,x) in Ω.
(ii) The (closed) set $\operatorname{epi}H(t,x,\,\cdot\,)$ depends continuously on (t,x) in Ω.
(iii) The function H is lower semicontinuous on $\Omega\times\mathbb{R}^n$, and epicontinuous in (t,x): that is, for any point (t,x,p) in $\Omega\times\mathbb{R}^n$ at which H is finite and for any sequence $(t_k,x_k)\to(t,x)$, there is a sequence $p_k\to p$ along which $H(t_k,x_k,p_k)\to H(t,x,p)$.

This equivalence reveals an appealing symmetry between the requirements on L and H.

Hypothesis (H4) is a subgradient sufficient condition for the multifunction E to display a type of Lipschitz continuity around the points $\big(t,\overline{x}(t),\dot{\overline{x}}(t),\overline{L}(t)\big)$. The continuity property we require was introduced for general multifunctions by Aubin [2]; in our context, we require that the graph of $\overline{x}$ have a neighborhood in which, for some constants R and K, one has

$$E(s,x)\cap\big(\big(\dot{\overline{x}}(t),\overline{L}(t)\big)+R\mathbb{B}\big)\subseteq E(t,y)+K|(s,x)-(t,y)|\mathbb{B}. \tag{1.5}$$

Mordukhovich [14] has shown that a necessary and sufficient condition for the validity of this "Aubin continuity property" near a given point $\big(t,\overline{x}(t),\dot{\overline{x}}(t),\overline{L}(t)\big)$ is that $|(u,w)|\le K|(p,-\lambda)|$ for all vectors $(u,w,p,-\lambda)$ normal to the set $\operatorname{gph}E=\operatorname{epi}L$ at points near $\big(t,\overline{x}(t),\dot{\overline{x}}(t),\overline{L}(t)\big)$. The subgradient inequality in (H4) arises from just such considerations (compare (1.1)), weakened somewhat thanks to the technical advantages of the Erdmann transform.

Another helpful perspective on (H4) is available if we consider the stronger hypothesis obtained by omitting the Hamiltonian term from the right side of the central

inequality. For any integrand L that satisfies the resulting condition with respect to $\overline{x}$, the same hypothesis holds (with a larger coefficient κ) for any integrand $L + G_1$ involving a function G_1 that is Lipschitz with respect to (t, x). This insensitivity to Lipschitzian perturbations reveals that the real job of (H4) is to regulate the integrand's non-Lipschitz dependence on (t, x), if any, near the nominal trajectory.

It is important to note that both (H3) and (H4) can be checked easily for several classes of integrand L. We describe some of these reductions in the discussion of special cases below, and we treat this matter more generally in section 3.

Local minimizers. Given an arc $\overline{x}$ and an open set Ω relative to which (H1)–(H4) hold, we will always consider problem (P) under the implicit constraint $x(t) \in \Omega$ for all t in $[a, b]$. Thus all of our results pertain to *strong local minimizers* in the basic problem. Here is the main theorem.

THEOREM 1.1. *Let the arc $\overline{x}$ with interval $[\overline{a}, \overline{b}]$ provide the minimum in problem* (P). *Assume* (H1)–(H4). *Then some arc (h, p) taking values in $\mathbb{R} \times \mathbb{R}^n$ satisfies either the normal conditions or the singular conditions below.*
Normal conditions:

(a) $\Big(\dot{h}(t), \dot{p}(t)\Big) \in \operatorname{co}\big\{(u, w) \,:\, (-u, w, p(t)) \in \partial L\big(t, \overline{x}(t), \dot{\overline{x}}(t)\big)\big\}$
$= \operatorname{co}\big\{(u, w) \,:\, \big(u, -w, \dot{\overline{x}}(t)\big) \in \partial H\big(t, \overline{x}(t), p(t)\big)\big\}$ *a.e.* $t \in [\overline{a}, \overline{b}]$.

(b) $h(t) = H(t, \overline{x}(t), p(t))$ *a.e.* $t \in [\overline{a}, \overline{b}]$ (*but see Proposition* 1.2);
$h(t) = \big\langle p(t)\,,\, \dot{\overline{x}}(t)\big\rangle - L\big(t, \overline{x}(t), \dot{\overline{x}}(t)\big)$ *a.e.* $t \in [\overline{a}, \overline{b}]$.

(c) $\big(-h(\overline{a}), p(\overline{a}), h(\overline{b}), -p(\overline{b})\big) \in \partial l\big(\overline{a}, \overline{x}(\overline{a}), \overline{b}, \overline{x}(\overline{b})\big)$.

Singular conditions: *One has $|(h(t), p(t))| > 0$ for all t in $[\overline{a}, \overline{b}]$, and*

(a^∞) $(\dot{h}(t), \dot{p}(t)) \in \operatorname{co}\big\{(u, w) \,:\, (-u, w, p(t)) \in \partial^\infty L(t, \overline{x}(t), \dot{\overline{x}}(t))\big\}$ *a.e.* $t \in [\overline{a}, \overline{b}]$.

(b^∞) $h(t) = \big\langle p(t)\,,\, \dot{\overline{x}}(t)\big\rangle$ *a.e.* $t \in [\overline{a}, \overline{b}]$.

(c^∞) $\big(-h(\overline{a}), p(\overline{a}), h(\overline{b}), -p(\overline{b})\big) \in \partial^\infty l\big(\overline{a}, \overline{x}(\overline{a}), \overline{b}, \overline{x}(\overline{b})\big)$.

In particular, if the only arc (h, p) satisfying conditions (a^∞)–(c^∞) *is identically zero, then the normal conditions must hold.*

The proof of Theorem 1.1 is given in section 4 below. Before discussing this technical development, we pause to emphasize some of the result's useful implications.

1.3. General time constraints. The formulation of (P) covers the full range of situations from the case of a fixed planning interval $[a, b] = [0, 1]$, through the situation where there is a required relationship between the final time and state (like the target condition $(b, x(b)) \in C$ for a given set C), to the case where the interval $[a, b]$ is completely unrestricted. For example, to treat a problem where the final time is fixed at T, it suffices to include the additive term $\Psi_{\{T\}}(b)$ in the endpoint cost l. Such extended-valued and discontinuous behavior is allowed by the hypotheses, but it will of course be reflected in conclusions (c) and (c^∞) of Theorem 1.1. In this instance, the third components of conclusions 1.1(c)/(c^∞) would give only the redundant statement "$h(T) \in \mathbb{R}$" because of the elementary identity $\partial\Psi_{\{T\}}(T) = \partial^\infty\Psi_{\{T\}}(T) = \mathbb{R}$; in variable-endtime problems the corresponding conclusions would be richer, providing nontrivial algebraic conditions linking the final value of h with other ingredients of the extremal system.

1.4. Conservation of the Hamiltonian. Conclusion (b) in the normal case of Theorem 1.1 generalizes the classical equation

$$\dot{h}(t) = H_t(t, \overline{x}(t), p(t))\,, \text{ where } h(t) = H(t, \overline{x}(t), p(t))\,. \tag{1.6}$$

In particular, if the integrand L is free of explicit t-dependence, then the Hamiltonian must be constant along normal extremal trajectories—a conclusion that recalls the law of conservation of energy in classical mechanics, or the second Weierstrass–Erdmann condition in the calculus of variations. This condition is useful in fixed-time problems, where the constant value of h is unknown, and indispensable in free-time problems, where the constant value of h is determined through the transversality inclusion (c) or (c^∞) of Theorem 1.1, as noted above.

1.5. Maximization conditions. Eliminating h between the two equations in Theorem 1.1(b) leads to the first of two equivalent maximization conditions:

$$\langle p(t)\,,\,\dot{\overline{x}}(t)\rangle - L\big(t,\overline{x}(t),\dot{\overline{x}}(t)\big) = \max_{v\in\mathbb{R}^n}\{\langle p(t)\,,\,v\rangle - L(t,\overline{x}(t),v)\}\ \text{ a.e. } t\in\big[\overline{a},\overline{b}\big]\,, \tag{1.7}$$

$$\big\langle p(t)\,,\,\dot{\overline{x}}(t)\big\rangle - H(t,\overline{x}(t),p(t)) = \max_{q\in\mathbb{R}^n}\big\{\big\langle q\,,\,\dot{\overline{x}}(t)\big\rangle - H(t,\overline{x}(t),q)\big\}\ \text{ a.e. } t\in\big[\overline{a},\overline{b}\big]\,. \tag{1.8}$$

(Here (1.7) simply restates $p(t)\in\partial_v L\big(t,\overline{x}(t),\dot{\overline{x}}(t)\big)$, while (1.8) is a transcription of $\dot{\overline{x}}(t)\in\partial_p H(t,\overline{x}(t),p(t))$; these inclusions are equivalent by convex analysis.) Condition (1.7) is the analogue in this theory of the maximization condition in Pontryagin's maximum principle and can be derived directly from the adjoint inclusion (a) of Theorem 1.1. To see this, notice that (a) implicitly asserts that for almost every t, there are points (u,w) such that

$$(-u,w,p(t))\in\partial L\big(t,\overline{x}(t),\dot{\overline{x}}(t)\big)\,. \tag{1.9}$$

Under our hypotheses on L, this inclusion implies the partial-subgradient relation $p(t)\in\partial_v L\big(t,\overline{x}(t),\dot{\overline{x}}(t)\big)$. (See [22, Prop. 2.2]; the Lipschitz case is in [3, Prop. 2.5.3].) The definition of the convex subgradient allows this inclusion to be expressed as shown in (1.7). Thus inclusion (1.9) implies

$$\big\langle p(t)\,,\,\dot{\overline{x}}(t)\big\rangle - L\big(t,\overline{x}(t),\dot{\overline{x}}(t)\big) = H(t,\overline{x}(t),p(t))\,. \tag{1.10}$$

This derivation of (1.10) from (1.9) proceeds equally well for any evaluation point in $\Omega\times\mathbb{R}^n$, and its more general form will be used repeatedly in what follows. To illustrate, we note that the subgradient inequality in (H4) is equivalent to

$$|(u,w)|\le\kappa[1+|p|+|\langle p\,,\,v\rangle - L(r,x,v)|]\quad\forall(u,w,p)\in\partial L(r,x,v). \tag{1.11}$$

1.6. Lipschitzian minimizers. Problem (P) involves minimization over the space of all absolutely continuous functions x. However, it often happens that the solution lies in the subspace of Lipschitzian arcs. (Sufficient conditions guaranteeing this desirable outcome may be found in [8, 9, 6, 1], for example.) In this situation, our results can be simplified: the hypotheses can be weakened by dropping the term $\big|\dot{\overline{x}}(t)\big|$ from the right-hand side of (H4(ii)), while the conclusions can be strengthened by demonstrating the identity in the first line of Theorem 1.1(b) for all t in $\big[\overline{a},\overline{b}\big]$ *without exception.* The latter assertion follows from our next result, in which m denotes Lebesgue measure, and we write $I^+=\bigcup_{R>0}I_R^+$ and $I^-=\bigcup_{R>0}I_R^-$, where

$$I_R^+ = \big\{t\in\big[\overline{a},\overline{b}\big)\ :\ \forall\varepsilon>0,\ m\{s\in[t,t+\varepsilon]\ :\ \big|\dot{\overline{x}}(s)\big|\le R\}>0\big\}\,,$$
$$I_R^- = \big\{t\in\big(\overline{a},\overline{b}\big]\ :\ \forall\varepsilon>0,\ m\{s\in[t-\varepsilon,t]\ :\ \big|\dot{\overline{x}}(s)\big|\le R\}>0\big\}\,.$$

(Note that for Lipschitzian $\overline{x}$, $I^+=\big[\overline{a},\overline{b}\big)$ and $I^-=\big(\overline{a},\overline{b}\big]$.)

PROPOSITION 1.2. *Let (h,p) be a function of bounded variation on $\big[\overline{a},\overline{b}\big]$ for which the two statements in Theorem* 1.1(b) *hold. Then*

(a) $h(t^+) = H\big(t, \overline{x}(t), p(t^+)\big)\ \ \forall t \in I^+$,
(b) $h(t^-) = H\big(t, \overline{x}(t), p(t^-)\big)\ \ \forall t \in I^-$.

In particular, if $\overline{x}$ is Lipschitzian and (h,p) is an arc, then $h(t) = H(t, \overline{x}(t), p(t))$ for all t in $\big[\overline{a}, \overline{b}\big]$ without exception.

Proof. (a) Fix any t in I^+. Since h and p have bounded variation on $\big[\overline{a}, \overline{b}\big]$, their right limits $h(t^+)$ and $p(t^+)$ exist finitely and can be realized along any sequence $s_k \to t^+$. Also, $t \in I^+_R$ for some $R > 0$, so each interval $[t, t + 1/k]$ contains a nonnull set on which $\big|\dot{\overline{x}}(s)\big| \le R$. Thus we may choose a sequence $s_k \to t^+$ along which $\dot{\overline{x}}(s_k)$ exists, the two equations in Theorem 1.1(b) hold, and $\sup_k \big|\dot{\overline{x}}(s_k)\big| \le R$.

Now H is lower semicontinuous on $\Omega \times \mathbb{R}^n$ by (H3), so

$$(*) \qquad h(t^+) = \liminf_{k\to\infty} H(s_k, \overline{x}(s_k), p(s_k)) \ge H\big(t, \overline{x}(t), p(t^+)\big)\,.$$

On the other hand, H is epicontinuous on Ω. Line $(*)$ shows that $(t, \overline{x}(t), p(t^+))$ is a point where H is finite, and we have a sequence $(s_k, \overline{x}(s_k))$ converging to $(t, \overline{x}(t))$. Thus there must be a sequence $q_k \to p(t^+)$ along which $H(s_k, \overline{x}(s_k), q_k) \to H(t, \overline{x}(t), p(t^+))$. The maximum condition (1.8) (a consequence of Theorem 1.1(b)) then supplies the inequality

$$\begin{aligned} h(t^+) &= \lim_{k\to\infty} H(s_k, \overline{x}(s_k), p(s_k)) \\ &\le \liminf_{k\to\infty} \big[H(s_k, \overline{x}(s_k), q_k) - \big\langle q_k - p(s_k)\,,\, \dot{\overline{x}}(s_k)\big\rangle\big] \qquad (**) \\ &= H\big(t, \overline{x}(t), p(t^+)\big) - 0. \end{aligned}$$

Combining $(*)$ and $(**)$ gives $h(t^+) \le H\big(t, \overline{x}(t), p(t^+)\big) \le h(t^+)$, as required.

(b) This proof is similar.

Now if (h,p) is continuous on $\big[\overline{a}, \overline{b}\big]$, then we have $h(t) = H(t, \overline{x}(t), p(t))$ for all t in $I^- \cup I^+$. Furthermore, $I^- \cup I^+ = \big[\overline{a}, \overline{b}\big]$ when $\overline{x}$ is Lipschitzian, so the upgraded form of Theorem 1.1(b) follows. □

2. Problems with explicit velocity constraints. The fully intrinsic formulation of (P) enjoys both a simple statement and a rich heritage of classical antecedents. On the other hand, its practical importance comes from its applicability to problems with velocity and endpoint constraints beyond the scope of its predecessors. Take, for example, the problem of choosing an interval $[a,b]$ and an arc x on $[a,b]$ in order to

$$\begin{aligned} &\text{minimize } \Gamma[a,b;x] := g_1(a, x(a), b, x(b)) + \int_a^b G_1(t, x(t), \dot{x}(t))\ dt \\ &\text{subject to } \dot{x}(t) \in F(t, x(t)) \text{ a.e. } t \in [a,b], \\ &\qquad (a, x(a), b, x(b)) \in S. \end{aligned} \tag{2.1}$$

This problem is an instance of (P) with endpoint cost $l = g_1 + \Psi_S$ and integrand $L = G_1 + \Psi_{\operatorname{gph} F}$. Suitable hypotheses of Lipschitz continuity on g_1 and G_1, together with mild conditions on the multifunction F, will not only establish (H1)–(H4) but also allow us to derive from the conclusions of Theorem 1.1 the usual dichotomy between the normal and abnormal forms of standard necessary conditions. To express this concisely, let us write for any $\lambda \ge 0$

$$\begin{aligned} l_\lambda(a,x,b,y) &:= \lambda g_1(a,x,b,y) + \Psi_S(a,x,b,y), \\ L_\lambda(t,x,v) &:= \lambda G_1(t,x,v) + \Psi_{\operatorname{gph} F}(t,x,v), \\ H_\lambda(t,x,p) &:= \sup_{v\in\mathbb{R}^n} \{\langle p\,,\, v\rangle - L_\lambda(t,x,v)\} \\ &\qquad = \sup\{\langle p\,,\, v\rangle - \lambda G_1(t,x,v)\ :\ v \in F(t,x)\}\,. \end{aligned} \tag{2.2}$$

The hypotheses in question are as follows. Again we state them in terms of a given arc $\overline{x}$ with associated interval $[\overline{a}, \overline{b}]$ assumed to solve problem (2.1), and a fixed open set Ω containing the graph of $\overline{x}$. We also assume that $\overline{x}$ is Lipschitzian.

(h_1) The target set S is closed, and the endpoint cost g_1 is Lipschitzian on the set

$$\left\{(a,x,b,y) \,:\, |(a,x)-(\overline{a},\overline{x}(\overline{a}))|<\rho,\ \left|(b,y)-(\overline{b},\overline{x}(\overline{b}))\right|<\rho\right\}.$$

(h_2) For each fixed (t,x) in Ω, both the function $v \mapsto G_1(t,x,v)$ and the set $F(t,x)$ are convex.

(h_3) The function G_1 is finite valued and continuous on $\Omega \times \mathbb{R}^n$. The multifunction F is continuous on Ω, in the sense that (1.4) holds for $E = F$, at every point (t,x) in Ω.

(h_4) There are positive constants δ and R for which almost all t in $(\overline{a}-\rho, \overline{b}+\rho)$ have this property: for every point (r,x,v) in $\Omega \times \mathbb{R}^n$ satisfying the three conditions

$$|(r,x)-(t,\overline{x}(t))|<\rho, \qquad \left|v-\dot{\overline{x}}(t)\right|<\delta, \qquad v \in F(r,x),$$

both subgradient estimates below are valid:

$$\begin{aligned} |(u_1,w_1,p_1)| &\le R && \forall (u_1,w_1,p_1) \in \partial G_1(r,x,v), \\ |(u_\nu,w_\nu)| &\le R[1+|p_\nu|] && \forall (u_\nu,w_\nu,p_\nu) \in N_{\operatorname{gph} F}(r,x,v). \end{aligned}$$

Notice that the conditions on F imposed by (h_3)–(h_4) amount to nothing more than continuity in general together with a sort of uniform Aubin property near the minimizing arc of interest. Since the arc $\overline{x}$ is Lipschitzian, the first subgradient inequality in (h_4) will be satisfied (for some R) by any locally Lipschitzian integrand G_1, while a sufficient condition for the second is the qualification condition

$$(u,w,0) \in \partial L_0\big(t,\overline{x}(t),\dot{\overline{x}}(t)\big) \implies (u,w)=(0,0) \qquad \text{a.e. } t \in [\overline{a},\overline{b}].$$

(This follows from Theorem 3.2 below and the special structure of L.) In particular, there is no requirement that F be bounded or compact valued—the next result applies even when $F \equiv \mathbb{R}^n$, so (2.1) reduces to a standard variational problem with Lipschitzian data and a general target set.

THEOREM 2.1. *Let the arc $\overline{x}$ with interval $[\overline{a},\overline{b}]$ provide the minimum in problem (2.1). Suppose that $\overline{x}$ is Lipschitzian and (h_1)–(h_4) hold. Then there exist a scalar $\lambda \in \{0,1\}$ and an absolutely continuous pair $(h,p)\colon [\overline{a},\overline{b}] \to \mathbb{R} \times \mathbb{R}^n$, with $\lambda + |(h(t),p(t))| > 0$ for all t in $[\overline{a},\overline{b}]$, such that*

(a) $$\Big(\dot{h}(t),\dot{p}(t)\Big) \in \operatorname{co}\left\{(u,w) \,:\, (-u,w,p(t)) \in \partial L_\lambda\big(t,\overline{x}(t),\dot{\overline{x}}(t)\big)\right\}$$
$$= \operatorname{co}\left\{(u,w) \,:\, (u,-w,\dot{\overline{x}}(t)) \in \partial H_\lambda(t,\overline{x}(t),p(t))\right\} \ \text{a.e. } t \in [\overline{a},\overline{b}].$$

(b) $h(t) = H_\lambda(t,\overline{x}(t),p(t))\ \forall t \in [\overline{a},\overline{b}]$;
$$h(t) = \big\langle p(t)\,,\,\dot{\overline{x}}(t)\big\rangle - \lambda G_1\big(t,\overline{x}(t),\dot{\overline{x}}(t)\big) \ \text{a.e. } t \in [\overline{a},\overline{b}].$$

(c) $\big(-h(\overline{a}),p(\overline{a}),h(\overline{b}),-p(\overline{b})\big) \in \partial l_\lambda\big(\overline{a},\overline{x}(\overline{a}),\overline{b},\overline{x}(\overline{b})\big)$.

Again, we note that condition (b) in the statement of Theorem 2.1 implies two (equivalent) maximization conditions valid for almost all t:

$$\begin{aligned} \dot{\overline{x}}(t) &\in \partial_p H_\lambda(t,\overline{x}(t),p(t)) = \arg\max_{v\in\mathbb{R}^n}\left\{\langle p(t)\,,\,v\rangle - L_\lambda(t,\overline{x}(t),v)\right\}, \\ p(t) &\in \partial_v L_\lambda(t,\overline{x}(t),\dot{\overline{x}}(t)) = \arg\max_{q\in\mathbb{R}^n}\left\{\big\langle q\,,\,\dot{\overline{x}}(t)\big\rangle - H_\lambda(t,\overline{x}(t),q)\right\}. \end{aligned} \tag{2.3}$$

Proof. We reduce to an application of Theorem 1.1, by choosing $l = l_1$ and $L = L_1$. For $i = 1,2,3$, (Hi) follows directly from (h_i); for $i = 4$, this statement can be justified

as follows. Using the constant R provided by (h_4), define $\kappa = R^2 + 2R$. Then consider any point (r, x, v) of the sort described in (H4(i)–(iii)) and any subgradient $(u, w, p) \in \partial L(r, x, v)$. Conditions (H4(i)–(ii)) imply that (h_4(i)–(ii)) hold for the evaluation point (r, x, v); condition (H4(iii)) certainly requires the finiteness of $L(r, x, v)$ so that (h_4(iii)) must follow. With the three prerequisites of (h_4) in place, we observe that the given subgradient has a decomposition as

$$(u, w, p) = (u_1, w_1, p_1) + (u_\nu, w_\nu, p_\nu)$$

for some $(u_1, w_1, p_1) \in \partial G_1(r, x, v)$ and some $(u_\nu, w_\nu, p_\nu) \in N_{\operatorname{gph} F}(r, x, v)$. The two estimates in ($h_4$) then give the second inequality in the estimate

$$\begin{aligned} |(u, w)| &\le |(u_1, w_1)| + |(u_\nu, w_\nu)| \\ &\le R + R[1 + |p_\nu|] \\ &\le R + R[1 + |p_1 + p_\nu| + |-p_1|] \\ &\le R + R[1 + |p|] + R^2 \\ &\le \kappa[1 + |p|]\,. \end{aligned}$$

In the last step we have used the choice $\kappa = R^2 + 2R$ mentioned above. The resulting inequality confirms (H4).

We may now apply Theorem 1.1 to $\overline{x}$ on $[\overline{a}, \overline{b}]$. If the conclusions of Theorem 1.1 hold in normal form, then conditions (a)–(c) follow immediately, with $\lambda = 1$. The nontriviality condition is evident. Suppose, therefore, that we have only the singular conditions of Theorem 1.1, satisfied by some nonvanishing pair (h, p). Again the nontriviality condition is immediate, but now the special form of l and L allows conclusions (a^∞)–(c^∞) to be simplified. Specifically, since the endpoint cost g_1 and integrand G_1 are locally Lipschitz, we have

$$\begin{aligned} \partial^\infty l\big(\overline{a}, \overline{x}(\overline{a}), \overline{b}, \overline{x}(\overline{b})\big) &= N_S\left(\overline{a}, \overline{x}(\overline{a}), \overline{b}, \overline{x}(\overline{b})\right) = \partial l_0\big(\overline{a}, \overline{x}(\overline{a}), \overline{b}, \overline{x}(\overline{b})\big)\,, \\ \partial^\infty L\big(t, \overline{x}(t), \dot{\overline{x}}(t)\big) &= N_{\operatorname{gph} F}\left(t, \overline{x}(t), \dot{\overline{x}}(t)\right) = \partial L_0\big(t, \overline{x}(t), \dot{\overline{x}}(t)\big)\,. \end{aligned}$$

Thus the transversality inclusion (c) with $\lambda = 0$ follows directly from condition (c^∞) of Theorem 1.1, while condition (a^∞) of that result gives

$$(*) \qquad \Big(\dot{h}(t), \dot{p}(t)\Big) \in \operatorname{co}\left\{(u, w) \,:\, (-u, w, p(t)) \in \partial L_0\big(t, \overline{x}(t), \dot{\overline{x}}(t)\big)\right\} \text{ a.e. } t \in [\overline{a}, \overline{b}]\,.$$

Apply to this inclusion the results of Rockafellar [22, Thm. 1.1] as they pertain to the function $L_0 = \Psi_{\operatorname{gph} F}$. Hypothesis ($h_3$) implies that L_0 is lower semicontinuous and has the required epicontinuity property. Furthermore, any point $(u, w, 0) \in \partial^\infty L_0\big(t, \overline{x}(t), \dot{\overline{x}}(t)\big) = N_{\operatorname{gph} F}\left(t, \overline{x}(t), \dot{\overline{x}}(t)\right)$ must satisfy

$$\alpha(u, w, 0) \in N_{\operatorname{gph} F}\left(t, \overline{x}(t), \dot{\overline{x}}(t)\right) \quad \forall \alpha > 0,$$

whereupon (h_4) requires $\alpha|(u, w)| \le R$ for all $\alpha > 0$, i.e., $|(u, w)| = 0$. Thus the calculus qualification of [22] is in force, and we may conclude that

$$\begin{aligned} &\operatorname{co}\left\{(u, w) \,:\, (-u, w, p(t)) \in \partial L_0\big(t, \overline{x}(t), \dot{\overline{x}}(t)\big)\right\} \\ &\qquad = \operatorname{co}\left\{(u, w) \,:\, \big(u, -w, \dot{\overline{x}}(t)\big) \in \partial H_0(t, \overline{x}(t), p(t))\right\} \text{ a.e. } t \in [\overline{a}, \overline{b}]\,. \end{aligned}$$

In conjunction with $(*)$ above, this equation establishes conclusion (a) with $\lambda = 0$.

Turning finally to the maximum condition, we note that inclusion $(*)$ implies $p(t) \in \partial_v L_0\big(t, \overline{x}(t), \dot{\overline{x}}(t)\big) = N_{F(t, \overline{x}(t))}\big(\dot{\overline{x}}(t)\big)$. In particular, $\langle p(t)\,, \dot{\overline{x}}(t)\rangle \ge \langle p(t)\,, v\rangle$ for all v in $F(t, \overline{x}(t))$, which can be restated as $\langle p(t)\,, \dot{\overline{x}}(t)\rangle = H_0(t, \overline{x}(t), p(t))$. We already have $h(t) = \langle p(t)\,, \dot{\overline{x}}(t)\rangle$ from conclusion (b^∞), so conclusion (b) holds for almost all t. The first line can be upgraded to an equation valid for all t by applying Proposition 1.2. ∎

2.1. Minimum-time problems. A further specialization of (2.1) is the problem of steering the state trajectory from the origin at time 0 to a moving target set C in least time, subject to given differential constraints. Here the termination condition is $x(b) \in C(b)$, and the following choices put this problem into the form (2.1):

$$G_1 \equiv 0, \qquad g_1(a,x,b,y) = b, \qquad S = \{(0,0)\} \times \operatorname{gph} C.$$

Here (h_1) holds for any target multifunction C whose graph is closed, (h_2) reduces to a convexity requirement on the velocity sets $F(t,x)$, (h_3) changes only slightly, and the first of the subgradient estimates in (h_4) becomes self-evident. In addition, one has $L_\lambda = \Psi_{\operatorname{gph} F} = L_0$, and hence $H_\lambda = H_0$, for all $\lambda \geq 0$. Thus a Lipschitzian minimizer $\overline{x}$ with interval $\big[0,\overline{b}\big]$ must have an associated scalar $\lambda \in \{0,1\}$ and arc $(h,p)\colon \big[0,\overline{b}\big] \to \mathbb{R} \times \mathbb{R}^n$, not both zero, such that

(a) $\Big(\dot{h}(t), \dot{p}(t)\Big) \in \operatorname{co}\big\{(u,w) \,:\, (u,-w,\dot{\overline{x}}(t)) \in \partial H_0(t,\overline{x}(t),p(t))\big\}$ a.e. $t \in \big[0,\overline{b}\big]$.

(b) $h(t) = H_0(t,\overline{x}(t),p(t))\ \forall t \in \big[\overline{a},\overline{b}\big]$, and $h(t) = \big\langle p(t)\,,\,\dot{\overline{x}}(t)\big\rangle$ a.e. $t \in \big[\overline{a},\overline{b}\big]$.

(c) $\big(-h(0), p(0), h(\overline{b}), -p(\overline{b})\big) \in \lambda(0,0,1,0) + \mathbb{R} \times \mathbb{R}^n \times N_{\operatorname{gph} C}\big(\overline{b}, \overline{x}(\overline{b})\big)$.

The cost multiplier λ appears only in (c), which reduces to

$$(\mathrm{c}')\ \big(h(\overline{b}) - \lambda, -p(\overline{b})\big) \in N_{\operatorname{gph} C}\big(\overline{b}, \overline{x}(\overline{b})\big).$$

In the case where the target set multifunction $t \rightrightarrows C(t)$ is single valued and moves smoothly with time, this conclusion coincides with the classical one, namely,

$$\big(h(\overline{b}) - \lambda, -p(\overline{b})\big) \in \big(1, C'(\overline{b})\big)^\perp, \quad \text{i.e.,} \quad h(\overline{b}) - \big\langle p(\overline{b})\,,\,C'(\overline{b})\big\rangle = \lambda.$$

Another possibility is that the target is stationary: $C(t) = C$ for all $t > 0$. Here $\operatorname{gph} C(\cdot) = \mathbb{R} \times C$, so (c') decouples to yield the well-known relations

$$h(\overline{b}) = \lambda, \qquad -p(\overline{b}) \in N_C\big(\overline{x}(\overline{b})\big).$$

In the further special case where the velocity sets $F(t,x)$ have no explicit t-dependence, conditions (a) and (b) imply that the Hamiltonian is constant along extremal trajectories, with the fixed value $\lambda = 1$ for normal problems and $\lambda = 0$ for singular ones.

3. On the continuity conditions. Both the epicontinuity condition (H3) and the Aubin continuity hypothesis (H4) follow from more elementary assumptions when the integrand L has suitable structure. We discuss some of these reductions in this section, paying particular attention to simplifications available when the function $L(t,x,v)$ is convex in (x,v) for each fixed t.

Our first result is a sufficient condition for the Aubin continuity property (H4). A simplified version appears as Theorem 3.2 below.

PROPOSITION 3.1. *Suppose that $\dot{\overline{x}}$ and $\overline{L}$ are essentially bounded. Then, upon reducing $\rho > 0$ if necessary, hypothesis* (H4) *holds whenever there exists a multifunction $\Gamma\colon \mathbb{R} \times \mathbb{R}^n \times \mathbb{R}^n \rightrightarrows \mathbb{R}$ with these three properties*:

(a) *The graph of Γ is a compact subset of* $\operatorname{epi} L$; *the images of Γ are compact convex sets.*

(b) *If $\Gamma(t,x,v) \neq \emptyset$, then $L(t,x,v) \in \Gamma(t,x,v)$; moreover, $\Gamma\big(t,\overline{x}(t),\dot{\overline{x}}(t)\big) \neq \emptyset$ for almost all t in $\big[\overline{a},\overline{b}\big]$.*

(c) *One has*

$$(3.1) \qquad \forall \gamma \in \Gamma(t,x,v), \qquad (u,w,0,0) \in N_{\operatorname{epi} L}\,(t,x,v,\gamma) \implies (u,w) = (0,0).$$

Conditions (a) and (b) require, among other things, that the graph of Γ be a compact superset of the "curve" $\big\{\big(t,\overline{x}(t),\dot{\overline{x}}(t),L\big(t,\overline{x}(t),\dot{\overline{x}}(t)\big)\big) \,:\, t \in [\overline{a},\overline{b}]\big\}$. This curve

will admit a compact superset if and only if both $\dot{\overline{x}}$ and $\overline{L}$ are essentially bounded, which we have emphasized by making it an explicit hypothesis in the statement above. In the case where both $\dot{\overline{x}}$ and $\overline{L}$ are continuous, the curve in question is itself compact, and conditions (a) and (b) are satisfied by the multifunction

$$\Gamma(t,x,v) = \begin{cases} \{L(t,x,v)\} & \text{if } (t,x,v) = \big(t,\overline{x}(t),\dot{\overline{x}}(t)\big),\ t \in \big[\overline{a},\overline{b}\big], \\ \emptyset & \text{otherwise.} \end{cases} \tag{3.2}$$

At any point where $\Gamma(t,x,v) = \{L(t,x,v)\}$, the key condition (3.1) can be rewritten in terms of singular subgradients using (1.1):

$$(u,w,0) \in \partial^\infty L(t,x,v) \implies (u,w) = (0,0). \tag{3.3}$$

The discrepancy between (3.1) and (3.3) shows up only when (3.1) refers to normals to $\operatorname{epi} L$ at points (r,x,v,γ) where $\gamma > L(r,x,v)$.

Proof of Proposition 3.1. The essential boundedness of $\dot{\overline{x}}$ and $\overline{L}$ allows the right sides in (H4(ii)–(iii)) to be replaced by δ. This is the form of (H4) that we will confirm.

First, fix any point $(t,\overline{x},\overline{v},\overline{\gamma})$ in $\operatorname{gph}\Gamma$. We claim that there exists $M > 0$ so large that the following geometrical relation holds near the given point:

$$(*) \qquad \left.\begin{array}{l} |(r,x,v) - (t,\overline{x},\overline{v})| < 1/M \\ \gamma < \overline{\gamma} + 1/M \end{array}\right\} \\ \implies |(u,w)| \le M[|p| + |q|] \quad \forall (u,w,p,q) \in N_{\operatorname{epi} L}\,(r,x,v,\gamma)\,.$$

Indeed, if this claim were false, then it would have to fail for every positive integer m. Each m would give rise to a point (r_m,x_m,v_m,γ_m) satisfying the antecedent inequalities in $(*)$ but associated with a normal vector (u_m,w_m,p_m,q_m) in $N_{\operatorname{epi} L}\,(r_m,x_m,v_m,\gamma_m)$ for which

$$(\dagger) \qquad \frac{1}{m}|(u_m,w_m)| > [|p_m| + |q_m|]\,, \qquad |(u_m,w_m,p_m,q_m)| = 1.$$

Now since $(r_m,x_m,v_m) \to (t,\overline{x},\overline{v})$ and $\gamma_m \ge L(r_m,x_m,v_m)$ as $m \to \infty$, $(\dagger)$ gives

$$\begin{aligned} \overline{\gamma} \ge \limsup_{m\to\infty} \gamma_m &\ge \limsup_{m\to\infty} L(r_m,x_m,v_m) \\ &\ge \liminf_{m\to\infty} L(r_m,x_m,v_m) \\ &\ge L(t,\overline{x},\overline{v}) \quad \text{(by lower semicontinuity).} \end{aligned}$$

Conditions (a) and (b) imply that all the limit points of the sequence γ_m lie in the compact interval $\Gamma(t,\overline{x},\overline{v})$; by passing to a subsequence if necessary, we may assume that $\gamma_m \to \widehat{\gamma}$, where $L(t,\overline{x},\overline{v}) \le \widehat{\gamma} \le \overline{\gamma}$ and $\widehat{\gamma} \in \Gamma(t,\overline{x},\overline{v})$. Along a further subsequence, the given normals converge to a unit vector (u,w,p,q) with the property that

$$0 \ge |p| + |q|, \qquad (u,w,p,q) \in N_{\operatorname{epi} L}\,(t,\overline{x},\overline{v},\widehat{\gamma})\,.$$

This contradicts (3.1), so the claim involving $(*)$ must hold.

Second, we use the compact-graph condition (a). Fix any point $(t,\overline{x},\overline{v},\overline{\gamma})$ in $\operatorname{gph}\Gamma$. Let $M = M(t,\overline{x},\overline{v},\overline{\gamma})$ be a constant with the properties specified in $(*)$; use it to define an open set in $\mathbb{R}^{1+n+n+1}$:

$$\Omega(t,\overline{x},\overline{v},\overline{\gamma}) = \{(r,x,v,\gamma)\ :\ |(r,x) - (t,\overline{x})| < 1/M,\ |v - \overline{v}| < 1/M,\ \gamma < \overline{\gamma} + 1/M\}\,.$$

Now as $(t,\overline{x},\overline{v},\overline{\gamma})$ runs through $\operatorname{gph}\Gamma$, the open sets $\Omega(t,\overline{x},\overline{v},\overline{\gamma})$ cover $\operatorname{gph}\Gamma$. Thus we can extract a finite list of points $(t_i,\overline{x}_i,\overline{v}_i,\overline{\gamma}_i)$ in $\operatorname{gph}\Gamma$, $i=1,\dots,N$, such that

$$(**)\qquad \operatorname{gph}\Gamma \subseteq \bigcup_{i=1}^{N} \Omega(t_i,\overline{x}_i,\overline{v}_i,\overline{\gamma}_i).$$

Since the left side is compact and the right side is open, there is a positive constant δ such that any point (r,x,v,γ) satisfying the three conditions below for some point $(t,\overline{x},\overline{v},\overline{\gamma})$ in $\operatorname{gph}\Gamma$ will lie in the right side of $(**)$:

$$|(r,x)-(t,\overline{x})|<\delta, \qquad |v-\overline{v}|<\delta, \qquad \gamma<\overline{\gamma}+\delta.$$

In particular, fix any time t in $[\overline{a},\overline{b}]$ at which the inclusion in (b) holds. Then the point $(t,\overline{x},\overline{v},\overline{\gamma})=\big(t,\overline{x}(t),\dot{\overline{x}}(t),L\big(t,\overline{x}(t),\dot{\overline{x}}(t)\big)\big)$ lies in $\operatorname{gph}\Gamma$, so the three hypotheses below are enough to situate the point $(r,x,v,L(r,x,v))$ in the right side of $(**)$:

(i) $|(r,x)-(t,\overline{x}(t))|<\delta$,
(ii) $\big|v-\dot{\overline{x}}(t)\big|<\delta$,
(iii) $L(r,x,v)<L(t,\overline{x}(t),\dot{\overline{x}}(t))+\delta$.

This means that there is some index i for which $(r,x,v,L(r,x,v))$ lies in $\Omega(t_i,\overline{x}_i,\overline{v}_i,\overline{\gamma}_i)$, and hence that any vector (u,w,p,q) in $N_{\operatorname{epi}L}(r,x,v,L(r,x,v))$ obeys

$$|(u,w)|\le M_i[|p|+|q|]\le \widehat{M}[|p|+|q|],$$

where $\widehat{M}=\max\{M_1,\dots,M_N\}$. In particular, if $(u,w,p)\in\partial L(r,x,v)$, then $|(u,w)|\le \widehat{M}[1+|p|]$. This establishes (H4), with constants δ and $\widehat{M}$. ∎

THEOREM 3.2. *Suppose that both $\dot{\overline{x}}$ and $\overline{L}$ are essentially bounded. Then, upon reducing $\rho>0$ if necessary, the following condition implies* (H4): *there exists $\delta>0$ so small that for almost all $t\in[\overline{a},\overline{b}]$, the three inequalities*

$$(3.4)\qquad |(r,x)-(t,\overline{x}(t))|<\delta, \qquad \big|v-\dot{\overline{x}}(t)\big|<\delta, \qquad \big|\gamma-\overline{L}(t)\big|<\delta$$

imply the geometrical condition

$$(3.5)\qquad (u,w,0,0)\in N_{\operatorname{epi}L}(t,x,v,\gamma) \implies (u,w)=(0,0).$$

Proof. It suffices to construct a multifunction Γ satisfying conditions (a) and (b) of Proposition 3.1 such that any quadruple (r,x,v,γ) satisfying the three inequalities above automatically lies in $\operatorname{gph}\Gamma$. We do this using the "essential value" multifunctions [7]

$$V(t):=\big\{v\in\mathbb{R}^n \,:\, \forall\varepsilon>0,\ 0<m\big\{s\in[t-\varepsilon,t+\varepsilon]\cap[\overline{a},\overline{b}] \,:\, \big|v-\dot{\overline{x}}(s)\big|<\varepsilon\big\}\big\},$$
$$I(t):=\big\{\gamma\in\mathbb{R} \,:\, \forall\varepsilon>0,\ 0<m\big\{s\in[t-\varepsilon,t+\varepsilon]\cap[\overline{a},\overline{b}] \,:\, \big|\gamma-\overline{L}(s)\big|<\varepsilon\big\}\big\}.$$

Evidently both $\operatorname{gph}V$ and $\operatorname{gph}I$ are compact, while $\dot{\overline{x}}(t)\in V(t)$ and $\overline{L}(t)\in I(t)$ for almost all t. Fix any $\delta_0\in(0,\delta)$, and let

$$G:=\{(r,x,v,\gamma) \,:\, \text{for some } t \text{ in } [\overline{a},\overline{b}], \text{ one has } v\in V(t)+\delta_0\mathbb{B},$$
$$\gamma\in\operatorname{co}I(t)+\delta_0\mathbb{B}, \text{ and } |(r,x)-(t,\overline{x}(t))|\le\delta_0\}.$$

This set is compact, contains almost all the points $(t,\overline{x}(t),\dot{\overline{x}}(t),\overline{L}(t))$ for t in $[\overline{a},\overline{b}]$, and has convex sections in the last variable. Thus the relation $\operatorname{gph}\Gamma=G$ defines a multifunction Γ satisfying Theorem 3.1(a) and (b). For almost all t, any quadruple (r,x,v,γ) obeying the three inequalities in (3.4) will satisfy $\gamma\in\Gamma(r,x,v)$. Thus condition (3.5) implies (3.1), and the result follows. ∎

3.1. The convex case. Necessary (and sufficient) conditions for the fixed-time case of problem (P) where the Lagrangian is jointly convex in (x, v) for each fixed t have been known for some time: see [17, 18, 19], for example. These early results have a role not only in the direct solution of applied problems but also as test cases for the correctness of further generalizations. Thus we seek to confirm that our results in this paper represent faithful generalizations of the convex theory. The key issue, of course, is the extent to which convex problems can be expected to satisfy our standing hypotheses (H3) and (H4). Here is a precise formulation of our joint convexity assumption.

(H2)$_c$ For each fixed t in $\big(\overline{a} - \rho, \overline{b} + \rho\big)$, the function $(x, v) \mapsto L(t, x, v)$ is convex on $\Omega_t \times \mathbb{R}^n$. Furthermore, one has $\operatorname{dom} L(t, x, \cdot) \neq \emptyset$ for each (t, x) in Ω.

We will show that for autonomous problems, both (H3) and (H4) follow from (H2)$_c$ and more generally, that assuming (H2)$_c$ allows (H3) and (H4) to be derived from their weakened analogues below, in which only the t-dependence is involved:

(H3)$_c$ The function L is lower semicontinuous on $\Omega \times \mathbb{R}^n$ and epicontinuous in t: that is, for any point (t, x, v) in $\Omega \times \mathbb{R}^n$ where $L(t, x, v)$ is finite and for any sequence $t_k \to t$, there exists a sequence $(x_k, v_k) \to (x, v)$ along which $L(t_k, x_k, v_k) \to L(t, x, v)$.

(H4)$_c$ Both $\dot{\overline{x}}$ and $\overline{L}$ are essentially bounded, and there are positive constants δ and κ such that for almost all t in $\big[\overline{a}, \overline{b}\big]$, every point (r, x, v, γ) in $\Omega \times \mathbb{R}^n \times \mathbb{R}$ obeying the three inequalities

$$|(r, x) - (t, \overline{x}(t))| < \rho, \qquad \big|v - \dot{\overline{x}}(t)\big| < \delta, \qquad \big|\gamma - \overline{L}(t)\big| < \delta,$$

satisfies the geometrical condition

$$(u, 0, 0, 0) \in N_{\operatorname{epi} L}\,(r, x, v, \gamma) \implies u = 0.$$

Both (H3)$_c$ and (H4)$_c$ hold trivially if L has no explicit dependence on t, provided that both $\dot{\overline{x}}$ and $\overline{L}$ are essentially bounded. (In the autonomous case, Lipschitz continuity of minimizers is a consequence of other modest hypotheses—see [1].) Notice that (H4)$_c$ is a geometrical sufficient condition for the uniform Aubin continuity of the multifunction $t \rightrightarrows \operatorname{epi} L(t, \cdot, \cdot)$ near the optimal trajectory.

We deal first with the epicontinuity conditions (H3) and (H3)$_c$, starting from a technical lemma.

LEMMA 3.3. *Let $\overline{x}_0, \ldots, \overline{x}_n$ be points of $\mathbb{R}^n$ such that*

$$0 \in \operatorname{int} \operatorname{co} \{\overline{x}_0, \ldots, \overline{x}_n\}. \tag{3.6}$$

For each $j = 0, \ldots, n$, let $\{\overline{x}_j^k\}_k$ be a sequence converging to $\overline{x}_j$. Given any sequence $w_k \to 0$ in $\mathbb{R}^n$, there exists for each index k sufficiently large a collection $\lambda_0^k, \lambda_1^k, \ldots, \lambda_n^k \geq 0$ such that

$$\text{(i) } w_k = \sum_{j=0}^n \lambda_j^k \overline{x}_j^k, \qquad\qquad \text{(ii) } \sum_{j=0}^n \lambda_j^k \to 0 \text{ as } k \to \infty.$$

Proof. Use (3.6) to fix $\sigma > 0$ so small that $\operatorname{co}\{\overline{x}_0, \ldots, \overline{x}_n\}$ contains $2\sigma\mathbb{B}$. Then for all k sufficiently large, the set $S^k := \operatorname{co}\{\overline{x}_0^k, \ldots, \overline{x}_n^k\}$ contains the smaller ball $\sigma\mathbb{B}$. In particular, $\sigma w_k/|w_k|$ lies in S^k for all such k and therefore has a representation in terms of scalars $\mu_j^k \geq 0$, $\sum_{j=0}^n \mu_j^k = 1$:

$$\sigma \frac{w_k}{|w_k|} = \sum_{j=0}^n \mu_j^k \overline{x}_j^k, \text{ i.e., } w_k = \sum_{j=0}^n \left(\frac{\mu_j^k |w_k|}{\sigma}\right) \overline{x}_j^k.$$

Choosing $\lambda_j^k = \sigma^{-1} \mu_j^k |w_k|$ gives (i), while (ii) holds because $\sum_{j=0}^n \lambda_j^k = |w_k|/\sigma$. ∎

Here is an abstract result concerning epicontinuity, phrased in notation that will facilitate its application to our problem.

THEOREM 3.4. *Let $L\colon \mathbb{R}\times\mathbb{R}^n\times\mathbb{R}^n \to \mathbb{R}\cup\{\infty\}$ be lower semicontinuous at every point in $\big\{(\overline{t},\overline{x})\big\}\times\mathbb{R}^n$, and suppose that for some $\eta>0$, the function $(x,v)\mapsto L(t,x,v)$ is convex for each fixed t obeying $\big|t-\overline{t}\big|\le\eta$. Assume further that*

(i) $\operatorname{dom} L(\overline{t},x,\cdot)\neq\emptyset$ *for all x where $|x-\overline{x}|\le\eta$;*

(ii) *L is epicontinuous in t at the point $t=\overline{t}$.*

Then L is epicontinuous in (t,x) at the point $(\overline{t},\overline{x})$.

Proof. Without loss of generality, take $\overline{t}=0$ and $\overline{x}=0$. Fix unit vectors $\widehat{u}_0,\dots,\widehat{u}_n$ such that $0\in\operatorname{int}\operatorname{co}\{\widehat{u}_0,\dots,\widehat{u}_n\}$. Then for each j, let $\overline{x}_j=\eta\widehat{u}_j$ $(j=0,\dots,n)$, note that $\operatorname{dom} L(0,\overline{x}_j,\cdot)\neq\emptyset$ by (i), and pick some $\overline{v}_j$ where $L(0,\overline{x}_j,\overline{v}_j)<\infty$.

Now fix any v in $\operatorname{dom} L(0,0,\cdot)$, and let any sequence $(t_k,x_k)\to(0,0)$ be given. We must construct a sequence $v_k\to v$ along which $L(t_k,x_k,v_k)\to L(0,0,v)$. To do this, we apply the epicontinuity property (ii) $n+2$ times: once to generate a sequence

$$(*)\qquad (x_k^*,v_k^*)\to(0,v) \text{ along which } L(t_k,x_k^*,v_k^*)\to L(0,0,v)$$

and $n+1$ more times to find for each $j=0,\dots,n$ a sequence

$$(**)\qquad (\overline{x}_j^k,\overline{v}_j^k)\to(\overline{x}_j,\overline{v}_j) \text{ along which } L(t_k,\overline{x}_j^k,\overline{v}_j^k)\to L(0,\overline{x}_j,\overline{v}_j).$$

Now $w_k=x_k-x_k^*$ is a sequence with limit 0. By Lemma 3.3, with moving simplex $\big\{\overline{x}_j^k-x_k^*\,:\,j=0,\dots,n\big\}$, there are sequences of scalars $\lambda_j^k\ge0$ $(j=0,\dots,n)$ such that both

$$x_k-x_k^*=\sum_{j=0}^n\lambda_j^k\big(\overline{x}_j^k-x_k^*\big)\,,\ \text{i.e., } x_k=\sum_{j=0}^n\lambda_j^k\overline{x}_j^k+\left(1-\sum_{j=0}^n\lambda_j^k\right)x_k^*$$

and $\sum_{j=0}^n\lambda_j^k\to0$ as $k\to\infty$. We use these scalars to define

$$v_k=\sum_{j=0}^n\lambda_j^k\overline{v}_j^k+\left(1-\sum_{j=0}^n\lambda_j^k\right)v_k^*.$$

For each k, the convexity of the function $L(t_k,\cdot,\cdot)$ now yields

$$\begin{aligned}L(t_k,x_k,v_k)&=L\left(t_k,\sum_{j=0}^n\lambda_j^k(\overline{x}_j^k,\overline{v}_j^k)+\left(1-\sum_{j=0}^n\lambda_j^k\right)(x_k^*,v_k^*)\right)\\&\le\sum_{j=0}^n\lambda_j^kL(t_k,\overline{x}_j^k,\overline{v}_j^k)+\left(1-\sum_{j=0}^n\lambda_j^k\right)L(t_k,x_k^*,v_k^*).\end{aligned}$$

On the right side of this estimate, each of the sequences $L(t_k,\overline{x}_j^k,\overline{v}_j^k)$ in the first term is bounded, by $(**)$. The sequence $L(t_k,x_k^*,v_k^*)$ in the second term converges to $L(0,0,v)$, by $(*)$. The construction of the coefficient sequences λ_j^k therefore implies that the right side above converges to $L(0,0,v)$. This gives the first inequality in the estimate

$$\begin{aligned}L(0,0,v)&\ge\limsup_{k\to\infty}L(t_k,x_k,v_k)\\&\ge\liminf_{k\to\infty}L(t_k,x_k,v_k)\ge L(0,0,v).\end{aligned}$$

The second inequality is obvious, and the third follows from the lower semicontinuity of L. Taken together, they show that $L(t_k, x_k, v_k) \to L(0, 0, v)$, as required. ∎

COROLLARY 3.5. *Together* (H2)$_c$ *and* (H3)$_c$ *imply* (H3).

The autonomous case of Theorem 3.4 is of independent interest.

COROLLARY 3.6. *Let* $g\colon \mathbb{R}^m \times \mathbb{R}^n \to \mathbb{R} \cup \{\infty\}$ *be lower semicontinuous and jointly convex. Let* $(\overline{x}, \overline{v})$ *be a point where* g *is finite. Then the following statements are equivalent:*

(a) *There exists* $\eta > 0$ *such that* $\operatorname{dom} g(x, \cdot) \neq \emptyset$ *for all* x *with* $|x - \overline{x}| < \eta$.

(b) g *is epicontinuous in* x *at the point* $x = \overline{x}$.

Now we turn from epicontinuity to the Aubin property.

THEOREM 3.7. *Together* (H2)$_c$–(H4)$_c$ *imply* (H4).

Proof. Write $\psi(t, x, v, \gamma) := \Psi_{\operatorname{epi} L}(t, x, v, \gamma)$. It follows from (H3)$_c$ that the function ψ is both lower semicontinuous and epicontinuous in t. Meanwhile, (H2)$_c$ implies that for each fixed t, the function $(x, v, \gamma) \mapsto \psi(t, x, v, \gamma)$ is convex. Thus we have [22, Prop. 2.2]

$$\begin{aligned} N_{\operatorname{epi} L}\,(t, x, v, \gamma) &= \partial\psi(t, x, v, \gamma) \\ &\subseteq \mathbb{R} \times \partial_{x,v,\gamma}\psi(t, x, v, \gamma) = \mathbb{R} \times N_{\operatorname{epi} L(t,\cdot,\cdot)}\,(x, v, \gamma)\,. \end{aligned}$$

It follows that any point $(u, w, 0, 0)$ normal to $\operatorname{epi} L$ at (t, x, v, γ) will have a projection $(w, 0, 0)$ normal to $\operatorname{epi} L(t, \cdot, \cdot)$ at (x, v, γ). But the latter relation concerns a closed convex set, for which normality has a simple characterization by inequalities:

$$0 \geq \langle (w, 0, 0)\,,\, (x', v', \gamma') - (x, v, \gamma) \rangle = \langle w\,,\, x' - x \rangle \qquad \forall (x', v', \gamma') \in \operatorname{epi} L(t, \cdot, \cdot).$$

Evidently any x' for which $\operatorname{dom} L(t, x', \cdot) \neq \emptyset$ must satisfy the inequality above; in view of (H2)$_c$, this includes all x' in the neighborhood Ω_t of x. Thus $w = 0$. It follows that the key condition (3.5) of Theorem 3.2 can be rewritten as

$$\forall \gamma \in \Gamma(t, x, v), \qquad (u, 0, 0, 0) \in N_{\operatorname{epi} L}\,(t, x, v, \gamma) \implies u = 0.$$

This is precisely the condition supplied by (H4)$_c$. ∎

3.2. Lipschitzian perturbations of convex problems. The various formulations of (H3) and (H4) described above are all designed to regulate non-Lipschitz behavior in the (t, x)-dependence of the integrand L. Perhaps the easiest way to see this is to note that if (H3) and (H4) hold for a given integrand L, then they also hold for the integrand $L + G_1$, under the sole hypothesis that G_1 is Lipschitzian of constant rank on some neighborhood of $\operatorname{gph}\big(\overline{x}, \dot{\overline{x}}\big)$. It follows that the simplified conditions outlined in Corollary 3.5 and Theorem 3.7 apply not just to convex integrands but equally well to Lipschitzian perturbations of convex integrands. An interesting family of such functions has the form $L(t, x, v) = G_1(t, x, v) + k(t, v - A(t)x)$, where G_1 is locally Lipschitzian on $\mathbb{R} \times \mathbb{R}^n \times \mathbb{R}^n$, and $k(t, \cdot)$ is a lower semicontinuous, proper convex function for each fixed t. If we deal with a Lipschitzian arc $\overline{x}$ along which $\overline{L}$ is essentially bounded, conditions (H3) and (H4) for L are consequences of (H3)$_c$ and (H4)$_c$ for the integrand $K(t, x, v) = k(t, v - A(t)x)$. Here Corollary 3.5 and Theorem 3.7 apply directly, under the additional assumption that the matrix-valued function A is Lipschitzian, to show that (H3) and (H4) hold whenever k is epicontinuous in t and one has the Aubin continuity condition below on a suitable neighborhood of the points $(t, \overline{x}(t), \dot{\overline{x}}(t), \overline{L}(t))$:

$$(u, 0, 0) \in N_{\operatorname{epi} k}\,(t, v - A(t)x, \gamma) \implies u = 0.$$

4. Proof of Theorem 1.1. The proof of Theorem 1.1 takes up this whole section; in particular, hypotheses (H1)–(H4) are in force throughout. The unifying idea is a classical one. We formulate an autonomous fixed-time problem for which a relative of $\overline{x}$ provides the solution, and then deduce the main result from the fixed-time theory of Loewen and Rockafellar [13].

4.1. The Erdmann transform. There is a standard method for transforming problem (P) and its solution $\overline{x}$ on $[\overline{a}, \overline{b}]$ into a fixed-time problem solved by an arc related to $\overline{x}$. (See, for example, Clarke [3, sect. 3.6].) We use Greek characters to describe states and costates in the latter problem, for which the underlying time interval is $[0,1]$ and the state is a vector (θ, ξ) in $\mathbb{R} \times \mathbb{R}^n$. Here ξ is a transformed version of the original state x, while θ is a new state which keeps track of the variable time in problem (P).

Choose any constant m such that $0 < m < \overline{b} - \overline{a}$. The transformed problem is to

$$(\Pi) \qquad \begin{aligned} &\text{minimize} \quad l(\theta(0), \xi(0), \theta(1), \xi(1)) + \int_0^1 \widetilde{L}(\theta(\tau), \xi(\tau), \theta'(\tau), \xi'(\tau))\, d\tau \\ &\text{subject to} \quad (\theta(\tau), \xi(\tau)) \in \widetilde{\Omega}_\tau \ \forall \tau \in [0,1]. \end{aligned}$$

Here the integrand $\widetilde{L}$ and domain $\widetilde{\Omega}$ are given by

$$(4.1) \qquad \begin{aligned} &\widetilde{L}(\theta, \xi, \theta', \xi') = \begin{cases} \theta' L(\theta, \xi, \xi'/\theta') & \text{if } \theta' \geq m, \\ +\infty, & \text{otherwise,} \end{cases} \\ &\widetilde{\Omega} := [0,1] \times \Omega. \end{aligned}$$

Comparing the objective functional in problem (Π) with that in problem (P) provides the key to understanding this reduction, as the following lemma reveals.

LEMMA 4.1. *The arc* $\overline{\theta}(\tau) = \overline{a} + \tau(\overline{b} - \overline{a})$, $\overline{\xi}(\tau) = \overline{x}\left(\overline{\theta}(\tau)\right)$ *solves* (Π).

Proof. Consider any pair $(x, [a,b])$ admissible for (P), with $b - a \geq m$. Define $\theta(\tau) := a + \tau(b-a)$ and $\xi(\tau) = x(\theta(\tau))$. Notice that $(\theta(\tau), \xi(\tau)) \in \widetilde{\Omega}_\tau = \Omega$, so this pair satisfies the localization constraint in (Π). On some subset of $[0,1]$ with full Lebesgue measure, $\xi'(\tau) = \dot{x}(\theta(\tau))\theta'(\tau)$, so the substitution $t = \theta(\tau)$ yields

$$(4.2) \qquad (a, x(a), b, x(b)) = (\theta(0), \xi(0), \theta(1), \xi(1)),$$

$$(4.3) \qquad \int_{t=a}^{b} L(t, x(t), \dot{x}(t))\, dt = \int_{\tau=0}^{1} L(\theta(\tau), \xi(\tau), \xi'(\tau)/\theta'(\tau))\, \theta'(\tau)\, d\tau.$$

It follows that the cost of $(x, [a,b])$ in problem (P) equals the cost of (θ, ξ) in problem (Π). Applying this transformation to the pair $(\overline{x}, [\overline{a}, \overline{b}])$ solving (P), we find

$$(4.4) \qquad \begin{aligned} \inf(\Pi) &\leq l(\overline{\theta}(0), \overline{\xi}(0), \overline{\theta}(1), \overline{\xi}(1)) + \int_0^1 \widetilde{L}(\overline{\theta}(\tau), \overline{\xi}(\tau), \overline{\theta}'(\tau), \overline{\xi}'(\tau))\, d\tau \\ &= l(\overline{a}, \overline{x}(\overline{a}), \overline{b}, \overline{x}(\overline{b})) + \int_{\overline{a}}^{\overline{b}} L(t, \overline{x}(t), \dot{\overline{x}}(t))\, dt = \inf \text{(P)}. \end{aligned}$$

Conversely, consider any arc (θ, ξ) admissible for (Π). One has $\theta'(\tau) \geq m$ for all $\tau \in [0,1]$, so the function θ is strictly increasing, hence invertible. Define $a = \theta(0)$, $b = \theta(1)$, and let $x(t) = \xi(\theta^{-1}(t))$ for $t \in [a,b]$. Evidently $x(t) \in \Omega_t$ for all t in $[a,b]$, while the same substitution as before gives both (4.2) and

$$\int_{\tau=0}^{1} \widetilde{L}(\theta(\tau), \xi(\tau), \theta'(\tau), \xi'(\tau))\, d\tau = \int_{t=a}^{b} L(t, x(t), \dot{x}(t))\, dt.$$

Thus the pair $(x,[a,b])$ is admissible for (P), where it is assigned the same objective value as the pair (θ,ξ) receives in (Π). Consequently $\inf(\mathrm{P}) \le \inf(\Pi)$. Thus equality holds throughout in (4.4) above, and the lemma follows. □

4.2. Verification of hypotheses. The fixed-time problem (Π), with its solution $(\overline{\theta},\overline{\xi})$, has the form to which the necessary conditions of Loewen and Rockafellar [13] may be applied. The first step is to confirm the hypotheses. It is evident that the endpoint cost in problem (Π) is lower semicontinuous and that the (autonomous) integrand $\widetilde{L}$ is lower semicontinuous, hence Borel measurable. Also, for each fixed (θ,ξ), the mapping $(\theta',\xi') \mapsto \widetilde{L}(\theta,\xi,\theta',\xi')$ is convex on the set $(0,\infty)\times\mathbb{R}^n$—the proof is a standard exercise in convex analysis. Only two of the requirements of [13] remain to check.

First, we must verify the epicontinuity of $\widetilde{L}$. To do this, fix τ in $[0,1]$, and choose any point $(\widehat{\theta},\widehat{\xi},\widehat{\theta}',\widehat{\xi}')$ in $\Omega\times\mathbb{R}^{1+n}$ where $\widetilde{L}$ is finite and $|(\widehat{\theta},\widehat{\xi}) - (\overline{\theta}(\tau),\overline{\xi}(\tau))| < \rho$. Consider any sequence $(\theta_k,\xi_k)\to(\widehat{\theta},\widehat{\xi})$. Then the finiteness of $\widetilde{L}$ implies that $\widehat{\theta}' \ge m$, so $L(\widehat{\theta},\widehat{\xi},\widehat{\xi}'/\widehat{\theta}')$ is finite; also $|(\widehat{\theta},\widehat{\xi}) - (t,\overline{x}(t))| < \rho$ for $t=\overline{\theta}(\tau)$. Taking $\widehat{v}=\widehat{\xi}'/\widehat{\theta}'$, we recognize this as a situation to which (H3) applies: that hypothesis provides a sequence $v_k\to\widehat{v}$ along which $L(\theta_k,\xi_k,v_k)\to L(\widehat{\theta},\widehat{\xi},\widehat{v})$. Upon defining $\widehat{\theta}'_k=\widehat{\theta}'$ and $\xi'_k=\widehat{\theta}'v_k$, we deduce that $\widetilde{L}(\theta_k,\xi_k,\theta'_k,\xi'_k)\to\widetilde{L}(\widehat{\theta},\widehat{\xi},\widehat{\theta}',\widehat{\xi}')$, as required.

Second, we must verify the differential inequality in [13, (H5)]. This requires us to produce nonnegative functions $\widetilde{\delta}$ and $\widetilde{\kappa}$ in $L^1[0,1]$, with $\widetilde{\kappa}/\widetilde{\delta}$ in $L^\infty[0,1]$, such that for almost all τ in $[0,1]$, the three conditions

(i) $\big|(\theta,\xi) - (\overline{\theta}(\tau),\overline{\xi}(\tau))\big| < \rho,$

(ii) $\Big|(\theta',\xi') - (\overline{\theta}'(\tau),\overline{\xi}'(\tau))\Big| < \widetilde{\delta}(\tau),$

(iii) $\Big|\widetilde{L}(\theta,\xi,\theta',\xi') - \widetilde{L}(\overline{\theta}(\tau),\overline{\xi}(\tau),\overline{\theta}'(\tau),\overline{\xi}'(\tau))\Big| < \widetilde{\delta}(\tau)$

imply the subgradient inequality

$$|(u,w)| \le \widetilde{\kappa}(\tau)[1+|p|+|q|] \qquad \forall (u,w,q,p)\in\partial\widetilde{L}(\theta,\xi,\theta',\xi'). \tag{4.5}$$

Upon expressing $\widetilde{L}(\theta,\xi,\theta',\xi') = \theta' L(\theta,\xi,\xi'/\theta') + \Psi_{[m,+\infty)}(\theta')$, we can estimate the subgradient set appearing in (4.5) using the calculus rules of Rockafellar [21, Cor. 7.1.2]: for evaluation points where $\theta' > m$, we obtain

$$\partial\widetilde{L}(\theta,\xi,\theta',\xi') \subseteq \Big\{(u\theta', w\theta', -H(\theta,\xi,p), p) \,:\, (u,w,p)\in\partial L(\theta,\xi,\xi'/\theta')\Big\}. \tag{4.6}$$

(A direct application of the cited chain rule produces a third component of the form $L(\theta,\xi,\xi'/\theta') - \langle p\,,\,\xi'/\theta'\rangle$ on the right side. However, this expression equals $-H(\theta,\xi,p)$, as explained in subsection 1.5.) Thus a sufficient condition for (4.5) (at least when $\theta' > m$) is

$$|(u,w)| \le \frac{\widetilde{\kappa}(\tau)}{\theta'}[1+|p|+|H(\theta,\xi,p)|] \quad \forall (u,w,p)\in\partial L(\theta,\xi,\xi'/\theta'). \tag{4.7}$$

Let us prove that under our assumption (H4), conditions (i)–(iii) imply both $\theta' > m$ and (4.7), using the constant functions

$$\widetilde{\kappa} = \Big(\overline{\theta}'(\tau) + m\Big)\,\kappa, \qquad \widetilde{\delta} = \min\Big\{m\delta,\, \overline{\theta}'(\tau) - m,\, m\Big\}. \tag{4.8}$$

Indeed, fix τ in $[0,1]$ and a point $(\theta,\xi,\theta',\xi')$ obeying (i)–(iii). By (ii) and our choice of $\widetilde{\delta}$,

$$m \le \overline{\theta}'(\tau) - \widetilde{\delta} < \theta' < \overline{\theta}'(\tau) + \widetilde{\delta} < \overline{\theta}'(\tau) + m, \tag{4.9}$$

so we have $\theta' > m$ as required. Next, using the triangle inequality with condition (ii), we have

$$\begin{aligned}\left|\frac{\xi'}{\theta'} - \frac{\overline{\xi}'(\tau)}{\overline{\theta}'(\tau)}\right| &\le \frac{1}{\theta'}\left|\xi' - \overline{\xi}'(\tau)\right| + \left|\frac{\overline{\xi}'(\tau)}{\overline{\theta}'(\tau)}\right|\frac{\left|\overline{\theta}'(\tau) - \theta'\right|}{\theta'} \\ &\le \frac{1}{m}\left[1 + \left|\frac{\overline{\xi}'(\tau)}{\overline{\theta}'(\tau)}\right|\right]\widetilde{\delta}.\end{aligned} \tag{4.10}$$

Likewise, consider condition (iii). Expanding the definition of $\widetilde{L}$, and using the shorthand $L = L(\theta, \xi, \xi'/\theta')$ and $\overline{L} = L(\overline{\theta}(\tau), \overline{\xi}(\tau), \overline{\xi}'(\tau)/\overline{\theta}'(\tau))$, we derive

$$\widetilde{\delta} > \left|\theta' L - \overline{\theta}'\overline{L}\right| \ge \theta'\left|L - \overline{L}\right| - \left|\left(\overline{\theta}'(\tau) - \theta'\right)\overline{L}\right| \ge m\left|L - \overline{L}\right| - \widetilde{\delta}\left|\overline{L}\right|.$$

This implies

$$\left|L - \overline{L}\right| \le \frac{\widetilde{\delta}}{m}\left[1 + \left|\overline{L}\right|\right]. \tag{4.11}$$

Now for $\widetilde{\kappa}$ and $\widetilde{\delta}$ as shown in (4.8) above, we have $\widetilde{\delta}/m \le \delta$. Thus, with the change of variable $t = \overline{\theta}(\tau)$, under which

$$\tau = \overline{\theta}^{-1}(t), \quad \overline{\xi}(\tau) = \overline{x}(t), \quad \overline{\xi}'(\tau)/\overline{\theta}'(\tau) = \dot{\overline{x}}(t), \tag{4.12}$$

and the parallel change of notation $r = \theta$, $\xi = x$, $\xi'/\theta' = v$, condition (i) states

(i′) $|(r, x) - (t, \overline{x}(t))| < \rho$.

Meanwhile, condition (ii) implies, through (4.10), that

(ii′) $\left|v - \dot{\overline{x}}(t)\right| \le \delta\left[1 + \left|\dot{\overline{x}}(t)\right|\right]$.

Finally, estimate (4.11)—a consequence of (iii)—becomes

(iii′) $\left|L(r, x, v) - L(t, \overline{x}(t), \dot{\overline{x}}(t))\right| \le \delta\left[1 + \left|L(t, \overline{x}(t), \dot{\overline{x}}(t))\right|\right]$.

These are precisely the conditions under which (H4) implies

$$|(u, w)| \le \kappa[1 + |p| + |H(r, x, p)|] \qquad \forall (u, w, p) \in \partial L(r, x, v).$$

Our choice of $\widetilde{\kappa}$ and inequality (4.9) imply that $\kappa = \widetilde{\kappa}/(\overline{\theta}'(\tau) + m) < \widetilde{\kappa}/\theta'$ for all points of interest, so inequality (4.7) follows. This establishes [13, (H5)].

4.3. Retrieval of conclusions. Having confirmed the hypotheses of [13], we may now use its conclusions. In the state-constraint-free case at hand, these are expressed most clearly in [13, Thm. 2.1]. They assert the existence of an absolutely continuous function $(\eta, \pi)\colon [0, 1] \to \mathbb{R} \times \mathbb{R}^n$ satisfying adjoint equations of either normal or singular type. We will rewrite these conclusions in terms of the absolutely continuous functions $h(t) = -\eta(\overline{\theta}^{-1}(t))$ and $p(t) = \pi(\overline{\theta}^{-1}(t))$.

In the normal conditions, [13] provides the transversality relation

$$(\eta(0), \pi(0), -\eta(1), -\pi(1)) \in \partial l\big(\overline{\theta}(0), \overline{\xi}(0), \overline{\theta}(1), \overline{\xi}(1)\big) \tag{4.13}$$

and the Euler–Lagrange inclusion

$$(\eta'(\tau), \pi'(\tau)) \in \operatorname{co}\left\{(\alpha, \beta) : (\alpha, \beta, \eta(\tau), \pi(\tau)) \in \partial\widetilde{L}(\overline{\theta}(\tau), \overline{\xi}(\tau), \overline{\theta}'(\tau), \overline{\xi}'(\tau))\right\} \tag{4.14}$$

for almost all τ in $[0, 1]$. The subgradient estimate (4.6) reveals that the inclusion involving $\partial\widetilde{L}$ in the latter condition implies that for some (u, w, p) in $\partial L(\overline{\theta}(\tau), \overline{\xi}(\tau),$

$\overline{\xi}'(\tau)/\overline{\theta}'(\tau)$), one has $\alpha = u\overline{\theta}'(\tau)$, $\beta = w\overline{\theta}'(\tau)$, $\eta(\tau) = -H\big(\overline{\theta}(\tau), \overline{\xi}(\tau), p\big)$, and $\pi(\tau) = p$. Consequently the Euler–Lagrange inclusion implies

$$(\eta'(\tau), \pi'(\tau)) \in \mathrm{co}\Big\{\Big(u\overline{\theta}'(\tau), w\overline{\theta}'(\tau)\Big) : (u, w, \pi(\tau)) \in \partial L\Big(\overline{\theta}(\tau), \overline{\xi}(\tau), \overline{\xi}'(\tau)/\overline{\theta}'(\tau)\Big), \\ \eta(\tau) = -H\big(\overline{\theta}(\tau), \overline{\xi}(\tau), \pi(\tau)\big)\Big\}.$$

In particular, the arcs h and p defined above obey, for almost all t in $[\overline{a}, \overline{b}]$,

$$\Big(-\dot{h}(t), \dot{p}(t)\Big) \in \mathrm{co}\{(u, w) : (u, w, p(t)) \in \partial L(t, \overline{x}(t), \dot{\overline{x}}(t)), h(t) = H(t, \overline{x}(t), p(t))\}.$$

The second condition defining the set on the right side places no restrictions on the points (u, w) in this set, but it does serve to make the set empty at every time t where the condition fails. Thus the indicated inclusion can be split apart to give the Euler–Lagrange inclusion stated in Theorem 1.1(a) and the identity

$$h(t) = H(t, \overline{x}(t), p(t)) \text{ a.e. } t \in \big[\overline{a}, \overline{b}\big]\,. \tag{4.15}$$

The Hamiltonian form of the set on the right in Theorem 1.1(a) is provided by Rockafellar [22]. Theorem 1.1(b) then follows from (4.15), as explained in subsection 1.5.

In the singular case, the transversality relation in [13] states

$$(\eta(0), \pi(0), -\eta(1), -\pi(1)) \in \partial^\infty l\big(\overline{\theta}(0), \overline{\xi}(0), \overline{\theta}(1), \overline{\xi}(1)\big) \tag{4.16}$$

and the Euler–Lagrange inclusion is replaced by

$$(\eta'(\tau), \pi'(\tau)) \in \mathrm{co}\Big\{(\alpha, \beta) : (\alpha, \beta, \eta(\tau), \pi(\tau)) \in \partial^\infty \widetilde{L}(\overline{\theta}(\tau), \overline{\xi}(\tau), \overline{\theta}'(\tau), \overline{\xi}'(\tau))\Big\}$$

for almost all τ in $[0, 1]$. The analysis of this statement parallels the developments in the normal case line by line, starting with an estimate of the singular subgradients of $\widetilde{L}$ again furnished by [21, Cor. 7.1.2]:

$$\partial^\infty \widetilde{L}(\theta, \xi, \theta', \xi') \subseteq \Big\{(u\theta', w\theta', -\langle p\,,\, \xi'/\theta'\rangle\,, p) : (u, w, p) \in \partial^\infty L(\theta, \xi, \xi'/\theta')\Big\}. \tag{4.17}$$

As before, the subgradient estimate leads to the conclusion that the arcs h and p defined above obey, for almost all t in $[\overline{a}, \overline{b}]$,

$$\Big(-\dot{h}(t), \dot{p}(t)\Big) \in \mathrm{co}\{(u, w) : (u, w, p(t)) \in \partial^\infty L(t, \overline{x}(t), \dot{\overline{x}}(t)), h(t) = \big\langle p(t)\,,\, \dot{\overline{x}}(t)\big\rangle\}.$$

The latter inclusion implies both the Euler–Lagrange inclusion stated as conclusion (a^∞) of Theorem 1.1 and the identity $h(t) = \big\langle p(t)\,,\, \dot{\overline{x}}(t)\big\rangle$ of conclusion (b^∞).

4.4. Nontriviality. A direct application of [13, Thm. 2.1] yields an apparently weaker nontriviality assertion than the one made in Theorem 1.1 in that the singular conditions of [13] refer to an adjoint arc that is *not the zero function,* whereas those in the present paper assert that it is a function that *never takes the value zero.* Under our hypotheses, these two properties are actually equivalent—although the second, being more explicit, is clearly preferable. We justify this claim in the context of our earlier paper [13] so as to sharpen the results in that work at the same time that we note their consequences for our current investigation. No generality is lost in working on the fixed time interval $[0, 1]$.

Assume [13, (H5)], which provides two positive-valued functions $\kappa, \delta \in L^1[0, 1]$ such that $\kappa/\delta \in L^\infty[0, 1]$ and the three inequalities

$$|x - \overline{x}(t)| < \rho, \quad \big|v - \dot{\overline{x}}(t)\big| < \delta(t), \quad \big|L(t, x, v) - L\big(t, \overline{x}(t), \dot{\overline{x}}(t)\big)\,\big| < \delta(t) \tag{4.18}$$

imply the subgradient inequality

$$|w| \le \kappa(t)[1+|p|] \quad \forall (w,p) \in \partial L(t,x,v). \tag{4.19}$$

(Here the subgradient is taken only with respect to the pair (x,v).) It is easy to deduce that the same three inequalities also imply the *singular* subgradient estimate

$$|w| \le \kappa(t)|p| \quad \forall (w,p) \in \partial^\infty L(t,x,v). \tag{4.20}$$

Now if an arc p satisfies the singular Euler–Lagrange inclusion of [13, Thm. 2.1], i.e.,

$$\dot p(t) \in \operatorname{co}\{w : (w,p(t)) \in \partial^\infty L\big(t,\overline{x}(t),\dot{\overline{x}}(t)\big)\} \text{ a.e. } t \in [0,1],$$

then for almost all t, the representation $\dot p(t) = \sum \lambda_j w_j$ for suitable scalars $\lambda_j \ge 0$, $\sum \lambda_j = 1$, and pairs $(w_j, p(t)) \in \partial^\infty L\big(t,\overline{x}(t),\dot{\overline{x}}(t)\big)$, leads via (4.20) to the inequality

$$|\dot p(t)| \le \sum \lambda_j |w_j| \le \sum \lambda_j [\kappa(t)|p(t)|] = \kappa(t)|p(t)|. \tag{4.21}$$

Under (4.21), the statements, "$p(t) \ne 0$ for *all* $t \in [0,1]$" and "$p(t) \ne 0$ for *some* $t \in [0,1]$" are known to be equivalent, in consequence of Gronwall's lemma.

5. Unilateral state constraints. The methods described above apply equally well when the basic problem (P) is augmented by a unilateral constraint of the form

$$x(t) \in X(t) \quad \forall t \in [a,b]. \tag{5.1}$$

Under the standing hypotheses (H1)–(H4), we can treat any such constraint for which

(H5) the multifunction $X\colon [\overline{a}-\rho, \overline{b}+\rho] \rightrightarrows \mathbb{R}^n$ has closed graph.

To state the corresponding extension of Theorem 1.1, we use the block-structured $(1+n)\times(1+n)$ matrix $A = \left[\begin{smallmatrix} -I & 0 \\ 0 & I \end{smallmatrix}\right]$ and the Clarke normal cone [3] to $\operatorname{gph} X$,

$$\overline{N}_{\operatorname{gph} X}(t,x) = \operatorname{cl}\operatorname{co} N_{\operatorname{gph} X}(t,x) \qquad \forall (t,x) \in \operatorname{gph} X.$$

We also rely on the terminology and conventions explained in [12] (see also [13]).

THEOREM 5.1. *Let the arc $\overline{x}$ with interval $[\overline{a},\overline{b}]$ provide the minimum in problem* (P) *under the additional constraint* (5.1). *If* (H1)–(H5) *hold and*

(CQ) *the normal cone $N_{\operatorname{gph} X}(t,\overline{x}(t))$ is pointed for all t in $[\overline{a},\overline{b}]$,*

then there is a function $(h,p)\colon [\overline{a},\overline{b}] \to \mathbb{R}\times\mathbb{R}^n$ of bounded variation satisfying either the normal conditions or the singular conditions below. In either case, the singular part of the measure $(-dh, dp)$ is $\overline{N}_{\operatorname{gph} X}(t,\overline{x}(t))$-valued, so its support is a subset of

$$\{t : \overline{N}_{\operatorname{gph} X}(t,\overline{x}(t)) \ne \{0\}\} = \{t \in [\overline{a},\overline{b}] : (t,\overline{x}(t)) \in \operatorname{bdy}\operatorname{gph} X(t)\}.$$

Normal conditions: *For almost every t in $[\overline{a},\overline{b}]$,*

(a) $(\dot h(t), \dot p(t)) \in \operatorname{co}\{(u,w) : (-u,w,p(t)) \in \partial L\big(t,\overline{x}(t),\dot{\overline{x}}(t)\big)\} + A\overline{N}_{\operatorname{gph} X}(t,\overline{x}(t))$
$= \operatorname{co}\{(u,w) : \big(u,-w,\dot{\overline{x}}(t)\big) \in \partial H\big(t,\overline{x}(t),p(t)\big)\} + A\overline{N}_{\operatorname{gph} X}(t,\overline{x}(t))$.

(b) $h(t) = H(t,\overline{x}(t),p(t))$,
$h(t) = \langle p(t), \dot{\overline{x}}(t)\rangle - L\big(t,\overline{x}(t),\dot{\overline{x}}(t)\big)$.

(c) $\big(-h(\overline{a}), p(\overline{a}), h(\overline{b}), -p(\overline{b})\big) \in \partial l\big(\overline{a},\overline{x}(\overline{a}),\overline{b},\overline{x}(\overline{b})\big)$.

Singular conditions: *The pair (h,p) is not identically zero, and for a.e. t in $[\overline{a},\overline{b}]$,*

(a^∞) $(\dot h(t), \dot p(t)) \in \operatorname{co}\{(u,w) : (-u,w,p(t)) \in \partial^\infty L(t,\overline{x}(t),\dot{\overline{x}}(t))\} + A\overline{N}_{\operatorname{gph} X}(t,\overline{x}(t))$.

(b^∞) $h(t) = \langle p(t)\,,\, \dot{\overline{x}}(t)\rangle$.

(c^∞) $(-h(a), p(a), h(b), -p(b)) \in \partial^\infty l(\overline{a}, \overline{x}(a), \overline{b}, \overline{x}(b))$.

In particular, if the only function pair (h,p) satisfying conditions (a^∞)–(c^∞) is identically zero, then the normal conditions are satisfied.

Remarks. (1) If the state constraint is inactive along the optimal arc, i.e., $(t, \overline{x}(t))$ is an interior point of $\operatorname{gph} X$ for all t, then the singular part of (dh, dp) must be $\{(0,0)\}$-valued. That is, (h,p) is actually an arc, and we recover the conclusions of Theorem 1.1.

(2) The times when the first equation in conclusion (b) holds can be estimated using Proposition 1.2, as explained in subsection 1.6.

REFERENCES

[1] L. Ambrosio, O. Ascenzi, and G. Buttazzo, *Lipschitz regularity for the minimizers of integral functionals with highly discontinuous integrands*, J. Math. Anal. Appl., 142 (1985), pp. 301–316.

[2] J.-P. Aubin, *Lipschitz behavior of solutions to convex minimization problems*, Math. Oper. Res., 9 (1984), pp. 87–111.

[3] F. H. Clarke, *Optimization and Nonsmooth Analysis*, John Wiley, New York, 1983.

[4] F. H. Clarke, *Hamiltonian analysis of the generalized problem of Bolza*, Trans. Amer. Math. Soc., 301 (1987), pp. 385–400.

[5] F. H. Clarke, *Methods of Dynamic and Nonsmooth Optimization*, CBMS-NSF Regional Conference Series, 57, SIAM, Philadelphia, 1989.

[6] F. H. Clarke and P. D. Loewen, *An intermediate existence theory in the calculus of variations*, Ann. Scuola Norm. Sup. Pisa (Fis. e Mat.), 16 (1989), pp. 487–526.

[7] F. H. Clarke, P. D. Loewen, and R. B. Vinter, *Differential inclusions with free time*, Ann. Inst. H. Poincaré Anal. Non Linéaire, 5 (1988), pp. 573–593.

[8] F. H. Clarke and R. B. Vinter, *Regularity properties of solutions of the basic problem in the calculus of variations*, Trans. Amer. Math. Soc., 289 (1985), pp. 73–98.

[9] F. H. Clarke and R. B. Vinter, *Existence and regularity in the small in the calculus of variations*, J. Differential Equations, 59 (1985), pp. 336–354.

[10] P. D. Loewen, *Optimal Control via Nonsmooth Analysis*, CRM-AMS Lecture Notes in Mathematics 2, AMS, Providence, RI, 1993.

[11] P. D. Loewen and R. T. Rockafellar, *The adjoint arc in nonsmooth optimization*, Trans. Amer. Math. Soc., 325 (1991), pp. 39–72.

[12] P. D. Loewen and R. T. Rockafellar, *Optimal control of unbounded differential inclusions*, SIAM J. Control Optim., 32 (1994), pp. 442–470.

[13] P. D. Loewen and R. T. Rockafellar, *New necessary conditions for the generalized problem of Bolza*, SIAM J. Control Optim., 34 (1996), pp. 1496–1511.

[14] B. S. Mordukhovich, *Complete characterization of openness, metric regularity, and Lipschitzian properties of multifunctions*, Trans. Amer. Math. Soc., 340 (1993), pp. 1–35.

[15] B. S. Mordukhovich, *Generalized differential calculus for nonsmooth and set-valued mappings*, J. Math. Anal. Appl., 183 (1994), pp. 250–288.

[16] B. S. Mordukhovich, *Discrete approximations and refined Euler–Lagrange conditions for nonconvex differential inclusions*, SIAM J. Control Optim., 33 (1995), pp. 882–915.

[17] R. T. Rockafellar, *Conjugate convex functions in optimal control and the calculus of variations*, J. Math. Anal. Appl., 23 (1970), pp. 174–222.

[18] R. T. Rockafellar, *Existence and duality theorems for convex problems of Bolza*, Trans. Amer. Math. Soc., 159 (1971), pp. 1–40.

[19] R. T. Rockafellar, *State constraints in convex problems of Bolza*, SIAM J. Control, 10 (1972), pp. 691–715.

[20] R. T. Rockafellar, *Dual problems of Lagrange for arcs of bounded variation*, in Calculus of Variations and Control Theory, D. L. Russell, ed., Academic Press, New York, 1976, pp. 155–192.

[21] R. T. Rockafellar, *Extensions of subgradient calculus with applications to optimization*, Nonlinear Anal., 9 (1985), pp. 665–698.

[22] R. T. Rockafellar, *Equivalent subgradient versions of Hamiltonian and Euler–Lagrange equations in variational analysis*, SIAM J. Control Optim., 34 (1996), pp. 1300–1314.

[23] R. B. Vinter and H. Zheng, *The extended Euler–Lagrange condition for nonconvex variational problems*, SIAM J. Control Optim., 35 (1997), pp. 56–79.

SIAM J. CONTROL OPTIM.
Vol. 35, No. 6, pp. 2070–2085, November 1997

010

ASYMPTOTIC OPTIMIZATION OF A NONLINEAR HYBRID SYSTEM GOVERNED BY A MARKOV DECISION PROCESS*

EITAN ALTMAN† AND VLADIMIR GAITSGORY‡

Abstract. We consider in this paper a continuous time stochastic hybrid control system with finite time horizon. The objective is to minimize a nonlinear function of the state trajectory. The state evolves according to a nonlinear dynamics. The parameters of the dynamics of the system may change at discrete times $l\epsilon$, $l = 0, 1, ...$, according to a controlled Markov chain which has finite state and action spaces. Under the assumption that ϵ is a small parameter, we justify an averaging procedure allowing us to establish that our problem can be approximated by the solution of some deterministic optimal control problem.

Key words. hybrid stochastic systems, asymptotic optimality, nonlinear dynamics, Markov decision processes, averaging

AMS subject classifications. 49B10, 49B50

PII. S0363012995279985

1. Introduction and statement of the problem.

Consider the following hybrid stochastic control system. The state $Z_t \in \mathbb{R}^n$ evolves according to the following dynamics:

$$\frac{d}{dt} Z_t = f(Z_t, Y_t), \qquad t \in [0,1], \qquad Z_0 = z, \tag{1}$$

where $Y_t \in \mathbb{R}^k$ is the "control" to be specified later and z is the initial state. f is assumed to be linear in the second argument (for each value of the first argument), i.e.,

$$f(z,y) = f^1(z) + f^2(z)y, \tag{2}$$

where f^1 is an n-dimensional vector and f^2 is an $n \times k$ matrix; $f^2(z)y$ is the multiplication between the matrix $f_2(z)$ and the vector y. The functions $f^1(z)$ and $f^2(z)$ are supposed to be bounded and to satisfy the Lipschitz condition

$$\left\| f^i(z) - f^i(z') \right\|_1 \le C_1 \left\| z - z' \right\|_1 \qquad \forall z, z', \tag{3}$$

$$\left\| f^i(z) \right\|_1 \le C_2, \tag{4}$$

where z, z' are from a sufficiently large domain which contains all possible trajectories of (1), C_1 and C_2 are constants, and $\|\cdot\|_1$ stands for the L_1 norm in the finite-dimensional space. That is, $\|q\|_1 = \max_{i=1,...,k} |q_i|$ for the vector $q = \{q_i\}, i = 1, ..., k$, and $\|A\|_1 = \max_{\|q\|_1 = 1} \|Aq\|_1$ for the matrix $A(n \times k)$.

It is assumed in what follows that there exists a bounded domain containing all the trajectories of (1), and, thus, (4), in fact, is implied by (3).

*Received by the editors January 13, 1995; accepted for publication (in revised form) September 11, 1996. The research undertaken in this paper was supported by the Australian Research Council (ARC).

http://www.siam.org/journals/sicon/35-6/27998.html

†INRIA, BP93, 2004 Route des Lucioles, 06902 Sophia Antipolis Cedex, France (altman@martingale.inria.fr).

‡School of Mathematics, University of South Australia, The Levels, Pooraka, South Australia 5095, Australia (mavg@lux.levels.unisa.edu.au).

Y_t is not chosen directly by the controller but is obtained as a result of controlling the following underlying stochastic discrete event system. Let ϵ be the basic time unit. Time is discretized; i.e., transitions occur at times $t = n\epsilon,\ n = 0, 1, 2, ..., \lfloor \epsilon^{-1} \rfloor$, where $\lfloor x \rfloor$ stands for the greatest integer which is smaller than or equal to x. There is a finite state space $\mathbf{X} = \{1, ..., N\}$ and a finite action space $\mathbf{A}$. If a state is v and an action a is chosen, then the next state is w with the probability P_{vaw}. A policy $u = \{u_0, u_1, ...\}$ in the set of policies U is a sequence of probability measures on $\mathbf{A}$; at each time $t = n\epsilon$ the controller chooses u_n based on the history of all previous states and actions, as well as the present state. Thus, u_n is a function that maps histories of the form $h_n = (x_0, a_0, x_1, a_1, ..., x_{n-1}, a_{n-1}, x_n)$ to probability measures on $\mathbf{A}$.

We shall be especially interested in the following classes of policies:

- the Markov policies, denoted by $\mathcal{M}$, i.e., policies for which u_t depends only on the current state and does not depend on previous states and actions.
- the stationary policies, denoted by $\mathcal{S}$, i.e., policies for which u_t depends only on the current state and does not depend on previous states and actions nor on the time.

The stochastic process $\{X_n, A_n\}$ is known as a controlled Markov chain, or Markov decision process (MDP); see Derman [11, pp. 2–4]. We assume throughout the paper that under any stationary policy, the state space forms an aperiodic Markov chain such that all states communicate (regular Markov chain). The results of the paper hold, in fact, under weaker ergodicity assumptions; however, the restricted assumption makes the presentation clearer.

Denote by $\mathbf{H}$ the set of all possible states and actions histories which can be observed until time $\lfloor \epsilon^{-1} \rfloor$:

$$\mathbf{H} = \bigcup \{h\}, \qquad h = \left\{ (x_n, a_n), n = 0, 1, ..., \lfloor \epsilon^{-1} \rfloor \right\}.$$

Let $\mathcal{F}$ be the σ-algebra of all subsets of $\mathbf{H}$. Each policy u and initial state x determines a probability measure on $\mathcal{F}$, on which the stochastic state and action process $H = \left\{ X_n, A_n, n = 0, 1, ..., \lfloor \epsilon^{-1} \rfloor \right\}$ is defined. Denote by P_x^u and E_x^u the probability measure and mathematical expectation that correspond to an initial state $X_0 = x$ and a policy u. Sometimes we shall assume an initial distribution ξ on X_0, instead of a fixed initial state. In that case P_ξ^u, E_ξ^u denote the corresponding probability measure and mathematical expectation.

Let $y : \mathbf{X} \times \mathbf{A} \to \mathbb{R}^k$, $j = 1, ..., k$, be some given vector-valued function. Then Y_t in (1) is given by

$$Y_t = y(X_{\lfloor t/\epsilon \rfloor}, A_{\lfloor t/\epsilon \rfloor}). \tag{5}$$

The system (1) with thus-defined Y_t is called hybrid, first, because Y_t changes its values via some random jumps whereas Z_t is a smooth (differentiable) function of time and, second, because, as follows from the consideration below, Y_t being controlled "statistically" through controlling the transition probabilities plays by itself the role of a "direct" control with respect to Z_t.

Let $g : \mathbb{R}^n \to \mathbb{R}$ be some operating cost related to the process Z_t. We assume that it is Lipschitz continuous; i.e.,

$$\|g(z) - g(z')\|_1 \le C_1 \|z - z'\|_1 .$$

We consider the following control problem with ϵ and x fixed.

$\mathbf{Q}_\epsilon$: find a policy u that achieves $F^\epsilon(z,x) = \inf_{u\in U} E_x^u g(Z_1)$, where Z_1 is obtained through (1).

Our model is characterized by the fact that ϵ is supposed to be a small parameter and our objective is to construct a policy (depending, in general, on ϵ) which is asymptotically optimal for $\mathbf{Q}_\epsilon$. That is, the difference between the cost under this policy and $F^\epsilon(z,x)$ converges to zero as $\epsilon \to 0$.

The type of model which we introduce is natural in the control of inventories or of production, where we deal with material whose quantity may "slowly" change in a continuous (linear) way. Breakdowns, repairs, and other control decisions yield the underlying MDP. Our model may also be used in the control of highly loaded queueing networks for which the fluid approximation holds (see Kleinrock [20, p. 56]). The slow variables Z_t may then represent the number of customers in the different queues, whereas the underlying MDP may correspond to routing, or flow control of, say, some long on/off traffic.

The fact that ϵ is chosen to be small means that the variables Y_t along with the MDP X_t can be considered to be fast with respect to the time scale t in which Z_t evolves. Indeed, Y_t and X_t may have large jumps between $t = m\epsilon$ and $t = (m+1)\epsilon$, whereas the corresponding change in Z_t in that period is of order ϵ. The problem is, thus, close in nature to stochastic singular perturbed control problems intensively studied in the literature (see, for example, [1], [5], [6], [7], [9], [10], [21], [23], [24], [25] and references therein). A common approach to this kind of problem is an application of singular perturbations or averaging techniques to the Hamilton–Jacobi–Bellman (HJB) equation for problems in continuous time (as in [5], [6], [21]) or to the dynamic programming equation for singularly perturbed MDPs [1], [7], [9], [10], [24], [25]. In contrast to this approach, we, as in [23], apply an averaging method directly to the "slow" stochastic equation. Our model differs, however, from the ones in [23] in many respects—mainly in the type of fast motions involved, which implies the differences in both the technique used and the results obtained.

In our previous paper [2], we considered the problem similar to $\mathbf{Q}_\epsilon$ for the case of linear dynamics f and cost g and showed that an asymptotically optimal policy can be constructed via maximization of the Hamiltonian of some linear deterministic system. The technique we used was, however, strongly related to the linearity of the model, and it is not applicable to the case when the dynamics and/or the cost are nonlinear. As opposed to the linear case, the consideration for the nonlinear case is much more involved and based on an ergodicity-type result for MDPs obtained in this paper (see Theorem 4.1 below). Using this result we establish that the trajectories of stochastic hybrid system (1) are approximated by the trajectories of some nonlinear deterministic control system, and the problem $\mathbf{Q}_\epsilon$ is approximated by the corresponding deterministic optimal control problem allowing us, in particular, to construct an asymptotically optimal policy for $\mathbf{Q}_\epsilon$. Notice that this result can be viewed as an extension of the averaging technique for deterministic singularly perturbed control systems (see, e.g., [15]) to the stochastic case under consideration. On the other hand, it can be viewed as an extention of results on uncontrolled motions establishing that the solution of the original stochastic system is approximated by the solution of some deterministic system obtained via averaging over the fast random dynamics [16], [19], [22] to the case when this random dynamics is defined by the controlled Markov chain.

The paper consists of four sections. Section 1 is this introduction; section 2 describes the main results about the approximation of the problem of optimal control

of the hybrid system by a deterministic optimal control problem. In section 3 we discuss ways that the solution of the deterministic optimal control problem can be characterized and how it can be used to obtain an asymptotically optimal policy. Section 4 contains the above-mentioned Theorem 4.1, as well as the proofs of some basic lemmas used in section 2.

2. Description of main results. Let

$$\mathbf{Y}(m,x) \stackrel{\text{def}}{=} \bigcup_{u\in U} \left\{ (m+1)^{-1} \sum_{t=0}^{m} E_x^u Y_t \right\},$$

where the union is taken over all policies. As follows from Theorem 3 in [2], the set $\mathbf{Y}(m,x)$ converges in the Haussdorff metric to a set $\mathbf{Y}$ defined below:

$$\lim_{m\to\infty} \mathbf{Y}(m,x) = \mathbf{Y} \stackrel{\text{def}}{=} \bigcup_{u\in\mathcal{S}} \left\{ \sum_{v,a} \eta(u;v,a) y(v,a) \right\}, \tag{6}$$

where the union is taken over all stationary policies, and $\eta(u) = \{\eta(u;v,a)\}$ is the vector of steady state probabilities of state-action pairs obtained when using a stationary policy u. That is,

$$\eta(u;v,a) = \lim_{n\to\infty} P_x^u(X_n = v, A_n = a). \tag{7}$$

Notice that due to the ergodicity assumption on our model, $\eta(u;v,a)$ does not depend on the initial distribution. Notice also that, since the set

$$W \stackrel{\text{def}}{=} \bigcup_{u\in\mathcal{S}} \{\eta(u)\} \tag{8}$$

is a polyhedron (see, for example, [11, pp. 93–95]), the set $\mathbf{Y}$ is a polyhedron as well.

Define now the averaged deterministic control system as

$$\frac{d}{dt} z_t = f(z_t, y_t), \qquad z_0 = z, \tag{9}$$

where y_t is a measurable function of t taking values in $\mathbf{Y}$. The set of such functions

$$y : [0,1] \to \mathbf{Y}$$

will be called the set of admissible controls.

Our claim is that the set of all random trajectories of (1) is approximated by the set of solutions of (9) obtained with all admissible controls. More specifically, we establish that there exists a function $\gamma(\epsilon)$ satisfying

$$\lim_{\epsilon\to 0} \gamma(\epsilon) = 0$$

such that the following holds.

LEMMA 2.1. *Corresponding to any admissible control $y = \{y_t, t \in [0,1]\}$, there exists a Markov policy $u_\epsilon(y)$ such that the random trajectory Z_t of (1), obtained with this policy $u_\epsilon(y)$, and the deterministic solution z_t^y of (9), obtained with y, satisfy the inequality*

$$\max_{t\in[0,1]} E_x^{u_\epsilon(y)} \left\| Z_t - z_t^y \right\|_1 \le \gamma(\epsilon). \tag{10}$$

LEMMA 2.2. *There exists a function* $\tilde{y}^\epsilon_t(h)$,

$$\tilde{y}^\epsilon : [0,1] \times \mathbf{H} \to \mathbf{Y},$$

such that (a) *for each* $h \in \mathbf{H}$, $\tilde{y}^\epsilon_t(h)$ *is a piecewise constant function of* t *and* (b) *for any policy* u,

$$\max_{t\in[0,1]} E^u_x \left\| Z_t - \tilde{z}^\epsilon_t(H) \right\|_1 \le \gamma(\epsilon), \tag{11}$$

where Z_t *is the solution of* (1), $\tilde{z}^\epsilon_t(H)$ *is the solution of* (9) *obtained with* $y_t = \tilde{y}^\epsilon_t(H)$, *and* H *is the random realization of the state-action trajectories.*

Notice that the quantity under the expectation sign in (11) is a random variable for any policy u since $\mathbf{H}$ is a finite set and $\mathcal{F}$ is the σ-algebra of all subsets of $\mathbf{H}$.

Notice also that a construction of a policy $u_\epsilon(y)$ which allows an estimate (10) in Lemma 2.1 is described below in section 3. This is just a stationary policy when the deterministic control y is a constant function of time, and it consists of a finite number of stationary policies (and thus is not stationary itself) when y is piecewise constant.

Define the "deterministic" optimal control problem $\mathbf{Q_0}$ as follows.

$\mathbf{Q_0}$: Find an admissible control y which minimizes the cost function

$$F^0(z) \stackrel{\text{def}}{=} \inf_y g(z_1)$$

over the trajectories z of system (9). The following theorem about approximation of $\mathbf{Q}_\epsilon$ by $\mathbf{Q_0}$ is then easily established on the basis of Lemmas 2.1 and 2.2.

THEOREM 2.1. *The values* $F^\epsilon(z,x)$ *of the original problem* $\mathbf{Q}_\epsilon$ *converge to the value* $F^0(z)$ *of the problem* $\mathbf{Q_0}$, *as* $\epsilon \to 0$. *More precisely,*

$$\left| F^\epsilon(z,x) - F^0_x(z) \right| \le C_1 \gamma(\epsilon).$$

If y^* *is an optimal control for* $\mathbf{Q_0}$, *then the Markov policy* $u_\epsilon(y^*)$ *allowing estimate* (10) *with* $y = y^*$ *satisfies the inequality*

$$\left| E^{u_\epsilon(y^*)}_x g(Z_1) - F^\epsilon(z,x) \right| \le C_1 \gamma(\epsilon).$$

That is, $u_\epsilon(y^*)$ *is asymptotically optimal for* $\mathbf{Q}_\epsilon$.

Remark 2.1. In the linear case studied in [2], γ can be chosen such that

$$\lim_{\epsilon\to 0} \epsilon^{-(1/2)} \gamma(\epsilon) = 0.$$

Hence, for the linear case, simple bounds on the rate of convergence are available for Lemmas 2.1 and 2.2 as well as for Theorem 2.1.

Proof of Theorem 2.1. Let u be an arbitrary policy and $\tilde{y}^\epsilon(h) \in \mathbf{Y}$ be the function defined in Lemma 2.2. Then

$$|E^u_x g(Z_1) - E^u_x g(\tilde{z}^\epsilon_1(H))| \le C_1 E^u_x \left\| Z_1 - \tilde{z}^\epsilon_1(H) \right\|_1 \le C_1 \gamma(\epsilon), \tag{12}$$

where C_1 is defined in (3). Being piecewise constant, the function $\tilde{y}^\epsilon$ is measurable in t. Hence,

$$g(\tilde{z}^\epsilon_1(h)) \ge F^0(z) \qquad \forall h \in \mathbf{H},$$

which implies

$$E_x^u g(\tilde{z}_1^\epsilon(H)) \geq F^0(z)$$

for any policy u. From the last inequality and (12), it follows that

$$E_x^u g(Z_1) \geq F^0(z) - C_1\gamma(\epsilon),$$

so that

$$F^\epsilon(z,x) = \inf_u E_x^u g(Z_1) \geq F^0(z) - C_1\gamma(\epsilon). \tag{13}$$

Now let y^* be an optimal control in $\mathbf{Q_0}$. By (10),

$$\left|E_x^{u_\epsilon(y^*)} g(Z_1) - F^0(z)\right| = \left|E_x^{u_\epsilon(y^*)} g(Z_1) - g(z_1^{y^*})\right| \leq C_1 E_x^{u_\epsilon(y)} \left\|Z_1 - z_1^{y^*}\right\|_1 \leq C_1\gamma(\epsilon).$$

Hence

$$E_x^{u_\epsilon(y^*)} g(Z_1) \leq F^0(z) + C_1\gamma(\epsilon). \tag{14}$$

Since $E_x^{u_\epsilon(y^*)} g(Z_1) \geq F^\epsilon(z,x)$, the inequalities (13) and (14) conclude the proof of the theorem. □

3. Construction of an asymptotically optimal policy. Let y be an arbitrary admissible control for $\mathbf{Q_0}$. We show below how to construct the policy $u_\epsilon(y)$ (appearing in Lemmas 2.1 and 2.2 and in Theorem 2.1). Choose a function $\Delta = \Delta(\epsilon)$ in such a way that

$$\lim_{\epsilon\to 0} \Delta(\epsilon) = 0, \qquad \lim_{\epsilon\to 0} \frac{\Delta(\epsilon)}{\epsilon} = \infty, \tag{15}$$

and set $\tau_l = \tau(l,\epsilon) := l\Delta(\epsilon),\ l = 0,1,2,...,\ell(\epsilon)$, where $\ell(\epsilon) := \lfloor \Delta(\epsilon)^{-1} \rfloor$. Let

$$r_l^\epsilon(y) \stackrel{\text{def}}{=} (\Delta(\epsilon))^{-1} \int_{\tau_l}^{\tau_{l+1}} y_t dt, \qquad l = 0,1,...,\ell(\epsilon) - 1. \tag{16}$$

Since $\mathbf{Y}$ is a convex set, $r_l^\epsilon(y) \in \mathbf{Y}$. Hence there exists a stationary policy $s_l^\epsilon(y)$ such that

$$r_l(\epsilon) = \sum_{v,a} \eta(s_l^\epsilon(y); v, a) y(v,a). \tag{17}$$

Now construct $u_\epsilon(y)$ as the Markov policy obtained by applying $s_l^\epsilon(y)$ during $n = \lfloor \tau_l/\epsilon \rfloor, \lfloor \tau_l/\epsilon \rfloor + 1, ..., \lfloor \tau_{l+1}/\epsilon \rfloor - 1$, where $l = 0,1,...,\ell(\epsilon) - 1$, and by applying an arbitrary stationary policy during $\lfloor \tau_{\ell(\epsilon)}/\epsilon \rfloor, \lfloor \tau_{\ell(\epsilon)}/\epsilon \rfloor + 1, ..., \lfloor \epsilon^{-1} \rfloor$. In the proof of Lemma 2.1 it is established that the policy $u_\epsilon(y)$ thus constructed satisfies inequality (10).

As follows from Theorem 2.1, the described procedure for obtaining the policy $u_\epsilon(y^*)$, on the basis of a control y_t^* which is optimal for the deterministic problem $\mathbf{Q_0}$, yields an asymptotically optimal policy for problems $\mathbf{Q}_\epsilon$. The optimal control y_t^* can by itself be characterized by necessary and sufficient optimality conditions. To formulate these, let us consider a parametrized set $\mathcal{L} = \{L(z,\lambda)\}$ of MDPs, $(z,\lambda) \in \mathbb{R}^n \times \mathbb{R}^n$, all of which have $\mathbf{X}$ and $\mathbf{A}$ as state and action spaces, and $\mathcal{P} = \{P_{vaw}, v, w \in$

$\mathbf{X}, a \in \mathbf{A}\}$ as transition probabilities. They differ by the immediate cost, which is given by

$$r(z,\lambda;v,a) = \lambda^T f(z, y(v,a)) = \lambda^T f^1(z) + \lambda^T f^2(z) y(v,a).$$

Consider the problem of minimization of the infinite horizon expected average cost related to an initial distribution ξ over $\mathbf{X}$:

(18)

$$J_\xi(z,\lambda) \stackrel{\text{def}}{=} \inf_u J_\xi(z,\lambda;u), \qquad J_\xi(z,\lambda;u) \stackrel{\text{def}}{=} \varliminf_{m\to\infty} \frac{1}{m+1} E_\xi^u \sum_{j=0}^{m} r(z,\lambda;X_j,A_j).$$

It is well known (see Derman [11, section 6]) that

a) The optimal value of the above problem does not depend on the initial distribution ξ, and it is equal to the optimal value of the following linear programming problem:

$$\begin{aligned} J_\xi(z,\lambda) = J(z,\lambda) &\stackrel{\text{def}}{=} \min_\eta \left\{ \sum_{v,a} r(z,\lambda;v,a)\eta(v,a) | \eta = \{\eta(v,a)\} \in W \right\} \\ &= \lambda^T f^1(z) + \min_\eta \left\{ \lambda^T f^2(z) \sum_{v,a} y(v,a)\eta(v,a) | \eta = \{\eta(v,a)\} \in W \right\}. \end{aligned} \tag{19}$$

b) There is a one-to-one correspondence between optimal stationary policies of $L(z,\lambda)$ and the optimal solutions of (19).

The following statement describes necessary optimality conditions for $\mathbf{Q_0}$.

THEOREM 3.1. *Let y_t^* be an optimal control in $\mathbf{Q_0}$ and let z_t^* be the solution of* (9) *obtained with y^*. That is,*

$$\frac{d}{dt} z_t^* = f(z_t^*, y_t^*), \qquad z_0 = z. \tag{20}$$

Then, for almost all $t \in [0,1]$,

$$y_t^* = \sum_{v,a} \eta(z_t^*, \lambda_t; v, a) y(v,a),$$

where $\eta(z,\lambda) = \{\eta(z,\lambda;v,a\}_{v,a}$ stands for a solution of (19) *and λ_t is the solution of the conjugate system*

$$\frac{d}{dt}\lambda_t = -f_z(z_t^*, y_t^*)\lambda_t, \qquad \lambda_1 = g_z(1); \tag{21}$$

f_z and g_z are $n\times n$ and $n\times 1$ matrices of the partial derivatives of f and g, respectively, over the components of z.

Proof. The proof follows from a direct application of the Pontryagin maximum principle [8, 13] to problem $\mathbf{Q_0}$. □

Notice that if the solution of (19) with $z = z_t^*$ and $\lambda = \lambda_t$ is unique for all $t \in [0,1]$ except for a finite number of switching points and, thus, for all these $t \in [0,1]$, the corresponding stationary policy $u(z_t^*, \lambda_t)$ achieving inf in (18) with $z = z_t^*$ and $\lambda = \lambda_t$ is unique, then an asymptotically optimal policy for $\mathbf{Q}_\epsilon$ can be defined by simply applying $u(z_{\tau_l}^*, \lambda_{\tau_l})$ during $\lfloor \tau_l/\epsilon \rfloor, \lfloor \tau_l/\epsilon \rfloor + 1, \lfloor \tau_{l+1}/\epsilon \rfloor - 1$, where $l = 0, 1, ..., \ell(\epsilon) - 1$.

Another way to characterize the optimal control in the problem $\mathbf{Q_0}$ is related to the HJB equation written for this problem in the form

$$B_t^0(z,t) + \min_{y\in\mathbf{Y}}\{(B_z^0(z,t))^T f(z,y)\} = 0, \qquad B^0(z,1) = g(z), \tag{22}$$

where $B_t^0(z,t)$, $B_z^0(z,t)$ stand for the partial derivatives of $B^0(z,t)$ over t and components of z, respectively. By (2), (6), (8), for any z and λ,

$$\min_{y\in\mathbf{Y}} \lambda^T f(z,y) = \lambda^T f^1(z) + \min_y\{\lambda^T f^2(z)y | y\in\mathbf{Y}\} = \lambda^T f^1(z)$$

$$+ \min_\eta \left\{ \sum_{v,a} \lambda^T f^2(z) y(v,a)\eta(v,a) | \eta = \{\eta(v,a)\} \in W \right\} = J(z,\lambda),$$

where $J(z,\lambda)$ is the optimal value of (19). Hence, HJB equation (22) can be rewritten in the form

$$B_t^0(z,t) + J(z, B_z^0(z,t)) = 0, \qquad B^0(z,1) = g(z). \tag{23}$$

This equation allows us to construct both necessary and sufficient conditions of optimality for $\mathbf{Q_0}$ and, in particular, to verify whether a given admissible control y_t and the corresponding solution z_t of (9) are optimal in $\mathbf{Q_0}$ (see details in [8]). On the other hand, the viscosity solution of (23) (see, e.g., [14]) defines the optimal value of the problem $\mathbf{Q_0}$ on the interval $[s,1]$ subject to the initial condition $z_s = z$, which provides an approximation for the optimal value $B^\epsilon(z,x,s)$ of the problem $\mathbf{Q}_\epsilon$ on the same interval $[s,1]$ subject to the same initial condition $z_s = z$ and with the initial state of the MDP being x. More precisely, since, by definition, $B^\epsilon(z,x,0) = F^\epsilon(z,x)$ and $B^0(z,0) = F^0(z)$, from Theorem 2.1 it follows that

$$\lim_{\epsilon\to 0} B^\epsilon(z,x,0) = B^0(z,0).$$

As in this theorem, one can also establish that

$$\lim_{\epsilon\to 0} B^\epsilon(z,x,s) = B^0(z,s),$$

with the convergence being uniform with respect to $s\in[0,1]$, $x\in\mathbf{X}$, and $z\in\mathcal{Z}$, where $\mathcal{Z}$ is a compact subset of $\mathbb{R}^n$.

Notice that the described approach has a decomposition structure. It consists of two phases. First is the optimization of the fast motions which is achieved via the solution of (18) with fixed "slow variables" z and λ. Second is the "slow optimization" achieved via the solution of HJB (23). Notice also that in a general case the solution of equation (23) can be quite complicated. If, however,

$$f(z,y) = Az + By, \qquad g(z) = c^T z, \tag{24}$$

where $A(n\times n)$, $B(n\times k)$, and $c(n\times 1)$ are matrices (that is, if as in [2], $\mathbf{Q_0}$ is a linear optimal control problem), then the solution of (23) is obvious:

$$B^0(z,s) = \lambda_s^T z + \int_s^1 J(\lambda(t))dt,$$

where $J(\lambda) \stackrel{\text{def}}{=} J(z,\lambda) - \lambda^T Az$ and λ_t is the solution of (21) under assumption (24).

4. Proof of Lemmas 2.1 and 2.2.

LEMMA 4.1. *Let $y_t^i(h)$, $i=1,2$, be functions of time t and state-action histories h. Let $z_t^i(h)$ be the solution of (9) obtained with $y_t^i(h)$ (h is fixed), $i=1,2$. Then there exists a constant L such that for any policy u and any initial state x,*

$$\max_{t\in[0,1]} E_x^u \left\|z_t^1(H)-z_t^2(H)\right\|_1 \le L\left(\Delta(\epsilon)+(\Delta(\epsilon))^{-1}\max_{l=0,\dots,\ell(\epsilon)-1} E_x^u\left\|\int_{\tau_l}^{\tau_{l+1}}[y_t^1(H)-y_t^2(H)]dt\right\|_1\right), \tag{25}$$

where H is the random realization of the state-action trajectories.

Proof. For the sake of brevity, we omit H from the notation below and write Δ and ℓ instead of $\Delta(\epsilon)$ and $\ell(\epsilon)$. By definition,

$$z_{\tau_{l+1}}^i = z_{\tau_l}^i + \int_{\tau_l}^{\tau_{l+1}} f(z_t^i, y_t^i)dt.$$

Hence, denoting

$$\delta_l := E_x^u \left\|z_{\tau_l}^1 - z_{\tau_l}^2\right\|_1$$

and taking into account (2), one can write

$$\begin{aligned}
\delta_{l+1} &\le \delta_l + \int_{\tau_l}^{\tau_{l+1}} E_x^u \left\|f(z_t^1,y_t^1)-f(z_{\tau_l}^1,y_t^1)\right\|_1 dt \\
&\quad + E_x^u\left\|\int_{\tau_l}^{\tau_{l+1}}\left[f(z_{\tau_l}^1,y_t^1)-f(z_{\tau_l}^1,y_t^2)\right]dt\right\|_1 \\
&\quad + \int_{\tau_l}^{\tau_{l+1}} E_x^u \left\|f(z_{\tau_l}^1,y_t^2)-f(z_{\tau_l}^2,y_t^2)\right\|_1 dt + \int_{\tau_l}^{\tau_{l+1}} E_x^u \left\|f(z_{\tau_l}^2,y_t^2)-f(z_t^2,y_t^2)\right\|_1 dt \\
&\le \delta_l + L_1\Delta E_x^u\left\|\frac{1}{\Delta}\int_{\tau_l}^{\tau_{l+1}}(y_t^1-y_t^2)dt\right\|_1 + L_3\Delta\delta_l + L_1\Delta^2,
\end{aligned}$$

where L_i are constants defined by C_1 and C_2 in (3) and (4) (and thus do not depend on H). Applying now Proposition 5.1 of Gaitsgory [15], one obtains that for any $K=0,1,...,\ell$,

$$\delta_K \le \tilde{L}\left(\Delta + \max_{l=0,\dots,\ell-1} E_x^u\left\|\frac{1}{\Delta}\int_{\tau_l}^{\tau_{l+1}}(y_t^1-y_t^2)dt\right\|_1\right), \tag{26}$$

where $\tilde{L}$ is a constant. Since

$$\left\|z_t^i - z_{\tau_l}^i\right\|_1 \le L_4\Delta \qquad \forall t\in[\tau_l,\tau_{l+1}]$$

for some constant L_4, (26) implies (25) with $L=\tilde{L}+2L_4$. □

We need another general result on MDPs that establishes the uniform convergence of the state-action frequencies to their limits. More precisely, consider arbitrary integers m and K, and define the random variables

$$\psi_m^K(v,a) = \psi_m^K(H;v,a) := \frac{1}{K}\sum_{n=m+1}^{m+K} 1\{X_n=v, A_n=a\}.$$

Let $\psi_m^K := \{\psi_m^K(v,a)\}_{v,a}$ denote the vector of state-action frequencies. Denote

$$d_K^1 = \operatorname{dist}\{\psi_0^K, W\} = \inf_{\eta \in W} \left\|\psi_0^K - \eta\right\|_1.$$

It follows from Derman [11, Chapter 8, p. 98] (see also [3, section 3]) that for any policy u and initial distribution ξ,

$$\lim_{K\to\infty} d_K^1 = 0, \qquad P_\xi^u \text{ a.s.} \tag{27}$$

This implies, by the bounded convergence theorem, that

$$\lim_{K\to\infty} E_\xi^u d_K^1 = 0. \tag{28}$$

For any stationary policy $u \in \mathcal{S}$ the limit

$$\psi_0 := \lim_{K\to\infty} \psi_0^K$$

exists (P_ξ^u a.s.), and it does not depend on the initial distribution ξ (in fact, $\psi_0(v,a) = \eta(u;v,a)$). Define

$$d_K^2 = \left\|\psi_0^K - \psi_0\right\|_1.$$

THEOREM 4.1. *The following holds:*

$$\lim_{K\to\infty} \sup_\xi \sup_{u\in U} E_\xi^u d_K^1 = 0, \tag{29}$$

$$\lim_{K\to\infty} \sup_\xi \sup_{u\in \mathcal{S}} E_\xi^u d_K^2 = 0. \tag{30}$$

Proof. In order to prove the theorem, we define some operations on policies. A k-shift $v = \Theta^k u$ of a policy u is defined to be a sequence $v = \{v_k, v_{k+1}, ...\}$, where

$$v_{n+k}(x_0, a_0, x_1, a_1, ..., x_{n+k-1}, a_{n+k-1}, x_{n+k})$$

$$= u_n(x_k, a_k, x_{k+1}, a_{k+1}, ..., x_{n+k-1}, a_{n+k-1}, x_{n+k}).$$

A policy w is defined to be a concatenation of u and v from time k if

$$w_n = \begin{cases} u_n, & n < k, \\ (\Theta^k v)_n, & n \geq k. \end{cases}$$

We then denote this policy by $w = [u\{k\}v]$. We similarly define a concatenation of a sequence of policies u^i with times t^i, and denote it by $[u^1\{t^1\}u^2\{t^2\}...]$ (where policy u^i is used for a duration of t^i time units).

Assume (29) does not hold. Then there exist sequences of initial distribution over the states $\xi(i) = \{\xi_1(i), ..., \xi_N(i)\}$, of strictly increasing times $t(i)$ and of policies $u(i)$, and a constant $\alpha_1 > 0$ such that for all i,

$$E_{\xi(i)}^{u(i)} d_{t(i)}^1 \geq \alpha_1. \tag{31}$$

It follows that there exist sequences of strictly increasing times $t'(i)$ and of policies $u'(i)$, and a constant $\alpha_2 > 0$ such that for all i,

$$E_{\xi'}^{u'(i)} d_{t'(i)}^1 \geq \alpha_2 \tag{32}$$

for any initial distribution ξ'. Indeed, fix $t'(i) = t(i) + N, i = 1, 2, ...$ (N is the number of states). Fix some stationary policy s and let $u'(i)$ be the policy $[s\{N\}\Theta^N u(i)]$, i.e., the policy obtained by using s during the first N steps, and then using a shifted policy $\Theta^N u(i)$. Due to the unichain and aperiodicity assumption, the Markov chain induced by the stationary policy s is regular, and it follows (see [18]) that there exists some $\alpha_3 > 0$ such that $P^s_{\xi'}(X_N = z) > \alpha_3$ for any z and ξ'. (31) then implies that (32) holds for all i sufficiently large and ξ' with $\alpha_2 = \alpha_1\alpha_3/2$. Indeed, let i be such that

$$t(i) \geq \frac{4N}{\alpha_1\alpha_3}.$$

It then follows that

$$|d^1_{t(i)} - d^1_{t'(i)}| \leq 2N/t(i) \leq \frac{\alpha_1\alpha_3}{2}.$$

This implies that

$$\begin{aligned}
E^{u'(i)}_{\xi'} d^1_{t'(i)} &= \sum_z P^{u'(i)}_{\xi'}(X_N = z)\left[E^{u'(i)}_{\xi'} d^1_{t'(i)} \middle| X_N = z\right] \\
&= \sum_z P^{s}_{\xi'}(X_N = z)\left[E^{u'(i)}_{\xi'} d^1_{t'(i)} \middle| X_N = z\right] \\
&\geq \sum_z P^{s}_{\xi'}(X_N = z) E^{u(i)}_z d^1_{t(i)} - \frac{\alpha_1\alpha_3}{2} \geq \alpha_3 \sum_z E^{u(i)}_z d^1_{t(i)} - \frac{\alpha_1\alpha_3}{2} \\
&\geq \alpha_3 \sum_z \xi_z(i) E^{u(i)}_z d^1_{t(i)} - \frac{\alpha_1\alpha_3}{2} = \alpha_3 E^{u(i)}_{\xi(i)} d^1_{t(i)} - \frac{\alpha_1\alpha_3}{2} \geq \frac{\alpha_1\alpha_3}{2}. \qquad (33)
\end{aligned}$$

Equation (33) is due to the following. Policy $u'(i)$ behaves like the stationary policy s during the first N steps. So, at time N, we reach state z with probability $P^s_{\xi'}(X_N = z)$. Then the behavior during the interval $[N, t'(i)]$, according to policy $u'(i)$, is that of the policy u during the interval $[0, t'(i) - N] = [0, t(i)]$.

Consider now some subsequence $t'(i)$ for which (32) holds and for which

$$\frac{\sum_{l=1}^i t'(l)}{t'(i+1)} \leq \frac{\alpha_2}{4}.$$

Consider the concatenated policy $\tilde{u}$ defined as $\tilde{u} = [u'(1)\{t'(1)\}u'(2)\{t'(2)\}...]$. (32) implies that

$$\overline{\lim_{K\to\infty}} E^{\tilde{u}}_{\xi'} d^1_K \geq \frac{\alpha_2}{2} > 0 \qquad (34)$$

for any initial distribution ξ'. Indeed, choose any integer n and define $K = \sum_{i=1}^n t'(i)$, $K' = \sum_{i=1}^{n+1} t'(i)$. Then

$$|E^{\bar{u}}_{\xi'}[d^1_{K'}|X(K) = z] - E^{u'(i+1)}_z d^1_{t'(i+1)}| \leq \frac{2\sum_{l=1}^i t'(l)}{t'(i+1)} \leq \frac{\alpha_2}{2},$$

which implies that

$$\begin{aligned}
E^{\tilde{u}}_{\xi'} d^1_{K'} &= \sum_z P^{\tilde{u}}_{\xi'}(X(K) = z) E^{\tilde{u}}_{\xi'}[d^1_{K'}|X(K) = z] \\
&\geq \sum_z P^{\tilde{u}}_{\xi'}(X(K) = z) E^{u'(i+1)}_z d^1_{t'(i+1)} - \frac{\alpha_2}{2} \geq \frac{\alpha_2}{2}. \qquad (35)
\end{aligned}$$

This, however, contradicts (28) for $u = \tilde{u}$. We thus conclude that the convergence in (28) is uniformly in ξ and $u \in U$.

Next, assume that (30) does not hold. Below, if u is stationary, we understand $u(a|x)$ to be the probability of choosing action a when in state x. The class of stationary policies is compact; i.e., for any sequence $u(i) \in \mathcal{S}$, there exists a subsequence $u(i_j)$ such that the policy $u^* = \lim_{j\to\infty} u(i_j)$ (i.e., the policy for which $u^*(a|x) = \lim_{j\to\infty} u(i_j)(a|x)$ for all a and x) is stationary.

It follows by arguments as in the first part of the proof that there exist sequences of times $t(i)$ and of stationary policies $s(i)$, and a constant $\alpha_4 > 0$ such that for all i,

$$E_\xi^{s(i)} d_{t(i)}^2 \geq \alpha_4 \tag{36}$$

for any initial distribution ξ. Moreover, due to the compactness of $\mathcal{S}$, $s(i)$ can be chosen to be a convergent sequence, with s^* its limit. It then follows that

$$\lim_{i\to\infty} \eta(s(i)) = \eta(s^*) \tag{37}$$

(see [17, p. 82]).

Consider now the Markov policy $\tilde{s}$ that follows policy $s(1)$ until time $t(1)$, then switches to $s(2)$ and uses that policy until $t(2)$, then switches to $s(3)$ and uses it until $t(3)$, and so on. Since for any initial distribution ξ and for any stationary policy $s(i)$, we have

$$\psi_0 = \eta(s(i)), \qquad P_\xi^{s(i)} \text{ a.s.}, \tag{38}$$

it follows by choosing the sequence of times $t(i)$ so that the intervals $t(i+1) - t(i)$ are sufficiently large, that (36) implies that

$$\overline{\lim_{i\to\infty}} E_\xi^{\tilde{s}} \left\| \psi_0^{t(i)} - \eta(s(i)) \right\|_1 > 0 \tag{39}$$

for any initial distribution ξ. It then follows from (37) and (39) that

$$\overline{\lim_{t\to\infty}} E_\xi^{\tilde{s}} \left\| \psi_0^{t} - \eta(s^*) \right\|_1 > 0 \tag{40}$$

for any initial distribution ξ.

Since $s(i)$ converges to s^*, it follows that $\tilde{s}$ is an asymptotically stationary policy (see (1.2) in [3]), and therefore,

$$\lim_{K\to\infty} \psi_0^K = \eta(s^*), \qquad P_\xi^{\tilde{s}} \text{ a.s.}$$

(see Lemma 6.3 in [3]; also see [4]). Hence

$$\lim_{K\to\infty} E_\xi^{\tilde{s}} \left\| \psi_0^K - \eta(s^*) \right\|_1 = 0 \tag{41}$$

for any initial distribution ξ. This contradicts (40), and thus (30) is established. □

Proof of Lemma 2.1. Let y_t be an admissible control for $\mathbf{Q_0}$ and let $u_\epsilon(y)$ be constructed as indicated in the beginning of section 3. Consider the policy $u_\epsilon(y)$ and a random realization of states and actions history $H \in \mathbf{H}$. The solution Z_t of (1) is the solution of (9) obtained with the random control

$$y_t(H) \stackrel{\text{def}}{=} y(X_{\lfloor t/\epsilon \rfloor}, A_{\lfloor t/\epsilon \rfloor}).$$

By Lemma 4.1, the mathematical expectation of the norm of the difference between Z_t and the solution z_t^y of (9) with the control y_t is bounded by

$$E_x^{u_\epsilon(y)} \|Z_t - z_t^y\|_1 \le L \left(\Delta + \max_{l=0,\dots,\ell-1} E_x^{u_\epsilon(y)} \left\| \frac{1}{\Delta} \int_{\tau_l}^{\tau_{l+1}} y_s(H) ds - \frac{1}{\Delta} \int_{\tau_l}^{\tau_{l+1}} y_s ds \right\|_1 \right)$$

for any $t \in [0,1]$. Hence, taking into account (16) and (17),

$$\begin{aligned} (42) \quad & \max_{t\in[0,1]} E_x^{u_\epsilon(y)} \|Z_t - z_t^y\|_1 \\ & \le L \left(\Delta + \max_{l=0,\dots,\ell-1} E_x^{u_\epsilon(y)} \left\| \frac{1}{\Delta} \int_{\tau_l}^{\tau_{l+1}} y_s(H) ds - \sum_{v,a} \eta(s_l^\epsilon(y); v, a) y(v,a) \right\|_1 \right). \end{aligned}$$

To bound the right-hand side in (42), consider the state-action frequencies ψ_m^K corresponding to the realization H. It follows from Theorem 4.1 that there exists some $\mu : \mathbb{N} \to \mathbb{R}$ with

$$\lim_{K\to\infty} \mu(K) = 0 \tag{43}$$

such that for any stationary policy s applied during $n = m+1, ..., m+K$, and any probability distribution ζ over X_m,

$$E_\zeta^s \left(\max_{v,a} |\psi_m^K(v,a) - \eta(s; v, a)| \right) \le \mu(K). \tag{44}$$

Denote

$$K(\epsilon) \stackrel{\text{def}}{=} \min_{l=0,1,\dots,\ell-1} \left(\lfloor \tau_{l+1}/\epsilon \rfloor - \lfloor \tau_l/\epsilon \rfloor \right),$$

and notice that

$$\begin{aligned} & 2 \ge \lfloor \tau_{l+1}/\epsilon \rfloor - \lfloor \tau_l/\epsilon \rfloor - K(\epsilon) \ge 0, \qquad \left| K(\epsilon) - \frac{\Delta(\epsilon)}{\epsilon} \right| \le 1 \\ (45) \quad & \Rightarrow \qquad \left| \frac{1}{K(\epsilon)} - \frac{\epsilon}{\Delta(\epsilon)} \right| \le \frac{\epsilon^2}{\Delta(\epsilon)^2} \left(\frac{1}{1 - \epsilon/\Delta(\epsilon)} \right). \end{aligned}$$

From (45) it follows that there exist constants L_1 and L_2 such that

$$\left\| \frac{1}{\Delta(\epsilon)} \int_{\tau_l}^{\tau_{l+1}} y_t(H) dt - \frac{\epsilon}{\Delta(\epsilon)} \sum_{n=\lfloor \tau_l/\epsilon \rfloor + 1}^{\lfloor \tau_l/\epsilon \rfloor + K(\epsilon)} y(X_n, A_n) \right\|_1 \le L_1 \frac{\epsilon}{\Delta(\epsilon)}, \tag{46}$$

$$\left\| \frac{\epsilon}{\Delta(\epsilon)} \sum_{n=\lfloor \tau_l/\epsilon \rfloor + 1}^{\lfloor \tau_l/\epsilon \rfloor + K(\epsilon)} y(X_n, A_n) - \frac{1}{K(\epsilon)} \sum_{n=\lfloor \tau_l/\epsilon \rfloor + 1}^{\lfloor \tau_l/\epsilon \rfloor + K(\epsilon)} y(X_n, A_n) \right\|_1 \le L_2 \frac{\epsilon}{\Delta(\epsilon)}. \tag{47}$$

Since

$$\frac{1}{K(\epsilon)} \sum_{n=\lfloor \tau_l/\epsilon \rfloor + 1}^{\lfloor \tau_l/\epsilon \rfloor + K(\epsilon)} y(X_n, A_n) = \sum_{v,a} \psi_{\lfloor \tau_l/\epsilon \rfloor}^{K(\epsilon)}(H; v, a) y(v,a), \tag{48}$$

one can obtain, using (44), (46), and (47),

$$
\begin{aligned}
&E_x^{u_\epsilon(y)}\left\|\frac{1}{\Delta(\epsilon)}\int_{\tau_l}^{\tau_{l+1}} y_t(H)dt-\sum_{v,a}\eta(s_l^\epsilon(y);v,a)y(v,a)\right\|_1\\
&\le (L_1+L_2)\frac{\epsilon}{\Delta(\epsilon)}+E_x^{u_\epsilon(y)}\left\{E^{s_l^\epsilon(y)}_{X_{\lfloor\tau_l/\epsilon\rfloor}}\sum_{v,a}\left(\left|\psi^{K(\epsilon)}_{\lfloor\tau_l/\epsilon\rfloor}(H;v,a)-\eta(s_l^\epsilon(y);v,a)\right|\|y(v,a)\|_1\right)\right\}\\
&\le (L_1+L_2)\frac{\epsilon}{\Delta(\epsilon)}+L_3\mu(K(\epsilon)),
\end{aligned}
$$

where

$$L_3=\sum_{v,a}\|y(v,a)\|_1.$$

Substituting the last inequality in (42), one obtains

$$\max_{t\in[0,1]}E_x^{u_\epsilon(y)}\|Z_t-z_t^y\|_1\le L\left[\Delta(\epsilon)+(L_1+L_2)\frac{\epsilon}{\Delta(\epsilon)}+L_3\mu(K(\epsilon))\right],$$

which, by (43), completes the proof of the lemma. □

Proof of Lemma 2.2. Let $h=\left\{x_0,a_0,...,x_{\lfloor\epsilon^{-1}\rfloor},a_{\lfloor\epsilon^{-1}\rfloor}\right\}\in\mathbf{H}$ be some state-action trajectory, and define

$$y_t(h)\overset{\text{def}}{=}y(x_{\lfloor t/\epsilon\rfloor},a_{\lfloor t/\epsilon\rfloor}).$$

As in (46)–(48), one obtains

$$\left\|\frac{1}{\Delta(\epsilon)}\int_{\tau_l}^{\tau_{l+1}} y_t(h)dt-\sum_{v,a}\psi^{K(\epsilon)}_{\lfloor\tau_l/\epsilon\rfloor}(h;v,a)y(v,a)\right\|_1\le(L_1+L_2)\frac{\epsilon}{\Delta(\epsilon)}. \tag{49}$$

Denote by $\sigma_l(H)$ the projection of $\psi^{K(\epsilon)}_{\lfloor\tau_l/\epsilon\rfloor}(H)$ on W; i.e., $\sigma_l(H):=\{\sigma_l(H;v,a)\}_{v,a}$ is the solution of

$$\min_\eta\left\{\left\|\psi^{K(\epsilon)}_{\lfloor\tau_l/\epsilon\rfloor}(H)-\eta\right\|_1\Big|\,\eta\in W\right\}. \tag{50}$$

It follows from Theorem 4.1 that there exists a function $\nu(K)$,

$$\lim_{K\to\infty}\nu(K)=0,$$

such that for any policy u,

$$E_x^u\text{dist}\left\{\psi_m^K(H),W\right\}\le\nu(K)$$

where

$$\text{dist}\left\{\psi_m^K(H),W\right\}\overset{\text{def}}{=}\min_\eta\left\{\left\|\psi_m^K(H)-\eta\right\|_1\Big|\,\eta\in W\right\}.$$

Hence,

$$E_x^u\left\{\max_{v,a}\left|\psi^{K(\epsilon)}_{\lfloor\tau_l/\epsilon\rfloor}(H;v,a)-\sigma_l(H;v,a)\right|\right\}\le\nu(K(\epsilon)). \tag{51}$$

Define the vectors $y_l : \mathbf{H} \to \mathbb{R}$ as

$$y_l(h) = \sum_{v,a} \sigma_l(h; v, a) y(v, a). \tag{52}$$

Since, by definition, $\sigma_l(h) \in W$, then

$$y_l(h) \in \mathbf{Y} \qquad \forall l = 0, 1, ..., \ell - 1.$$

Define now the piecewise constant function $\tilde{y}^\epsilon_t(h)$ as follows: for $t \in [0, \ell\Delta)$, set $\tilde{y}^\epsilon_t(h) := y_l(h)$ for $t \in [\tau_l, \tau_{l+1})$, $l = 0, 1, ..., \ell - 1$. For $t \in [\ell\Delta, 1]$, set $\tilde{y}^\epsilon(h) = \overline{y}$ where $\overline{y}$ is an arbitrary element of $\mathbf{Y}$. Let u be an arbitrary policy. Taking into account (49), (51), and (52), one obtains

$$\begin{aligned}
&E^u_x \left\| \frac{1}{\Delta(\epsilon)} \int_{\tau_l}^{\tau_{l+1}} y_t(H) dt - \frac{1}{\Delta(\epsilon)} \int_{\tau_l}^{\tau_{l+1}} \tilde{y}^\epsilon_t(H) dt \right\|_1 \\
&\leq (L_1 + L_2) \frac{\epsilon}{\Delta(\epsilon)} + E^u_x \max_{v,a} \left| \psi^{K(\epsilon)}_{\lfloor \tau_l / \epsilon \rfloor}(H; v, a) - \sigma_l(H; v, a) \right| \sum_{v,a} \|y(v, a)\|_1 \\
&\leq (L_1 + L_2) \frac{\epsilon}{\Delta(\epsilon)} + L_3 \nu(K(\epsilon)).
\end{aligned}$$

Applying (25) one obtains

$$\max_{t \in [0,1]} E^u_x \left\| Z_t - \tilde{z}^\epsilon_t(H) \right\|_1 \leq L \left[\Delta(\epsilon) + (L_1 + L_2) \frac{\epsilon}{\Delta(\epsilon)} + L_3 \nu(K(\epsilon)) \right],$$

which completes the proof. □

REFERENCES

[1] M. Abbad and J. A. Filar, *Perturbation and stability theory for Markov control problems*, IEEE Trans. Automat. Control, 37 (1992), pp. 1415–1420.

[2] E. Altman and V. A. Gaitsgory, *Control of a hybrid stochastic system*, Systems Control Lett., 20 (1993), pp. 307–314.

[3] E. Altman and A. Shwartz, *Adaptive control of constrained Markov chains: Criteria and policies*, Ann. Oper. Res., 28 (1991), Special issue on "Markov Decision Processes," O. Hernandez-Lerma and J. B. Lasserre, eds., pp. 101–134.

[4] E. Altman and O. Zeitouni, *Rate of convergence of empirical measures and costs in controlled Markov chains and transient optimality*, Math. Oper. Res., 19 (1994), pp. 955–974.

[5] A. Bensoussan, *Perturbation Methods in Optimal Control Problems*, John Wiley, New York, 1989.

[6] A. Bensoussan and G.L. Blankenship, *Singular perturbations in stochastic control*, in Singular Perturbations and Asymptotic Analysis in Control Systems, P. Kokotovic, A. Bensoussan, and G. Blankenship, eds., Lecture Notes in Control and Inform. Sciences 90, Springer-Verlag, New York, 1987, pp. 171–260.

[7] D. Bielecki and J.A. Filar, *Singularly perturbed Markov control problem: Limiting average cost*, Ann. Oper. Res., 28 (1991), pp. 153–168.

[8] F. H. Clarke, *Optimization and Nonsmooth Analysis*, John Wiley, New York, 1983.

[9] F. Delebecque and J. Quadrat, *Optimal control of Markov chains admitting strong and weak interactions*, Automatica J. IFAC, 17 (1980), pp. 281–296.

[10] F. Delebecque and J. Quadrat, *Contribution of stochastic control singular perturbation averaging and team theories to an example of large scale systems: Management and hydropower production*, IEEE Trans Automat. Control, AC-23 (1978), pp. 209–222.

[11] C. Derman, *Finite State Markovian Decision Processes*, Academic Press, New York, 1970.

[12] C. Derman and R. E. Strauch, *A note on memoryless rules for controlling sequential control processes*, Ann. Math. Stat., 37 (1966), pp. 276–278.

[13] W. H. Fleming and R. W. Rishel, *Deterministic and Stochastic Optimal Control*, Springer-Verlag, Berlin, Heidelberg, New York, 1975.
[14] W. H. Fleming and H. M. Soner, *Controlled Markov Processes and Viscosity Solutions*, Springer-Verlag, New York, 1993.
[15] V. Gaitsgory, *Suboptimization of singularly perturbed control systems*, SIAM J. Control Optim., 30 (1992), pp. 1228–1249.
[16] I. I. Gikhman, *Po povodu odnoi teoremi N.I. Bogolubova*, Ukrain. Mat. Zh., 4 (1952), pp. 215–218.
[17] A. Hordijk, *Dynamic Programming and Markov Potential Theory*, 2nd ed., Mathematical Centre Tracts 51, Mathematisch Centrum, Amsterdam, 1977.
[18] J. G. Kemeny and J. L. Snell, *Finite Markov Chains*, D. Van Nostrand, New York, 1960.
[19] R. Z. Khas'minsky, *Ustoichivost' Sistem Differencial'nih Uravnenii Pri Sluchainih Vozmuscheniah Ikh Parametrov*, Nauka, Moskva, 1969. English translation: *Stochastic Stability of Differential Equations*, 2nd ed., Sijthoff and Noordhoff, Alphen aan den Rijn, the Netherlands, Rockville, MD, 1980.
[20] L. Kleinrock, *Queuing Systems, Volume* II*: Computer Applications*, John Wiley, New York, 1976.
[21] P. V. Kokotovic, H. Khalil, and J. O'Reilly, *Singular Perturbations in Control Analysis and Design*, Academic Press, New York, 1986.
[22] M. A. Krasnosel'skii and S.G. Krein, *O principe usrednenia v nelineinoi mechanike*, Uspekhi Mat. Nauk, 10 (1955), pp. 147–152.
[23] H. Kushner, *Weak Convergence and Singularly Perturbed Stochastic Control and Filtering Problems*, Birkhäuser, Boston, 1990.
[24] R. G. Philips and P.V. Kokotovic, *A singular perturbation approach to modeling and control of Markov chains*, IEEE Trans. Automat. Control, AC-26 (1981), pp. 1087–1094.
[25] A. A. Pervozvansky and V. Gaitsgory, *Theory of Suboptimal Decisions*, Kluwer, Dordrecht, the Netherlands, 1988.

SIAM J. CONTROL OPTIM.
Vol. 35, No. 6, pp. 2086–2100, November 1997

011

PERTURBATION FORMULA FOR REGULAR FREE BOUNDARIES IN ELLIPTIC AND PARABOLIC OBSTACLE PROBLEMS*

SRDJAN STOJANOVIC†

Abstract. The perturbation formula for regular free boundaries with respect to various parameters of the obstacle problem, which has been known in the case of elliptic problems, is proved by a new method. The method introduced here extends easily to parabolic problems as well. This enables a study of optimal control of regular free boundaries in elliptic and parabolic obstacle problems.

Key words. obstacle problem, free boundary, perturbation formula

AMS subject classifications. 35J85, 35R35

PII. S0363012995282796

1. Introduction.

In the past 20 years or so, there have been many works on so-called optimal control theory of variational inequalities (see [18, 10, 1, 11, 12, 9]). One could say that the problem of distributed control, i.e., the problem where the objective is to control a solution of the variational inequality (more precisely, a solution of the obstacle problem), is well understood. Also, various attempts were made to handle the issue of controlling the free boundary via different types of regularizations or relaxations (see, e.g., [1, 2] and references given there).

More appropriately, optimal control of the free boundary in the elliptic obstacle problem was considered in [17] via a perturbation formula for free boundaries. In such an approach no regularization or relaxation of the optimal control problem is needed. The proof of the perturbation formula in [17] applies to the elliptic problem. Moreover, in [17], authors assume differentiability (of the obstacle map).

In this paper we introduce a new proof of the perturbation formula, which, contrary to [17], naturally extends to the parabolic case, providing a unified theory for various optimal control problems for regular free boundaries in elliptic and parabolic obstacle problems. Furthermore, as opposed to [17], we discuss the relationship between the issue of regularity of the free boundary and the issue of differentiability of the obstacle map and of the cost functional.

We shall state the perturbation formula for the free boundary in the case when the perturbation of data is in the right-hand side of the obstacle problem. Many other kinds of data perturbations can be considered by the same method.

In this paper, by optimal control, we mean only characterizing the gradient of the cost functional. This can be used either for numerical minimization or for further analysis, to characterize possible minimizers.[1] Neither is in the scope of the present paper.

In several crucial places[2] in this paper, we shall exploit the viewpoint that the obstacle problem is a semilinear *equation*. Namely, as was discussed in detail in [14] by the present author, the obstacle problem (2.10) is equivalent to the semilinear

*Received by the editors January 25, 1995; accepted for publication (in revised form) September 5, 1996. This research was supported in part by the Taft Memorial Foundation.
http://www.siam.org/journals/sicon/35-6/28279.html

†Department of Mathematical Sciences, University of Cincinnati, Cincinnati, OH 45221-0025 (srdjan@math.uc.edu).

[1] That appears possible, but not in full generality. One has to consider particular examples, as in the study of regularity of the free boundary; see, e.g., [8].

[2] See the proofs of Lemma 1, Theorem 4, and Theorem 5.

equation

$$Az + uI_{\{z^u>0\}} = 0, \qquad z \in W^{2,2}(\Omega), \tag{1.1}$$

with appropriate boundary condition (for the notation, see (2.1), (2.8), and (2.49)). In [14] the variational principle for equation (1.1) is given, as well.

Some of the results of this paper were announced in [15]. Some further developments can be found in [16].

2. Perturbation formula for regular free boundaries in elliptic obstacle problems. Let Ω be a bounded domain in R^n such that $\partial\Omega$ is locally in $C^{2,\alpha}$ for some $\alpha > 0$. Let A be an elliptic operator defined by

$$Az = -\sum_{i,j=1}^{n} a_{ij}\frac{\partial^2 z}{\partial x_i \partial x_j} + \sum_{i=1}^{n} b_i \frac{\partial z}{\partial x_i} + cz. \tag{2.1}$$

We assume that

$$a_{ij} = a_{ji}, \quad a_{ij} \in C^{2,\alpha}(\Omega) \cap C^{1,\alpha}(\bar{\Omega}), \tag{2.2}$$

$$b_i, c \in C^{\alpha}(\bar{\Omega}), \tag{2.3}$$

$$\sum_{i,j=1}^{n} a_{ij}(x)\xi_i\xi_j \geq \lambda_0|\xi|^2 \quad \forall x \in \Omega,\ \xi \in R^n \ \ (\lambda_0 > 0), \tag{2.4}$$

$$c \geq 0. \tag{2.5}$$

Functions a_{ij}, b_i, and c are going to be referred to as *data*. Consider also the associated bilinear form

$$a(z,\zeta) = \int_\Omega \left\{ \sum_{i,j=1}^{n} a_{ij}\frac{\partial z}{\partial x_i}\frac{\partial \zeta}{\partial x_j} + \sum_{i=1}^{n}\left(b_i + \sum_{j=1}^{n}\frac{\partial a_{ij}}{\partial x_j}\right)\frac{\partial z}{\partial x_i}\zeta + cz\zeta \right\} dx, \tag{2.6}$$

and assume that a is coercive, that is, that

$$a(z,z) \geq \beta\|z\|^2_{H_0^1(\Omega)} \ \ \forall z \in H_0^1(\Omega) \ \ (\beta > 0). \tag{2.7}$$

Let $u \in C^{0,1}(\bar{\Omega})$ and $g \in C^{2,\alpha}(\bar{\Omega})$ be given. We also assume here that

$$u \geq \lambda_1 > 0 \ \text{ in } \bar{\Omega}, \tag{2.8}$$

$$g > 0 \ \text{ in } \bar{\Omega}. \tag{2.9}$$

Consider the following obstacle problem: find $z \in K$, such that

$$a(z, \zeta - z) \geq -\int_\Omega u(\zeta - z)dx \ \ \forall \zeta \in K, \tag{2.10}$$

where K is the following closed convex set in $H^1(\Omega)$:

$$K = \left\{\zeta \in H^1(\Omega); \zeta - g \in H_0^1(\Omega), \zeta \geq 0 \text{ a.e.}\right\}. \tag{2.11}$$

Remark 1. We have taken the obstacle to be equal to 0. It is trivial to extend everything that follows to the case when the obstacle is any $C^{2,\alpha}(\bar{\Omega})$ function ψ, such that $\psi < g$ on $\partial\Omega$ (see Remark 3).

It is very well known (see, e.g., [8]; see also [14] for a new proof of $W^{2,p}$-regularity) that under the above assumptions, the obstacle problem has a unique solution z, and

$$z \in W^{2,p}(\Omega) \cap W^{2,\infty}_{loc}(\Omega)\ \forall p < \infty. \tag{2.12}$$

Moreover, by the fundamental theorem of Caffarelli [3, 4, 5], (see also [13]). In some cases, it is possible to claim smoothness of the free boundary $\partial\{z > 0\} \cap \Omega$ (see also Chapter 2 of [8]). For the convenience of the reader, we state those results here. To this end we need some notation.

DEFINITION 1. *For any bounded set $D \subset R^n$ the* minimum diameter *of D, $m.d.(D)$, is the infimum of the distances between pairs Π_1, Π_2 of parallel planes such that D is contained in the strip determined by Π_1, Π_2. Let $B_r(x) = \{y \in R^n; |x-y| < r\}$.*

NOTATION 1.

$$\Gamma^u = \partial\{z^u > 0\} \cap \Omega, \tag{2.13}$$

$$D_{-\epsilon} = \{x \in D \subset R^n; \operatorname{dist}(x, R^n \setminus D) > \epsilon\}, \tag{2.14}$$

$$\eta_r(D; x) = \frac{m.d.(D \cap B_r(x))}{r}, \tag{2.15}$$

$$\eta_r(D) = \eta_r(D; 0). \tag{2.16}$$

THEOREM 1 (Caffarelli). *There exists a positive nondecreasing function $\sigma(r)(0 < r < r_0)$ with $\sigma(0^+) = 0$ such that, if for some $0 < r < r_0$,*

$$\eta_r(\{z^u = 0\}) > \sigma(r), \tag{2.17}$$

then for some $\tilde{r} > 0$, $\Gamma^u \cap B_{\tilde{r}}(0)$ is a C^1 surface given by

$$x_i = k(x_1, ..., x_{i-1}, x_{i+1}, ..., x_n) \qquad (k \in C^1) \tag{2.18}$$

for $\sqrt{x_1^2 + \cdots + x_{i-1}^2 + x_{i+1}^2 + \cdots + x_n^2} < \tilde{r}$, and for some $i \in \{1, ..., n\}$, and moreover

$$z^u \in C^2\left(\left(\{z^u > 0\} \cup \Gamma^u\right) \cap B_{\tilde{r}}(0)\right). \tag{2.19}$$

The nonexplicit nature of function σ prevents us from applying Theorem 1 directly. Nevertheless, from the properties of σ and η_r, one can see that the following holds.

COROLLARY 1. *If*

$$\limsup_{r \to 0} \frac{\mathcal{L}^n(\{z^u > 0\} \cap B_r(0))}{r^n} > 0, \tag{2.20}$$

then the assertions of Theorem 1 are valid.

COROLLARY 2. *If $\Gamma^u \subset \Omega$ is a Lipschitz continuous surface, then it is a C^1 surface, and z^u is C^2 in the closure of the noncoincidence set $\{z^u > 0\}$.*

DEFINITION 2. *If the assumptions of Theorem 1 are (not) fulfilled at each free boundary point $x \in \Gamma^u$, we shall say either that the free boundary Γ^u is* (not) regular

(or that it is irregular*) or that the coincidence set* $\{z^u = 0\}$ *is* not too thin *(is too thin).*

Example 1. Let $\Omega = (-1, 1)$, $g = 1$, $Az = -\frac{1}{2}z''$, $u = v = 1$. Then it is elementary to compute

$$z^{u+\lambda v}(x) = \left[(1+\lambda)x^2 - 2(1+\lambda-\sqrt{1+\lambda})|x| + 2 + \lambda - 2\sqrt{1+\lambda}\right] \cdot \left(1 - I_{\left(-1+\frac{1}{\sqrt{1+\lambda}}, 1-\frac{1}{\sqrt{1+\lambda}}\right)}(x)\right) \tag{2.21}$$

if $\lambda \geq 0$, and

$$z^{u+\lambda v}(x) = (1+\lambda)x^2 - \lambda \tag{2.22}$$

if $\lambda \leq 0$. So there is no free boundary if $\lambda < 0$, and otherwise, if $\lambda \geq 0$, then

$$\eta_r(\{z^{u+\lambda v} = 0\}) = \frac{m.d.(\{z^{u+\lambda v} = 0\} \cap B_r)}{r} = \begin{cases} 1, & \lambda > 0, \\ 0, & \lambda = 0, \end{cases} \tag{2.23}$$

if r is small enough. According to Definition 2, the free boundary is regular if $\lambda > 0$, but it is *not* regular if $\lambda = 0$, i.e., the coincidence set *is* too thin.[3]

THEOREM 2 (Caffarelli). *If*

$$\|z^u - z^v\|_{L^\infty(\Omega)} < \epsilon^2, \tag{2.24}$$

then

$$\{z^u = 0\}_{-C\epsilon} \subset \{z^v = 0\}, \tag{2.25}$$

where C *depends on the data, on the* $C^{1,1}$ *norm of* z^u *and* z^v*, and on* λ_1*.*

THEOREM 3 (Caffarelli). *Let* K', K *be domains with* C^1 *boundary such that* $K' \subset K, \bar{K} \subset \Omega, c_0 = \operatorname{dist}(K', \partial K) > 0$*. If* $\Gamma^u \cap K$ *is a* C^1 *surface, then* $\Gamma^v \cap K'$ *is a* C^1 *surface provided that*

$$\|z^u - z^v\|_{L^\infty(\Omega)} < \epsilon_0, \tag{2.26}$$

where ϵ_0 *is sufficiently small, depending on the data and on the* $C^{0,1}$ *norms of* u *and* v*, the* $C^{1,1}$ *norms of* z^u *and* z^v*,* λ_1*,* c_0*, and a* C^1 *bound on* $\Gamma^u \cap K$*.*

PROPOSITION 1. *The obstacle map* $u \mapsto z^u$ *is Lipschitz continuous in the following sense:*

$$\|z^u - z^v\|_{H^1(\Omega)} \leq c\|u - v\|_{H^{-1}(\Omega)} \tag{2.27}$$

and, for $p > \frac{n}{2}$*,*

$$\|z^u - z^v\|_{L^\infty(\Omega)} \leq c\|u - v\|_{L^p(\Omega)}. \tag{2.28}$$

[3]Indeed, it is a single point: $\{z^u = 0\} = \{0\}$. So, we notice that, conceptually, the statement about regularity of the free boundary is more about the thickness of the coincidence set than about the smoothness of the free boundary. Indeed, in dimension one, like in the above example, both regular and irregular free boundaries are isolated points, so that the issue of smoothness of the free boundary does not make sense, while the issue of thickness of the coincidence set does.

Proof. Estimate (2.27) is well known. We shall prove (2.28). To this end we prove the following lemma.

LEMMA 1. *If* $u - v \le \epsilon \in L^p(\Omega), p > \frac{n}{2}, \epsilon \ge 0$, *then*

$$z^v - z^u \le \delta \in L^\infty(\Omega), \tag{2.29}$$

where $0 \le \delta \in W^{2,p}(\Omega) \cap H_0^1(\Omega)$ *solves*

$$A\delta = \epsilon. \tag{2.30}$$

Proof. Recall $u, v \ge 0$.[4] Let

$$\bar{u} = uI_{\{z^u > 0\}}, \qquad \bar{v} = \min\{\bar{u}, v\}. \tag{2.31}$$

Then

$$z^{\bar{u}} = z^u, \qquad z^{\bar{v}} \ge z^v. \tag{2.32}$$

Notice also

$$(\bar{u} - \bar{v})I_{\{z^u > 0\}} \le \epsilon \tag{2.33}$$

and

$$\bar{v}I_{\{z^u = 0\}} = 0. \tag{2.34}$$

So it suffices to show that

$$(z^{\bar{v}} - (z^{\bar{u}} + \delta))^+ = 0. \tag{2.35}$$

We have

$$Az^{\bar{v}} + \bar{v}I_{\{z^{\bar{v}} > 0\}} = 0 \tag{2.36}$$

and

$$A(z^{\bar{u}} + \delta) + \bar{u}I_{\{z^{\bar{u}} > 0\}} - \epsilon = 0. \tag{2.37}$$

Subtracting, we get

$$A(z^{\bar{v}} - (z^{\bar{u}} + \delta)) + \bar{v}(I_{\{z^{\bar{v}} > 0\}} - I_{\{z^{\bar{u}} > 0\}}) = (\bar{u} - \bar{v})I_{\{z^{\bar{u}} > 0\}} - \epsilon \le 0. \tag{2.38}$$

Multiplying (2.38) by $(z^{\bar{v}} - (z^{\bar{u}} + \delta))^+$ and using the fact that

$$\bar{v}(I_{\{z^{\bar{v}} > 0\}} - I_{\{z^{\bar{u}} > 0\}})(z^{\bar{v}} - z^{\bar{u}}) \ge 0, \tag{2.39}$$

we get

$$\lambda_0 \int_\Omega |\nabla(z^{\bar{v}} - (z^{\bar{u}} + \delta))^+|^2 - \int_{\{z^{\bar{v}} - (z^{\bar{u}} + \delta) > 0\}} \bar{v}(I_{\{z^{\bar{v}} > 0\}} - I_{\{z^{\bar{u}} > 0\}})\delta \le 0. \tag{2.40}$$

[4] This assumption can be removed.

But, using (2.34),

$$\int_{\{z^{\bar v}-(z^{\bar u}+\delta)>0\}} \bar v(I_{\{z^{\bar v}>0\}}-I_{\{z^{\bar u}>0\}})\delta$$

$$= \int_{\{z^{\bar v}-(z^{\bar u}+\delta)>0\}} \bar v I_{\{z^{\bar v}>0\}} I_{\{z^{\bar u}=0\}}\delta = 0. \qquad \square \tag{2.41}$$

To prove the theorem, one needs only to apply the L^p-estimate and the imbedding theorem for problem (2.30). $\square$

From Theorems 2 and 3 and from Proposition 1, it is not difficult to conclude (cf. [9]) the following proposition.

PROPOSITION 2. *If u is such that the corresponding free boundary Γ^u is regular, then the obstacle map is differentiable at u in the following sense:*

$$\frac{z^{u+\lambda v}-z^u}{\lambda} \rightharpoonup w^{u;v} \quad \textit{weakly in } H_0^1(\Omega) \textit{ and weakly* in } L^\infty(\Omega) \tag{2.42}$$

as $\lambda \to 0$, where

$$w^{u;v} = \begin{cases} \delta^{u;v} & \textit{in } \{z^u>0\}, \\ 0 & \textit{in } \{z^u=0\} \end{cases} \tag{2.43}$$

and where $\delta = \delta^{u;v}$ is the unique solution of the elliptic equation

$$\begin{aligned} A\delta &= -v \quad \textit{in } \{z^u>0\}, \\ \delta &= 0 \quad \textit{on } \partial\{z^u>0\}. \end{aligned} \tag{2.44}$$

Example 2. We continue Example 1. It is elementary to compute that

$$\frac{z^{u+\lambda v}-z^u}{\lambda} \longrightarrow \begin{cases} x(x+1), & x\in(-1,0), \\ x(x-1), & x\in(0,1), \end{cases} \tag{2.45}$$

as $\lambda \downarrow 0$, and

$$\frac{z^{u+\lambda v}-z^u}{\lambda} = (x-1)(x+1) \tag{2.46}$$

as $\lambda \uparrow 0$. Comparing (2.45) and (2.46) we see that (2.42) does not hold.

So, if the coincidence set corresponding to the control u is too thin, then differentiability of the obstacle map may fail at u.

NOTATION 2. *Let $d\Gamma^u$ denote the measure (cf.* [6])

$$d\Gamma^u = \mathcal{H}^{n-1}\lfloor \Gamma^u, \tag{2.47}$$

the restriction of the $(n-1)$-dimensional Hausdorff measure $\mathcal{H}^{n-1}$ on the set $\Gamma^u \subset R^n$.

Remark 2. If $\Gamma^u \subset \Omega$ is a Lipschitz continuous surface, then by the trace theorem,

$$d\Gamma^u \in H^{-1}(\Omega). \tag{2.48}$$

Let ν^u be the unit normal to Γ^u exterior to $\{z^u>0\}$. Also, let ν_A^u be a conormal, i.e., $(\nu_A^u)_i = \sum_{j=1}^n a_{ij}(\nu^u)_j$.

NOTATION 3.

$$I_D(x) = \begin{cases} 1, & x \in D, \\ 0, & x \notin D. \end{cases} \tag{2.49}$$

We have the following theorem.

THEOREM 4 (perturbation formula for regular free boundaries in elliptic obstacle problems). *Suppose that z^u has a regular free boundary. Then the following perturbation formula holds:*

$$\frac{I_{\{z^{u+\lambda v}>0\}} - I_{\{z^u>0\}}}{\lambda} \rightharpoonup -\frac{1}{u}\frac{\partial \delta^{u;v}}{\partial \nu_A^u} d\Gamma^u \tag{2.50}$$

weakly in $H^{-1}(\Omega)$, as $\lambda \to 0$. Also,

$$-\frac{1}{u}\frac{\partial \delta^{u;v}}{\partial \nu_A^u} d\Gamma^u = -\frac{\dfrac{\partial \delta^{u;v}}{\partial \nu^u}}{\dfrac{\partial^2 z^u}{(\partial \nu^u)^2}} d\Gamma^u. \tag{2.51}$$

Proof. z^u solves the semilinear equation.

$$\begin{aligned} Az^u + uI_{\{z^u>0\}} &= 0 \ \text{ a.e. in } \Omega, \\ z^u &= g \ \text{ on } \partial\Omega. \end{aligned} \tag{2.52}$$

So, writing (2.52) in the weak form for $z^{u+\lambda v}$ and z^u, subtracting and dividing by λ, we get, for every $\varphi \in H_0^1(\Omega)$,

$$\begin{aligned} &-a\left(\frac{z^{u+\lambda v} - z^u}{\lambda}, \varphi\right) \\ &\quad = \int_\Omega v I_{\{z^{u+\lambda v}>0\}} \varphi dx + \int_\Omega u \frac{1}{\lambda}\left(I_{\{z^{u+\lambda v}>0\}} - I_{\{z^u>0\}}\right)\varphi dx. \end{aligned} \tag{2.53}$$

We can pass the limit $\lambda \to 0$ in (2.53) to conclude

$$\begin{aligned} &-a(w, \varphi) \\ &\quad = \int_\Omega v I_{\{z^u>0\}} \varphi dx + \lim_{\lambda\to 0} \int_\Omega u \frac{1}{\lambda}\left(I_{\{z^{u+\lambda v}>0\}} - I_{\{z^u>0\}}\right)\varphi dx. \end{aligned} \tag{2.54}$$

Now, from (2.43) and (2.54), we conclude that

$$\lim_{\lambda\to 0} \int_\Omega u \frac{1}{\lambda}\left(I_{\{z^{u+\lambda v}>0\}} - I_{\{z^u>0\}}\right)\varphi dx = -\int_{\Gamma^u} \frac{\partial \delta^{u;v}}{\partial \nu_A^u} \varphi d\sigma, \tag{2.55}$$

and hence, since $\varphi \in H_0^1(\Omega)$ if and only if $\frac{\varphi}{u} \in H_0^1(\Omega)$,

$$\lim_{\lambda\to 0} \int_\Omega \frac{1}{\lambda}\left(I_{\{z^{u+\lambda v}>0\}} - I_{\{z^u>0\}}\right)\varphi dx = -\int_{\Gamma^u} \frac{1}{u}\frac{\partial \delta^{u;v}}{\partial \nu_A^u} \varphi d\sigma, \tag{2.56}$$

and hence (2.50) is proved.

Notice that, since on Γ^u,

$$z^u = |\nabla z^u| = 0, \tag{2.57}$$

we have

$$u(x) = -Az^u(x) = \sum_{i,j=1}^{n} a^{ij} z_{x_i x_j}(x) \ \ \forall x \in \Gamma^u. \tag{2.58}$$

Notice that (2.58) makes sense, since Γ^u is regular, and hence (2.19) holds. Furthermore, fix $x^0 \in \Gamma^u$ and notice that, without loss of generality, we can suppose that $\nu^u(x^0) = e_n$. Then it is easy to see, using (2.57), that

$$\sum_{i,j=1}^{n} a_{ij} z_{x_i x_j}(x^0) = a_{nn} z_{x_n x_n}(x^0) \tag{2.59}$$

and, since $\delta^{u;v} = 0$ on Γ^u, we deduce also

$$\sum_{i,j=1}^{n} a_{ij} (\delta^{u;v})_{x_i} (\nu^u)_j = a_{nn} (\delta^{u;v})_{x_n}. \tag{2.60}$$

We conclude that, in general,

$$\frac{\displaystyle\sum_{i,j=1}^{n} a_{ij} (\delta^{u;v})_{x_i} (\nu^u)_j}{\displaystyle\sum_{i,j=1}^{n} a_{ij} z_{x_i x_j}} = \frac{(\delta^{u;v})_{\nu^u}}{(z^u)_{\nu^u \nu^u}} \ \text{ on } \ \Gamma^u, \tag{2.61}$$

and hence (2.51) follows. □

Remark 3. It is easy to see that in the case of the general obstacle $\psi \in C^{2,\alpha}(\bar{\Omega})$, formula (2.50) becomes

$$\frac{I_{\{z^{u+\lambda v}>0\}} - I_{\{z^u>0\}}}{\lambda} \rightharpoonup -\frac{1}{u + A\psi} \frac{\partial \delta^{u;v}}{\partial \nu_A^u} d\Gamma^u \tag{2.62}$$

weakly in $H^{-1}(\Omega)$, as $\lambda \to 0$.

DEFINITION 3. *If there exists a function $s^{u;v}$ on Γ^u, such that*

$$\frac{I_{\{z^{u+\lambda v}>0\}} - I_{\{z^u>0\}}}{\lambda} \rightharpoonup s^{u;v} d\Gamma^u \tag{2.63}$$

weakly in $H^{-1}(\Omega)$, as $\lambda \to 0$, then $s^{u;v}$ is called the perturbation function.

We can rephrase Theorem 4 now as follows.

PERTURBATION FORMULA 1.

$$s^{u;v} = -\frac{1}{u} \frac{\partial \delta^{u;v}}{\partial \nu_A^u}, \tag{2.64}$$

where $\delta^{u;v}$ is defined by (2.44).

3. Optimal control of free boundaries. Again, we emphasize that, in this paper, by optimal control we mean only characterizing the gradient of the cost functional.

3.1. Linear growth functional. Let $\mathcal{O}$ be a closed set such that $\mathcal{O} \subset \Omega$. The set of controls is

$$\mathcal{C}_1 = \{u \in C^{0,1}(\bar{\Omega}); u \geq \lambda_1 > 0\}. \tag{3.1}$$

Consider a problem of finding $u \in \mathcal{C}_1$ such that the corresponding coincidence set $\{z^u = 0\}$ is, in some sense, as close as possible to $\mathcal{O}$. For example, one can try to minimize the following functional:

$$\begin{aligned} \Phi_1(u) &= \int_\Omega (I_{\{z^u>0\}} - I_{\mathcal{O}^c})^2 dx \\ &= \mathcal{L}^n\left((\{z^u > 0\} \setminus \mathcal{O}^c) \cup (\mathcal{O}^c \setminus \{z^u > 0\})\right). \end{aligned} \tag{3.2}$$

Various regularizations of this functional have been considered (see, e.g., [1]). On the contrary, using formula (2.50), we can consider (3.2) without any regularization.

Compute the directional derivative $\Phi_1'(u; v)$:

$$\begin{aligned} \Phi_1'(u;v) &= \lim_{\lambda\to 0} \frac{\Phi_1(u+\lambda v) - \Phi_1(u)}{\lambda} \\ &= \lim_{\lambda\to 0} \int_\Omega \frac{1}{\lambda}\left[(I_{\{z^{u+\lambda v}>0\}} - I_{\mathcal{O}^c})^2 - (I_{\{z^u>0\}} - I_{\mathcal{O}^c})^2\right] dx \\ &= \lim_{\lambda\to 0} \int_\Omega \frac{1}{\lambda}(I_{\{z^{u+\lambda v}>0\}} - I_{\{z^u>0\}})(I_{\{z^{u+\lambda v}>0\}} + I_{\{z^u>0\}} - 2I_{\mathcal{O}^c})dx \\ &= \lim_{\lambda\to 0} \int_\Omega \frac{1}{\lambda}(I_{\{z^{u+\lambda v}>0\}} - I_{\{z^u>0\}})(I_{\mathcal{O}} - I_{\mathcal{O}^c})dx \\ &= -\int_{\Gamma^u} \frac{1}{u}\frac{\partial \delta^{u;v}}{\partial \nu_A^u}(I_{\mathcal{O}} - I_{\mathcal{O}^c})d\sigma \end{aligned} \tag{3.3}$$

provided, say,[5]

$$\partial\{z^u > 0\} \cap \partial\mathcal{O} \text{ consists of finitely many points.} \tag{3.4}$$

Define the adjoint operator A^* by

$$A^* z = -\sum_{i,j=1}^n \frac{\partial^2}{\partial x_i \partial x_j}(a_{ij} z) - \sum_{i=1}^n \frac{\partial}{\partial x_i}(b_i z) + cz. \tag{3.5}$$

Define the (adjoint) function $p = p^u$ as a solution of

$$\begin{aligned} A^* p &= 0 \text{ in } \{z^u > 0\}, \\ p &= \frac{1}{u} \text{ on } \Gamma^u \cap \mathcal{O}, \\ p &= -\frac{1}{u} \text{ on } \Gamma^u \cap \mathcal{O}^c, \\ p &= 0 \text{ on } \partial\Omega, \\ p &= 0 \text{ in } \{z^u = 0\}. \end{aligned} \tag{3.6}$$

Then

$$\begin{aligned} -\int_{\Gamma^u} \frac{1}{u}\frac{\partial \delta^{u;v}}{\partial \nu_A^u}(I_{\mathcal{O}^c} - I_{\mathcal{O}})d\sigma &= \int_{\Gamma^u} p^u \frac{\partial \delta^{u;v}}{\partial \nu_A^u} d\sigma \\ &= \int_{\{z^u>0\}} (\delta^{u;v} A^* p^u - p^u A \delta^{u;v})\, dx = \int_\Omega p^u v dx, \end{aligned} \tag{3.7}$$

and we have proved the following proposition.

[5] See Remark 5.

PROPOSITION 3. *Let $u \in \mathcal{C}_1$ be such that the corresponding free boundary Γ^u is regular. Also, let* (3.4) *hold. Then Φ_1 is differentiable at u, and*

$$\Phi_1'(u;v) = \int_\Omega p^u v dx, \tag{3.8}$$

where p^u is defined in (3.6).

Remark 4. Under the above assumption, if $\nabla\Phi_1(u)$ is identified as an element of $L^2(\Omega)$, then

$$\nabla\Phi_1(u) = p^u. \tag{3.9}$$

This can be used for numerical calculations (e.g., steepest descent method) or for further analysis.

Remark 5. It does not seem possible to state reasonable a priori conditions to ensure the fulfillment of the assumption (3.4). The value of (3.3) and (3.4) is then in numerical computations: for given u, one computes z^u and checks (3.4); if (3.4) holds, one proceeds as explained. If, on the other hand, $\mathcal{H}^{n-1}(\partial\{z^u > 0\} \cap \partial\mathcal{O}) \neq 0$, then Φ_1 is not differentiable at u. In that case, formally, the best selection (for numerical calculations) of an element of "$\partial\Phi_1(u)$" would be achieved by imposing $p = 0$ on $\partial\{z^u > 0\} \cap \partial\mathcal{O}$ in (3.6).

Remark 6 (minimizing the noncoincidence set). Let

$$\Phi_2(u) = \int_\Omega \left[I_{\{z^u > 0\}} + \frac{\epsilon}{2} u^2 \right] dx \tag{3.10}$$

for $\epsilon > 0$.[6] The first term in Φ_2 corresponds to the case when $\mathcal{O} = \emptyset$, i.e., $\mathcal{O}^c = \Omega$. Then, if Γ^u is regular, Φ_2 is differentiable at u, and

$$\Phi_2'(u;v) = \int_\Omega (p^u + \epsilon u) v dx, \tag{3.11}$$

where $p = p^u$ is a solution of

$$\begin{aligned} A^* p &= 0 \ \text{ in } \{z^u > 0\}, \\ p &= -\frac{1}{u} \ \text{ on } \Gamma^u, \\ p &= 0 \ \text{ on } \partial\Omega, \\ p &= 0 \ \text{ in } \{z^u = 0\}. \end{aligned} \tag{3.12}$$

Remark 7 (minimizing a solution). We want to compare Φ_2 from Remark 6 with a "distributed" functional, say,

$$\Psi(u) = \int_\Omega \left[z^u + \frac{\epsilon}{2} u^2 \right] dx. \tag{3.13}$$

Then, if Γ^u is regular, Ψ is differentiable at u, and it is not difficult to see that

$$\Psi'(u;v) = \int_\Omega (q^u + \epsilon u) v dx, \tag{3.14}$$

[6] Of course, this L^2-penalization is insufficient to claim existence of a minimizer, since the control set imposes $C^{0,1}$-regularity on control functions. Its purpose is only to avoid monotonicity of Φ_2, which may be viewed as trivial. Indeed, if $\epsilon = 0$, then by the monotonicity of the obstacle problem, $u \geq v \Rightarrow z^u \leq z^v \Rightarrow \Phi_2(u) \leq \Phi_2(v)$ and, formally, $\Phi_2(+\infty) = 0$.

where $q = q^u$ is the solution of

$$\begin{aligned} A^*q &= -1 \ \text{ in } \{z^u > 0\}, \\ q &= 0 \ \text{ on } \partial\{z^u > 0\}, \\ q &= 0 \ \text{ in } \{z^u = 0\}. \end{aligned} \tag{3.15}$$

Notice that p^u in (3.12) is discontinuous on Γ_u, so the global regularity of p^u is, say, $L^2(\Omega)$, while q^u in (3.15) is continuous on Γ^u, and then the global regularity of q^u is one derivative more, say, $H^1(\Omega)$.

3.2. Quadratic functional. Without dwelling on details, now change the boundary conditions and introduce other conditions that ensure that z^u is monotone, say, $z^u_{x_n} \le 0$ for some class of admissible controls u. So the corresponding free boundaries Γ^u are graphs: $x_n = \gamma^u(x_1, ..., x_{n-1}) = \gamma^u(x')$. Let Ω_1 be an appropriate domain in R^{n-1}, and consider the functional

$$\Phi_3(u) = \frac{1}{2}\int_{\Omega_1} (\gamma^u - \gamma^*)^2 dx'. \tag{3.16}$$

Then, if Γ^u is regular, Φ_3 is differentiable (in suitable directions) and

$$\begin{aligned} &\Phi_3'(u; v) \\ &\quad = -\int_{\Omega_1} \frac{(\gamma^u(x') - \gamma^*(x'))\dfrac{\partial \delta^{u;v}}{\partial \nu^u_A}(x', \gamma^u(x'))}{u(x', \gamma^u(x'))\sqrt{1 + |\nabla_{x'}\gamma^u(x')|^2}} dx' \\ &\quad = \int_{\Omega_u} p^u v dx, \end{aligned} \tag{3.17}$$

where $p = p^u$ is defined as a solution of

$$\begin{aligned} &A^*p = 0 \ \text{ in } \{z^u > 0\}, \\ &\quad p(x', \gamma^u(x')) = -\frac{(\gamma^u(x') - \gamma^*(x'))}{u(x', \gamma^u(x'))(1 + |\nabla_{x'}\gamma^u(x')|^2)} \ \text{ for } x' \in \Omega_1, \\ &\quad p = 0 \ \text{ on the rest of } \Gamma^u, \\ &\quad p = 0 \ \text{ in } \{z^u = 0\}, \end{aligned} \tag{3.18}$$

with appropriate conditions on $\partial\Omega \cap \partial\{z^u > 0\}$.

4. Extension to parabolic problems.

4.1. Perturbation formula for regular free boundaries in parabolic obstacle problems. For simplicity, we shall consider the heat operator. Results can be extended to the case of general parabolic operator (cf. [7]).

As before, let Ω be a bounded domain in R^n such that $\partial\Omega$ is locally in $C^{2,\alpha}$ for some $\alpha > 0$. Let $Q_T = \Omega \times \{0 < t < T\}$. We assume

$$\begin{aligned} g, D_x g, D_x^2 g, D_t g &\in C^{\alpha}(\bar{Q_T}), \\ u &\in C^{0,1}(\bar{Q_T}). \end{aligned} \tag{4.1}$$

We also assume

$$u \ge \lambda_1 > 0, \qquad g > 0 \ \text{ in } \ \bar{Q_T}. \tag{4.2}$$

Consider the following obstacle problem: find $z \in K$ such that

$$\begin{aligned} &\int_\Omega [z_t(\zeta - z) + \nabla z \cdot \nabla(\zeta - z)]\, dx \\ &\geq -\int_\Omega u(\zeta - z)dx \quad \text{for a.e. } t \ \ \forall \zeta \in K, \end{aligned} \tag{4.3}$$

where K is the following closed convex set in $H^1(Q_T)$:

$$K = \{\zeta \in H^1(Q_T); \zeta = g \ \text{ on } \ \partial_p Q_T, \ \zeta \geq 0 \text{ a.e.}\}. \tag{4.4}$$

∂_p represents the parabolic boundary, in the above case $\partial_p Q_T = \Omega \times \{0\} \cup \partial\Omega \times \{0 < t < T\}$, and the boundary value is taken in the trace sense.

It is well known (see, e.g., [8]) that the obstacle problem (4.3) has a unique solution z, and

$$\begin{aligned} z, D_x z, D_x^2 z, D_t z \in L^p(\Omega) \quad &\text{for a.e. } t \quad \forall p < \infty, \\ z, D_x z, D_x^2 z, D_t z \in L^\infty_{loc}(Q_T). \end{aligned} \tag{4.5}$$

Moreover, by Caffarelli's theorem [3], in some cases, it is possible to claim smoothness of the free boundary $\partial\{z > 0\} \cap Q_T$.

The stability of the free boundaries in the parabolic case is discussed in [8]. The result is similar as in the case of an elliptic obstacle problem.

As before, we denote z^u as the solution of the obstacle problem (4.3) corresponding to the right-hand side u. It is known (see, e.g., [1]) that the map $u \mapsto z^u$ is

$$\text{Lipschitz from } L^2(Q_T) \text{ to } C([0,T]; L^2(\Omega)) \cap L^2(0,T; H^1(\Omega)). \tag{4.6}$$

As in (2.28) in Proposition 1, we can prove the following.

PROPOSITION 4. *The (parabolic) obstacle map $u \mapsto z^u$ is Lipschitz in the following sense:*

$$\|z^u - z^v\|_{L^\infty(Q_T)} \leq c\|u - v\|_{L^p(Q_T)} \tag{4.7}$$

for $p > n$.

Hence, it is not difficult to prove the following proposition.

PROPOSITION 5. *Let u be such that the corresponding free boundary $\partial\{z > 0\} \cap Q_T$ is regular, and that the problem* (4.10) *admits a unique solution. Then*

$$\frac{z^{u+\lambda v} - z^u}{\lambda} \rightharpoonup w^{u;v} \tag{4.8}$$

weakly in $L^2(0,T; H_0^1(\Omega))$ and weakly in $L^\infty(Q_T)$, as $\lambda \to 0$, where*

$$w^{u;v} = \begin{cases} \delta^{u;v} & \text{in } \{z^u > 0\}, \\ 0 & \text{in } \{z^u = 0\}, \end{cases} \tag{4.9}$$

and where $\delta = \delta^{u;v}$ is the unique solution of the parabolic equation

$$\begin{aligned} \delta_t - \Delta\delta &= -v \ \text{ in } \{z^u > 0\}, \\ \delta &= 0 \ \text{ on } \partial_p\{z^u > 0\}. \end{aligned} \tag{4.10}$$

Let $\Gamma^u = \partial\{z^u > 0\} \cap Q_T$, and let, for every $t \in (0,T)$, $\Gamma^u(t) = \partial\{z^u(\cdot,t) > 0\} \cap \Omega$. Also let $\{z^u > 0\}(t) = \{z^u(\cdot,t) > 0\} \cap \Omega$.

NOTATION 4. *Let*

$$X = \{\varphi \in H^1(0,T; H_0^1(\Omega)); \varphi|_{t=T} = 0\} \tag{4.11}$$

and let X^* be the dual of X.

We have the following theorem.

THEOREM 5 (perturbation formula for regular free boundaries; parabolic obstacle problem case). *Under the previous assumptions the following perturbation formula holds:*

$$\frac{I_{\{z^{u+\lambda v}>0\}}-I_{\{z^u>0\}}}{\lambda}\rightharpoonup -\frac{1}{u}\frac{\partial\delta^{u;v}}{\partial\nu_x^u}d\Gamma^u(t)dt \tag{4.12}$$

weakly in X^*, *as* $\lambda\to 0$.

Proof. It is easy to see that z^u solves the semilinear equation

$$\begin{aligned} z_t^u-\Delta z^u+uI_{\{z^u>0\}}&=0 \ \text{ a.e. in } Q_T,\\ z^u&=g \ \text{ on } \partial_pQ_T. \end{aligned} \tag{4.13}$$

So, writing (4.13) in the weak form for $z^{u+\lambda v}$ and z^u, subtracting and dividing by λ, we get, for any $\varphi\in X$,

$$\begin{aligned} \int_{Q_T}\frac{z^{u+\lambda v}-z^u}{\lambda}\varphi_t dxdt-\int_{Q_T}\nabla\left(\frac{z^{u+\lambda v}-z^u}{\lambda}\right)\cdot\nabla\varphi dxdt&\\ =\int_{Q_T}vI_{\{z^{u+\lambda v}>0\}}\varphi dxdt&\\ +\int_{Q_T}u\frac{1}{\lambda}\left(I_{\{z^{u+\lambda v}>0\}}-I_{\{z^u>0\}}\right)\varphi dxdt.& \end{aligned} \tag{4.14}$$

Using (4.8) and the stability theorem of [8], we can pass the limit $\lambda\to 0$ in (2.53) to conclude

$$\begin{aligned} \int_{Q_T}w^{u;v}\varphi_t dxdt-\int_{Q_T}\nabla w^{u;v}\cdot\nabla\varphi dxdt&\\ =\int_{Q_T}vI_{\{z^u>0\}}\varphi dxdt&\\ +\lim_{\lambda\to 0}\int_{Q_T}u\frac{1}{\lambda}\left(I_{\{z^{u+\lambda v}>0\}}-I_{\{z^u>0\}}\right)\varphi dxdt.& \end{aligned} \tag{4.15}$$

Now, from (4.9) and (4.15), we conclude that

$$\begin{aligned} \lim_{\lambda\to 0}\int_0^T\int_\Omega u\frac{1}{\lambda}\left(I_{\{z^{u+\lambda v}>0\}}-I_{\{z^u>0\}}\right)\varphi dxdt&\\ =-\int_0^T\int_{\Gamma^u(t)}\frac{\partial\delta^{u;v}}{\partial\nu_x^u}\varphi d\sigma_x dt& \end{aligned} \tag{4.16}$$

and hence

$$\begin{aligned} \lim_{\lambda\to 0}\int_0^T\int_\Omega\frac{1}{\lambda}\left(I_{\{z^{u+\lambda v}>0\}}-I_{\{z^u>0\}}\right)\varphi dxdt&\\ =-\int_0^T\int_{\Gamma^u(t)}\frac{1}{u}\frac{\partial\delta^{u;v}}{\partial\nu_x^u}\varphi d\sigma_x dt,& \end{aligned} \tag{4.17}$$

which proves the theorem. □

DEFINITION 4. *If there exists a function* $s^{u;v}$ *on* Γ^u, *such that*

$$\frac{I_{\{z^{u+\lambda v}>0\}}-I_{\{z^u>0\}}}{\lambda}\rightharpoonup s^{u;v}d\Gamma^u(t)dt \tag{4.18}$$

weakly in X^*, *as* $\lambda\to 0$, *then* $s^{u;v}$ *is called the* perturbation function.

We can rephrase Theorem 5 now as the following formula.

PERTURBATION FORMULA 2.

$$s^{u;v} = -\frac{1}{u}\frac{\partial \delta^{u;v}}{\partial \nu_x^u}, \tag{4.19}$$

where $\delta^{u;v}$ *is defined by* (4.10).

Remark 8. If, instead of the heat operator $\frac{\partial}{\partial t} - \Delta$, we are dealing with the general parabolic operator $\frac{\partial}{\partial t} + A$, then the differentiation formula (4.12) becomes

$$\frac{I_{\{z^u+\lambda v>0\}} - I_{\{z^u>0\}}}{\lambda} \rightharpoonup -\frac{1}{u}\frac{\partial \delta^{u;v}}{\partial \nu_A^u} d\Gamma^u(t)dt \tag{4.20}$$

weakly in X^*, as $\lambda \to 0$.

4.2. An optimal control problem. We shall consider here only the problem of minimizing the noncoincidence set. So let

$$\Phi_4(u) = \int_{Q_T} \left[I_{\{z^u>0\}} + \frac{\epsilon}{2} u^2 \right] dxdt \tag{4.21}$$

for $\epsilon > 0$. The set of controls is

$$\mathcal{C}_2 = \{u \in C^{0,1}(\bar{Q_T}); u \geq \lambda_1 > 0\}. \tag{4.22}$$

We compute the directional derivative

$$\begin{aligned} \Phi_4'(u;v) &= \lim_{\lambda\to 0} \int_{Q_T} \frac{1}{\lambda} \left[I_{\{z^{u+\lambda v}>0\}} - I_{\{z^u>0\}} \right] dxdt + \int_{Q_T} \epsilon uv dxdt \\ &= -\int_0^T \int_{\Gamma^u(t)} \frac{1}{u}\frac{\partial \delta^{u;v}}{\partial \nu_x^u} d\sigma_x dt + \int_{Q_T} \epsilon uv dxdt. \end{aligned} \tag{4.23}$$

Let $p = p^u$ be defined by

$$\begin{aligned} p_t + \Delta p = 0 \ \text{ in } \{z^u > 0\}, \\ p = -\frac{1}{u} \ \text{ on } \Gamma^u, \\ p = 0 \ \text{ on } \partial_{bp}\{z^u > 0\} \setminus \Gamma^u, \\ p = 0 \ \text{ in } \{z^u = 0\}. \end{aligned} \tag{4.24}$$

Here ∂_{bp} represents the backward parabolic boundary. Then

$$\begin{aligned} \Phi_4'(u;v) &= \int_0^T \int_{\Gamma^u(t)} p^u \frac{\partial \delta^{u;v}}{\partial \nu_x^u} d\sigma_x dt + \int_{Q_T} \epsilon uv dxdt \\ &= \int_0^T \int_{\{z^u>0\}(t)} [(\Delta \delta^{u;v}) p^u - \delta^{u;v} \Delta p^u] \, dxdt + \int_{Q_T} \epsilon uv dxdt \\ &= \int_0^T \int_{\{z^u>0\}(t)} [(\delta_t^{u;v} + v) p^u + \delta^{u;v} p_t^u] \, dxdt + \int_{Q_T} \epsilon uv dxdt \\ &= \int_{Q_T} (p^u + \epsilon u) v dxdt, \end{aligned} \tag{4.25}$$

and we have proved the following proposition.

PROPOSITION 6. *Let $u \in \mathcal{C}_2$ be such that the corresponding free boundary Γ^u is regular and that problems* (4.10) *and* (4.24) *admit unique solutions. Then Φ_4 is differentiable at u, and*

$$\Phi_4'(u;v) = \int_\Omega (p^u + \epsilon u)v dx, \tag{4.26}$$

where p^u is defined in (4.24).

Acknowledgment. The author wishes to thank very much an anonymous referee for very useful remarks regarding the previous version of this paper.

REFERENCES

[1] V. BARBU, *Optimal Control of Variational Inequalities*, Research Notes in Mathematics 100, Pitman, London, 1984.
[2] V. BARBU AND S. STOJANOVIC, *Controlling the moving boundary of the parabolic obstacle problem*, Appl. Math. Optim., 27 (1993), pp. 213–230.
[3] L. A. CAFFARELLI, *Regularity of free boundaries in higher dimensions*, Acta Math., 139 (1977), pp. 155–184.
[4] L. A. CAFFARELLI, *Compactness methods in free boundary problems*, Comm. Partial Differential Equations, 5 (1980), pp. 427–448.
[5] L. A. CAFFARELLI, *A remark on the Hausdorff measure of a free boundary, and the convergence of coincidence sets*, Boll. Un. Mat. Ital. (5), 18-A (1981), pp. 109–113.
[6] L. C. EVANS AND R. F. GARIEPY, *Measure Theory and Fine Properties of Functions*, CRC Press, Boca Raton, FL, 1992.
[7] A. FRIEDMAN, *Partial Differential Equations of Parabolic Type*, Prentice-Hall, Englewood Cliffs, NJ, 1964.
[8] A. FRIEDMAN, *Variational Principles and Free-Boundary Problems*, John Wiley, New York, 1982.
[9] A. FRIEDMAN, *Optimal control for variational inequalities*, SIAM J. Control Optim., 24 (1986), pp. 439–451.
[10] F. MIGNOT, *Contrôle dans les inéquationes variationnelles elliptiques*, J. Funct. Anal., 22 (1976), pp. 130–185.
[11] F. MIGNOT AND J. P. PUEL, *Optimal control in some variational inequalities*, SIAM J. Control Optim., 22 (1984), pp. 466–476.
[12] J. P. PUEL, *Some results on optimal control for unilateral problems*, in Control of Partial Differential Equations, A. Bermudez, ed., Springer-Verlag, New York, 1989.
[13] D. G. SCHAEFFER, *A stability theorem for the obstacle problem*, Adv. Math., 16 (1975), pp. 34–47.
[14] S. STOJANOVIC, *Remarks on $W^{2,p}$-solutions of bilateral obstacle problems*, IMA Preprint 1318, University of Minnesota, Minneapolis, 1995.
[15] S. STOJANOVIC, *Optimal control of free boundaries*, in Control of Partial Differential Equations and Applications, E. Casas, ed., Marcel Dekker, New York, 1996, pp. 277–285.
[16] S. STOJANOVIC, *Modeling and minimization of extinction in Volterra-Lotka type equations with free boundaries*, J. Differential Equations, 134 (1977), pp. 320–422.
[17] J. SOKOLOWSKI AND J. P. ZOLESIO, *Introduction to Shape Optimization*, Springer-Verlag, Berlin, 1992.
[18] J. P. YVON, *Contrôle Optimal de Systèmes Gouvernés par des Inéquationes Variationnelles*, Thèse, Paris, 1973.

SIAM J. CONTROL OPTIM.
Vol. 35, No. 6, pp. 2101–2117, November 1997

012

APPROXIMATIONS IN DYNAMIC ZERO-SUM GAMES II*

MABEL M. TIDBALL†, ODILE POURTALLIER‡, AND EITAN ALTMAN‡

Abstract. We pursue in this paper our study of approximations of values and ϵ-saddle-point policies in dynamic zero-sum games. After extending the general theorem for approximation, we study zero-sum stochastic games with countable state space and unbounded immediate reward. We focus on the expected average payoff criterion. We use some tools developed in [M. M. Tidball and E. Altman, *SIAM J. Control Optim.*, 34 (1996), pp. 311–328] to obtain the convergence of the values as well as the convergence of the ϵ saddle-point policies in various approximation problems. We consider several schemes of truncation of the state space (e.g., finite state approximation) and approximations of games with discount factor close to one for the game with expected average cost. We use the extension of the general theorem for approximation to study approximations in stochastic games with complete information. Finally, we consider the problem of approximating the sets of policies. We obtain some general results that we apply to a pursuit evasion differential game.

Key words. zero-sum games, approximations, stochastic games, pursuit evasion games

AMS subject classifications. 90D05, 93E05

PII. S0363012994272460

1. Introduction.

We pursue in this paper our study of approximations of values and saddle-point policies in dynamic zero-sum games. In a previous paper [34], we developed some tools for approximating zero-sum games and applied them to stochastic games with discounted payoff criterion. In this paper we extend the general theory for approximation to handle cases where a value does not exist for the limit game, and we apply the general theorems for approximation to some dynamic zero-sum games.

We first consider approximation problems arising in stochastic games with expected average cost: finite state approximation of stochastic games with a countable state space, and convergence of stochastic games with discounted cost to the stochastic game with average cost. We then consider approximations in stochastic games with complete information, and problems in dynamic games related to discretization of the strategy sets.

There is a rich literature on finite state approximation in the context of a single controller. The discounted reward was extensively studied; see [2, 13, 21, 22, 35, 36] and [29, 37, 38] for related discretization results. For the expected average cost, there exist only few works on state approximations in the context of control, and none in the context of stochastic games. Even if existing schemes could be extended to the setting of a stochastic game, they are still quite restricted since their convergence (in the setting of control) was established under conditions that seem very strong, and are quite often nonapplicable. Thomas and Stengos [33] obtained several schemes for finite state approximations. They impose some scrambling conditions which should hold uniformly in the states. They do not seem to hold for queueing applications, such as the models in [3, 4, 7]. Altman introduced several finite state approximation schemes [1, 2] for constrained control. They do not require the scrambling conditions, but

*Received by the editors August 5, 1994; accepted for publication (in revised form) September 11, 1996.

http://www.siam.org/journals/sicon/35-6/27246.html

†Facultad de Ciencias Exactas, Ingeniera y Agrimensura, Universidad Nacional de Rosario, Pellegrini 250, 2000 Rosario, Argentina (mabel@unrctu.edu.ar).

‡INRIA, Centre Sophia-Antipolis, 2004 Route des Lucioles, B. P. 93, 06902 Sophia-Antipolis Cedex, France (pourtali@sophia.inria.fr, altman@sophia.inria.fr).

have other restrictive conditions: the scheme in [1] requires some monotone structure on the immediate cost, and holds for immediate costs that are only functions of the state, not of the actions. The scheme in [2] has the "finite neighbor" restriction; i.e., from each state, only finitely many states are accessible within one step.

The two approximation schemes that we introduce in the current paper relax the above restrictions and are thus also useful and new in the case of a single controller. In addition to the convergence of the value, which is the question studied in most of the papers on state approximations, we obtain (i) the convergence of the policies and (ii) the robustness of policies; i.e., an equilibrium point for the limiting (infinite state) stochastic game $G = G_\infty$ is shown to be an ϵ-equilibrium for the approximating games G_n for n large enough. On the other hand, for any ϵ, the equilibrium policies for G_n are almost optimal for the limiting game for all n large enough.

In the previous paper [34] we focused on approximations of stochastic games with discounted cost and bounded reward, and mentioned that standard techniques can be used to transform problems with unbounded reward to problems with bounded ones. This is, however, not the case for the expected average payoff criterion. The question of existence of value and of equilibrium stationary policies (under some recurrence conditions) for the case of unbounded reward was solved recently in [7, 8, 12, 30]. The growing interest in stochastic games with unbounded cost in recent years was partly driven by applications of stochastic games in telecommunications systems in general, and in queueing systems in particular. Although queues are always finite in practice (which results in a finite state space description), models of infinite queues are frequently more useful, since they are usually easier to solve. Indeed, several dynamic games arising in such applications were explicitly solved [7, 9] or at least reduced to the search for equilibrium policies among small classes of policies [4, 5, 7]. The scheduling problem described in [7, 9], the problem of routing into two queues [4, 7], and the flow and service control problem in [5] have not been solved for the case of finite state space, since there is an effect of the boundaries due to the finiteness of the queues that destroys the nice structure of the problem with infinite state space. In all the above problems, it is unnatural to consider bounded costs. Since costs represent queue lengths or waiting times, these typically grow to infinity as the number of "customers" in the queues grows to infinity. The theory developed in this paper allows the use of the equilibrium policies obtained for the infinite queues to construct ϵ-equilibrium policies for the corresponding problems with finite queues, provided that they are sufficiently large.

A second issue in this paper is the convergence of stochastic games in the discount factor. The convergence of the value and equilibrium policies for discounted cost stochastic games to those of the average cost game are well known; see, e.g., [18]. These were extended recently to unbounded cost (see [8, 30]). We obtain not only an alternative proof for the above convergence of the values and policies but also new robustness results.

When the players are restricted to using pure strategies in a stochastic game, the game in general does not have a value anymore. Using an extension of the general approximation theorems, we study approximations under that restriction. This yields approximation theorems for stochastic games with complete information (where player 2 knows at time t the action taken by player 1 at time t).

Finally, we consider the problem of approximating the set of policies by other sets. We obtain a general approximating theorem for the case when the strategy sets are endowed with the Hausdorff metric. We apply the theorem to a zero-sum pursuit evasion differential game introduced in [11, 31].

The structure of the paper is as follows. We begin in section 2 by citing and extending the general theory for approximations developed in [34]. We then introduce in section 3 the model, notation, and assumptions for the stochastic game. We present two schemes for state approximation in section 4. The convergence in the discount factor is established in section 5. In section 6 we discuss approximations for stochastic games with complete information. The approximation of the strategy sets is finally presented in section 7 together with the application to the pursuit evasion game.

2. Key theorems for approximations. We consider the following sequence $G_n = (S_n, U_n, V_n),\ n = 1, 2, ..., \infty$, of generic zero-sum games where U_n is the set of strategies (or policies) of player 1 and V_n is the set of strategies of player 2 for the nth game. We assume that both U_n and V_n are endowed with some topology. $S_n : U_n \times V_n \to \mathbb{R}$ is a measurable function for all n. We define the upper (lower) value of the game:

$$\overline{R_n} = \inf_{v\in V_n} \sup_{u\in U_n} S_n(u,v) \quad \left(\underline{R_n} = \sup_{u\in U_n} \inf_{v\in V_n} S_n(u,v)\right). \tag{2.1}$$

$G = (S, U, V) \stackrel{\text{def}}{=} (S_\infty, U_\infty, V_\infty)$ will be called the limit game. It will first be assumed (Theorem 2.1) that it has a value $R \stackrel{\text{def}}{=} R_\infty = \mathbf{Val}\{S(u,v)\}_{u,v}$. This assumption will be relaxed in Theorem 2.5.

A strategy $u^* \in U_n$ is said to be ϵ-optimal for player 1 in game n if

$$\inf_{v\in V_n} S_n(u^*, v) \geq \inf_{v\in V_n} S_n(u,v) - \epsilon \qquad \forall u \in U_n, \tag{2.2}$$

which is equivalent to $\inf_{v\in V_n} S_n(u^*, v) \geq \underline{R}_n - \epsilon$. It is said to be strong ϵ-optimal for player 1 in game n if it satisfies

$$\inf_{v\in V_n} S_n(u^*, v) \geq \overline{R}_n - \epsilon.$$

A strategy $v^* \in V_n$ is said to be ϵ-optimal for player 2 in game n if

$$\sup_{u\in U_n} S_n(u, v^*) \leq \sup_{u\in U_n} S_n(u,v) + \epsilon \qquad \forall v \in V_n, \tag{2.3}$$

which is equivalent to $\sup_{u\in U_n} S_n(u, v^*) \leq \overline{R}_n + \epsilon$. It is said to be strong ϵ-optimal if

$$\sup_{u\in U_n} S_n(u, v^*) \leq \underline{R}_n + \epsilon.$$

Note that strong ϵ-optimality implies ϵ-optimality. If a game has a value $\underline{R}_n = \overline{R}_n$, then strong ϵ-optimality is equivalent to ϵ-optimality.

Assume that (S_n, U_n, V_n) converge (in some sense) to (S, U, V). We are interested in the following questions:

(Q1) Convergence of the values: does $\underline{R}_n$ (or $\overline{R}_n$) converge to R?

(Q2) Convergence of policies: fix some $\epsilon \geq 0$. Let ϵ_n be a sequence of positive real numbers such that $\overline{\lim}_{n\to\infty} \epsilon_n \leq \epsilon$. Assume that u_n^* and v_n^* are ϵ_n-optimal policies for the nth game. Are u_n^* and v_n^* "almost" optimal for the limit game, for all n large enough?

(Q3) Let $\overline{u} \in U$ (resp., $\overline{v} \in V$) be some limit point of u_n^* (resp., v_n^*), defined above. Is $\overline{u}$ (resp., $\overline{v}$) ϵ-optimal for the limit game?

(Q4) Robustness of the optimal policy: if u^* (resp., v^*) is an ϵ-optimal policy for the limit game, can we derive from it an "almost" (strong) optimal policy for the nth approximating game for all n large enough?

A straightforward generalization of Theorem 2.1 in [34] yields our Theorem 2.1.

THEOREM 2.1. *Assume that for any $\epsilon_1 > 0$ there exists a sequence of functions, $\pi_n^1 : U_n \to U$, $\pi_n^2 : V_n \to V$, $\sigma_n^1 : U \to U_n$, $\sigma_n^2 : V \to V_n$, $n = 1, 2, ...$, such that*

($\mathcal{A}1$) $\overline{\lim}_{n\to\infty}[S_n(u, \sigma_n^2(v)) - S(\pi_n^1(u), v)] \le \epsilon_1$ *uniformly in $u \in U_n$ for each $v \in V$.*

($\mathcal{A}2$) $\underline{\lim}_{n\to\infty}[S_n(\sigma_n^1(u), v) - S(u, \pi_n^2(v))] \ge -\epsilon_1$ *uniformly in $v \in V_n$ for each $u \in U$.*

Then

(1) $\lim_{n\to\infty} \underline{R}_n = \lim_{n\to\infty} \overline{R}_n = R$.

(2) *For any $\epsilon' > \epsilon + 3\epsilon_1$, there exists N such that $\pi_n^1(u_n^*)$ (resp., $\pi_n^2(v_n^*)$, see definitions in* (Q2)) *is ϵ'-optimal for the limit game for all $n \ge N$.*

(3) *Let u^* (resp., v^*) be ϵ-optimal for the limit game. Then for all $\epsilon' > \epsilon + 3\epsilon_1$, there exists $N(\epsilon')$ such that $\sigma_n^1(u^*)$ (resp., $\sigma_n^2(v^*)$) is strong ϵ'-optimal for the nth approximating game for all $n \ge N(\epsilon')$.*

(4) *Suppose that*

($\mathcal{A}3$) *$S(u, v)$ is a lower semicontinuous function in u for each v.*

($\mathcal{A}4$) *$S(u, v)$ is an upper semicontinuous function in v for each u.*

Suppose that $\bar{u} \in U$ (resp., $\bar{v} \in V$) is a limit point of $\pi_n^1(u_n^*)$ (resp., $\pi_n^2(v_n^*)$). Then $\bar{u}$ (resp., $\bar{v}$) is $(\epsilon + 5\epsilon_1)$-optimal for the limit game.

Remark 2.2. (i) Whenever $U_n = U$ and $V_n = V$ do not depend on n, π_n and σ_n will be chosen as the identity maps.

(ii) It follows from the proof of part (1) in the above theorem that if for every G_n, $n = 1, 2, ..., \infty$, there exist optimal policies for both players, and if $U_n = U$ and $V_n = V$ do not depend on n, then

$$|\overline{R}_n - R| \le \sup_{u,v} |S_n(u, v) - S(u, v)|, \qquad |\underline{R}_n - R| \le \sup_{u,v} |S_n(u, v) - S(u, v)|.$$

Remark 2.3. Some results related to those in Theorem 2.1 can be found in [10, 17, 20, 26] and in [27], whose authors also consider additional constraints on the policies studied there.

We now relax the assumption that the limit game has a value $\underline{R}_\infty \ne \overline{R}_\infty$. We show that Theorem 2.1 still holds, by appropriately enlarging the policy spaces and redefining the cost, so that the upper (or lower) value becomes a real value of a new game.

We consider the convergence of the upper values (and corresponding optimal or almost optimal policies) of the approximating games to those of the limit game. The corresponding convergence for the lower values are obtained in a similar way. Define $\mathcal{U}_n = \{$the class of functions $V_n \to U_n\}$. Define the cost $\hat{\mathcal{S}}_n : \mathcal{U}_n \times V_n \to \mathbb{R}$ by $\hat{\mathcal{S}}_n(\psi, v) = S_n(\psi(v), v)$.

LEMMA 2.4. (i) *For all n, the new game $\mathcal{G}_n = (\hat{\mathcal{S}}_n, \mathcal{U}_n, V_n)$ has a value $\mathcal{R}_n$, and $\mathcal{R}_n = \overline{R}_n$.*

(ii) *v^* is ϵ-optimal for player 2 in game $\mathcal{G}_n$ if and only if it is ϵ-optimal in game G_n.*

Proof.

$$\inf_{v\in V_n} \sup_{\psi\in\mathcal{U}_n} \hat{\mathcal{S}}_n(\psi, v) = \inf_{v\in V_n} \sup_{\psi\in\mathcal{U}_n} S_n(\psi(v), v) = \inf_{v\in V_n} \sup_{u\in U_n} S_n(u, v) = \overline{R}_n. \tag{2.4}$$

On the other hand,

$$\inf_{v\in V_n}\sup_{\psi\in\mathcal{U}_n} S_n(\psi(v),v)=\sup_{\psi\in\mathcal{U}_n}\inf_{v\in V_n} S_n(\psi(v),v)=\sup_{\psi\in\mathcal{U}_n}\inf_{v\in V_n}\hat{\mathcal{S}}_n(\psi,v). \tag{2.5}$$

The first equality in (2.5) is due to the following. Clearly,

$$\inf_{v\in V_n}\sup_{\psi\in\mathcal{U}_n} S_n(\psi(v),v)\geq\sup_{\psi\in\mathcal{U}_n}\inf_{v\in V_n} S_n(\psi(v),v),$$

and we have to show that the reverse inequality also holds. Fix some $\epsilon'>0$ and let ψ^* be such that for all v and all $\psi\in\mathcal{U}_n$,

$$S_n(\psi^*(v),v)\geq S_n(\psi(v),v)-\epsilon'.$$

Hence,

$$\inf_{v\in V_n} S_n(\psi^*(v),v)\geq\inf_{v\in V_n}\sup_{\psi\in\mathcal{U}_n} S_n(\psi(v),v)-\epsilon'.$$

We conclude that

$$\sup_{\psi\in\mathcal{U}_n}\inf_{v\in V_n} S_n(\psi(v),v)\geq\inf_{v\in V_n}\sup_{\psi\in\mathcal{U}_n} S_n(\psi(v),v)-\epsilon',$$

which establishes the first equality in (2.5). (i) is obtained by combining (2.4) and (2.5). (ii) follows, since for any $v\in V_n$,

$$\sup_{u\in U_n} S_n(u,v)=\sup_{\psi\in\mathcal{U}_n}\hat{\mathcal{S}}_n(\psi,v). \qquad \square$$

By using the new games for which the values exist, and applying Theorem 2.1, we may conclude the following convergence properties of the original games.

THEOREM 2.5. *Assume that the functions π_n and σ_n exist as in Theorem* 2.1*, and that conditions* ($\mathcal{A}$1) *and* ($\mathcal{A}$2) *hold. Then*

(1) $\lim_{n\to\infty}\overline{R}_n=\overline{R}$, $\lim_{n\to\infty}\underline{R}_n=\underline{R}$.

(2) *For any $\epsilon'>\epsilon+3\epsilon_1$, there exists N such that $\pi_n^1(u_n^*)$ (resp., $\pi_n^2(v_n^*)$; see definitions in* (Q2)*) is ϵ'-optimal for the limit game for all $n\geq N$.*

(3) *Let u^* (resp., v^*) be ϵ-optimal for the limit game. Then for all $\epsilon'>\epsilon+3\epsilon_1$, there exists $N(\epsilon')$ such that $\sigma_n^1(u^*)$ (resp., $\sigma_n^2(v^*)$) is ϵ'-optimal for the n approximating game for all $n\geq N(\epsilon')$.*

Proof. Consider the new games $\mathcal{G}_n$ defined above. We show that the assumptions of Theorem 2.1 hold also for $\mathcal{G}_n$. The mapping $\tilde{\pi}_n^2,\tilde{\sigma}_n^2$ for the new games are unchanged:

$$\tilde{\pi}_n^2=\pi_n^2,\qquad \tilde{\sigma}_n^2=\sigma_n^2.$$

The mappings $\tilde{\pi}_n^1:\mathcal{U}_n\to\mathcal{U}$ and $\tilde{\sigma}_n^1:\mathcal{U}\to\mathcal{U}_n$ for the new games are defined as

$$[\tilde{\pi}_n^1(\psi)](v)=\pi_n^1(\psi(v))\qquad\forall v\in V,\qquad [\tilde{\sigma}_n^1(\psi)](v)=\sigma_n^1(\psi(v))\qquad\forall v\in V_n.$$

With these definitions, as well as the definition of the costs $\hat{\mathcal{S}}_n$, it follows that ($\mathcal{A}$1) and ($\mathcal{A}$2) hold for $\mathcal{G}_n$. The proof now follows by Lemma 2.4. $\square$

We may further obtain convergence results for the optimal (or ϵ-optimal) *responses* (in case the value of the limit game does not exist). To simplify the formulation, this is done below in terms of the new games $\mathcal{G}_n$.

THEOREM 2.6. *Consider the new games $\mathcal{G}_n$, and let ψ_n^*, v_n^* be defined as u_n^* in* (Q2) *(above Theorem* 2.1*). Under the conditions of Theorem* 2.5*,*

(1) $\lim_{n\to\infty} \mathcal{R}_n = \mathcal{R} = \overline{R}$.

(2) *For any $\epsilon' > \epsilon + 3\epsilon_1$, there exists N such that $\tilde{\pi}_n^1(\psi_n^*)$ (resp., $\tilde{\pi}_n^2(v_n^*)$) is ϵ'-optimal for $\mathcal{G}_\infty$ for all $n \geq N$.*

(3) *Let ψ^* be ϵ-optimal for player* 1 *in the limit game $\mathcal{G}_\infty$. Then, for all $\epsilon' > \epsilon + 3\epsilon_1$, there exists $N(\epsilon')$ such that $\tilde{\sigma}_n^1(\psi^*)$ (resp., $\tilde{\sigma}_n^2(v^*)$) is ϵ'-optimal for the n approximating game $\mathcal{G}_n$ for all $n \geq N(\epsilon')$.*

Next, we consider the result corresponding to statement (4) in Theorem 2.1.

THEOREM 2.7. *Assume that the conditions of Theorem* 2.6 *hold, that the set of response strategies for player* 2 *in games $\mathcal{G}_n$ is endowed with some topology, and that* ($\mathcal{A}$3) *and* ($\mathcal{A}$4) *hold for game $\mathcal{G}_\infty$. Then statement* (4) *of Theorem* 2.1 *holds for games $\mathcal{G}_n$.*

3. Stochastic games with expected average payoff. We consider the two-person, zero-sum stochastic game defined by the objects $\{\mathbf{I}, \mathbf{A}, \mathbf{B}, P, r\}$, where

- $\mathbf{I}$ is a countable state space;
- $\mathbf{A}$ and $\mathbf{B}$ are sets of actions for player 1 and player 2, respectively; at each state $j \in \mathbf{I}$, the available actions for the players are $\mathbf{A}_j$ and $\mathbf{B}_j$, respectively. These sets are assumed to be compact metric sets.
- $P(a,b) = [p(i,a,b,j)]_{i,j}$, $a \in \mathbf{A}$, $b \in \mathbf{B}$, $i,j \in \mathbf{I}$, are the transition probabilities, so that $p(i,a,b,j)$ is the probability of moving from i to j if the players use actions a and b.
- $r : \mathbf{I} \times \mathbf{A} \times \mathbf{B} \to \mathbb{R}$ is an immediate reward function.

The game is played in stages $t = 0, 1, 2, \ldots$. If at some stage t, the state is i, then the players independently choose actions $a \in \mathbf{A}_i$, $b \in \mathbf{B}_i$. Player 2 then pays player 1 the amount $r(i,a,b)$, and at stage $t+1$, the new state is chosen according to the transition probabilities $p(i,a,b,\bullet)$. The game continues at this new state.

Let U and V be the set of behavioral strategies for both players. A strategy $u \in U$ is a sequence $u = (u_0, u_1, \ldots)$, where u_t is a probability measure over the available actions, given the whole history of previous states and of previous actions of both players as well as the current state.

A Markov policy $q = \{q_0, q_1, \ldots\}$ is a policy (for either player 1 or 2) where q_t is allowed to depend only on t and on the state at time t.

A *stationary (mixed) policy* g for player 1 is characterized by a conditional distribution $p^g(\bullet \mid j)$ over $\mathbf{A}_j$, so that $p^g(\mathbf{A}_j \mid j) = 1$, which is interpreted as the distribution over the actions available at state j which player 1 uses when it is in state j. With some abuse of notation, we shall set $g(\bullet \mid j) = p^g(\bullet \mid j)$ for stationary g. Let U_S be the set of stationary policies for player 1, and define similarly the stationary policies V_S for player 2. If both players use stationary policies, say u and v, then $\{X_t\}$ becomes a Markov chain with stationary transition probabilities, given by

$$p(j,u,v,k) := \int_{\mathbf{A}_j} \int_{\mathbf{B}_j} p(j,a,b,k) u(da|j) v(db|j). \tag{3.1}$$

The expected immediate reward at state j becomes

$$r(j,u,v) := \int_{\mathbf{A}_j} \int_{\mathbf{B}_j} r(j,a,b) u(da|j) v(db|j).$$

Denote by $P(u,v)$ the (infinite-dimensional) matrix whose (j,k)th component equals $p(j,u,v,k)$. Similarly, denote by $r(u,v)$ the column vector whose jth component equals $r(j,u,v)$.

Next, we introduce a topology on the sets of stationary policies. For any compact metric set Γ, let $M_1(\Gamma)$ denote the set of probability measures on the Borel subsets of Γ endowed with the weak topology $\xi(\Gamma)$ (see [28]). The class of stationary policies for player 1 (and similarly for player 2) can be identified with the set $\prod_{i\in\mathbf{I}} M_1(\mathbf{A}_i) \times M_1(\mathbf{B}_i)$; moreover, it is compact with respect to the product topology $\prod_{i\in\mathbf{I}}\xi(\mathbf{A}_i) \times \xi(\mathbf{B}_i)$, and it is metrizable (by virtue of Theorem 4.14 in Kelley [23]).

Let (u, v) be a pair of strategies and let $i \in \mathbf{I}$ be a fixed initial state. Let $I_t, A_t, B_t, t = 0, ...,$ be the resulting stochastic process of the states and actions of the players. Let $E_i^{u,v}$ denote the expectation with respect to the measure defined by u, v, i.

Let $\mu : \mathbf{I} \to \mathbb{R}$ be some positive function. Following Dekker and Hordijk [16] and Spieksma [32], define the μ-norm of any vector $x \in \mathbb{R}^{\mathbf{I}}$ as

$$\| x \|_\mu = \sup_{i\in\mathbf{I}} \frac{|x_i|}{\mu_i}.$$

In a similar way, we will use the μ-J-norm, for any finite subset J of the state space I, defined by

$$\| x \|_\mu^J = \sup_{i\in J} \frac{|x_i|}{\mu_i}.$$

Define the μ-norm of matrices $Q \in \mathbb{R}^{\mathbf{I}\times\mathbf{I}}$ as

$$\|Q\|_\mu = \sup_{i\in\mathbf{I}} \mu_i^{-1} \sum_{j\in\mathbf{I}} |Q_{ij}|\mu_j.$$

We denote by $\bigvee_\mu = \{x : \| x \|_\mu < \infty\}$ the space of all vectors that are μ-bounded.

We introduce the following assumptions:

- ($\mathcal{B}1$)
  - i) The instantaneous reward $r(i, a, b)$ is continuous and μ-bounded, i.e.,

$$\sup_{i\in\mathbf{I}} \sup_{a,b} \frac{|r(i, a, b)|}{\mu_i} \le M < +\infty.$$

    This condition can be rewritten as $\|r(\cdot, u, v)\|_\mu \le M < +\infty$ for all pure stationary policies u and v.
  - ii) The transition probabilities are μ-continuous; i.e., if $a(n) \to a$, $b(n) \to b$ when $n \to +\infty$, then

$$\lim_{n\to\infty} \sum_{j\in\mathbf{I}} |p(i, a(n), b(n), j) - p(i, a, b, j)|\, \mu_j = 0 \qquad \forall i \in I.$$

- ($\mathcal{B}2$)
  - i) Under any pure stationary policies for the players, the state space does not contain more than one ergodic class.
  - ii) There exists a finite set $\mathcal{M} \subset E$ and a constant $\beta < 1$ such that

$$\sum_{j\in I} {}_{\mathcal{M}}p(i, a, b, j)\, \mu_j \le \beta\, \mu_i \qquad \forall a, b,\ \forall i \in \mathbf{I}, \tag{3.2}$$

    where ${}_{\mathcal{M}}p(i, a, b, j) = p(i, a, b, j)$ if j does not belong to the set $\mathcal{M}$, and is null otherwise. (3.2) can be rewritten as $\sum_{j\notin\mathcal{M}} p(i, a, b, j)\, \mu_j \le \beta\, \mu_i \ \ \forall a, b, \forall i \in \mathbf{I}$, or as $\|{}_{\mathcal{M}}P(u, v)\|_\mu \le \beta$ for all pure stationary policies u and v.

Remark 3.1. If assumptions ($\mathcal{B}1$) and ($\mathcal{B}2$) hold for some $\mathcal{M}$ and μ, then one can choose a state $0 \in \mathcal{M}$ and another μ' such that these assumptions hold with the set $\mathcal{M}' = \{0\}$ replacing $\mathcal{M}$, and μ' replacing μ (see [32]). Therefore, we assume in the sequel, without loss of generality, that $\mathcal{M} = \{0\}$ for some state 0.

Define

$$S(i,u,v) = \overline{\lim_{t\to\infty}} \frac{1}{t} E_i^{u,v} \sum_{s=0}^{t-1} r(I_s, A_s, B_s). \tag{3.3}$$

THEOREM 3.2 (see [8]). *Suppose that assumptions* ($\mathcal{B}1$) *and* ($\mathcal{B}2$) *hold. Then*

(i) *The stochastic game with expected average payoff criterion has a value.*

(ii) *There exists a unique solution pair* (g, v), $g \in \mathbb{R}$, $v \in R^{\mathbf{I}}$, *to the functional equation*

$$v(i) = \mathbf{Val}\left\{ r(i,a,b) - g + \sum_{j\in\mathbf{I}} p(i,a,b,j) v(j) \right\}_{a,b}, \qquad i \in \mathbf{I}, \tag{3.4}$$

such that $\|v\|_\mu$ *is finite and* $v(0) = 0$.

(iii) $g = R(i) = \mathbf{Val}\{S(i,u,v)\}_{u,v}$ *is the unique value of the stochastic game.*

(iv) *Let* (u, v) *be stationary policies such that* $u(i), v(i)$ *are optimal for the dummy game in the curly brackets in* (3.4). *Then they are optimal for the stochastic game.*

4. State truncation and approximation. In the following approximating schemes, we modify the "limit" stochastic game in the following way. We consider an increasing sequence of sets of states $\mathbf{I}_1, \mathbf{I}_2, \ldots$ converging to $\mathbf{I}$, such that $0 \in \mathbf{I}_1$. The nth stochastic game is restricted to the set $\mathbf{I}_n$. In the game G_n, we modify the transition probabilities so as to eliminate all transitions outside the set $\mathbf{I}_n$. The two schemes will differ by the way that such transitions will be replaced. Introduce the following assumption:

- ($\mathcal{B}3$) $\quad \delta(r,n) = \sup_{\substack{i\in\mathbf{I}_r \\ a\in\mathbf{A}, b\in\mathbf{B}}} \sum_{j\notin\mathbf{I}_n} p(i,a,b,j)\mu(j) \to 0 \text{ as } n \to +\infty \quad \forall r.$

Under the assumptions of our model (i.e., ($\mathcal{B}1$)–($\mathcal{B}2$)), ($\mathcal{B}3$) holds if $\mathbf{I}_n$ are finite sets $\forall n$ (the proof is similar to the one in [34]).

4.1. Scheme I. In the game G_n, we modify the transition probabilities so as to eliminate all transitions outside the set $\mathbf{I}_n$; we replace transitions outside of $\mathbf{I}_n$ by transitions to state 0. Hence, $p^n(i,a,b,j)$ is defined by

$$p^n(i,a,b,j) = \begin{cases} p(i,a,b,0) + \sum_{l\notin\mathbf{I}_n} p(i,a,b,l), & j = 0, i \in \mathbf{I}_n, \\ p(i,a,b,j), & j \neq 0, i,j \in \mathbf{I}_n, \\ 0, & j \notin \mathbf{I}_n, \\ 1\{j=0\}, & i \notin \mathbf{I}_n. \end{cases} \tag{4.1}$$

For $i \in \mathbf{I}_n$, $S_n(i,u,v)$ is defined as (3.3), where the expectation is taken with respect to the signed measure generated by the new transition probabilities (4.1). $S_n(i,u,v)$ is defined to be 0 for $i \notin \mathbf{I}_n$, for all u and v.

For $n \in \mathbb{N}$, let the pair $(g_n, v_n \in \bigvee_\mu)$ be the solutions of the dynamic programming equation

$$\begin{array}{ll} v_n(i) = \mathbf{Val}\left\{r(i,a,b) - g_n + \sum_{j\in\mathbf{I}} p^n(i,a,b,j)v_n(j)\right\}_{a,b}, & i \in \mathbf{I}_n, \quad i \neq 0, \\ v_n(i) = 0, & i \notin \mathbf{I}_n \text{ or } i = 0. \end{array} \tag{4.2}$$

One can show using Theorem 3.2 that for all n, g_n and v_n indeed exist and are unique. Moreover, we have that

$$R_n(i) = \mathbf{Val}\left\{S_n(i,u,v)\right\}_{u,v} = g_n \qquad \forall i \in \mathbf{I}.$$

In order to prove the convergence of the state approximation scheme we introduce the following quantities:

- $\tau := \inf\{t \geq 1, I_t = 0\}$ is the time to reach state zero (with the convention that $\inf\{t : t \in \emptyset\} = \infty$).
- $w(i,u,v)$:= the total cost to reach zero from state i when policies u and v are used:

$$w(i,u,v) = E_i^{u,v} \sum_{s=0}^{\tau} r(I_s, A_s, B_s),$$

  which can be rewritten in vector form as

$$w(u,v) = \sum_{s=0}^{\infty} [{}_0P(u,v)]^s r(u,v).$$

- $w^n(i,u,v)$:= the total cost to reach zero from state i when policies u and v are used, when the transition probabilities are replaced by (4.1).
- $\overline{\tau}(i,u,v), \overline{\tau}^n(i,u,v)$:= the expectations of τ when using the original transition probabilities, and when using those in (4.1), respectively. For $i \notin \mathbf{I}_n$ we have $\overline{\tau}^n(i,u,v) = 1$ for all policies u and v.

We note that $w(\cdot,u,v)$, $w^n(\cdot,u,v)$, $\overline{\tau}(\cdot,u,v)$, and $\overline{\tau}^n(\cdot,u,v)$ are uniformly μ bounded. Indeed,

$$\left\|w(u,v)\right\|_\mu \leq \sum_{s=0}^{\infty} \left[\left\|{}_0P(u,v)\right\|_\mu\right]^s \left\|r(u,v)\right\|_\mu \leq \frac{M}{1-\beta} \tag{4.3}$$

with the same bound for $w^n(u,v)$. Similarly, $\left\|\overline{\tau}(u,v)\right\|_\mu$ and $\left\|\overline{\tau}^n(u,v)\right\|_\mu$ are bounded by $(1-\beta)^{-1}$. It is easily seen that $w(\cdot,u,v)$, $w^n(\cdot,u,v)$, $\overline{\tau}(\cdot,u,v)$, and $\overline{\tau}^n(\cdot,u,v)$ are the unique solutions in $\bigvee_\mu$ of the fixed point equations

$$w(i,u,v) = r(i,u,v) + \sum_{j\neq 0} p(i,u,v,j)w(j,u,v), \tag{4.4}$$

$$w^n(i,u,v) = \begin{cases} r(i,u,v) + \sum_{j\neq 0} p^n(i,u,v,j)w^n(j,u,v) \text{ for } i \in I_n, \\ 0 \text{ for } i \notin I_n, \end{cases} \tag{4.5}$$

$$\overline{\tau}(i,u,v) = 1 + \sum_{j\neq 0} p(i,u,v,j)\overline{\tau}(j,u,v), \tag{4.6}$$

$$\overline{\tau}^n(i,u,v) = \begin{cases} 1 + \sum_{j\neq 0} p^n(i,u,v,j)\overline{\tau}^n(j,u,v) \text{ for } i \in I_n, \\ 1 \text{ for } i \notin I_n. \end{cases} \tag{4.7}$$

The uniqueness follows from the fact that the above equations are contracting due to ($\mathcal{B}2$). Note that functions $w(.,u,v)$ and $w^n(.,u,v)$ are μ-bounded for all pairs (u,v) on every subset J of $\mathbf{I}$. Since both $\overline{\tau}^n(0,u,v)$ and $\overline{\tau}(0,u,v)$ are both nonzero and finite, it follows (see Chung [14, pp. 91–92]) that the expected average cost is given by the following ratio between the total cost and the expected hitting time of state zero:

$$S(i,u,v)=\frac{w(0,u,v)}{E\tau(0,u,v)} \quad \text{and} \quad S_n(i,u,v)=\frac{w^n(0,u,v)}{E\tau^n(0,u,v)}. \tag{4.8}$$

THEOREM 4.1. *Assume* ($\mathcal{B}1$)–($\mathcal{B}3$). *All statements of Theorem* 2.1 *hold, where* S_n *and* S *are the expected average payoffs defined in* (3.3)*, with the transition probabilities* p *and* p^n (*defined in* (4.1))*, respectively.*

Proof. Fix some initial state i. We use Theorem 2.1. We begin by establishing conditions ($\mathcal{A}1$) and ($\mathcal{A}2$). Since $U=U_n$ and $V_n=V$ for each n, it suffices to show that $S_n(u,v):=S_n(i,u,v)$ converges to $S(u,v):=S(i,u,v)$ uniformly on $\mathbf{I}$. Hence, we set π_n^1, π_n^2, σ_n^1, and σ_n^2 to be identical.

Let J be a given subset of $\mathbf{I}$, and (u,v) a pair of strategies. To avoid cumbersome notations we will write $w(.)$ (resp., $w^n(.)$) instead of $w(.,u,v)$ (resp., $w^n(.,u,v)$). We first want to prove that

$$\lim_{n\to+\infty} \parallel w^n-w \parallel_\mu^J = 0.$$

Once we show that, one obtains in the same way that $\lim_{n\to+\infty} \parallel E\tau-E\tau^n \parallel_\mu^J = 0$, and the uniform convergence of $S_n(u,v)$ to $S(u,v)$ now follows from (4.8).

We use an idea introduced by Cavazos-Cadena [13] and used in [34] for a similar problem. Fix ϵ arbitrarily small and define the sequence g_k in the following way. $g_0=\min\{m : J\subset \mathbf{I}_m\}$ and, recursively,

$$g_k=g(\epsilon,g_{k-1}), \qquad g(\epsilon,r)=\min\{m:\delta(r,m)\le\epsilon\},$$

where δ is defined in ($\mathcal{B}3$). Due to assumption ($\mathcal{B}3$), this sequence is well defined, and for all k, g_k is finite. Let ν be a given integer; define also

$$m^\nu(\epsilon)=\max\{g_m, m=0,1,...,\nu\}.$$

Let $n\ge m^\nu(\epsilon)$, $i\in J$. Let us now compute $\parallel w^n-w\parallel_\mu^J$. We obviously have that $\parallel w^n-w\parallel_\mu^J \le \parallel w^n-w\parallel_\mu^{I_{g_0}}$, since $J\subset I_{g_0}$, and for $i\in\mathbf{I}_{g_0}$,

$$\begin{aligned}
\frac{1}{\mu_i}|w^n(i)-w(i)| &= \frac{1}{\mu_i}\left|\sum_{j\neq 0} p^n(i,u,v,j)w^n(j)-p(i,u,v,j)w(j)\right| \\
&\le \frac{1}{\mu_i}\sum_{j\in\mathbf{I}_{g_1}\setminus\{0\}} |p^n(i,u,v,j)w^n(j)-p(i,u,v,j)w(j)| \\
&\quad +\frac{1}{\mu_i}\sum_{j\in\mathbf{I}_n\setminus\mathbf{I}_{g_1}} |p^n(i,u,v,j)w^n(j)-p(i,u,v,j)w(j)| \\
&\le \sum_{j\in\mathbf{I}_{g_1}\setminus\{0\}} \frac{p(i,u,v,j)\mu_j}{\mu_i}\frac{|w^n(j)-w(j)|}{\mu_j} \\
&\quad +\frac{1}{\mu_i}\sum_{j\in\mathbf{I}_n\setminus\mathbf{I}_{g_1}} p(i,u,v,j)|w^n(j)-w(j)|.
\end{aligned}$$

In the last inequality the first term can be bounded by $\beta \parallel w - w^n \parallel_\mu^{\mathbf{I}_{g_1}}$ because of assumption ($\mathcal{B}2$(ii)), and the second by $2\frac{M}{1-\beta}\epsilon$, since

$$\sum_{j\in\mathbf{I}_n\setminus\mathbf{I}_{g_1}} p(i,u,v,j)|w^n(j)-w(j)| \le \sum_{j\in\mathbf{I}_n\setminus\mathbf{I}_{g_1}} \mu_j p(i,u,v,j)\left[\frac{|w^n(j)|}{\mu_j}+\frac{|w(j)|}{\mu_j}\right].$$

This is due to the definition of the sequence I_{g_k}, since i belongs to I_{g_0} due to ($\mathcal{B}3$), and since w^n and w are bounded by $\frac{M}{1-\beta}$; see (4.3). We obtain

$$\parallel w^n - w \parallel_\mu^{\mathbf{I}_{g_0}} \;\le\; \beta \parallel w^n - w \parallel_\mu^{\mathbf{I}_{g_1}} + 2\frac{M}{1-\beta}\epsilon.$$

In exactly the same way we get for $k \le m^\nu(\epsilon)$,

$$\parallel w^n - w \parallel_\mu^{\mathbf{I}_{g_k}} \;\le\; \beta \parallel w^n - w \parallel_\mu^{\mathbf{I}_{g_{k+1}}} + 2\frac{M}{1-\beta}\epsilon,$$

and finally,

$$\parallel w^n - w \parallel_\mu^{J} \;\le\; \beta^\nu \frac{2M}{1-\beta} + \frac{2M\epsilon}{1-\beta}\left(\frac{1-\beta^\nu}{1-\beta}\right). \tag{4.9}$$

Since ν can be chosen arbitrarily large when n tends to infinity, and β is strictly lower than 1, this bound can be as small as needed for n large enough. This establishes ($\mathcal{A}1$)–($\mathcal{A}2$) in Theorem 2.1. It follows from [8] that S is a continuous function of u and v, which implies ($\mathcal{A}3$)–($\mathcal{A}4$) of Theorem 2.1. This completes the proof. □

4.2. Scheme II. In the previous scheme, we replaced transitions outside of $\mathbf{I}_n$ by transitions to state 0. In some applications this may be undesirable; this is the case when the games with truncated space describe real problems that we wish to approximate by some game with an infinite state space. To illustrate this, consider a queue with a finite length L, and assume that the state is the number of customers in the queue. Then, typically, if a transition from state L to state $L+1$ were possible in the case of an infinite queue, then in the problem with truncated state space, which corresponds to a finite queue, it is replaced by a transition from L to L. In the previous scheme, it would be replaced by a transition to state 0. This would be especially undesirable, since in queueing problems, we usually have the property of transitions to closest neighbors: from each state, only finitely many neighboring states can be reached in one step. So, having a transition from state L to 0 does not describe a realistic model of a finite queue.

Let $\{q^n(i,a,b,j), i,j \in \mathbf{I}, a \in \mathbf{A}, b \in \mathbf{B}\}$ be sequences of measures such that for all n, $i \in \mathbf{I}$, $a \in \mathbf{A}, b \in \mathbf{B}$,

$$q^n(i,a,b,j) \le 0 \text{ for } j \in \mathbf{I}_n, \qquad q^n(i,a,b,j) = 0 \text{ for } j \notin \mathbf{I}_n,$$

$$\sum_{j\in\mathbf{I}_n} (p(i,a,b,j) + q^n(i,a,b,j)) = 1.$$

The transitions for the approximating problems are then given by

$$p^n(i,a,b,j) = \begin{cases} p(i,a,b,j) + q^n(i,a,b,j), & i,j \in \mathbf{I}_n, \\ 0 & \text{otherwise.} \end{cases} \tag{4.10}$$

It follows that

$$\sum_{j\in\mathbf{I}_n} q^n(i,a,b,j) = \sum_{j\notin\mathbf{I}_n} p(i,a,b,j). \tag{4.11}$$

We make the following assumption on μ and on $\mathbf{I}_n$:

$$\text{for any } n>m \text{ and } i\in\mathbf{I}_n\setminus\mathbf{I}_m,\ \mu(i)\geq \sup_{j\in\mathbf{I}_m}\mu(j).$$

For $i\in\mathbf{I}_n$, $S_n(i,u,v)$ is defined as (3.3), where the expectation is taken with respect to the measure generated by the new transition probabilities (4.10). We set $S_n(i,u,v)=0$ for all $i\notin\mathbf{I}_n$.

THEOREM 4.2. *Assume* ($\mathcal{B}1$)–($\mathcal{B}3$)*, and consider the above finite approximation scheme. Then all statements of Theorem* 2.1 *hold.*

Proof. We consider, as in the previous section, the total expected cost and total expected time between consecutive epochs until state 0 is reached. By similar arguments to those in the previous scheme, one then establishes that $S_n(u,v)$ converge to $S(u,v)$ uniformly in all stationary policies (the exact proof can be found in [6, Chapter 8]). This implies assumptions ($\mathcal{A}1$) and ($\mathcal{A}2$). Assumptions ($\mathcal{A}3$) and ($\mathcal{A}4$) relate only to the limit game, and therefore the proof is the same as in the previous section. The theorem now follows from Theorem 2.1. □

5. Convergence of the discounted cost to the average cost. Conditions for the convergence of the value and equilibrium policies for discounted cost stochastic games to those of the average cost game are well known; see, e.g., [18]. These were extended recently to unbounded cost (see [8, 30]). Theorem 2.1 enables us not only to obtain an alternative proof for the above convergence of the values and policies but also to obtain new robustness results, as in Theorem 5.2 below.

Define the *β-discounted game payoff*

$$S_\beta(i,u,v) = (1-\beta)E_i^{u,v}\sum_{t=0}^{\infty}\beta^t r(I_t,A_t,B_t). \tag{5.1}$$

The following was proved in [8, Theorem 3.4], (see also [15]).

THEOREM 5.1. *Assume* ($\mathcal{B}1$) *and* ($\mathcal{B}2$)*. Then*

(1) *A value $R_\beta(i)$ exists for the discounted cost.*

(2) *Optimal stationary policies exist for both players for any discount factor $0<\beta<1$ (they are said to be β-optimal).*

(3) *Any limit-point (as β tends to 1) of β-optimal stationary policies is expected to be average optimal; moreover, the value of the discounted games converges to the value of the expected average game.*

THEOREM 5.2. (1) *Let (u^*,v^*) be any stationary policy pair which is expected to be average optimal. Then for any $\epsilon>0$, (u^*,v^*) is ϵ-optimal for the β-discounted cost for all β sufficiently close to* 1*, and for all $u\in U$, $v\in V$,*

$$\overline{\lim_{\beta\to1}}[S_\beta(i,u,v^*)-S_\beta(i,u^*,v^*)]\leq 0, \qquad \underline{\lim_{\beta\to1}}[S_\beta(i,u^*,v)-S_\beta(i,u^*,v^*)]\geq 0.$$

(2) *For any $\epsilon>0$ there exists some $\beta_0<1$ such that for any $\beta_0\leq\beta<1$ and any stationary pair u^β,v^β which is β-optimal, (u^β,v^β) is ϵ-optimal for the expected average game.*

Proof. It is sufficient to prove that ($\mathcal{A}1$) and ($\mathcal{A}2$) in Theorem 2.1 hold. This follows indeed from the fact that

$$\lim_{\beta \to 1} \|S_\beta(\cdot, u, v) - S(\cdot, u, v)\|_\mu = 0$$

uniformly over all stationary policies u and v; see [2, p. 166]. □

6. Stochastic games with complete information. In sections 4 and 5 and in [34] we described several approximation problems in stochastic games, where we had a value in the limit game. In all those cases, we considered (without loss of optimality) the (randomized) stationary or the (randomized) Markov policies. Consider now the game $G^0 = (S, U_D, V_D)$, where S is given in (3.3) for some fixed initial state i, and where U_D and V_D are the classes of the pure stationary policies. Then the game will generally not have any value: it will have typically a lower and an upper value.

Since we established conditions ($\mathcal{A}1$) and ($\mathcal{A}2$) for all the problems considered in sections 4 and 5 (and also for the discounted cost, in [34]), they hold in particular if we restrict to purely stationary strategies. Therefore, the convergence of $\underline{R}_n^D$ and $\overline{R}_n^D$ to the upper and lower values of the limit game ($\underline{R}^D$ and $\overline{R}^D$) as well as the convergence of the policies in Theorems 2.5 and 2.6 hold for all these problems. (G_n^0 is the nth approximating game, and $\overline{R}_n^D$ and $\underline{R}_n^D$, its upper and lower values.)

Applying Theorem 2.7 is more delicate, since only in special cases can we define a topology over the space of responses of player 2 such that assumptions ($\mathcal{A}3$) and ($\mathcal{A}4$) hold (as opposed to standard stochastic games, where ($\mathcal{A}3$) and ($\mathcal{A}4$) need to hold for policies, and not for responses). When the action space available to player 1 is finite, one may identify the class of pure stationary response strategies of player 2 (corresponding to purely stationary policies of player 1) with the set of functions $\mathbf{I} \times \mathbf{A} \to \mathbf{B}$, endowed with topology of weak convergence of measures. The continuity assumptions ($\mathcal{A}3$) and ($\mathcal{A}4$) can now be established using arguments as in Remark 3.1 in [34] and [8].

We shall now show the usefulness of the above results for stochastic games with complete information in the sense used, e.g., by Küenle [24, 25]. There, one of the players—say, player 2—has at each time t the additional information of the action chosen by player 1 at time t (i.e., in addition to the information of all past states and actions of both players, plus the current state). The information structure for the other player is unchanged. We thus define the class of policies V^* to be the set of policies of the form $v = (v_0, v_1, ...)$, where v_t is a probability measure over $\mathbf{B}$ conditioned on the history $(x_0, a_0, b_0, x_1, a_1, b_1, ..., x_{t-1}, a_{t-1}, b_{t-1}, x_t, a_t)$.

Consider the following game: $G^1 = (S, U, V^*)$, where S is given by (3.3) for some fixed initial state i. (The results below will also hold for the expected discounted cost with infinite horizon, considered in [34].) We shall be interested in approximating

$$\underline{R}^1 = \sup_{u \in U} \inf_{v \in V^*} S(u, v).$$

Under quite general conditions [18, 24, 25], there exists optimal policies in U_D for player 1 in the sense that there exists some $u^1 \in U_D$ such that

$$\underline{R}^1 = \sup_{u \in U} \inf_{v \in V^*} S(u, v) = \sup_{u \in U_D} \inf_{v \in V^*} S(u, v) = \inf_{v \in V^*} S(u^1, v).$$

(This property does not hold for non-zero-sum games, for which there are counter-examples; see [18].)

On the other hand, under the conditions in section 3, we have for any $u \in U_D$,

$$\inf_{v \in V^*} S(u,v) = \inf_{v \in V} S(u,v) = \inf_{v \in V_D} S(u,v).$$

The first equality follows from the fact that $u \in U_D$, so that for any policy $v^* \in V^*$, we have $S(u,v) = S(u,v^*)$, where $v = (v_0, v_1, ...)$ is given by

$$v_t(x_0, a_0, b_0, ..., x_{t-1}, a_{t-1}, b_{t-1}, x_t) = v_t^*(x_0, a_0, b_0, ..., x_{t-1}, a_{t-1}, b_{t-1}, x_t, u(x_t)).$$

The second equality follows, since for any $u \in U_D$, player 2 is now faced with a standard Markov decision process (see, e.g., [32]). We thus conclude that $\underline{R}^1 = \underline{R}^D$. Thus, the convergence of $\underline{R}_n^D$ to $\underline{R}^D$ implies convergence of lower values for the game with complete information. Similarly, the convergence of the policies in Theorems 2.5 and 2.6 for games G_n^0 implies the convergence of policies for player 1 in the games with complete information that approximate G^1.

Remark 6.1. An alternative approach to obtaining convergence of values and policies for games with complete information is to transform these into standard stochastic games [25]. Note, however, that if the action space is uncountably infinite (e.g., an interval), then the transformation results in an uncountable state space.

7. A finite approximation of strategy sets. Another type of approximation that arises in dynamic games is the countable or finite approximation of strategy sets. This step is necessary when we want to perform numerical computations, and when the strategy sets are infinite or continuous, or both.

Let $\mathcal{U}$ and $\mathcal{V}$ be metric sets of policies for players 1 and 2, and let $S(u,v)$ correspond to the cost associated to the pair of strategies $u \in \mathcal{U}, v \in \mathcal{V}$. Introduce the following sets of strategies: $U \subset \mathcal{U}$ and $V \subset \mathcal{V}$, and the sequences $\{U_n\}_{n \in \mathbb{N}} \subset \mathcal{U}$ and $\{V_n\}_{n \in \mathbb{N}} \subset \mathcal{V}$. U_n and V_n are assumed to be countable or finite sets of policies.

THEOREM 7.1. *Suppose that*

($\mathcal{A}'1$) $\lim_{n \to +\infty} U_n = U$ *and* $\lim_{n \to +\infty} V_n = V$, *in the Hausdorff topology sense,*

($\mathcal{A}'3$) $S(u,v)$ *is a lower semicontinuous function in* $u \in U$ *uniformly in* $v \in V$,

($\mathcal{A}'4$) $S(u,v)$ *is an upper semicontinuous function in* $v \in V$ *uniformly in* $u \in U$.

Then the conclusions of Theorem 2.1 *hold (where* $\overline{R}_n$ *and* $\underline{R}_n$ *are defined in* (2.1) *and* R *is defined with respect to the policy sets* U *and* V *for* $\epsilon_1 = 0$*).*

Proof. We need to prove that under the set of assumptions ($\mathcal{A}'1$), ($\mathcal{A}'3$), and ($\mathcal{A}'4$) the hypotheses of Theorem 2.1 are satisfied. ($\mathcal{A}'3$) and ($\mathcal{A}'4$) directly imply ($\mathcal{A}3$) and ($\mathcal{A}4$). We shall prove only that ($\mathcal{A}1$) holds, since the proof of ($\mathcal{A}2$) is identical. Choose any ϵ_0 and introduce some sequence of functions:

$$\pi_n^1 : U_n \longrightarrow U \text{ such that } d(u_n, \pi_n^1(u_n)) < \inf_{u \in U} d(u_n, u) + \epsilon_0$$

and

$$\sigma_n^2 : V \longrightarrow V_n \text{ such that } d(\sigma_n^2(v), v) < \inf_{v_n \in V_n} d(v_n, v) + \epsilon_0.$$

By ($\mathcal{A}'1$), for all ϵ_2, there exists $N_1 = N_1(\epsilon_2)$ such that for all $n > N_1$,

$$\max\left(\sup_{u_n \in U_n} \inf_{u \in U} d(u_n, u) \; ; \; \sup_{u \in U} \inf_{u_n \in U_n} d(u_n, u) \right) \leq \epsilon_2;$$

that is, for all $u_n \in U_n$,

$$\inf_{u \in U} d(u_n, u) \leq \epsilon_2,$$

so for all $u_n \in U_n$,

$$d(u_n, \pi_n^1(u_n)) \leq \epsilon_0 + \epsilon_2. \tag{7.1}$$

Similarly, for all ϵ_3, there exists $N_3 = N_3(\epsilon_3)$ such that for all $n > N_3$ and for all $v \in V$,

$$d(v, \sigma_n^2(v)) \leq \epsilon_0 + \epsilon_3. \tag{7.2}$$

To prove $(\mathcal{A}1)$ we shall show that for all ϵ, there exists $N = N(\epsilon)$ such that for all $n > N$, and for all $u_n \in U_n$, $v \in V$, we have

(7.3)

$$S(u_n, \sigma_n^2(v)) - S(\pi_n^1(u_n), v) = S(u_n, \sigma_n^2(v)) - S(u_n, v) + S(u_n, v) - S(\pi_n^1(u_n), v) < \epsilon.$$

Since S is upper semicontinuous in v uniformly in u, there exists η_1 such that if $d(\sigma_n^2(v), v) \leq \eta_1$, $S(u_n, \sigma_n^2(v)) - S(u_n, v) \leq \epsilon/2$. Choose ϵ_0 and ϵ_3 such that $\epsilon_0 + \epsilon_3 = \eta_1$ in (7.2); there exists N_1 such that for all $n > N_1$,

$$S(u_n, \sigma_n^2(v)) - S(u_n, v) \leq \frac{\epsilon}{2}. \tag{7.4}$$

Similarly, it follows from (7.1) and the fact that S is lower semicontinuous in u uniformly in v that there exists N_2 such that for all $n > N_2$,

$$S(u_n, v) - S(\pi_n^1(u_n), v) \leq \frac{\epsilon}{2}. \tag{7.5}$$

Equations (7.4) and (7.5) imply (7.3) by choosing $N = \sup(N_1, N_2)$, which concludes the proof. □

Remark 7.2. In many applications (e.g., [11, 31]), the strategy sets are compact. Hence it suffices to require in $(\mathcal{A}'3)$ and $(\mathcal{A}'4)$ the semicontinuity properties; the uniform semicontinuity is then a consequence of the compactness of the strategy sets.

As an application of Theorem 7.1, we present the following continuous-time differential pursuit evasion game by Bernhard and Shinar [11, 31]. We shall use the same notation as in [11, 31]. The game is governed by a differential equation

$$\frac{dx}{dt} = f(x, a, b), \quad x \in \mathbf{I}, \quad a \in \mathbf{A}, \quad b \in \mathbf{B},$$

where $\mathbf{A}$ and $\mathbf{B}$ are compact subsets of $\mathbb{R}^{m_1}$ and $\mathbb{R}^{m_2}$, respectively, and $\mathbf{I}$ is a domain of $\mathbb{R}^n$. Some regularity and growth conditions on f ensure the existence of the solution of the differential equation over $(0, \infty)$ for every pair of measurable functions $a(\cdot)$ (and $b(\cdot)$) from $(0, \infty)$ to $\mathbf{A}$ (resp., $\mathbf{B}$). The players have access to noisy partial information

$$y_a = h_a(x, w), \qquad y_b = h_b(x, w),$$

where w is a noise and h_a, h_b are globally Lipschitz over $\mathbf{I}$. They are restricted to using feedback strategies $(a(t) = \delta_1(y_a(t)), b(t) = \delta_2(y_b(t)))$ Lipschitz continuous; then the set of strategies is compact in the topology of the uniform convergence. It is assumed that the noise model and the solution concept of the differential equations are such that the payoff P (the expected value of a continuous function of closest approach) is a continuous function of the strategies for the topology of uniform convergence.

If Ω_1 and Ω_2 are compact metric strategy spaces, and Δ_1 and Δ_2 are closed subsets of Ω_1 and Ω_2, respectively, and $U = M_1(\Delta_1)$ and $V = M_1(\Delta_2)$ are the sets of probability measures over Δ_1 and Δ_2, we know that there exist optimal mixed strategies that achieve the value

$$V(\Delta_1, \Delta_2) = \min_{u \in U} \max_{v \in V} J(u, v) = \int_{\Delta_1} \int_{\Delta_2} P(\delta_1, \delta_2) du(\delta_1) dv(\delta_2).$$

Bernhard and Shinar establish the convergence of the values of some approximating problems to the value of the original one.

We show that, in fact, all the statements concerning the convergence of policies in Theorem 2.1 also hold (with $\epsilon_1 = 0$). In [11], the continuity of $V(.,.)$ is proved; i.e., $(\mathcal{A}'3)$ and $(\mathcal{A}'4)$ are established. They present a finite approximation of this problem by considering finite subsets of Δ_i $i = 1, 2$, that converge to Δ_i in the Hausdorff topology (and thus, $(\mathcal{A}'1)$ holds).

REFERENCES

[1] E. Altman, *Denumerable constrained Markov decision problems and finite approximations*, Math. Oper. Res., 19 (1994), pp. 169–191.

[2] E. Altman, *Asymptotic properties of constrained Markov decision processes*, Z. Oper. Res., 37 (1993), pp. 151–170.

[3] E. Altman, *Flow control using the theory of zero-sum Markov games*, IEEE Trans. Automat. Control, 39 (1994), pp. 814–818.

[4] E. Altman, *A Markov game approach for optimal routing into a queueing network*, INRIA Report 2178, Sophia-Antipolis, France; Advances of Dynamic Games and Applications, to appear.

[5] E. Altman, *Non zero-sum stochastic games in admission, service and routing control in queueing systems*, QUESTA, 23 (1996), pp. 259–279.

[6] E. Altman, *Constrained Markov Decision Processes*, INRIA Report 2574, Sophia-Antipolis, France, May 1995.

[7] E. Altman and A. Hordijk, *Zero-sum Markov games and worst-case optimal control of queueing systems*, QUESTA, 21 (1995) (Special Issue on Optimization of Queueing Systems, S. Stidham, ed.), pp. 415–447.

[8] E. Altman, A. Hordijk, and F. M. Spieksma, *Contraction conditions for average and α-discounted optimality in countable state Markov games with unbounded rewards*, Math. Oper. Res., 1997, to appear.

[9] E. Altman and G. Koole, *Stochastic scheduling games with Markov decision arrival processes*, J. Comput. Math. Appl., 26 (1993) (third special issue on Differential Games), pp. 141–148.

[10] H. Attouch and J. B. Wets, *A convergence theory for saddle functions*, Trans. Amer. Math. Soc., 280 (1983), pp. 1–41.

[11] P. Bernhard and J. Shinar, *On finite approximation of a game solution with mixed strategies*, Appl. Math. Lett., 3 (1990), pp. 1–4.

[12] V. Borkar and M. K. Ghosh, *Denumerable state stochastic games with limiting average payoff*, J. Optim. Theory Appl., 76 (3) (1993), pp. 539–560.

[13] R. Cavazos-Cadena, *Finite-state approximations for denumerable state discounted Markov decision processes*, J. Appl. Math. Optim., 14 (1986), pp. 27–47.

[14] K. L. Chung, *Markov Chains with Stationary Transition Probabilities*, 2nd ed., Springer-Verlag, New York, 1967.

[15] H. A. M. Couwenbergh, *Stochastic games with metric state space*, Internat. J. Game Theory, 9 (1980), pp. 25–36.

[16] R. Dekker and A. Hordijk, *Average, sensitive and Blackwell optimal policies in denumerable Markov decision chains with unbounded rewards*, Math. Oper. Res., 13 (1988), pp. 395–421.

[17] M. C. Delfour and J. Morgan, *Derivative of minmax and saddle points with respect to a parameter*, Preprint 29, Departimento di Matematica e Applicazioni, Università degli Studi di Napoli "Federico II," Naples, 1992.

[18] A. Federgruen, *On N-person stochastic games with denumerable state space*, Adv. Appl. Prob., 10 (1978), pp. 452–471.
[19] D. Gillette, *Stochastic Games with Zero Stop Probabilities*, M. Dresher, A. W. Tucker, and P. Wolfe, eds., Princeton University Press, Princeton, NJ, 1957, pp. 179–187.
[20] J. Guillerme, *Convergence of approximate saddle point*, J. Math. Anal. Appl., 137 (1989), pp. 297–311.
[21] O. Hernandez-Lerma, *Finite state approximations for denumerable multidimensional - state discounted Markov decision processes*, J. Math. Anal. Appl., 113 (1986), pp. 382–389.
[22] O. Hernandez-Lerma, *Adaptive Control of Markov Processes*, Springer-Verlag, New York, 1989.
[23] J. L. Kelley, *General Topology*, Springer-Verlag, New York, 1955.
[24] H. W. Küenle, *Stochastiche Spiele und Entscheidungesmodelle*, Teubner-Texte, Band 89, Leipzig, Germany, 1986.
[25] H. W. Küenle, *On Nash equilibrium solutions in nonzero-sum stochastic games with complete information*, Internat. J. Game Theory, 23 (1994), pp. 303–324.
[26] M. B. Lignola and J. Morgan, *Existence and approximation for min-sup problems*, Proceedings of the International Conference of Operation Research 90, W. Bühler, G. Feichtinger, R. F. Hartl, F. J. Radermacher, and P. Stähly, eds., Springer-Verlag, Berlin, 1992, pp. 157–164.
[27] J. Morgan and R. Raucci, *Continuity properties of ϵ-solutions for generalized parametric saddle point problems*, Preprint 35, Departimento di Matematica e Applicazioni, Università degli Studi di Napoli "Federico II," Naples, 1994.
[28] A. S. Nowak, *On zero-sum stochastic games with general state space*, I, Probab. Math. Statist., IV (1984), pp. 13–32.
[29] A. S. Nowak, *Existence of equilibrium stationary strategies in discounted noncooperative stochastic games with uncountable state space*, J. Optim. Theory Appl., 45 (1985), pp. 592–602.
[30] L. I. Sennott, *Zero-sum stochastic games with unbounded costs: Discounted and average cost cases*, Z. Oper. Res., 40 (1994), pp. 145–162.
[31] J. Shinar and I. Forte, *On the optimal pure strategy sets for a missile guidance law synthesis*, in Proceedings of the 25th IEEE Conference on Decision and Control, Athens, Greece, IEEE, Piscataway, NJ, 1986.
[32] F. M. Spieksma, *Geometrically Ergodic Markov Chains and the Optimal Control of Queues*, Ph.D. Thesis, Leiden University, Leider, The Netherlands, 1990 (available on request from the author).
[33] L. C. Thomas and D. Stengos, *Finite state approximation algorithms for average cost denumerable state Markov decision processes*, OR Spektrum, 7 (1985), pp. 27–37.
[34] M. Tidball and E. Altman, *Approximations in dynamic zero-sum games*, I, SIAM J. Control Optim., 34 (1996), pp. 311–328.
[35] D. J. White, *Finite state approximations for denumerable state infinite horizon discounted Markov decision processes*, J. Math. Anal. Appl., 74 (1980), pp. 292–295.
[36] D. J. White, *Finite state approximations for denumerable state infinite horizon discounted Markov decision processes with unbounded rewards*, J. Math. Anal. Appl., 86 (1982), pp. 292–306.
[37] W. Whitt, *Approximations of dynamic programs*, I, Math. Oper. Res., 3 (1978), pp. 231–243.
[38] W. Whitt, *Representation and approximation of noncooperative sequential games*, SIAM J. Control Optim., 18 (1980), pp. 33–43.

SIAM J. CONTROL OPTIM.
Vol. 35, No. 6, pp. 2118–2127, November 1997

013

NP-HARDNESS OF SOME LINEAR CONTROL DESIGN PROBLEMS*

VINCENT BLONDEL† AND JOHN N. TSITSIKLIS‡

Abstract. We show that some basic linear control design problems are NP-hard, implying that, unless P=NP, they cannot be solved by polynomial time algorithms. The problems that we consider include simultaneous stabilization by output feedback, stabilization by state or output feedback in the presence of bounds on the elements of the gain matrix, and decentralized control. These results are obtained by first showing that checking the existence of a stable matrix in an interval family of matrices is NP-hard.

Key words. control design, complexity, linear control

AMS subject classifications. 93D09, 93D15, 68Q25, 15A18

PII. S0363012994272630

1. Introduction.

Consider the following three problems; the first was mentioned as a "major open problem in systems and control theory" in a recent survey [5] of experts in the systems and control field, and the other two were mentioned indirectly.

Stabilization by static output feedback. This is perhaps the most basic problem in control theory. We are given a linear system

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

and we consider a static feedback control law of the form

$$u(t) = Ky(t).$$

The resulting closed loop system is

$$\dot{x}(t) = (A + BKC)x(t).$$

The problem is to find necessary and sufficient conditions on the triplet of real matrices (A, B, C) under which there exists a feedback gain matrix K such that $A + BKC$ is stable. In the case of state feedback ($C = I$), a necessary and sufficient stabilizability condition is given by the stabilizability of the pair (A, B) [17]. However, if C is not invertible, no general necessary and sufficient conditions are known.

Simultaneous stabilization by static output or state feedback. (This problem should not be confused with what is usually referred to as the "simultaneous stabilization problem" [16, 4], in which dynamic—instead of static—compensation is sought.) Our second problem is a generalization of the static output feedback problem. Suppose

*Received by the editors August 10, 1994; accepted for publication (in revised form) September 23, 1996. This research was supported by AFOSR grant AFOSR-91-0368 and by ARO grant DAAL03-92-G0115.

http://www.siam.org/journals/sicon/35-6/27263.html

†Institute of Mathematics, University of Liège, B-4000 Liège, Belgium (blondel@math.ulg.ac.be). The research of this author was partially supported by KTH, Stockholm.

‡Laboratory for Information and Decision Systems, Massachusetts Institute of Technology, Cambridge, MA 02139 (jnt@mit.edu).

that for each $i = 1, \ldots, k$ we are given a linear system

$$\dot{x}(t) = A_i x(t) + B_i u(t),$$
$$y(t) = C_i x(t).$$

Under the feedback control law,

$$u(t) = Ky(t),$$

the ith closed loop system is

$$\dot{x}(t) = (A_i + B_i K C_i) x(t).$$

The problem is to find conditions on the triplets of real matrices (A_i, B_i, C_i), $i = 1, \ldots, k$, under which there exists a matrix K such that $A_i + B_i K C_i$ is stable for each i. This problem is unsolved, even if $C_i = I$ for all i (simultaneous stabilization by state feedback).

Stabilization by decentralized static output feedback. We now impose some structure on the feedback gains. Consider a linear system of the form

$$\dot{x}(t) = Ax(t) + \sum_{i=1}^{k} B_i u_i(t),$$
$$y_i(t) = C_i x(t), \qquad i = 1, \ldots, k,$$

and suppose that we are interested in a static decentralized controller of the form

$$u_i(t) = K_i y_i(t), \qquad i = 1, \ldots, k.$$

The closed loop system is

$$\dot{x}(t) = \left(A + \sum_{i=1}^{k} B_i K_i C_i \right) x(t),$$

which is of the same form as in stabilization by static output feedback, except that several of the entries of K are forced to zero. This leads us to the problem of finding conditions on the triplet of real matrices (A, B, C) under which there exists a matrix K with a given structure such that $A + BKC$ is stable. The problem can be further constrained by requiring the matrix structure to be block diagonal, the blocks to have a bounded norm, or the blocks to be identical (we discuss all of these cases later).

The reader is referred to [3, p. 420], where the above three problems are presented and motivated and where references can be found. A common feature of these three problems is that, although they are easy to state, neither closed form nor efficient algorithmic solutions are known. It is rather improbable that closed form solutions to these problems are possible. On the other hand, algorithmic solutions do exist, as we now argue.

All of the problems that we have described are finitely parametrized. They all involve the search for a controller—the (possibly partitioned) matrix K—which can be specified in terms of finitely many real parameters. In theory, it is thus possible to apply the following methodology: (a) parametrize the gain matrix K in terms of finitely many real coefficients; (b) express the matrix stability condition(s) in terms

of the coefficients of the system(s) and of the controller; (c) use the Routh–Hurwitz test on the resulting characteristic polynomial(s). One is then left with a (large) set of multivariable polynomial inequalities that have to be simultaneously satisfied for some choice of the controller coefficients. As explained in [1], checking the existence of controller coefficients that satisfy this system of multivariable inequalities can be performed using the Tarski–Seidenberg elimination theory. The Tarski–Seidenberg elimination method leads, after a finite number of rational operations, to a yes-no answer regarding the existence of a solution. The method is systematic and amenable to computer implementation. Thus, *all three problems described above are algorithmically solvable*.

The advantage of the Tarski–Seidenberg method is its generality; its drawback is the fact that its computational complexity increases at least exponentially. The examples that can be worked on paper are very small (the example given in [1] involves only two parameters), and computer algorithms cannot digest more than five or six parameters in reasonable time.

In this paper we show that some of the above problems and their variations are very unlikely to allow for *efficient* algorithmic solutions. We adhere to the general consensus in computer science that identifies algorithmic efficiency with polynomial time computability. We then show that some of the above problems are NP-hard [8, 13], meaning that every problem in NP can be reduced to them. Thus, unless P=NP, these problems are not polynomial time solvable.

Our results are as follows (see later for precise definitions):

1. The static output feedback stabilization problem is NP-hard if one constrains the coefficients of the controller K to lie in prespecified intervals. The same is true in the case of static *state* feedback ($C = I$). We have not been able to establish the complexity of the problem in the absence of constraints on K, but we conjecture that it is also NP-hard.
2. Simultaneous stabilization by output feedback is NP-hard.
3. Stabilization by decentralized static output feedback is NP-hard if one imposes a bound on the norm of the controller or if the blocks are constrained to be identical.

These results will be proved as corollaries of the following main theorem: testing for the presence of a stable matrix in a family of matrices whose members have entries that are either fixed to some given real number or vary in the closed unit interval $[-1, 1]$ is an NP-hard problem. This latter result complements a recent theorem of Nemirovskii [11], who showed that testing for the stability of *all* elements of such a family of matrices is an NP-hard problem. Our proof is in fact inspired from his. This general research direction was initiated by Poljak and Rohn, who showed that checking nonsingularity of an interval family of matrices is NP-hard [14]. In other related research, NP-hardness of the computation of the structured singular value μ was shown by Braatz et al. [6] for the case where some perturbations are complex. (NP-completeness for the case of real perturbations was a corollary of the results of Poljak and Rohn.) Also, Coxson and DeMarco show that approximating the minimal perturbation scaling to achieve instability in an interval matrix is MAX-SNP-hard [7]. See also [15] for a review of other complexity results for problems in control theory.

In the next section, we prove the main result and derive some general corollaries. In the last section we link these results with the linear control design problems mentioned in this introduction.

2. Checking the existence of a stable matrix in an interval family of matrices is NP-hard. In this section we show that checking the existence of a stable matrix in a unit interval family of matrices is an NP-hard problem (a unit interval family of matrices is a family of matrices whose members have entries that are either fixed to some given real number or vary in the closed unit interval $[-1, 1]$). We prove this result by means of a polynomial time reduction from the following problem, which is already known to be NP-complete [10, 8].

PARTITION

Instance: A positive integer l, a set of l integers $a_i \in \mathcal{Z}$.

Question: Do there exist $t_1, \ldots, t_l \in \{-1, +1\}$ such that $\sum_{i=1}^{l} a_i t_i = 0$?

We now formally define the problem of interest.

STABLE MATRIX IN UNIT INTERVAL FAMILY

Instance: A positive integer n, a partition of $I = \{(i, j) : 1 \leq i, j \leq n\}$ into disjoint sets I_1 and I_2, rational numbers a_{ij}^* for $(i, j) \in I_1$.

Question: Does the set $\mathcal{A}$ of $n \times n$ matrices defined by

$$\mathcal{A} = \{A = (a_{ij}) : a_{ij} = a_{ij}^* \text{ for } (i, j) \in I_1,\ a_{ij} \in [-1, 1] \text{ for } (i, j) \in I_2\}$$

contain a stable matrix?

Remark. Throughout this paper, when writing "stable" we actually mean "asymptotically stable," i.e., "all eigenvalues have a negative real part." A slightly different problem is obtained if we are interested in marginal stability ("all eigenvalues have a nonpositive real part"). We call this second problem MARGINALLY STABLE MATRIX IN UNIT INTERVAL FAMILY.

The main result of this paper is as follows.

THEOREM 1. STABLE MATRIX IN UNIT INTERVAL FAMILY *and* MARGINALLY STABLE MATRIX IN UNIT INTERVAL FAMILY *are NP-hard.*

Proof. We prove NP-hardness of STABLE MATRIX IN UNIT INTERVAL FAMILY. NP-hardness of MARGINALLY STABLE MATRIX IN UNIT INTERVAL FAMILY can be shown in a similar way; we make a comment on this at the end of the proof.

Since PARTITION is NP-complete, it suffices to show that any instance of PARTITION can be transformed in polynomial time into an equivalent instance of STABLE MATRIX IN UNIT INTERVAL FAMILY.

Let $a_i \in \mathcal{Z}$ $(i = 1, \ldots, l)$ be an instance of PARTITION. We construct a unit interval matrix as follows. Let m be a positive integer such that $l < m = k^2$ for some positive integer k, and define the m-dimensional vector a by $a^T = (a_1, a_2, \ldots, a_l, 0, \ldots, 0) \in \mathcal{Z}^m$ (the superscript ^{T} denotes matrix transposition). Let $\gamma = a^T a$, $\beta = 1 - 1/(2m(1+\gamma))$, and

$$A(x, y) = \begin{pmatrix} -k(I_m + aa^T) & y \\ x^T & k\beta \end{pmatrix}, \tag{1}$$

with I_m the identity matrix of size m and $x, y \in \Re^m$ (note that $\gamma > 0$ and $0 < \beta < 1$).

The set of matrices

$$\mathcal{A} = \{A(x, y) : x, y \in [-1, 1]^m\} \tag{2}$$

forms an instance of STABLE MATRIX IN UNIT INTERVAL FAMILY and is constructed in polynomial time from the initial instance of PARTITION. It remains thus to show that $\mathcal{A}$ contains a stable matrix if and only if there exist $t_i \in \{-1, +1\}$ such that $\sum_{i=1}^{l} a_i t_i = 0$. We prove this in two steps.

Assume first that $t_i \in \{-1, +1\}$ satisfy $\sum_{i=1}^{l} a_i t_i = 0$. Define $x_0^T = (t_1, t_2, \ldots, t_l, 1, \ldots, 1) \in \mathcal{Z}^m$, $y_0 = -x_0$, and note that $a^T x_0 = x_0^T a = 0$. We claim that the matrix $A_0 = A(x_0, y_0) \in \mathcal{A}$ is stable. Indeed, A_0 can be decomposed as

$$A_0 = A_1 + A_2 + A_3 \tag{3}$$

$$= -kI_{m+1} + \begin{pmatrix} -kaa^T & 0 \\ 0 & 0 \end{pmatrix} + \begin{pmatrix} 0 & -x_0 \\ x_0^T & k(1+\beta) \end{pmatrix}. \tag{4}$$

The spectrum of A_0 is the spectrum of $A_2 + A_3$ shifted to the left by k. The matrix A_0 will thus be stable, provided that the real part of every eigenvalue of $A_2 + A_3$ is strictly less than k.

The matrix A_2 has rank one; it has one eigenvalue at $-k\gamma$ and m eigenvalues at the origin. The characteristic polynomial of the matrix A_3 is

$$s^{m-1}(s^2 - k(1+\beta)s + k^2), \tag{5}$$

whose roots are either at the origin or have a real part equal to $k(1+\beta)/2$ which is always strictly less than k, since we have already observed that $0 < \beta < 1$.

Due to the fact that $a^T x_0 = x_0^T a = 0$, we have $A_2 A_3 = A_3 A_2 = 0$. Let λ and w be an eigenvalue and an eigenvector, respectively, of $A_2 + A_3$. Thus, $(A_2 + A_3)w = \lambda w$. Multiplying by A_2, we obtain $A_2^2 w = \lambda A_2 w$. If $A_2 w \neq 0$, then λ is an eigenvalue of A_2. If $A_2 w = 0$, then λ is an eigenvalue of A_3. Consequently, every eigenvalue of $A_2 + A_3$ is either an eigenvalue of A_2 or of A_3. These eigenvalues have a real part which is smaller than k, and by our earlier comment, the matrix $A_0 \in \mathcal{A}$ is stable.

For the reverse implication, assume that $\mathcal{A}$ contains a stable matrix and let $x_0, y_0 \in [-1, 1]^m$ be such that $A_0 = A(x_0, y_0) \in \mathcal{A}$ is stable. Consider then the parametrized family of matrices

$$B(\theta) = A(\theta x_0, \theta y_0)/k. \tag{6}$$

We now study the dependence of the stability of $B(\theta)$ on the variable $\theta \in [0, 1]$. When $\theta = 0$, we have

$$B(0) = \begin{pmatrix} -(I_m + aa^T) & 0 \\ 0 & \beta \end{pmatrix}. \tag{7}$$

The matrix $-(I_m + aa^T)$ is negative definite, hence stable, and thus $B(0)$ has a single unstable eigenvalue (at $\beta > 0$). When $\theta = 1$, we have $B(1) = A_0/k$, and so $B(1)$ is stable since A_0 is.

The eigenvalues of $B(\theta)$ are symmetric with respect to the real axis (complex conjugate), and they vary continuously with θ. When moving from $\theta = 0$ to $\theta = 1$, we move from a configuration where there is exactly one unstable eigenvalue to a configuration with no unstable eigenvalues. When a conjugate pair of eigenvalues crosses the $j\omega$ axis, the number of unstable eigenvalues changes by an even number. Thus, for the number of unstable eigenvalues to change from one to zero, some eigenvalue must cross the $j\omega$ axis at the origin. Therefore, there exists some $\theta_0 \in (0, 1)$ for which $B(\theta_0)$ has an eigenvalue at the origin and $B(\theta_0)$ is singular. Elementary matrix manipulations show that the singularity condition for $B(\theta_0)$ is equivalent to

$$\theta_0^2 x_0^T (I_m + aa^T)^{-1} y_0 = -k^2 \beta. \tag{8}$$

A standard inversion formula [9, p. 19] gives

$$\theta_0^2 x_0^T (I_m - aa^T/(1+\gamma)) z_0 = k^2 \beta, \tag{9}$$

where we have defined $z_0 = -y_0$ and used the definition $\gamma = a^T a$. Remembering that $m = k^2$ and $\theta_0 \in (0,1)$, we finally obtain

$$x_0^T (I_m - aa^T/(1+\gamma)) z_0 > m\beta. \tag{10}$$

The matrix $(I_m - aa^T/(1+\gamma))$ is symmetric and positive definite. Using also the fact that the maximum of a convex function over a bounded polyhedron is attained at an extreme point, we obtain

$$\max_{x,y \in [-1,1]^m} x^T (I_m - aa^T/(1+\gamma)) y = \max_{x \in [-1,1]^m} x^T (I_m - aa^T/(1+\gamma)) x \tag{11}$$

$$= \max_{x \in \{-1,1\}^m} x^T (I_m - aa^T/(1+\gamma)) x \tag{12}$$

$$= m - \min_{x \in \{-1,1\}^m} (x^T a)^2/(1+\gamma). \tag{13}$$

In particular, this shows that

$$m - \min_{x \in \{-1,1\}^m} (x^T a)^2/(1+\gamma) \geq x_0^T (I - aa^T/(1+\gamma)) z_0. \tag{14}$$

Combining inequalities (10) and (14), we obtain

$$m - \min_{x \in \{-1,1\}^m} (x^T a)^2/(1+\gamma) > m\beta. \tag{15}$$

Using the definition of β, we finally arrive at

$$\min_{x \in \{-1,1\}^m} (x^T a)^2 < 1/2. \tag{16}$$

The left-hand side in this inequality is a nonnegative integer; we are thus forced to the conclusion

$$\min_{x \in \{-1,1\}^m} (x^T a)^2 = 0. \tag{17}$$

Assume that the minimum in (17) is obtained for $x^T = (x_1, x_2, \ldots, x_l, \ldots, x_m)$; we conclude the proof by setting $t_i = x_i$ for $i = 1, \ldots, l$.

Let us now briefly comment on the case where we are interested in marginal stability. NP-hardness for this case can be obtained by a small adaptation of the preceding proof. Let, as before, $a_i \in \mathcal{Z}$ $(i = 1, \ldots, l)$ be an instance of PARTITION. We construct an interval matrix as follows. Let m be a positive integer such that $l < m = k^2$ for some positive integer k and define the m-dimensional vector a by $a^T = (a_1, a_2, \ldots, a_l, 0, \ldots, 0) \in \mathcal{Z}^m$ and

$$A(x,y) = \begin{pmatrix} -k(I_m + aa^T) & y \\ x^T & k \end{pmatrix}. \tag{18}$$

The set of matrices $\mathcal{A} = \{A(x,y) : x, y \in [-1,1]^m\}$ forms an instance of MARGINALLY STABLE MATRIX IN UNIT INTERVAL FAMILY and is constructed in polynomial time from the initial instance of PARTITION. Moreover, by the same argument as above, it is clear that $\mathcal{A}$ contains a marginally stable matrix if and only if there exist $t_i \in \{-1, +1\}$ such that $\sum_{i=1}^{l} a_i t_i = 0$. This shows the equivalence between the instances and hence proves the second part of the theorem. ∎

Suppose now that we change the problem by including the additional requirement that the matrix A must be symmetric. Consider the problem of minimizing λ subject to $\lambda I - A$ being a positive semidefinite symmetric matrix and subject to the interval constraints on A. This is a semidefinite programming problem and can be solved, within any desired accuracy ϵ, in time which is polynomial in the size of the problem and the "size" $\log(1/\epsilon)$ of ϵ. Furthermore, the optimal cost in this minimization problem is less than or equal to zero (respectively, negative) if and only if there exists a marginally stable (respectively, stable) matrix A in the family. This argument, brought to our attention by M. Overton [12], comes close but does not quite establish polynomiality of the problem STABLE MATRIX IN UNIT INTERVAL FAMILY for the symmetric case; that would require an exact (as opposed to approximate) polynomial time solution of the semidefinite programming problem. If the symmetric problem is indeed polynomial time solvable, this would be in contrast to the results of Nemirovskii [11], who showed that deciding the stability of all elements of the interval family is NP-hard even if one restricts to symmetric matrices.

As a direct application of our main theorem, we introduce a few matrix and polynomial stability problems and show that they are NP-hard.

STABLE MATRIX IN INTERVAL FAMILY
Instance: A positive integer n, rational numbers $\underline{a}_{ij}, \overline{a}_{ij}$ for $1 \le i, j \le n$.
Question: Does there exist a stable matrix $A = (a_{ij})$ with $\underline{a}_{ij} \le a_{ij} \le \overline{a}_{ij}$?

STABLE MATRIX IN RANK ONE PERTURBED MATRIX
Instance: Positive integers n, k, and $k+1$ real $n \times n$ matrices $A_0, A_1, \ldots, A_k$ with rational entries, all of which have rank one, with the exception of A_0.
Question: Do there exist real values $q_i^* \in [-1, 1]$ such that $A = A_0 + q_1^* A_1 + \cdots + q_k^* A_k$ is stable?

STABLE POLYNOMIAL IN FAMILY OF BILINEAR POLYNOMIALS
Instance: A positive integer r, a multivariable polynomial $p(x, q_1, \ldots, q_r)$ with rational coefficients whose dependence on the real variables q_i is bilinear.
Question: Do there exist real values $q_i^* \in [-1, 1]$ for which the polynomial $p(x, q_1^*, \ldots, q_r^*)$ is stable?

COROLLARY 1. *The above three problems are all NP-hard.*

Proof. STABLE MATRIX IN INTERVAL FAMILY is NP-hard because it is a generalization of STABLE MATRIX IN UNIT INTERVAL FAMILY.

A matrix A in the unit interval family defined by I_1 and a_{ij}^*, $(i, j) \in I_1$, can be written in the form

$$A = A_0 + \sum_{(i,j) \notin I_1} q_{ij} A_{ij},$$

where A_0 has entries

$$\begin{aligned} a_{ij}^0 &= a_{ij}^* \quad \text{if } (i, j) \in I_1, \\ &= 0 \quad \text{if } (i, j) \notin I_1. \end{aligned}$$

A_{ij} is a matrix with all entries equal to zero except for the (i, j)th entry, which is equal to 1, and $q_{ij} \in [-1, 1]$; note that A_{ij} has rank one. This reduces STABLE MATRIX IN UNIT INTERVAL FAMILY to STABLE MATRIX IN RANK ONE PERTURBED MATRIX and shows that the latter problem is NP-hard.

In order to prove that STABLE POLYNOMIAL IN FAMILY OF BILINEAR POLYNOMIALS is NP-hard, we argue as in the proof of Theorem 1. Let $a_i \in \mathcal{Z}$ $(i = 1, \ldots, l)$ be an instance of PARTITION. Let m be a positive integer such that $l < m = k^2$ for some

positive integer k and define $\beta = 1 - 1/(2m(1 + \sum_{i=1}^{l} a_i^2))$ and

$$A(q_1, \ldots, q_k, q_{k+1}, \ldots, q_{2k}) = \begin{pmatrix} -k(I_m + aa^T) & (q_{k+1}, \ldots, q_{2k})^T \\ (q_1, \ldots, q_k) & k\beta \end{pmatrix}.$$

From the proof of Theorem 1, we know that the set of matrices $\mathcal{A} = \{A(q_1, \ldots, q_k, q_{k+1}, \ldots, q_{2k}) : q_i \in [-1, 1]\}$ contains a stable matrix if and only if there exist $t_i \in \{-1, +1\}$ such that $\sum_{i=1}^{l} a_i t_i = 0$. The set of matrices $\mathcal{A}$ contains a stable matrix if and only if the multivariable polynomial $p(x, q_1, \ldots, q_{2k}) = \det(xI_{2k} - A(q_1, \ldots, q_k, q_{k+1}, \ldots, q_{2k}))$ is stable for some choice of $q_i \in [-1, 1]$. The latter polynomial is bilinear in the variables q_i. We therefore have an instance of STABLE POLYNOMIAL IN FAMILY OF BILINEAR POLYNOMIALS which is equivalent to the original instance of PARTITION. □

Remarks.

1. All three problems addressed by Corollary 1 remain NP-hard if "stability" is replaced by "marginal stability"; the proof is similar.

2. By a similar proof, both Theorem 1 and Corollary 1 remain valid if the interval constraints $a_{ij} \in [-1, 1]$ are replaced by the open interval constraints $a_{ij} \in (-1, 1)$.

3. The decision problem for the existential theory of the reals is solvable in $s^{k+1} d^{O(k)}$ arithmetic operations where k denotes the number of variables, s is the number of polynomial (in)equalities, and d is the highest polynomial degree [2]. This shows that for fixed k, a polynomial time algorithm is possible. In particular, STABLE MATRIX IN INTERVAL FAMILY becomes polynomial time solvable if an a priori bound is given on the size of the matrix. The problems discussed in Corollary 1 also become polynomial time solvable when suitably constrained.

3. Application to linear control design problems. As explained in the introduction, our initial motivation for this work was to address the computational complexity of linear control design problems. We now introduce some such problems and show that they are NP-hard.

STATE FEEDBACK STABILIZATION BY BOUNDED CONTROLLER

Instance: A positive integer n, $n \times n$ matrices A and B with rational coefficients, rational numbers $\underline{k}_{ij}, \overline{k}_{ij}$ for $1 \le i, j \le n$.

Question: Does there exist a real matrix $K = (k_{ij})$ satisfying $\underline{k}_{ij} \le k_{ij} \le \overline{k}_{ij}$ and such that $A + BK$ is stable?

SIMULTANEOUS STABILIZATION BY OUTPUT FEEDBACK

Instance: Positive integers n, m, p, k, a collection of k triplets of matrices (A_i, B_i, C_i) with rational coefficients of respective sizes $n \times n, n \times m, p \times n$.

Question: Does there exist a real $m \times p$ matrix K such that $A_i + B_i K C_i$ is stable for all $i = 1, \ldots, k$?

DECENTRALIZED OUTPUT FEEDBACK STABILIZATION BY NORM BOUNDED CONTROLLER

Instance: Positive integers n and k with $n \ge k$, $n \times n$ matrices A, B and C with rational coefficients. A partition of n into k positive integers $n = n_1 + n_2 + \cdots + n_k$.

Question: Does there exist a $n \times n$ block-diagonal matrix K with blocks K_i of successive sizes $n_i \times n_i$ and $||K_i|| < 1$ such that $A + BKC$ is stable?

DECENTRALIZED STABILIZATION WITH IDENTICAL CONTROLLERS

Instance: Positive integers n_1, n_2, three $(n_1 n_2) \times (n_1 n_2)$ matrices A, B and C with rational coefficients.

Question: Does there exist a $n_1 \times n_1$ matrix M such that the $(n_1 n_2 \times n_1 n_2)$ block diagonal matrix K constructed with n_2 identical blocks M is such that $A + BKC$ is stable?

COROLLARY 2. *The above four problems are all NP-hard.*

Proof. (a) STATE FEEDBACK STABILIZATION BY BOUNDED CONTROLLER: Let n and $\underline{a}_{ij}$, $\overline{a}_{ij}$, for $1 \leq i, j \leq n$ be an instance of STABLE MATRIX IN INTERVAL FAMILY. An equivalent instance of STATE FEEDBACK STABILIZATION BY BOUNDED CONTROLLER is given by n, $A = 0$, $B = I_n$, $\underline{k}_{ij} = \underline{a}_{ij}$, and $\overline{k}_{ij} = \overline{a}_{ij}$ for $1 \leq i, j \leq n$.

(b) SIMULTANEOUS STABILIZATION BY OUTPUT FEEDBACK:

We prove NP-hardness for the case of marginal stability. Let n and $\underline{a}_{kl}, \overline{a}_{kl}$ $(1 \leq i, j \leq n)$ be an instance of STABLE MATRIX IN INTERVAL FAMILY. Define the $n \times n$ matrices A_{ij}^+, A_{ij}^-, B_i, and C_j by

$A_{ij}^+ = (a_{kl})$ with

$$a_{kl} = -\overline{a}_{ij} \text{ if } (k,l) = (1,1),$$
$$= 0 \text{ otherwise;}$$

$A_{ij}^- = (a_{kl})$ with

$$a_{kl} = \underline{a}_{ij} \text{ if } (k,l) = (1,1),$$
$$= 0 \text{ otherwise;}$$

$B_i = (b_{kl})$ with

$$b_{kl} = 1 \text{ if } (k,l) = (1,i),$$
$$= 0 \text{ otherwise;}$$

and $C_j = (c_{kl})$ with

$$c_{kl} = 1 \text{ if } (k,l) = (j,1),$$
$$= 0 \text{ otherwise.}$$

It is immediate to see that $(A_{ij}^+ + B_i K C_j)$ is marginally stable if and only if $k_{ij} \leq \overline{a}_{ij}$, and similarly, $(A_{ij}^- + B_i K C_j)$ is marginally stable if and only if $k_{ij} \geq \underline{a}_{ij}$. Thus, if we require the simultaneous stabilization of the $2n^2 + 1$ triplets $(0, I, I), (A_{ij}^+, B_i, C_j)$, and (A_{ij}^-, B_i, C_j) for $1 \leq i, j \leq n$, we have constructed an equivalent instance of SIMULTANEOUS STABILIZATION BY OUTPUT FEEDBACK.

(c) DECENTRALIZED OUTPUT FEEDBACK STABILIZATION BY NORM BOUNDED CONTROLLER: We prove that the problem is NP-hard even for the special case where all blocks are of size 1×1, in which case $A + BKC$ can be written as $A + \sum_{i=1}^n k_i b_i c_i^T$, where b_i is the ith column of B, c_i^T is the ith row of C, and k_i is the ith diagonal entry of K. Given that an arbitrary rank one matrix can be expressed in the form bc^T for some vectors b and c, it follows that every instance of STABLE MATRIX IN RANK ONE PERTURBED MATRIX can be expressed as an instance of DECENTRALIZED OUTPUT FEEDBACK STABILIZATION BY NORM BOUNDED CONTROLLER

(d) DECENTRALIZED STABILIZATION WITH IDENTICAL CONTROLLERS: We prove NP-hardness for the case of marginal stability. Consider k triplets of $n \times n$ matrices (A_i, B_i, C_i) that form an instance of SIMULTANEOUS STABILIZATION BY OUTPUT FEEDBACK. We define an equivalent instance of DECENTRALIZED STABILIZATION WITH IDENTICAL CONTROLLERS by letting $n_1 = n, n_2 = k, A = A_1 \oplus A_2 \oplus \cdots \oplus A_k, B = B_1 \oplus B_2 \oplus \cdots \oplus B_k$ and $C = C_1 \oplus C_2 \oplus \cdots \oplus C_k$, where $\oplus$ denotes direct sum of matrices. ∎

Remarks.

1. For some of the problems, we provided the proof for the case of stability; for others, we dealt with marginal stability. With little work and using the remarks at the end of the preceding section, it is easily shown that all problems are NP-hard for the case of either stability or marginal stability.

2. STATE FEEDBACK STABILIZATION BY BOUNDED CONTROLLER is easily shown to remain NP-hard even if the bounds $\underline{k}_{ij}, \overline{k}_{ij}$ are constrained to be either 0 or 1. We have assumed that we are dealing with square systems; the more general case of rectangular systems is at least as hard and is therefore also NP-hard. Finally, the problem of *output* feedback stabilization by a bounded controller is at least as hard as that of *state* feedback and is thus also NP-hard.

3. Our proof shows that SIMULTANEOUS STABILIZATION BY OUTPUT FEEDBACK remains NP-hard even if all the matrices involved are of the same size ($n = m = p$). The degenerate case $m = p = 1$ corresponds to simultaneous stabilization of single-input, single-output systems by proportional feedback and can be solved in polynomial time. (An argument for this follows from footnote 1 on p. 54 of [1].) For a priori fixed n, m, and p, the problem can also be solved in polynomial time (see Remark 3 in section 2). We do not know whether the state feedback formulation of this problem is NP-hard.

REFERENCES

[1] B. D. O. Anderson, N. K. Bose, and E. I. Jury, *Output feedback stabilization and related problems — solutions via decision methods*, IEEE Trans. Automat. Control, 20 (1975), pp. 53–66.

[2] S. Basu, R. Pollack, and M.-F. Roy, *On the combinatorial and algebraic complexity of quantifier elimination*, J. Assoc. Comput. Mach. 43 (1996), pp. 1002–1046.

[3] D. S. Bernstein, *Some open problems in matrix theory arising in linear systems and control*, Linear Algebra Appl., (1992), pp. 409–432.

[4] V. Blondel, *Simultaneous Stabilization of Linear Systems*, Lecture Notes in Control and Inform. Sci. 191 (1994), Springer-Verlag, London.

[5] V. Blondel, M. Gevers, and A. Lindquist, *Survey on the state of systems and control*, European J. Control, 1 (1995), pp. 5–23.

[6] R. Braatz, P. Young, J. Doyle, and M. Morari, *Computational complexity of mu calculation*, IEEE Trans. Automat. Control, 39 (1994), pp. 1000–1002.

[7] G. E. Coxson and C. L. DeMarco, *The computational complexity of approximating the minimal perturbation scaling to achieve instability in an interval matrix*, Math. Control Signals Systems, 7 (1994), pp. 279–291.

[8] M. R. Garey and D. S. Johnson, *Computers and Intractability: A Guide to the Theory of NP-completeness*, W. H. Freeman, New York, 1979.

[9] R. A. Horn and C. R. Johnson, *Matrix Analysis*, Cambridge University Press, Cambridge, UK, 1985.

[10] R. M. Karp, *Reducibility among combinatorial problems*, in Complexity of Computer Computations, R. E. Miller and J. W. Thatcher, eds., Plenum Press, New York, 1972, pp. 85–103.

[11] A. Nemirovskii, *Several NP-hard problems arising in robust stability analysis*, Math. Control Signals Systems, 6 (1993), pp. 99–105.

[12] M. Overton, *Personal communication*, 1994.

[13] C. H. Papadimitriou and K. Stieglitz, *Combinatorial Optimization: Algorithms and Complexity*, Prentice-Hall, Englewood Cliffs, NJ, 1982.

[14] S. Poljak and J. Rohn, *Checking robust nonsingularity is NP-hard*, Math. Control Signals Systems, 6 (1993), pp. 1–9.

[15] J. N. Tsitsiklis, *Complexity Theoretic Aspects of Problems in Control Theory*, Technical Report LIDS-P-2203, Laboratory for Information and Decision Systems, MIT, Cambridge, MA, 1993.

[16] M. Vidyasagar, *Control System Synthesis*, MIT Press, Cambridge, MA, 1986.

[17] W. M. Wonham, *Linear Multivariable Control: A Geometric Approach*, Springer-Verlag, New York, 1979.

SIAM J. CONTROL OPTIM.
Vol. 35, No. 6, pp. 2128–2136, November 1997

014

ON APPROXIMATE SOLUTIONS IN CONVEX VECTOR OPTIMIZATION*

SIEN DENG†

Abstract. Necessary and sufficient conditions are obtained for the existence of ϵ-weak minima for constrained convex vector optimization problems. The characterization of ϵ-weak minima is given in terms of ϵ-optimal solutions of the associated scalar optimization problems and ϵ-directional derivatives of objective functions. The Lipschitzian continuity of ϵ-weak minima is proved under mild conditions.

Key words. convex vector optimization, ϵ-weak minima, ϵ-directional derivatives, error bounds, Lipschitzian stability

AMS subject classifications. 90C29, 90C25, 90C31

PII. S0363012995292561

1. Introduction.

Consider the following convex vector optimization problem:

$$(\mathcal{P}) \qquad \begin{aligned} &\text{minimize} \quad F(x) \\ &\text{subject to} \quad x \in C, \end{aligned}$$

where $F(x) = (f_1(x), \ldots, f_m(x))^T : \mathbb{R}^n \to \mathbb{R}^m$, each component f_i of F is a finite convex function on $\mathbb{R}^n$, and $C \subset \mathbb{R}^n$ is a nonempty closed convex set.

Much has been said about the existence and stability of solutions of problem $(\mathcal{P})$ [1, 2, 3, 18].

The main theme of this paper is to study approximate solutions of problem $(\mathcal{P})$. Various concepts of approximate solutions were introduced in [4]. In this paper we focus on the existence, the characterization, and the Lipschitzian stability of ϵ-weak minima of problem $(\mathcal{P})$. In section 2, we provide necessary and sufficient conditions for the existence of ϵ-weak minima of problem $(\mathcal{P})$, and characterize ϵ-weak minima of problem $(\mathcal{P})$ in terms of ϵ-optimal solutions of the associated scalar optimization problems. In section 3, we consider a family of parametrized convex vector optimization problems and investigate the behavior of ϵ-weak minima with respect to perturbations on the problem data. Using the results in section 2 and an error bound result of Robinson for convex inequality systems, we obtain the main result of this paper: the Lipschitzian continuity of ϵ-weak minima, which generalizes a similar result in scalar optimization.

We will use the following notation throughout this paper. All vectors are column vectors. The superscript T denotes the transpose of vectors. The vector inequalities $\leq$ or $<$ are assumed to hold pointwise. We denote by $[1, k]$ the set $\{1, 2, \ldots, k\}$. We denote by $\mathbb{B}(x, \rho)$ the closed Euclidean ball in $\mathbb{R}^n$ of radius ρ around a point x. We denote by $\|\cdot\|$ and $\|\cdot\|_\infty$ the Euclidean norms on $\mathbb{R}^n$, $\mathbb{R}^m$, and $\mathbb{R}^l$ and the ∞-norms on $\mathbb{R}^n$, $\mathbb{R}^m$, and $\mathbb{R}^l$, respectively. For a vector x and a nonempty closed set U in some Euclidean space, we denote by $d(x, U) = \min\{\|x - y\| \mid y \in U\}$ the Euclidean distance from any x to U. Given nonempty closed sets U_1 and U_2 in some Euclidean space,

*Received by the editors September 29, 1995; accepted for publication (in revised form) September 24, 1996.

http://www.siam.org/journals/sicon/35-6/29256.html

†Department of Mathematical Sciences, Northern Illinois University, DeKalb, IL 60115 (deng@math.niu.edu).

we define the Hausdorff distance between them as

$$\operatorname{haus}(U_1, U_2) = \max\left\{\sup_{x\in U_1} d(x, U_2), \sup_{x\in U_2} d(x, U_1)\right\}.$$

We denote by S^m the unit-simplex of $\mathbb{R}^m$; that is,

$$S^m = \left\{\mu \in \mathbb{R}^m \mid \mu_i \geq 0, i \in [1, m], \sum_{i=1}^{m} \mu_i = 1\right\}.$$

For a proper lower semicontinuous (l.s.c.) function ϕ on $\mathbb{R}^n$, its *effective* domain is the set

$$\operatorname{dom} \phi = \{x \in \mathbb{R}^n \mid \phi(x) < +\infty\}.$$

For a proper l.s.c. convex function ϕ, we denote by ϕ^* the convex conjugate function of ϕ. Given $\epsilon \geq 0$, for a proper l.s.c. convex function ϕ, and $x \in \operatorname{dom} \phi$, we define the ϵ-subdifferential of ϕ at x as

$$\partial_\epsilon \phi(x) = \left\{y \in \mathbb{R}^n \mid \phi(z) \geq \phi(x) + y^T(z - x) - \epsilon \quad \text{for all } z \in \mathbb{R}^n\right\}. \tag{1}$$

When $\epsilon = 0$, $\partial_0 \phi(x) = \partial\phi(x)$, the usual subdifferential of ϕ at x. Given $\epsilon \geq 0$, for a finite convex function ϕ on $\mathbb{R}^n$, we define its ϵ-directional derivative at x in the direction d as

$$\phi'_\epsilon(x; d) = \inf_t \{(\phi(x + td) - \phi(x) + \epsilon)/t \mid t > 0\}. \tag{2}$$

When $\epsilon = 0$, $\phi'_0 = \phi'$, the usual directional derivative of ϕ. For more about ϵ-subdifferentials, ϵ-directional derivatives, and their applications in scalar optimization, see [5, 6].

2. ϵ-weak minima. We first introduce the definition of ϵ-weak minima of problem $(\mathcal{P})$ and then study basic properties of ϵ-weak minima.

Let $W = \mathbb{R}^m \backslash (-\operatorname{int} \mathbb{R}^m_+)$, where int denotes interior. A vector $y \in C$ is a weak minimum of problem $(\mathcal{P})$ if and only if

$$F(x) - F(y) \in W \quad \text{for all } x \in C.$$

We denote by E the set of all weak minima of problem $(\mathcal{P})$. For various characterizations of the nonemptiness and compactness of E, see [13].

Let $\epsilon \geq 0$, and let $\mathbf{1} \in \mathbb{R}^m$ be the vector all of whose coordinates are 1. A vector $y \in C$ is an ϵ-weak minimum of problem $(\mathcal{P})$ if and only if

$$F(x) + \epsilon\mathbf{1} - F(y) \in W \quad \text{for all } x \in C.$$

We denote by E_ϵ the set of all ϵ-weak minima of problem $(\mathcal{P})$. Note that $E_0 = E$.

PROPOSITION 1. *Consider problem $(\mathcal{P})$. The following statements hold.*

(a) *E_ϵ is a closed set for $\epsilon \geq 0$.*

(b) *Let $\epsilon_k \downarrow 0$. Suppose that $\{x(\epsilon_k)\}$ is a sequence such that $x(\epsilon_k) \in E_{\epsilon_k}$. Then any cluster point of $\{x(\epsilon_k)\}$ is a weak minimum of problem $(\mathcal{P})$.*

Proof. (a) The result is trivially true when E_ϵ is empty. Suppose that $\{x^k\} \subset E_\epsilon$ and $x^k \to y$. Then $y \in C$ by the closedness of C, and

$$F(x) + \epsilon\mathbf{1} - F(x^k) \in W \quad \text{for all } x \in C.$$

The vector $y \in E_\epsilon$ follows from the continuity of F and the closedness of W.

(b) Since $x(\epsilon_k) \in E_{\epsilon_k}$, $F(x) + \epsilon_k \mathbf{1} - F(x(\epsilon_k)) \in W$ for all $x \in C$. Without loss of generality, suppose that $x(\epsilon_k) \to z$. Then, it follows from the closedness of C that z is in C. Since $\epsilon_k \downarrow 0$, by the continuity of F and the closedness of W, we have

$$F(x) - F(z) \in W \quad \text{for all } x \in C.$$

This is what we wanted to prove. □

For $\mu \in S^m$, and $\epsilon \geq 0$, we define the ϵ-optimal solution set of minimizing $\mu^T F(x)$ over all $x \in C$ as

$$\epsilon\text{-argmin}_x \left\{\mu^T F(x) \mid x \in C\right\} = \left\{y \in C \mid \mu^T F(y) \leq \inf_x \left\{\mu^T F(x) \mid x \in C\right\} + \epsilon\right\}.$$

Given below is the first main result of this section, which characterizes the set E_ϵ in terms of ϵ-optimal solutions of scalar optimization problems.

THEOREM 2.1. *Consider problem* $(\mathcal{P})$. *Let* $\epsilon \geq 0$. *A vector* $y \in E_\epsilon$ *if and only if there is some* $\mu \in S^m$ *such that* $y \in \epsilon\text{-argmin}_x \left\{\mu^T F(x) \mid x \in C\right\}$.

Proof. (Necessity) Let $y \in E_\epsilon$. Then $F(x) + \epsilon\mathbf{1} - F(y) \in W$ for all $x \in C$. It follows that

$$\left(F(C) + \mathbb{R}^m_+ + \epsilon\mathbf{1} - F(y)\right) \cap \left(-\text{int } \mathbb{R}^m_+\right) = \emptyset.$$

Since $\left(F(C) + \mathbb{R}^m_+ + \epsilon\mathbf{1} - F(y)\right)$ and $\left(-\text{int } \mathbb{R}^m_+\right)$ are two nonempty disjoint convex sets, by the separation theorem (see Theorem 11.3 of [5]), there is some nonzero $\mu \in \mathbb{R}^m$ such that

$$\mu^T(-d') \leq 0 \quad \text{for all } d' \in \text{int } \mathbb{R}^m_+, \text{ and} \tag{3}$$

$$\mu^T(F(x) + d + \epsilon\mathbf{1} - F(y)) \geq 0 \quad \text{for all } x \in C \text{ and for all } d \in \mathbb{R}^m_+. \tag{4}$$

It follows from (3) that $\mu \in \mathbb{R}^m_+$. Normalizing μ if necessary, we can assume that $\mu \in S^m$. Thus by (4), $\mu^T F(x) \geq \mu^T F(y) - \epsilon$ for all $x \in C$. Therefore, $\inf_x \left\{\mu^T F(x) \mid x \in C\right\}$ is finite and $y \in \epsilon\text{-argmin}_x \left\{\mu^T F(x) \mid x \in C\right\}$.

(Sufficiency) Suppose that $y \in \epsilon\text{-argmin}_x \left\{\mu^T F(x) \mid x \in C\right\}$ for some $\mu \in S^m$, but $y \notin E_\epsilon$. Then there is a vector $z \in C$ such that $F(z) + \epsilon\mathbf{1} - F(y) \in -\text{int } \mathbb{R}^m_+$. Consequently, $\mu^T(F(z) + \epsilon\mathbf{1} - F(y)) < 0$, which implies that $y \notin \epsilon\text{-argmin}_x \left\{\mu^T F(x) \mid x \in C\right\}$. The contradiction completes the proof. □

To obtain necessary and sufficient conditions for the existence of ϵ-weak minima of problem $(\mathcal{P})$, we quote Theorem 1.1.2 and Proposition 1.2.1 in Chapter XI of [6] as the following proposition.

PROPOSITION 2. *Let* ϕ *be a proper l.s.c. convex function. Then the following statements hold.*

(a) *For all* $x \in \text{dom } \phi$, $\partial_\epsilon \phi(x) \neq \emptyset$ *whenever* $\epsilon > 0$.

(b) *For all* $x \in \text{dom } \phi$, *and* $\epsilon \geq 0$, *a vector* $y \in \partial_\epsilon \phi(x)$ *if and only if* $x \in \partial_\epsilon \phi^*(y)$.

THEOREM 2.2. *Consider problem* $(\mathcal{P})$. *For* $\epsilon > 0$, E_ϵ *is nonempty if and only if there is some* $\bar{\mu} \in S^m$ *such that* $\inf_x \left\{\bar{\mu}^T F(x) \mid x \in C\right\}$ *is finite.*

Proof. (Necessity) Since E_ϵ is nonempty, there is a y in E_ϵ. By Theorem 2.1, there is some $\bar{\mu} \in S^m$ such that $y \in \epsilon\text{-argmin}_x \left\{\bar{\mu}^T F(x) \mid x \in C\right\}$. It follows that $\inf_x \left\{\bar{\mu}^T F(x) \mid x \in C\right\}$ is finite.

(Sufficiency) Suppose that there is some $\bar{\mu} \in S^m$ such that $\inf_x \left\{\bar{\mu}^T F(x) \mid x \in C\right\}$ is finite. Let $\psi(x) = \bar{\mu}^T F(x) + \delta_C(x)$, where $\delta_C(\cdot)$ is the indicator function of C. Then ψ is a proper l.s.c. convex function and $\{x \mid 0 \in \partial_\epsilon \psi(x)\} = \epsilon\text{-argmin}_x \left\{\bar{\mu}^T F(x) \mid x \in C\right\}$

(see (1)). Furthermore, for $\epsilon > 0$, by Theorem 2.1, ϵ-$\operatorname{argmin}_x \{\bar{\mu}^T F(x) \mid x \in C\} \subset E_\epsilon$. By Proposition 2, $\{x \mid 0 \in \partial_\epsilon \psi(x)\} = \partial_\epsilon \psi^*(0)$, and $\partial_\epsilon \psi^*(0)$ is nonempty whenever $0 \in \operatorname{dom} \psi^*$. Thus we only have to show that $0 \in \operatorname{dom} \psi^*$. This is true because

$$\begin{aligned}\psi^*(0) &= \sup_x \{0^T x - \psi(x) \mid x \in \mathbb{R}^n\} = \sup_x \{-\bar{\mu}^T F(x) \mid x \in C\} \\ &= -\inf_x \{\bar{\mu}^T F(x) \mid x \in C\} < +\infty.\end{aligned}$$

This completes the proof of the desired result. □

We denote by $(f_i)'_\epsilon(x; d)$ the ϵ-directional derivative of f_i at x in the direction d, where f_i is the ith component of F. The next theorem uses ϵ-directional derivatives of f_i to describe the set E_ϵ.

THEOREM 2.3. *Consider problem* $(\mathcal{P})$. *Let* $\epsilon \geq 0$. *If*

$$((f_1)'_\epsilon(y; x-y), \ldots, (f_m)'_\epsilon(y; x-y))^T \in W \quad \text{for all } x \in C, \tag{5}$$

then $y \in E_\epsilon$. *When* $\epsilon = 0$, *the converse is also true; that is, if* $y \in E$, *then* (5) *holds with* $\epsilon = 0$.

Proof. For $\epsilon \geq 0$, by the definition of ϵ-directional derivative of a convex function (see (2)), we have

$$(F(x) + \epsilon \mathbf{1} - F(y)) - ((f_1)'_\epsilon(y; x-y), \ldots, (f_m)'_\epsilon(y; x-y))^T \in \mathbb{R}^m_+ \quad \text{for all } x \in C.$$

Since $((f_1)'_\epsilon(y; x-y), \ldots, (f_m)'_\epsilon(y; x-y))^T \in W$ for all $x \in C$, it follows that

$$F(x) + \epsilon \mathbf{1} - F(y) \in \mathbb{R}^m_+ + W \subset W \quad \text{for all } x \in C.$$

This proves the first part of the theorem.

When $\epsilon = 0$, by Theorem 23.1 of [5], $(f_i)'_0(y; d) = \lim_{t \downarrow 0} \{(f_i(y + td) - f_i(y))/t\}$ for all $i \in [1, m]$. Suppose that $y \in E$. For any $x \in C$, let $0 < \alpha < 1$. Then $y + \alpha(x - y) \in C$ by the convexity of C. Consequently, we have

$$(F(y + \alpha(x-y)) - F(y))/\alpha \in W \quad \text{for all } \alpha \in (0, 1).$$

The result thus follows from the fact that W is closed and each component f_i of F has one-sided directional derivatives. □

Remark 2.1. When $\epsilon = 0$, Theorem 2.3 is an extension of Theorem 2.1 of [1], where each f_i is differentiable.

3. Lipschitzian stability of ϵ-weak minima. We consider a family of parametrized convex vector optimization problems,

$$(\mathcal{P}(u)) \qquad \begin{aligned} &\text{minimize} \quad F(x) \\ &\text{subject to} \quad G(x) \leq u, \end{aligned}$$

where F is the same as in problem $(\mathcal{P})$, $u \in \mathbb{R}^l$, and $G(x) = (g_1(x), \ldots, g_l(x))^T$ with each component g_i of G being a finite convex function on $\mathbb{R}^n$. For $\epsilon \geq 0$, we denote by $E_\epsilon(u)$ all ϵ-weak minima of problem $(\mathcal{P}(u))$. In this section, our focus is on Lipschitzian stability of the multifunction $E_\epsilon : \mathbb{R}^l \rightrightarrows \mathbb{R}^n$. Recall [7] that a multifunction $\Gamma : \mathbb{R}^l \rightrightarrows \mathbb{R}^n$ is Lipschitzian relative to V, a subset of $\mathbb{R}^l$, if $\Gamma(v)$ is nonempty and compact for every $v \in V$, and there is a positive scalar λ such that

$$\operatorname{haus}(\Gamma(v'), \Gamma(v'')) \leq \lambda \|v' - v''\| \quad \text{for all } v', v'' \in V.$$

We denote by $C(u)$ the solution set of the convex system $G(x) \leq u$, which is a closed convex set. We make the following assumption regarding the convex inequality system $G(x) \leq 0$ throughout this section.

Assumption 3.1. For the convex inequality system $G(x) \leq 0$, suppose that there are some $x^*, \hat{x} \in \mathbb{R}^n$, and some positive scalars δ, Δ such that $\max_{1\leq i\leq l}\{g_i(x^*)\} \leq -2\delta$ (the Slater condition) and $C(0) \subset \mathbb{B}(\hat{x}, \Delta/2)$ (the boundedness condition).

Assumption 3.1 amounts to ensuring that the multifunction $C(\cdot)$ is Lipschitzian near 0. To derive the Lipschitzian continuity of E_ϵ near 0, we need the following form of Robinson's result on error bounds for convex inequality systems (see [8]). For related error bound results, see [12, 15, 16, 17].

PROPOSITION 3. *Consider an inequality system $H(x) \leq 0$, where $H(x) = (h_1(x), \ldots, h_k(x))^T$ and each component h_i of H is a finite convex function on $\mathbb{R}^n$. Let $S = \{x \mid H(x) \leq 0\}$. Suppose that there are some $z^*, \hat{z}$ of $\mathbb{R}^n$, and some positive scalars θ, Θ such that $\max_{1\leq i\leq k}\{h_i(z^*)\} \leq -\theta$ and $S \subset \mathbb{B}(\hat{z}, \Theta/2)$. Then*

$$d(z, S) \leq \theta^{-1}\Theta\|[H(z)]_+\| \quad \text{for all } z \in \mathbb{R}^n,$$

where $[\cdot]_+$ is the positive part of a vector.

Consider problem $(\mathcal{P}(u))$. For $\mu \in \mathbb{R}^m$, $u \in \mathbb{R}^l$, and $\epsilon \geq 0$, define

$$\begin{aligned} p(\mu, u) &= \inf_x \left\{\mu^T F(x) \mid x \in C(u)\right\}, \\ P(\mu, u) &= \operatorname{argmin}_x \left\{\mu^T F(x) \mid x \in C(u)\right\}, \\ P_\epsilon(\mu, u) &= \left\{x \in C(u) \mid \mu^T F(x) \leq p(\mu, u) + \epsilon\right\}. \end{aligned}$$

For $\epsilon > 0$ and $u \in \mathbb{R}^l$, $E_\epsilon(u) = \cup_{\mu\in S^m} P_\epsilon(\mu, u)$ by Theorem 2.1. We will prove that E_ϵ is Lipschitzian near 0 by showing that $P_\epsilon(\mu, \cdot)$ is Lipschitzian near 0 with a uniform Lipschitzian modulus for all $\mu \in S^m$. For this purpose, we need three lemmas.

LEMMA 3.1. *Suppose that Assumption* 3.1 *holds. Let $\mathcal{N}_1 = \{u \in \mathbb{R}^l \mid \|u\| \leq \delta\}$. Then for every $u \in \mathcal{N}_1$, the Slater condition holds for system $G(x) \leq u$, and $C(u) \subset \mathbb{B}(\hat{x}, \Delta)$. In particular, for $u \in \mathcal{N}_1$ and $\epsilon \geq 0$, $E_\epsilon(u)$ is nonempty, and $E_\epsilon(u) \subset \mathbb{B}(\hat{x}, \Delta)$.*

Proof. We first show that for every $u \in \mathcal{N}_1$, the Slater condition holds for the convex inequality system $G(x) \leq u$. Indeed, since $\|u\| \leq \delta$, $\max_{1\leq i\leq l}\{|u_i|\} \leq \delta$. Thus $g_i(x^*) - u_i \leq -\delta$ for $i \in [1, l]$. By applying Proposition 3 to the convex inequality system $G(x) \leq 0$, which has $C(0)$ as the solution set, we have, for any $z \in C(u)$,

$$\begin{aligned} d(z, C(0)) &\leq 1/2\delta^{-1}\Delta\|[G(z)]_+\| \\ &\leq 1/2\delta^{-1}\Delta\|u\| \leq 1/2\Delta. \end{aligned}$$

A straightforward calculation shows that $C(u) \subset \mathbb{B}(\hat{x}, \Delta)$ whenever $u \in \mathcal{N}_1$. This completes the first part of the proof. For $u \in \mathcal{N}_1$ and $\mu \in S^m$, since $C(u)$ is nonempty and compact, $P_\epsilon(\mu, u)$ is nonempty and $P_\epsilon(\mu, u) \subset C(u)$. It follows that $E_\epsilon(u)$ is nonempty, and $E_\epsilon(u) \subset C(u) \subset \mathbb{B}(\hat{x}, \Delta)$. □

For $u \in \mathcal{N}_1$, by the nonemptiness and the compactness of $C(u)$, $p(\mu, u)$ is finite for every $\mu \in S^m$. In view of Lemma 3.1, and by invoking Theorem 3.1 of [7], we obtain the following lemma about the Lipschitzian continuity of the optimal value function p. The proof is given in the appendix.

LEMMA 3.2. *Suppose that Assumption* 3.1 *holds and that $\mathcal{N}_1$ is the same as in Lemma* 3.1. *Then for any $(\mu, u) \in S^m \times \mathcal{N}_1$, the function p is finite and locally Lipschitzian relative to some neighborhood $V(\subset \mathbb{R}^m \times \mathbb{R}^l)$ of (μ, u).*

Thanks to the convexity and compactness of $S^m \times \mathcal{N}_1$, it follows from Lemma 3.2 that p is Lipschitzian on $S^m \times \mathcal{N}_1$; that is, there is a constant $L > 0$, which is independent of (μ, u) in $S^m \times \mathcal{N}_1$, such that

$$|p(\mu,u) - p(\mu',u')| \le L(\|\mu - \mu'\| + \|u - u'\|) \quad \text{for all } (\mu,u),(\mu',u') \in S^m \times \mathcal{N}_1. \tag{6}$$

Also, by the compactness of $S^m \times \mathcal{N}_1$, there is some positive scalar M_1 such that

$$\sup_{(\mu,u)\in S^m\times\mathcal{N}_1} |\mu^T F(x^*) - p(\mu,u)| \le M_1, \tag{7}$$

where x^* is given by Assumption 3.1.

The following lemma, which plays the crucial role for deriving the Lipschitzian continuity of E_ϵ, establishes a "uniform" Slater condition on $P_\epsilon(\mu,u)$.

LEMMA 3.3. *Suppose that Assumption* 3.1 *holds and that* ϵ *is given with* $0 < \epsilon < 2M_1$, *where* M_1 *is given by* (7). *Let* $\mathcal{N}(\epsilon) = \mathcal{N}_1 \cap \mathcal{N}_2$, *where* $\mathcal{N}_2 = \left\{u \in \mathbb{R}^l \mid \|u\|_\infty \le (\epsilon\delta)/(2M_1)\right\}$ *and* $\mathcal{N}_1$ *is the same as in Lemma* 3.1. *Then for any* $(\mu,u) \in S^m \times \mathcal{N}(\epsilon)$, *there is a* $y(\mu,u) \in P_\epsilon(\mu,u)$ *such that*

$$\mu^T F(y(\mu,u)) - p(\mu,u) - \epsilon \le -\tilde{\delta}(\epsilon),\ g_1(y(\mu,u)) - u_1 \le -\tilde{\delta}(\epsilon), \ldots, g_l(y(\mu,u)) - u_l \le -\tilde{\delta}(\epsilon),$$

where $\tilde{\delta}(\epsilon) = \min\{(\epsilon\delta)/(2M_1), \epsilon/2\}$.

Proof. Suppose that ϵ is given with $0 < \epsilon < 2M_1$. For $\mu \in S^m$ and $u \in \mathcal{N}(\epsilon)$, $P(\mu,u)$ is nonempty since $C(u)$ is nonempty and compact. Choose a vector $x(\mu,u) \in P(\mu,u)$, and let x^* be given by Assumption 3.1. Let $0 < \alpha < 1$. Then $\alpha x(\mu,u) + (1-\alpha)x^* \in C(u)$. By the convexity of g_i and $\mu^T F$, we have, for $i \in [1,l]$,

$$\begin{aligned} g_i(\alpha x(\mu,u) + (1-\alpha)x^*) - u_i &\le \alpha g_i(x(\mu,u)) + (1-\alpha)g_i(x^*) - u_i \\ &\le \alpha\left(g_i(x(\mu,u)) - u_i\right) + (1-\alpha)\left(g_i(x^*) - u_i\right) \\ &\le (1-\alpha)(-2\delta + \delta) = (1-\alpha)(-\delta) \end{aligned} \tag{8}$$

and

$$\begin{aligned} \mu^T F(\alpha x(\mu,u) + (1-\alpha)x^*) - p(\mu,u) - \epsilon &\le \alpha\mu^T F(x(\mu,u)) + (1-\alpha)\mu^T F(x^*) \\ &\quad - p(\mu,u) - \epsilon \\ &= (1-\alpha)\left(\mu^T F(x^*) - p(\mu,u)\right) - \epsilon \\ &\le (1-\alpha)M_1 - \epsilon \qquad \text{(by (7))}. \end{aligned} \tag{9}$$

Letting $\alpha = 1 - \epsilon/(2M_1)$ and $y(\mu,u) = (1-\epsilon/(2M_1))\,x(\mu,u) + (\epsilon/(2M_1))\,x^*$, we have, in view of (8) and (9),

$$\begin{gathered} g_i(y(\mu,u)) - u_i \le (1-\alpha)(-\delta) = -(\epsilon\delta)/(2M_1) \le -\tilde{\delta}(\epsilon) \quad \text{for } i \in [1,l], \text{ and} \\ \mu^T F(y(\mu,u)) - p(\mu,u) - \epsilon \le (1-\alpha)M_1 - \epsilon \le -\epsilon/2 \le -\tilde{\delta}(\epsilon). \end{gathered}$$

This completes the proof. ∎

We observe that the positive scalar $\tilde{\delta}(\epsilon)$ in Lemma 3.3 is independent of μ. This enables us to obtain the main result of this paper.

THEOREM 3.4. *Suppose that Assumption* 3.1 *holds and that* ϵ *is given with* $0 < \epsilon < 2M_1$, *where* M_1 *is given by* (7). *Let* $\mathcal{N}(\epsilon)$ *be the same as in Lemma* 3.3 *according to the given* ϵ. *Then the multifunction* $E_\epsilon(\cdot)$ *is Lipschitzian relative to* $\mathcal{N}(\epsilon)$, *a closed*

neighborhood of 0, *with the Lipschitzian modulus* $M = 2(\tilde{\delta}(\epsilon))^{-1}\Delta\sqrt{L^2+1}$, *where* $\tilde{\delta}(\epsilon)$ *is the same as in Lemma* 3.3 *and* L *is given by* (6). *That is,*

$$\text{haus}(E_\epsilon(u), E_\epsilon(v)) \le M\|u-v\| \quad \textit{for all } u, v \in \mathcal{N}(\epsilon).$$

Proof. Throughout the proof, suppose that ϵ is given with $0 < \epsilon < 2M_1$. For $u \in \mathcal{N}(\epsilon)$, $\mu \in S^m$, $\mathcal{P}_\epsilon(\mu, u)$ is nonempty since $C(u)$ is nonempty and compact. By Theorem 2.1, $E_\epsilon(u) = \cup_{\mu\in S^m} P_\epsilon(\mu, u)$ for every $u \in \mathcal{N}(\epsilon)$. The result follows if we can show that, for every $\mu \in S^m$,

$$\text{haus}(P_\epsilon(\mu, v), P_\epsilon(\mu, u)) \le M\|u-v\| \quad \text{for all } u, v \in \mathcal{N}(\epsilon).$$

Indeed, for given $u, v \in \mathcal{N}(\epsilon)$, let $z \in P_\epsilon(\mu, v)$. Applying Proposition 3 along with Lemmas 3.1 and 3.3 to the convex inequality system $\mu^T F(x) \le p(\mu, u) + \epsilon,\ g_1(x) \le u_1, \ldots, g_l(x) \le u_l$, which has $P_\epsilon(\mu, u)$ as the solution set, we obtain an upper bound for $d(z, P_\epsilon(\mu, u))$.

$$\begin{aligned}
d(z, P_\epsilon(\mu, u)) &\le 2(\tilde{\delta}(\epsilon))^{-1}\Delta \left\| \left[\left(\mu^T F(z) - p(\mu, u) - \epsilon, g_1(z) - u_1, \ldots, g_l(z) - u_l\right)^T \right]_+ \right\| \\
&\le 2(\tilde{\delta}(\epsilon))^{-1}\Delta \left\| \left[\left(p(\mu, v) - p(\mu, u), v_1 - u_1, \ldots, v_l - u_l\right)^T \right]_+ \right\| \\
&\le 2(\tilde{\delta}(\epsilon))^{-1}\Delta \left\| \left[\left(L\|u-v\|, v_1 - u_1, \ldots, v_l - u_l\right)^T \right]_+ \right\| \qquad \text{(by (6))} \\
&\le 2(\tilde{\delta}(\epsilon))^{-1}\Delta\sqrt{L^2+1}\|u-v\|.
\end{aligned}$$

Thus $\sup_{z\in P_\epsilon(\mu,v)} d(z, P_\epsilon(\mu, u)) \le M\|u-v\|$. Since u, v are any two vectors in $\mathcal{N}(\epsilon)$, we establish the theorem. □

Remark 3.1. A similar result in scalar optimization can be found in [9] and [14]. The difficulties in convex vector optimization are to show that the constant M is independent of μ, which is trivial in the scalar optimization case.

4. Appendix. For a multifunction $\Gamma : \mathbb{R}^d \rightrightarrows \mathbb{R}^n$, we shall use the notation

$$\Gamma(V) = \cup_{v\in V}\Gamma(v) \quad \text{for any } V \subset \mathbb{R}^d.$$

We shall say that Γ is locally bounded at $\bar{v}$ if there is a neighborhood V of $\bar{v}$ such that the set $\Gamma(V)$ is bounded (see [7]). To prove Lemma 3.2, Theorem 3.1 of [7] is quoted.

THEOREM 4.1. *Let* $h : \mathbb{R}^d \times \mathbb{R}^n \to \mathbb{R}$ *and* $H : \mathbb{R}^d \times \mathbb{R}^n \to \mathbb{R}^l$ *be locally Lipschitzian functions. Let*

$$\begin{aligned}
\psi(v) &= \inf_x \left\{ h(v, x) \mid H(v, x) \in \tilde{C},\ (v, x) \in \tilde{D} \right\}, \\
\Psi(v) &= \text{argmin}_x \left\{ h(v, x) \mid H(v, x) \in \tilde{C},\ (v, x) \in \tilde{D} \right\},
\end{aligned}$$

where $\tilde{C}$ *and* $\tilde{D}$ *are closed. Let* $\bar{v}$ *be a point at which* $\psi(\bar{v})$ *is finite and suppose that for some* $\beta > \psi(\bar{v})$ *the multifunction*

$$\Psi_\beta(v) = \left\{ x \in \mathbb{R}^n \mid h(v, x) \le \beta,\ H(v, x) \in \tilde{C},\ (v, x) \in \tilde{D} \right\} \quad \textit{is locally bounded at } \bar{v}. \tag{10}$$

Suppose further that the following constraint qualification holds for every $\bar{x} \in \Psi(\bar{v})$: *the only vectors* $y \in \mathbb{R}^l$ *and* $z \in \mathbb{R}^d$ *such that*

$$y \in N_{\tilde{C}}(H(\bar{v},\bar{x})) \quad \textit{and} \quad (z,0) \in y^T \partial H(\bar{v},\bar{x}) + N_{\tilde{D}}(\bar{v},\bar{x}) \quad \textit{are } y = 0, z = 0, \tag{11}$$

where the set $\partial H(\bar{v},\bar{x})$ *denotes the Clarke generalized Jacobian of* H *at* $(\bar{v},\bar{x})$ (*see p.* 69 *of* [10]).

Then, relative to some neighborhood V *of* $\bar{v}$, ψ *is finite and locally Lipschitzian.*

Proof of Lemma 3.2. Let $v = (\mu, u)$, $h(v,x) = \mu^T F(x)$, $H(v,x) = G(x) - u$, $d = m + l$, $\tilde{C} = \mathbb{R}^l_-$, and $\tilde{D} = \mathbb{R}^m \times \mathbb{R}^l \times \mathbb{R}^n$. Then

$$p(\mu,u) = \inf_x \left\{ \mu^T F(x) \mid G(x) - u \in \mathbb{R}^l_- \right\} = \psi(v), \quad \text{and}$$
$$P(\mu,u) = \operatorname{argmin}_x \left\{ \mu^T F(x) \mid G(x) - u \in \mathbb{R}^l_- \right\} = \Psi(v).$$

For $(\bar{\mu},\bar{u}) \in S^m \times \mathcal{N}_1$, we will show that p is locally Lipschitzian at $(\bar{\mu},\bar{u})$. In view of Lemma 3.1, for $(\bar{\mu},\bar{u}) \in S^m \times \mathcal{N}_1$, it is easy to see that $p(\bar{\mu},\bar{u})$ is finite and (10) holds (since $C(\cdot)$ is locally bounded at $\bar{u}$). Thus we only have to verify that (11) holds. Indeed, for every $\bar{x} \in P(\bar{\mu},\bar{u})$, the fact that $y \in N_{\mathbb{R}^l_-}(G(\bar{x}) - \bar{u})$ implies for $i \in [1,l]$

$$y_i = \begin{cases} 0 & \text{if } g_i(\bar{x}) - \bar{u}_i < 0, \\ \geq 0 & \text{if } g_i(\bar{x}) - \bar{u}_i = 0. \end{cases} \tag{12}$$

By Proposition 2.6 of [11], $y^T \partial H(\bar{v},\bar{x}) = \partial_{(v,x)} \left(y^T H(\bar{v},\bar{x}) \right) = \left(0, -y, \partial_x \left(y^T G(\bar{x}) \right) \right)$. Since $N_{\mathbb{R}^m \times \mathbb{R}^l \times \mathbb{R}^n}(\bar{v},\bar{x}) = \{0\}$, we only have to show that the only vector $y \in \mathbb{R}^l$ such that

$$y \in N_{\mathbb{R}^l_-}(G(\bar{x}) - \bar{u}) \text{ and } 0 \in \partial_x \left(y^T G(\bar{x}) \right) \tag{13}$$

is $y = 0$. Suppose to the contrary that there is a nonzero y satisfying (13). By (12), $y^T G(\cdot)$ is convex, and $y_i \geq 0$ if $g_i(\bar{x}) - \bar{u}_i = 0$ for $i \in [1,l]$. Let $I_0 = \{ i \in [1,l] \mid g_i(\bar{x}) = \bar{u}_i \}$ and x^* be given by Assumption 3.1. By Lemma 3.1, $g_i(x^*) - \bar{u}_i < 0$ for all $i \in [1,l]$. It follows that

$$y^T (G(x^*) - G(\bar{x})) = \sum_{i \in I_0} y_i (g_i(x^*) - \bar{u}_i) < 0,$$

which contradicts $0 \in \partial_x \left(y^T G(\bar{x}) \right)$. Therefore, by Theorem 4.1, p is locally Lipschitzian at $(\bar{\mu},\bar{u})$. This completes the proof.

Acknowledgment. The author thanks an anonymous referee for making a careful reading of this paper and offering many insightful comments on the presentation.

REFERENCES

[1] G. Y. Chen and B. D. Craven, *Existence and continuity of solutions for vector optimization,* J. Optim. Theory Appl., 81 (1994), pp. 459–467.
[2] T. Tanino, *Stability and sensitivity analysis in convex vector optimization,* SIAM J. Control Optim., 26 (1988), pp. 521–536.
[3] Y. Sawaragi, H. Nakayama, and T. Tanino, *Theory of Multiobjective Optimization,* Academic Press, New York, 1985.
[4] D. J. White, *Epsilon efficiency,* J. Optim. Theory Appl., 49 (1986), pp. 319–337.
[5] R. T. Rockafellar, *Convex Analysis,* Princeton University Press, Princeton, NJ, 1970.

[6] J.-B. Hiriart-Urruty and C. Lemarechal, *Convex Analysis and Minimization Algorithms,* Springer-Verlag, Berlin, Heidelberg, 1993.
[7] R. T. Rockafellar, *Lipschitzian properties of multifunctions,* Nonlinear Anal., 9 (1985), pp. 867–885.
[8] S. M. Robinson, *An application of error bounds for convex programming in a linear space,* SIAM J. Control, 13 (1975), pp. 271–273.
[9] A. A. Auslender and J.-P. Crouzeix, *Global regularity theorems,* Math. Oper. Res., 13 (1988), pp. 243–253.
[10] F. H. Clarke, *Optimization and Nonsmooth Analysis,* John Wiley, New York, 1983.
[11] R. T. Rockafellar, *Extensions of subgradient calculus with applications to optimization,* Nonlinear Anal., 9 (1985), pp. 665–698.
[12] J. V. Burke and P. Tseng, *A unified analysis of Hoffman's bound via Fenchel duality*, SIAM J. Optim., 6 (1996), pp. 265–282.
[13] S. Deng, *Characterizations of the nonemptiness and compactness of solution sets in convex vector optimization,* J. Optim. Theory Appl., to appear.
[14] D. Klatte and B. Kummer, *Stability properties of infima and optimal solutions of parametric optimization problems,* in Proc. IIASA Workshop on Nondifferentiable Optimization, Sopron, Hungary, 1984.
[15] S. M., Robinson, *Regularity and stability for convex multivalued functions,* Math. Oper. Res., 1 (1976), pp. 131–143.
[16] Z. Q. Luo and J. S. Pang, *Error bounds for analytic systems and their applications,* Math. Programming, 67 (1995), pp. 1–28.
[17] Z. Q. Luo and X. D. Luo, *Extension of Hoffman's error bound to polynomial systems,* SIAM J. Optim., 4 (1994), pp. 383–392.
[18] P. L. Yu, *Multiple-Criteria Decision Making: Concepts, Techniques, and Extensions,* Plenum Press, New York, 1985.

SIAM J. CONTROL OPTIM.
Vol. 35, No. 6, pp. 2137–2156, November 1997

015

OPTIMAL CONTROL OF LINEAR PERIODIC RESONANT SYSTEMS IN HILBERT SPACES*

VIOREL BARBU†

Abstract. This work concerns existence, the maximum principle, and synthesis for optimal control problems with linear periodic dynamics in Hilbert spaces in the resonant case. Applications to distributed and boundary periodic control problems are given.

Key words. distributed optimal control problems, maximum principle, closed range, p-stabilizable

AMS subject classifications. 93B50, 93C35, 93C05

PII. S036301299529478X

1. Introduction.

In this work we shall study the optimal control problem

$$\text{minimize} \quad \int_0^T (g(Cy(t)) + h(u(t)))dt \tag{1.1}$$

subject to $u \in L^2(0,T;U)$ and $y \in C([0,T];H)$ satisfying the state system

$$\begin{aligned} &\frac{dy}{dt} + Ay = Bu + f, \qquad t \in (0,T), \\ &y(0) = y(T). \end{aligned} \tag{1.2}$$

Here H, U, and Z are real Hilbert spaces, $-A$ is the infinitesimal generator of a C_0 semigroup e^{-At} on H, $B \in L(U,H)$, $C \in L(H,Z)$, $g : Z \to \bar{R} = (-\infty, +\infty]$, and $h : U \to \bar{R}$ are lower semicontinuous convex functions. The solution y to state system (1.1) is considered in the "mild sense," i.e.,

$$y(t) = e^{-At}y(T) + \int_0^t e^{-A(t-s)}(Bu(s) + f(s))ds \quad \forall t \in [0,T]. \tag{1.2}'$$

This is the general framework for representing distributed optimal control problems. In the resonant case, i.e., when the null space of the operator $\frac{d}{dt}+A$ with periodic condition is not trivial, this is a singular optimal control problem which so far was studied under the key assumptions that the pair (A,B) is stabilizable and (A,C) is detectable (see, e.g., [5], [7], [8]). However, though these assumptions are natural for infinite horizon problems, they are too restrictive for periodic control problems. It is our aim here to relax them to a closed range hypothesis which will be discussed in some details in section 2 below. It will be in this framework that we shall treat in sections 3 and 5 the existence, the Pontryagin maximum principle, and the synthesis of optimal controllers. In section 4 we shall briefly discuss the corresponding boundary control problem, while section 5 is devoted to construction of optimal feedback controllers for the linear quadratic problem. In section 6 we shall study the optimal control problem (1.1) with the dynamics

$$y'' + Ay = Bu + f, \quad t \in (0,T); \quad y(0) = y(T), \quad y'(0) = y'(T), \tag{1.3}$$

where A is a self-adjoint, positive definite operator on H and $B \in L(U,H)$.

*Received by the editors November 15, 1995; accepted for publication (in revised form) September 24, 1996.

http://www.siam.org/journals/sicon/35-6/29478.html

†University of Iaşi and Institute of Mathematics of Romanian Academy, University of Iaşi, 6600 Iaşi, Romania (barbu@uaic.ro).

We shall use the standard notations for the spaces of vector-valued functions on the interval [0,T]. In particular, $W^{1,2}([0,T];H)$ is the space of absolutely continuous functions $y:[0,T]\to H$ such that $y'=\frac{dy}{dt}\in L^2(0,T;H)$. The norms and the scalar products of H, U, Z are denoted by $|\cdot|$, $|\cdot|_U$, $|\cdot|_Z$ and $(\cdot,\cdot)$, $(\cdot,\cdot)_U$, $(\cdot,\cdot)_Z$, respectively. Given the lower semicontinuous, convex function φ on the Hilbert space X, we shall denote by $\partial\varphi$ the subdifferential of φ and by φ^* the conjugate of φ, i.e.,

$$\partial\varphi(y)=\{z\in H;\varphi(y)\le\varphi(p)+(z,y-p)\quad\forall p\in H\},$$

$$\varphi^*(p)=\sup\{(y,p)-\varphi(y);y\in H\}.$$

By φ_ε we denote the Moreau–Brezis regularization of φ, i.e.,

$$\varphi_\varepsilon(x)=\inf\{(2\varepsilon)^{-1}|x-y|^2+\varphi(y);\ y\in X\}. \tag{1.4}$$

(We refer to [1] and [4] for basic results on convex analysis relevant to this paper.) Given a linear, densely defined operator W on a Banach space we shall denote by $D(W)$ the domain of W and by $R(W)$ its range. The dual operator will be denoted by W^*. In a slightly different form part of the results of this paper were presented in [3].

2. Weak solutions and the closed range property. Let $\mathcal{A}$ be the linear operator defined in $L^2(0,T;H)$ as

$$\mathcal{A}y=f \tag{2.1}$$

if and only if

$$\int_0^T((y(t),\varphi'(t)-A^*\varphi(t))+(f(t),\varphi(t)))dt=0$$

for all $\varphi\in W^{1,2}([0,T];H)$ such that $A^*\varphi\in L^2(0,T;H);\varphi(0)=\varphi(T)$. A function $y\in L^2(0,T;H)$ satisfying (2.1) is called a *weak solution* to the periodic problem

$$\frac{dy}{dt}+Ay=f;\qquad y(0)=y(T). \tag{2.2}$$

It is readily seen that the operator $\mathcal{A}$ is closed and densely defined in $L^2(0,T;H)$. Moreover, the dual operator $\mathcal{A}^*$ is defined as

$$\mathcal{A}^*z=g \tag{2.3}$$

if and only if

$$\int_0^T((z(t),\varphi'(t)+A\varphi(t))-(\varphi(t),g(t)))dt=0 \tag{2.4}$$

for all $\varphi\in W^{1,2}([0,T];H)$ such that $A\varphi\in L^2(0,T;H),\varphi(0)=\varphi(T)$.

Let $N(\mathcal{A})$ and $N(\mathcal{A}^*)$ be the null spaces of $\mathcal{A}$ and $\mathcal{A}^*$, respectively. If $R(\mathcal{A})$ (the range of $\mathcal{A}$) is closed in $L^2(0,T;H)$, then by virtue of the closed range theorem (see, e.g., [19, p. 205]), so is $R(\mathcal{A}^*)$ and

$$L^2(0,T;H)=R(\mathcal{A})\oplus N(\mathcal{A}^*)=R(\mathcal{A}^*)\oplus N(\mathcal{A}). \tag{2.5}$$

This means that for each $f \in R(\mathcal{A})$ the solutions y to equation $\mathcal{A}y = f$ are expressed as $y = y_1 + N(\mathcal{A})$ where $y_1 \in R(\mathcal{A}^*)$ is uniquely defined. We define $\mathcal{A}^{-1}f = y_1$ and note that by the closed graph theorem, $\mathcal{A}^{-1} \in L(R(\mathcal{A}), L^2(0,T;H))$. The operator $(\mathcal{A}^*)^{-1} \in L(R(\mathcal{A}^*), L^2(0,T;H))$ is similarly defined.

Proposition 1 below is related to some earlier result of Prüss [15] (see also Haraux [9]).

PROPOSITION 1. *Assume that for each $m \in Z$, the range Y_m of $\mu_m iI + A$ is closed in H and*

$$\sup\{\|(\mu_m iI + A)^{-1}\|_{L(Y_m,H)}; m \in Z\} < \infty, \tag{2.6}$$

where $\mu_m = 2m\pi/T$. Then $R(\mathcal{A})$ is closed in $L^2(0,T;H)$.

Here we have again denoted A the realization of operator A into the complexified space H.

Proof. If $f \in R(\mathcal{A})$, then there is a $y \in L^2(0,T;H)$ such that

$$y(t) = \sum_m y_m \exp(\mu_m it), \qquad t \in (0,T), \tag{2.7}$$

where

$$y_m = (\mu_m i + A)^{-1} f_m, \qquad f_m = T^{-\frac{1}{2}} \int_0^T \exp(-\mu_m it) f(t)dt,$$

and so by (2.6) and Parseval's identity we get

$$\|y\|_{L^2(0,T;H)} \le C\|f\|_{L^2(0,T;H)}$$

where $\mathcal{A}y = f$. This implies that $R(\mathcal{A})$ is closed in $L^2(0,T;H)$, as claimed.

Let $\mathcal{A}_0 : D(\mathcal{A}_0) \subset L^2(0,T;H) \to L^2(0,T;H)$ be the linear operator defined as

$$\mathcal{A}_0 y = f \tag{2.8}$$

if and only if

$$y(t) = e^{-At}y(T) + \int_0^t e^{-A(t-s)} f(s)ds, \qquad t \in (0,T).$$

In other words, $\mathcal{A}_0 y = f$ if and only if y is continuous and it is a "mild" periodic solution to (2.2). It is easily seen that $\mathcal{A}_0$ is itself closed and densely defined in $L^2(0,T;H)$. Moreover, a simple integration by parts shows that $\mathcal{A}_0 \subset \mathcal{A}$. As a matter of fact we have the following proposition.

PROPOSITION 2. $\mathcal{A}_0 = \mathcal{A}$.

Proof. Since, as noticed earlier, the inclusion $\mathcal{A}_0 \subset \mathcal{A}$ is immediate, we confine ourselves to checking that $\mathcal{A} \subset \mathcal{A}_0$. Let $(y,f) \in \mathcal{A}$. We have

$$y(t) = \sum_{m \in Z} y_m \exp(\mu_m it) \text{ in } L^2(0,T;H); \qquad (\mu_m i + A)y_m = f_m. \tag{2.9}$$

Then the sequence

$$y_N(t) = \sum_{|m| \le N} y_m \exp(i\mu_m t)$$

is convergent to y in $L^2(0,T;H)$, and for each N, y_N is a mild solution to (2.2) where $f = f_N = \sum_{|m|\le N} f_m \exp(i\mu_m t)$. Hence

$$y_N(t) = e^{-A(t-s)}y_N(s) + \int_s^t e^{-A(t-r)}f_N(r)dr, \qquad 0 < s < t < T, \tag{2.10}$$

$$y_N(0) = y_N(T).$$

Since $y_N \to y$ and $f_N \to f$ in $L^2(0,T;H)$ and a.e. on $(0,T)$ (on some subsequence), we infer by (2.10) that $\{y_N(T)\}$ is strongly convergent in H to some y_1 and therefore $y_N(t)$ is uniformly convergent to $y(t) \in C([0,T];H)$ and $\mathcal{A}y = f$ as claimed.

By Proposition 2 we have

$$R(\mathcal{A}) = \left\{ f \in L^2(0,T;H); \int_0^T e^{-A(T-t)}f(t)dt \in R(I - e^{-AT}) \right\}, \tag{2.11}$$

$$N(\mathcal{A}) = \{y \in L^2(0,T;H); y(t) = e^{-At}y_0, (I - e^{-AT})y_0 = 0\}. \tag{2.12}$$

Moreover, the dual operator $\mathcal{A}^*$ is given by $\mathcal{A}^*z = g$ if and only if

$$z(t) = e^{-A^*(T-t)}z(0) + \int_t^T e^{-A^*(s-t)}g(s)ds \quad \forall t \in [0,T]. \tag{2.13}$$

PROPOSITION 3. *$R(\mathcal{A})$ is closed in $L^2(0,T;H)$ if and only if $R(I - e^{-AT})$ is closed in H.*

Proof. If $R(I - e^{-AT})$ is closed in H, then by (2.11) we see that $R(\mathcal{A})$ is closed in $L^2(0,T;H)$. Assume now that $R(\mathcal{A})$ is closed and consider the linear subspace of H,

$$X = \{x \in H; (e^{-At}x) \in R(\mathcal{A})\}.$$

(Here we have denoted by $(e^{-At}x)$ the function $t \longrightarrow e^{-At}x$.) We have

$$X = R(I - e^{-AT}). \tag{2.14}$$

Here is the argument. If $x \in R(I - e^{-AT})$, then $Te^{-AT}x \in R(I - e^{-AT})$, and so the equation

$$(I - e^{-AT})y_0 = Te^{-AT}x$$

has at least one solution $y_0 \in H$. Then the function

$$y(t) = e^{-At}y_0 + \int_0^t e^{-A(t-s)}e^{-As}xds = e^{-At}y_0 + te^{-At}x$$

is a solution to $\mathcal{A}y = e^{-At}x$, i.e., $x \in X$. Now let x be in X and let $y(t) = e^{-At}y(0) + te^{-At}x$ be a solution to $\mathcal{A}y = e^{-At}x$. Since $y(0) = y(T)$ the latter implies that $e^{-AT}x \in R(I - e^{-AT})$, and therefore $x \in R(I - e^{-AT})$. Since X is closed it follows from (2.14) that so is $R(I - e^{-AT})$.

COROLLARY 1. *If $R(\mathcal{A})$ is closed in $L^2(0,T;H)$, then $\mathcal{A}^{-1}f \in C([0,T];H)$ for each $f \in R(\mathcal{A})$ and*

$$\|\mathcal{A}^{-1}f\|_{C([0,T];H)} \le C\|f\|_{L^1(0,T;H)} \quad \forall\ f \in R(\mathcal{A}). \tag{2.15}$$

Proof. Since $R(\mathcal{A})$ is closed, so is $R(I - e^{-AT})$, and we have therefore

$$\mathcal{A}^{-1}f(t) = e^{-At}(I - e^{-AT})^{-1}\int_0^T e^{-A(T-t)}f(t)dt + \int_0^t e^{-A(t-s)}f(s)ds$$

$\forall t \in [0,T]$. Recalling that $(I - e^{-AT})^{-1}$ is continuous on $R(I - e^{-AT})$ the latter implies (2.15), as desired.

By the Riesz–Fredholm theory we also have the following corollary.

COROLLARY 2. *If e^{-AT} is compact, then $R(\mathcal{A})$ is closed and $N(\mathcal{A})$, $N(\mathcal{A}^*)$ are finite-dimensional.*

Given $F \in L(H,U)$ we shall denote by $\mathcal{A}_F$ the operator $\mathcal{A} + BF$ defined from $L^2(0,T;H)$ to itself and by $\mathcal{A}_F^* = \mathcal{A}^* + F^*B^*$ its dual.

DEFINITION 1. *The pair (A,B) is said to be p-stabilizable if there is an $F \in L(H,U)$ such that $R(\mathcal{A}_F)$ is closed in $L^2(0,T;H)$ and $N(\mathcal{A}_F^*)$ is finite-dimensional.*

By virtue of Proposition 3 and of (2.12), the pair (A,B) is p-stabilizable if and only if there is an $F \in L(H,U)$ such that $R(I - e^{-(A+BF)T})$ is closed in H and dim $N(I - e^{-(A^*+F^*B^*)T}) < \infty$. In particular, this happens if either e^{-AT} is compact in H or if the pair (A,B) is stabilizable, i.e., there is an $F \in L(H,U)$ such that $A+BF$ generates an exponentially stable semigroup.

DEFINITION 2. *The pair (A,C) is said to be p-detectable if there is a $K \in L(Z,H)$ such that $R(\mathcal{A}_K)$ is closed in $L^2(0,T;H)$ and* dim $N(\mathcal{A}_K) < \infty$.

Here $\mathcal{A}_K = \mathcal{A} + KC$.

Throughout this paper, by solution y to the state equation (1.2), we mean a weak solution, i.e., $\mathcal{A}y = Bu + f$.

3. Existence and the maximum principle. We shall study first the existence in problem (1.1) under the following assumptions.

(i) The pair (A,C) is p-detectable.

(ii) $g : Z \to \bar{R}, h : U \to \bar{R}$ are convex and lower semicontinuous and

$$g(z) \geq \alpha|z|_Z + \beta \quad \forall z \in Z, \tag{3.1}$$

$$h(u) \geq \omega|u|_U^2 + \gamma \quad \forall u \in U, \tag{3.2}$$

where $\alpha, \omega > 0$ and $\beta, \gamma \in R$.

THEOREM 1. *Assume that there is at least one admissible pair (y,u) in problem* (1.1). *Then, under hypotheses* (i), (ii), *problem* (1.1) *has at least one solution, $(y^*,u^*) \in C([0,T];H) \times L^2(0,T;U)$.*

Proof. Let $(y_n,u_n) \in C([0,T];H) \times L^2(0,T;U)$ be such that $\mathcal{A}y_n = Bu_n + f$ and

$$\inf(1.1) = d \leq \int_0^T (g(Cy_n(t)) + h(u_n(t)))dt \leq d + n^{-1}. \tag{3.3}$$

By (3.1), (3.2) we have

$$\|Cy_n\|_{L^1(0,T;H)} + \|u_n\|_{L^2(0,T;U)} \leq C_1. \tag{3.4}$$

By (i) there is a $K \in L(Z,H)$ such that $R(\mathcal{A}_K)$ is closed ($\mathcal{A}_K = \mathcal{A} + KC$) and dim $N(\mathcal{A}_K) < \infty$. We have

$$\mathcal{A}_K y_n = Bu_n + KCy_n + f \tag{3.5}$$

and set $y_n = y_n^1 + y_n^2$, where $y_n^1 = \mathcal{A}_K^{-1}(Bu_n + KCy_n + f) \in R(\mathcal{A}_K^*)$ and $y_n^2 \in N(\mathcal{A}_K)$. Then by (2.15) and (3.4) we have

$$\|y_n^1\|_{C([0,T];H)} \leq C_2 \quad \forall n \in N. \tag{3.6}$$

On the other hand, by the closed range theorem we know that

$$N(\mathcal{A}_K) = N(\mathcal{C}_K) \oplus R(\mathcal{C}_K^*).$$

We have denoted by $\mathcal{C}_K \in L(N(\mathcal{A}_K), L^2(0,T;Z))$ the operator $y \longrightarrow Cy$ restricted to $N(\mathcal{A}_K)$. Since $N(\mathcal{A}_K)$ is finite-dimensional, $\mathcal{C}_K$ has closed range in $L^2(0,T;Z)$, and because $\{Cy_n^2\}$ if bounded in $L^1(0,T;Z)$, it is bounded in $L^2(0,T;Z)$ as well. We have, therefore,

$$y_n^2 = z_n^1 + z_n^2,$$

where $\{z_n^1\}$ is bounded in $L^2(0,T;H)$ and $Cz_n^2 = 0$ a.e. in $(0,T)$. We may assume, therefore, that the sequence $\{y_n^1 + z_n^1\}$ is weakly compact in $L^2(0,T;H)$ and on a subsequence again denoted $\{n\}$ we have

$$\begin{array}{lcll} u_n & \longrightarrow & u^* & \text{weakly in } L^2(0,T;U), \\ y_n^1 + z_n^1 & \longrightarrow & y^* & \text{weakly in } L^2(0,T;H). \end{array}$$

Recalling that $\mathcal{A}(y_n^1 + z_n^1) = Bu_n + f$ we infer that $\mathcal{A}y^* = Bu^* + f$, and since the convex integrand is weakly lower semicontinuous we get

$$d = \int_0^T (g(Cy^*(t)) + h(u^*(t)))dt; \tag{3.7}$$

i.e., (y^*, u^*) is optimal in problem (1.1). This completes the proof.

In order to get the maximum principle for problem (1.1) we shall use the following assumptions.

(j) The pair (A, B) is p-stabilizable.

(jj) The function $g : Z \to R$ is convex and continuous, $h : U \to \bar{R}$ is convex and lower semicontinuous, int $D(h) \neq \emptyset$.

(jjj) The function f is in $C([0,T];H)$ and one of the following two conditions hold:

$(\text{jjj})_1$ $D(h) = U$ and h is bounded on every bounded subset of U.

$(\text{jjj})_2$ $f(t) = Bf_0(t)$ where $f_0 \in C([0,T];U)$ and $-f_0(t) \in \text{int}D(h)\, \forall t \in [0,T]$.

Here $D(h) = \{u \in U; h(u) < +\infty\}$ is the effective domain of h and int stands for interior.

THEOREM 2. *Assume that hypotheses* (j), (jj), (jjj) *hold. Then the pair* $(y^*, u^*) \in C([0,T];H) \times L^2(0,T;U)$ *is optimal in problem* (1.1) *if and only if there are* $p \in C([0,T];H)$ *and* $\eta \in L^\infty(0,T;Z)$ *such that*

$$\frac{dy^*}{dt} + Ay^* = Bu^* + f \text{ in } (0,T); \qquad y^*(0) = y^*(T), \tag{3.8}$$

$$\frac{dp}{dt} - A^*p = C^*\eta \text{ in } (0,T); \qquad p(0) = p(T), \tag{3.9}$$

$$\eta(t) \in \partial g(Cy^*(t)) \quad \text{a.e. } t \in (0,T), \tag{3.10}$$

$$u^*(t) \in \partial h^*(B^*p(t)) \quad \text{a.e. } t \in (0,T). \tag{3.11}$$

The system (3.8), (3.9) is considered, of course, in the weak sense:

$$\mathcal{A}y = Bu^* + f; \qquad \mathcal{A}^*p = -C^*\eta. \tag{3.8}'$$

Proof. It is readily seen that equations (3.8)–(3.11) are sufficient for optimality. To prove necessity we fix an optimal pair (y^*, u^*) and consider the approximating control problem

$$\text{Min}\left\{\int_0^T (g_\varepsilon(Cy) + h(u) + 2^{-1}(|y - y^*|^2 + |u - u^*|_U^2 + \varepsilon^{-1}|v|^2))\right\} dt \tag{3.12}$$

subject to

$$\mathcal{A}y = Bu + v + f, \quad u \in L^2(0,T;U), \quad v \in L^2(0,T;H), \quad y \in C([0,T];H).$$

Here $g_\varepsilon \in C^1(Z)$ is defined as in (1.4).

Arguing as above, it is easily seen that problem (3.12) has a unique solution $(y_\varepsilon, u_\varepsilon, v_\varepsilon) \in C([0,T];H) \times L^2(0,T;U) \times L^2(0,T;H)$, and by a standard device (see, e.g., [1], [4]) we have

$$u_\varepsilon \longrightarrow u^* \text{ strongly in } L^2(0,T;U), \tag{3.13}$$

$$y_\varepsilon \longrightarrow y^* \text{ strongly in } L^2(0,T;H).$$

We also have that

$$v_\varepsilon \longrightarrow 0 \text{ strongly in } L^2(0,T;H).$$

Next we have

$$\int_0^T ((C^*\nabla g_\varepsilon(Cy_\varepsilon), z) + (y_\varepsilon - y^*, z) + (u_\varepsilon - u^*, w)_U + h'(u_\varepsilon, w) + \varepsilon^{-1}(v_\varepsilon, v))dt \geq 0 \tag{3.14}$$

$\forall (z, w, v) \in C([0,T];H) \times L^2(0,T;U) \times L^2(0,T;H)$ such that $\mathcal{A}z = Bw + v$. We set $p_\varepsilon = \varepsilon^{-1}v_\varepsilon$. Then (3.14) yields

$$\int_0^T (C^*\nabla g_\varepsilon(Cy_\varepsilon) + y_\varepsilon - y^*, z) + (u_\varepsilon - u^*, w)_U + h'(u_\varepsilon, w) + (p_\varepsilon, \mathcal{A}z - Bw))dt \geq 0 \tag{3.14}'$$

$\forall z \in D(\mathcal{A})$, $\forall w \in L^2(0,T;U)$. (Here h' is the directional derivative of h.) For $w = 0$ the latter yields

$$\mathcal{A}^*p_\varepsilon = -C^*\nabla g_\varepsilon(Cy_\varepsilon) + y^* - y_\varepsilon. \tag{3.15}$$

Substituting the latter into (3.14)′, we get

$$\int_0^T (B^*p_\varepsilon + u^* - u_\varepsilon, w)_U dt \leq \int_0^T h'(u_\varepsilon, w)dt \quad \forall w \in U.$$

This yields

$$B^*p_\varepsilon \in \partial h(u_\varepsilon) + u_\varepsilon - u^* \quad \text{a.e.} \in (0,T). \tag{3.16}$$

We note also that

$$\mathcal{A}y_\varepsilon = Bu_\varepsilon + \varepsilon p_\varepsilon + f. \tag{3.17}$$

We are going to let ε tend to 0 in equations (3.15), (3.16) in order to get (3.9)–(3.11). To this aim, some a priori estimates on p_ε are necessary. Assume first that condition $(jjj)_2$ holds. Then by (3.16) and by the definition of ∂h, we have

$$\begin{aligned}(B^*p_\varepsilon(t) + u^*(t) - u_\varepsilon(t), u_\varepsilon(t) + f_0(t) - \rho w)_U \\ \geq h(u_\varepsilon(t)) - h(\rho w - f_0(t)) \quad \text{a.e. } t \in (0,T)\end{aligned} \tag{3.18}$$

$\forall w \in U, |w|_U = 1$, and ρ positive and sufficiently small. This yields

$$\rho \int_0^T |B^*p_\varepsilon(t)|_U dt \leq \int_0^T (p_\varepsilon(t), \mathcal{A}y_\varepsilon(t) - \varepsilon p_\varepsilon(t))dt + C_3.$$

Finally,

$$\begin{aligned}\rho \int_0^T (|B^*p_\varepsilon(t)|_U + \varepsilon |p_\varepsilon(t)|^2)dt \\ \leq -\int_0^T ((\nabla g_\varepsilon(Cy_\varepsilon(t)), Cy_\varepsilon(t))_Z + (y_\varepsilon(t) - y^*(t), y_\varepsilon(t)))dt \leq C_4\end{aligned}$$

because ∇g_ε is monotone. On the other hand, it follows that $\{y_\varepsilon\}$ is strongly convergent to y^* in $C([0,T];H)$. Indeed, by (3.17), we see that

$$\begin{aligned}y_\varepsilon(t) = e^{-\mathcal{A}_F(t-s)}y_\varepsilon(s) + \int_s^t e^{-\mathcal{A}_F(t-r)}(Bu_\varepsilon(r) + \varepsilon p_\varepsilon(r) \\ + BFy_\varepsilon(r))dr, \qquad 0 \leq s \leq t \leq T,\end{aligned}$$

and the conclusion follows by (3.13). (Here and everywhere in the following, $F \in L(H,U)$ is chosen as in Definition 1.) Since ∂g is locally bounded in H and

$$\nabla g_\varepsilon(z) \in \partial g((I + \varepsilon \partial g)^{-1}z) \quad \forall z \in Z; \qquad \int_0^T g_\varepsilon(Cy_\varepsilon)dt \leq C_5,$$

we have

$$|\nabla g_\varepsilon(Cy_\varepsilon(t))|_Z \leq C_6 \quad \forall \varepsilon > 0, \quad t \in [0,T]. \tag{3.19}$$

We may rewrite equation (3.15) as

$$\mathcal{A}_F^* p_\varepsilon = -C^* \nabla g_\varepsilon(Cy_\varepsilon) - (y_\varepsilon - y^*) + F^*B^*p_\varepsilon. \tag{3.20}$$

Then by (3.19) and Corollary 1 it follows that

$$|p_\varepsilon^1(t)| \leq C_7 \quad \forall t \in [0,T], \tag{3.21}$$

where $p_\varepsilon = p_\varepsilon^1 + p_\varepsilon^2$ and

$$p_\varepsilon^1 \in R(\mathcal{A}_F) = N(\mathcal{A}_F^*)^\perp, \qquad p_\varepsilon^2 \in N(\mathcal{A}_F^*).$$

Denote by $\mathcal{B}_F^*$ the operator $y \longrightarrow B^*y$ defined from $N(\mathcal{A}_F^*)$ to $L^2(0,T;U)$. Recalling that the space $N(\mathcal{A}_F^*)$ is finite-dimensional, we infer that $\mathcal{B}_F^*$ has closed range in $L^2(0,T;U)$, and so by the closed range theorem, it has a bounded inverse on its range. Since $N(\mathcal{A}_F^*) \subset C([0,T];H)$ and $\{\mathcal{B}_F^* p_\varepsilon^2\}$ is bounded in $L^1(0,T;U)$, it is bounded in $L^2(0,T;U)$ too, and we have $p_\varepsilon^2 = q_\varepsilon^1 + q_\varepsilon^2$ where $\{q_\varepsilon^1\}$ is bounded in $L^2(0,T;H)$ and $B^*q_\varepsilon^2 = 0$ a.e. in $(0,T)$. We conclude, therefore, that the sequence $\{p_\varepsilon^1 + q_\varepsilon^1\}$ is weakly compact in $L^2(0,T;H)$. Moreover, we may write (3.20) as

$$\mathcal{A}_F^*(p_\varepsilon^1 + q_\varepsilon^1) = -C^*\nabla g_\varepsilon(Cy_\varepsilon) - (y_\varepsilon - y^*) + F^*B^*(p_\varepsilon^1 + q_\varepsilon^1). \tag{3.20$'$}$$

Selecting further subsequences, if necessary, we may assume that

$$\begin{array}{lcll} p_\varepsilon^1 + q_\varepsilon^1 & \longrightarrow & p & \text{weakly in } L^2(0,T;H), \\ \nabla g_\varepsilon(Cy_\varepsilon) & \longrightarrow & \eta & \text{weak star in } L^\infty(0,T;Z). \end{array}$$

Since ∂g and ∂h are maximal monotone (and therefore weakly/strongly closed), we may pass to the limit in (3.15) and (3.20)$'$ to get the optimality system (3.8)–(3.11).

Assume now that condition $(\mathrm{jjj})_1$ is satisfied. We set $p_\varepsilon = p_\varepsilon^1 + q_\varepsilon^1$. Then by (3.16) we have

$$h(u_\varepsilon) - h(\rho w) \le (B^*p_\varepsilon + u^* - u_\varepsilon, u_\varepsilon - \rho w)$$

for all $w \in H, \rho > 0$. This yields

$$\begin{aligned} \rho \int_0^T |B^*p_\varepsilon(t)|_U dt &\le C_8 + Th(\rho w) + \int_0^T (p_\varepsilon(t), \mathcal{A}y_\varepsilon(t) - f(t))dt \\ &\le Th(\rho w) + C_9\left(1 + \int_0^T |p_\varepsilon(t)|dt\right). \end{aligned}$$

Finally,

$$\int_0^T |B^*p_\varepsilon(t)|_U dt \le C_\rho + C_{10}\rho^{-1}\int_0^T |p_\varepsilon(t)|dt \tag{3.22}$$

for all $\rho > 0$. Choosing ρ sufficiently large it follows by (3.20)$'$ and (3.2) that $\{p_\varepsilon^1 + q_\varepsilon^1\}$ is bounded in $L^2(0,T;H)$, and so we may conclude the proof as in the previous case.

Remark 1. The proof reveals that Theorem 2 remains true if assumption (j) is relaxed to: there is an $F \in L(H,U)$ such that $\mathcal{A}_F$ has closed range in $L^2(0,T;H)$ and $\mathcal{B}_F^*$ has closed range in $L^1(0,T;U)$.

The optimal control problem

$$\begin{aligned} \mathrm{Min}\Bigg\{ &\int_0^T (g_*(v(t)) + h^*(B^*p(t)) + (f(t), p(t)))dt; \\ &p \in C([0,T];H), v \in L^2(0,T;H)\Bigg\} \end{aligned} \tag{3.23}$$

subject to

$$\mathcal{A}^*p = -v \tag{3.24}$$

is the dual of (1.1) in the sense of Rockafellar (see, e.g., [17]). Here h^* is the conjugate of h, and g_* is the conjugate of the function $y \to g(Cy)$.

THEOREM 3. *Under the assumptions of Theorem* 2 *the pair* (y^*, u^*) *is optimal in problem* (1.1) *if and only if the dual problem* (3.23) *has a solution* (p^*, v^*) *and*

$$\int_0^T (g(Cy^*(t) + h(u^*(t))dt + \int_0^T (g_*(v^*(t)) + h^*(B^*p^*(t))) + (f(t), p^*(t)))dt = 0. \tag{3.25}$$

Proof. The argument is standard (see, e.g., [4], [16]), so only the proof will be sketched. If (y^*, u^*) is optimal in (1.1), then by Theorem 2 the optimality system (3.8)–(3.11) has a solution $(p^*, v^* = C^*\eta)$, and in virtue of the conjugacy relation, we have

$$h(u^*) + h^*(B^*p^*) = (B^*p^*, u^*)_U, \tag{3.26}$$

$$g(Cy^*) + g_*(v^*) = (y^*, v^*).$$

Integrating from 0 to T we get (3.25). On the other hand, for all $(p, v) \in C([0,T]; H) \times L^2(0,T;H), \mathcal{A}^*p = -v$, we have

$$h(u^*) + h^*(B^*p) \geq (B^*p, u^*)_U \quad \text{a.e. on } (0,T), \tag{3.27}$$

$$g(Cy^*) + g_*(v) \geq (y^*, v) \quad \text{a.e. on } (0,T),$$

which imply that the pair (p^*, v^*) is optimal in problem (3.23). Conversely, if (3.25) holds, then by (3.26) and (3.27), we see that y^*, p^*, u^* satisfy the optimality system (3.8)–(3.11), and therefore (y^*, u^*) is optimal in problem (1.1).

We end this section with a few examples of linear control systems of the form (1.2) for which the previous theorems are applicable.

1. *Parabolic control problems.* Consider the system

$$\begin{aligned} &\tfrac{\partial y}{\partial t} - \Delta y + b(x) \cdot \nabla y + c(x)y = Bu + f(x,t), \qquad (x,t) \in \Omega \times R, \\ &y(x,t) = 0 \quad \forall (x,t) \in \partial\Omega \times R, \\ &y(x,t+T) = y(x,t) \quad \forall (x,t) \in \Omega \times R, \end{aligned} \tag{3.28}$$

where $b \in W^{1,\infty}(\Omega; R^n)$, $c \in L^\infty(\Omega)$, $f \in L^2_{\text{loc}}(R; L^2(\Omega))$ is T-periodic in t, while $B \in L(L^2(\Omega), L^2(\Omega))$. Here Ω is a bounded and open subset of R^n with a sufficiently smooth boundary $\partial\Omega$. We may write (3.28) under the form (1.2) where $H = U = L^2(\Omega)$ and

$$Ay = -\Delta y + b \cdot \nabla y + cy, \qquad D(A) = H_0^1(\Omega) \cap H^2(\Omega). \tag{3.29}$$

Since the semigroup e^{-At} generated by $-A$ on $L^2(\Omega)$ is compact, it follows that $R(\mathcal{A})$ is closed in $L^2(0,T;H)$ and $N(\mathcal{A}^*)$ is finite-dimensional (Corollary 2). Hence the p-stabilizability hypothesis (j) is satisfied in the present situation with $F = 0$. A similar conclusion can be reached for the p-detectability hypothesis (i).

2. *Linear delay control systems.* Consider the control system governed by the delay system

$$\begin{aligned} &y'(t) + A_0y(t) + A_1y(t-h) = B_0u(t) + f(t), \\ &y(t) = y(t+T) \quad \forall t \in R, \end{aligned} \tag{3.30}$$

where A_0, A_1 are $n\times n$ matrices, B_0 is an $n\times l$ matrix, $f \in L^2_{\rm loc}(R; R^n)$, $f(t+T) = f(t)$, $u \in L^2_{\rm loc}(R; R^l)$, and $u(t) = u(t+T)$. It is well known that this system can be written in the form (1.2), where $H = M_2 = R^n \times L^2(-h, 0; R^n)$, $U = R^l$, $B = (B_0, 0)$, and

$$A(y_0, y^0) = \left\{ A_0 y_0 + A_1 y^0(-h), -\frac{dy^0}{ds} \right\},$$

$$D(A) = \{(y_0, y^0) \in R^n \times W^{1,2}([-h, 0]; R^n), y_0 = y^0(0)\}.$$

For each $m \in Z$, we may rewrite the equation $(\mu_m iI + A)y = (f_0, f_1)$ as

$$(i\mu_m I + A_0 + e^{-i\mu_m h} A_1) y_0 = f_0 + \int_0^{-h} e^{-i\mu_m (h+s)} A_1 f_1(s) ds, \tag{3.31}$$

$$y^0(s) = e^{-i\mu_m s} y_0 - \int_0^s e^{i\mu_m (s-t)} f_1(t) dt,$$

where $\mu_m = 2m\pi T^{-1}$. Moreover, after some calculation, we see that

$$N(\mathcal{A}^*) = \left\{ \sum_m (y_m e^{i\mu_m t}, -A_1^* y_m e^{i\mu_m (t-s-h)}); \right. \tag{3.32}$$

$$\left. (i\mu_m I + A_0^* + e^{-i\mu_m h} A_1^*) y_m = 0 \right\}.$$

By (3.31) we see that $R(i\mu_m I + A)$ is closed and condition (2.5) in Proposition 1 is satisfied. We conclude, therefore, that the corresponding operator $\mathcal{A}$ has closed range in $L^2(0, T; H)$. Moreover, by (3.31) and (3.32), it follows that $N(\mathcal{A})$ and $N(\mathcal{A}^*)$ are finite-dimensional.

3. *First-order hyperbolic systems.* Consider the control system governed by the linear system

$$y_t(x,t) - z_x(x,t) = u(x,t) + f(x,t), \quad x \in (0,1), \quad t \in (0,T), \tag{3.33}$$

$$z_t(x,t) - y_x(x,t) = B_0 v(x,t) + g(x,t), \quad x \in (0,1), \quad t \in (0,T),$$

$$y(0,t) = y(1,t) = 0; \quad y(x,T) = y(x,0), \quad z(x,T) = z(x,0) \quad \forall x \in (0,1).$$

Here $B_0 \in L(L^2(0,1), L^2(0,1))$ and $f, g \in C([0,T]; L^2(0,1))$ are given functions. System (3.33) can be written in the form (1.2) where $H = U = L^2(0,1) \times L^2(0,1)$ $B(u,v) = (u, B_0 v)$, and $A(y,z) = (-z_x, -y_x)$, $D(A) = \{y, z \in H^1(0,1), y(0) = y(1) = 0\}$. Consider the feedback control $F(y,z) = (-y, 0)$. Then it is easily seen that the corresponding operator $\mathcal{A}_F$ has closed range in $L^2(0,T;H)$, $N(\mathcal{A}_F) = N(\mathcal{A}_F^*) = \{(0,C); C \in R\}$, and therefore the pair (A,B) is p-stabilizable. This simple example extends to linear control hyperbolic systems in $R^n \times R^n$, and it is instructive to notice that in this case if T is irrational, then $R(\mathcal{A})$ is not closed and so assumption (j) does not hold with $F = 0$.

4. Periodic boundary control problems. Here we shall extend the previous results to boundary control problem

$$\text{Min} \left\{ \int_0^T (g(Cy(t)) + h(u(t)))dt, u \in L^2(0,T;U) \right\} \tag{4.1}$$

subject to

$$z' + Az = Du + (\lambda I + A)^{-1} f \quad \text{in } (0,T), \tag{4.2}$$

$$z(0) = z(T), \quad y = (\lambda I + A)z.$$

Here $-A$ is the infinitesimal generator of a C_0 semigroup in H, $\lambda \in \rho(-A)$, and $D \in L(U,H)$. We set $B = (\lambda I + A)D$. Clearly $B \in L(U, (D(A^*)'))$ and the dual operator B^* is in $L(D(A^*), U)$. Dealing with system (4.2), we shall use one of the following two conditions:

$$\|B^* e^{-A^* t}\|_{L(H,H)} \le C_1 t^{\gamma - 1} \quad \forall t \in (0,T) \tag{H_1}$$

where $\gamma \in (0,1)$.

$$\int_0^T |B^* e^{-A^* t} x|_U^2 dt \le C_2 |x|^2 \quad \forall x \in H. \tag{H_2}$$

It is well known that condition (H_1) is satisfied by parabolic equations with Dirichlet boundary control, while condition (H_2), due to Lasiecka and Triggiani, is specific to hyperbolic boundary control systems (see [10]). The state equation (4.2) is considered in the weak sense, i.e.,

$$\mathcal{A} z = Du + (\lambda I + A)^{-1} f. \tag{4.3}$$

We shall study problem (4.1) under the assumption

$$R(\mathcal{A}) \text{ is closed in } L^2(0,T;H) \text{ and } \dim N(\mathcal{A}^*) < +\infty. \tag{4.4}$$

Since, as seen earlier (Proposition 3), this implies that $(I - e^{-AT})^{-1} \in L(R(I - e^{-AT}); H)$, it follows that under assumption (H_2) the weak solution $y = (\lambda I + A)z$ is in $C([0,T];H)$, while under assumption (H_1), $y \in L^2(0,T;H)$.

THEOREM 4. *Assume that $R(\mathcal{A})$ is closed in $L^2(0,T;H)$ and $N(\mathcal{A})$ is finite-dimensional. Then problem* (4.1) *has at least one solution $(y^*, u^*) \in L^2(0,T;H) \times L^2(0,T;U)$.*

Proof. Let y_n, u_n be as in the proof of Theorem 1 where $y_n = (\lambda I + A)z_n$, $\mathcal{A} z_n = Du_n + (\lambda I + A)^{-1} f$. We have

$$\|Cy_n\|^2_{L^1(0,T;Z)} + \|u_n\|^2_{L^2(0,T;U)} \le C_3.$$

We set $y_n = y_n^1 + y_n^2$, where $y_n^1 \in R(\mathcal{A}^*)$ and $y_n^2 \in N(\mathcal{A})$. Since $N(\mathcal{A})$ is finite-dimensional we may write $y_n^2 = z_n^1 + z_n^2$ where $\{z_n^1\}$ is compact in $L^2(0,T;H)$ and $z_n^2 \in N(\mathcal{C}_K) \cap N(\mathcal{A})$. Hence $\{y_n^1 + z_n^1\}$ is weakly compact in $L^2(0,T;H)$, and so we may pass to the limit in (3.4) to conclude the proof.

As regards the maximum principle, we have the following theorem.

THEOREM 5. *Assume that hypotheses* (4.4) *and* (jj), (jjj) *hold. Then the pair (y^*, u^*) is optimal in problem* (4.1) *if and only if the system* (3.8)–(3.11) *is satisfied.*

Proof. The proof is essentially the same as that of Theorem 2, and so we shall omit it. We notice only that though B and B^* are not bounded (or closed), equations (3.15), (3.16) make sense in this case because

$$B^* p(t) = B^* e^{-A^*(T-t)} (I - e^{-A^* T})^{-1} \int_0^T e^{-A^* s} g(s) ds$$

$$+ \int_t^T B^* e^{-A^*(s-t)} g(s) ds \quad \forall t \in [0,T]$$

for all solutions p to equation $\mathcal{A}^*p = g$. The latter is also used to pass to the limit in equations (3.15), (3.16).

As an example, consider the optimal control problem with state dynamics

$$\begin{array}{ll} y_t - \Delta y + b(x)\cdot\nabla y + c(x)y = f(x,t), & (x,t)\in\Omega\times R, \\ y(x,t) = a(x)u(x,t) & \forall (x,t)\in\partial\Omega\times R, \\ y(x,t+T) = y(x,t) & \forall (x,t)\in\Omega\times R, \end{array} \tag{4.5}$$

where $b\in W^{1,\infty}(\Omega;R^n)$, $c\in L^\infty(\Omega)$, and $a\in C(\partial\Omega)$ is not identically 0.

Define $Du = y^0$, where y^0 is the solution to boundary value problem

$$\lambda y^0 - \Delta y^0 + b\cdot\nabla y^0 + cy^0 = 0 \text{ in } \Omega,$$

$$y^0 = au \text{ in } \partial\Omega,$$

and

$$Ay = -\Delta y + b\cdot\nabla y + cy, \qquad D(A) = H_0^1(\Omega)\cap H^2(\Omega).$$

We are in the case described by (H_1), where $H = L^2(\Omega)$, $U = L^2(\partial\Omega)$, and $B^*p = a\frac{\partial p}{\partial\nu}$ $\forall p\in D(A)$.

5. Synthesis of periodic optimal controller. Here we shall study the existence of optimal feedback controllers for problem (1.1) in the linear quadratic case,

$$g(y) = 2^{-1}|Cy|_Z^2, \qquad h(u) = 2^{-1}|u|_U^2,$$

i.e.,

$$\text{Min}\left\{2^{-1}\int_0^T (|Cy|_Z^2 + |u|_U^2)dt; \mathcal{A}y = Bu + f\right\}, \tag{5.1}$$

where $B\in L(U,H), C\in L(H,Z)$, and the pair (A,C) is detectable, i.e., there is a $K\in L(Z,H)$ such that $A+KC$ generates an exponentially stable semigroup. It is readily seen that under these assumptions problem (5.1) has a unique optimal pair (y^*,u^*). The existing results on feedback representation of optimal controller of problem (5.1) require the stabilizability of pair (A,B) (see, e.g., [7], [8]). Here we shall obtain such a representation under the following weaker hypothesis.

(j)′ The pair (A,B) is p-stabilizable and

$$N(\mathcal{A}^*)\cap\{p\in L^2(0,T;H); B^*p(t) = 0, \text{ a.e. } t\in(0,T)\} = \{0\}.$$

THEOREM 6. *Assume that the pair (A,C) is detectable and that hypothesis* (j)′ *holds. Then the optimal controller u^* of problem* (5.1) *is given by the feedback formula*

$$u^*(t) = B^*P^{-1}(y^*(t) - r(t) + z(t)) \quad \forall t\in[0,T], \tag{5.2}$$

where $P\in L(H,H)$ is a self-adjoint and positive solution to the algebraic Riccati equation

$$AP + PA^* + PC^*CP = BB^*, \tag{5.3}$$

$r\in C([0,T];H)$ is a weak solution to the periodic problem

$$r' + (A + PC^*C)r = f \quad \text{in } (0,T); \qquad r(0) = r(T), \tag{5.4}$$

and $z \in C([0,T];H)$ *is such that* $z(t) \in N(C^*C), \forall t \in [0,T]$. *In* (5.2) P^{-1} *is the generalized inverse of* P.

Proof. By Theorem 3 it follows that the solution (y^*, u^*) to problem (5.1) is given by

$$u^* = B^*p^*, \qquad C^*Cy^* = -v^*, \tag{5.5}$$

where $(p^*, v^*) \in C([0,T];H) \times L^2(0,T;H)$ is a solution to the optimal control problem

$$\text{Min}\left\{\int_0^T (2^{-1}|B^*p|_U^2 + H(v))dt + \int_0^T (f,p)dt;\ \mathcal{A}^*p = v\right\}, \tag{5.6}$$

where $H(v) = \sup\{(y,v) - 2^{-1}|Cy|_Z^2; y \in H\}$. Since the function H is strictly convex, the optimal control v^* is unique and can be expressed as

$$v^*(t) = -C^*CPp^*(t) - C^*Cr(t), \tag{5.7}$$

where $P \in L(H,H)$ is a self-adjoint, positive solution to the algebraic Riccati equation (5.3) and $r \in C([0,T];H)$ is a weak solution to (5.4). The solution P to (5.3) is considered in the following weak sense:

$$2(Px, A^*x) + |CPx|^2 = |B^*x|^2 \quad \forall x \in D(A).$$

In fact, it is easily seen, via the optimality system, that the feedback controller v^* given by (5.7) is optimal in problem (5.6). The existence of a positive, self-adjoint solution P to Riccati equation (5.3) follows from the general theory of linear quadratic infinite horizon control problems (see, e.g., [7, p. 265]) because the pair (A^*, C^*) is stabilizable. We shall prove now that with a such a P equation, (5.4) has at least one weak solution, $r \in C([0,T];H)$.

Let $\mathcal{A}_1$ be the operator defined in $L^2(0,T;H)$ by

$$\mathcal{A}_1 = \mathcal{A} + PC^*C.$$

Then equation (5.4) can be written as

$$\mathcal{A}_1 r = f. \tag{5.8}$$

We shall prove that $R(\mathcal{A}_1)$ is closed in $L^2(0,T;H)$ and $N(\mathcal{A}_1^*) = \{0\}$. Indeed, if $\mathcal{A}_1^*p = 0$, then

$$p' - (A^* + C^*CP)p = 0; \qquad p(0) = p(T).$$

After some calculation involving this latter equation and (5.3), we get

$$\int_0^T (|B^*p(t)|_U^2 + |CPp(t)|_Z^2)dt = 0.$$

Hence $B^*p = 0, CPp = 0$. In particular, this implies that $p \in N(\mathcal{A}^*)$, and so by assumption (j)$'$, we infer that $p(t) = 0\ \forall t \in (0,T)$. It remains to check that $R(\mathcal{A}_1)$ is closed. To this end it suffices to show that $R(\mathcal{A}_1^*)$ is closed in $L^2(0,T;H)$. Let $\{f_n\} \subset L^2(0,T;H)$ and $\{y_n\} \subset C([0,T];H)$ be such that $\mathcal{A}_1^*y_n = f_n$ and

$$f_n \longrightarrow f \quad \text{strongly in} \quad L^2(0,T;H).$$

Then, by equation (5.3), we get

$$\int_0^T (|B^*y_n|_U^2 + |CPy_n|_Z^2 - 2(f_n, Py_n))dt = 0,$$

and this yields

$$\|B^*y_n\|^2_{L^2(0,T;U)} + \|Py_n\|^2_{L^2(0,T;H)} \leq C_1. \tag{5.9}$$

On the other hand, we have

$$\mathcal{A}_F^* y_n = f_n - C^*CPy_n + F^*B^*y_n$$

where F is as in Definition 1. Then by (2.15) and (5.9) we get the estimate

$$\|y_n^1\|_{C([0,T];H)} + \|B^*y_n^2\|_{L^2(0,T;U)} \leq C_2, \tag{5.10}$$

where

$$y_n = y_n^1 + y_n^2, \quad y_n^1 \in R(\mathcal{A}_F), \quad y_n^2 \in N(\mathcal{A}_F^*).$$

Since $N(\mathcal{A}_F^*)$ is finite-dimensional, it follows by (5.10) and assumption (j)′ that$\{y_n^2\}$ is bounded in $C([0,T];H)$. Hence $\{y_n\}$ is bounded in $L^2(0,T;H)$, and therefore $\mathcal{A}_1^* y = f$ where y is a weak limit point of $\{y_n\}$ in $L^2(0,T;H)$. Hence equation (5.4) has at least one weak solution, and so the feedback controller (5.7) is well defined. Now, by (5.5) and (5.7), we get (5.3), thereby completing the proof.

Let us check hypothesis (j)′ in the case of control system (3.30), where $Bu = au$, $a \in C(\overline{\Omega})$. We shall assume also that a is not identically 0. If $p \in N(\mathcal{A}^*)$ and $B^*p = ap = 0$ a.e. in $Q = \Omega \times (0,T)$, then p is the solution to the parabolic boundary value problem

$$\begin{aligned} \frac{\partial p}{\partial t} + \Delta p + \operatorname{div}(bp) - cp = 0 \quad &\text{in } Q, \\ p = 0 \quad &\text{in } \partial\Omega \times (0,T). \end{aligned} \tag{5.11}$$

Then, by the unique continuation property of solutions to parabolic equations (see, e.g., [18]), we infer that $p \equiv 0$, and therefore hypothesis (j)′ is satisfied.

Coming back to the second example, delay control problems, we note that by (3.32) it follows that if

$$N(i\mu_m I + A_0^* + e^{-\mu_m ih}A_1^*) \cap N(B_0^*) = \{0\} \quad \forall m \in Z, \tag{5.12}$$

then hypothesis (j)′ holds. It should be mentioned, however, that the latter condition does not imply the stabilizability of the pair (A,B), as the following example shows:

$$x_1'(t) = x_1(t) + x_3(t-h), \quad x_2'(t) = x_2(t) + x_3(t), \quad x_3'(t) = u(t).$$

It is readily seen that condition (5.12) is satisfied, though the system is not stabilizable [13].

6. The optimal control of the wave equation. We shall study here the optimal control problem

$$\text{minimize} \quad \int_0^T (2^{-1}|Cy(t)|_Z^2 + h(u(t)))dt \tag{6.1}$$

subject to $u \in L^2(0,T;H), y \in L^2(0,T;U)$,

$$\begin{aligned} &y'' + Ay = Bu + f, \qquad t \in (0,T), \\ &y(0) = y(T), \qquad y'(0) = y'(T), \end{aligned} \tag{6.2}$$

where A is a self-adjoint, linear, and positively defined operator in H, $B \in L(U,H)$, $C \in L(H,Z)$, and h is a lower semicontinuous convex function on U. By *weak solution* to equation (6.2), we mean a function $y \in L^2(0,T;H)$ such that

$$\int_0^T (y(t), \varphi''(t) + A\varphi(t))dt = \int_0^T (f(t) + Bu(t), \varphi(t))dt, \tag{6.3}$$

for all $\varphi \in Y = \{\varphi \in C^2([0,T];H) \cap C([0,T];D(A));\ \varphi(0) = \varphi(T),\ \varphi'(0) = \varphi'(T)\}$. Equivalently,

$$\mathcal{W}y = Bu + f, \tag{6.2}'$$

where $\mathcal{W} : D(\mathcal{W}) \subset L^2(0,T;H) \to L^2(0,T;H)$ is the linear operator defined by

$$\mathcal{W}y = f \ \text{ iff } \ \int_0^T (y(t), \varphi''(t) + A\varphi(t))dt = \int_0^T (f(t), \varphi(t))dt \quad \forall \varphi \in Y. \tag{6.4}$$

It is readily seen that $\mathcal{W}$ is densely defined and closed in $L^2(0,T;H)$.

Writing equation (6.2) as a first-order differential equation on the product space $D(A^{\frac{1}{2}}) \times H$, we may apply the general results obtained in the previous section to problem (6.1). However, a direct treatment of such a problem requires less restrictive conditions in specific examples. On the other hand, for the sake of simplicity, we shall not put the results of this section in the general framework of the p-stabilizability condition; we shall confine ourselves to assuming that $R(\mathcal{W})$ is closed in $L^2(0,T;H)$. By virtue of the closed range theorem, this assumption implies that

$$L^2(0,T;H) = R(\mathcal{W}) \oplus N(\mathcal{W}); \quad \mathcal{W}^{-1} \in L(R(\mathcal{W}), L^2(0,T;H)).$$

Arguing as in the proof of Theorem 1, it follows that if $R(\mathcal{W})$ is closed in $L^2(0,T;H)$ and $N(\mathcal{W})$ is finite-dimensional, then problem (6.1) has at least one solution $(y,u) \in L^2(0,T;H) \times L^2(0,T;U)$. As regards the maximum principle, we have the following theorem.

THEOREM 7. *Assume that $R(\mathcal{W})$ is closed,* dim $N(\mathcal{W}) < \infty$, *and h, f satisfy hypotheses* (jj), (jjj). *Then the pair $(y^*, u^*) \in L^2(0,T;H) \times L^2(0,T;U)$ is optimal in problem* (6.1) *if and only if there is a $p \in L^2(0,T;H)$ such that*

$$\mathcal{W}p = -C^*Cy, \tag{6.5}$$

$$u^*(t) \in \partial h^*(B^*p(t)), \ \textit{a.e. } t \in (0,T). \tag{6.6}$$

We omit the proof because it is identical with that of Theorem 2. Since in most applications the null space $N(\mathcal{W})$ is infinite-dimensional (the state equation is highly resonant), we shall relax this condition as follows.

(k) $R(\mathcal{W})$ is closed and the operator $y \longrightarrow B^*y$ defined from $N(\mathcal{W})$ to $L^2(0,T;H)$ has closed range.

THEOREM 8. *Assume that hypotheses* (jj), (k) *hold, $f \in C([0,T];H)$, and h has quadratic growth, i.e.,*

$$h(u) \le \alpha_1 |u|_U^2 + \beta_1 \quad \forall u \in U. \tag{6.7}$$

Then the pair $(y^*, u^*) \in L^2(0,T;H) \times L^2(0,T;U)$ is optimal in problem (6.1) *if and only if it satisfies system* (6.5), (6.6).

Proof. Let $(y_\varepsilon, u_\varepsilon, v_\varepsilon)$ be the solution to the approximating problem (see (3.12))

$$\text{Min}\left\{\int_0^T (2^{-1}|Cy|_Z^2 + h(u) + 2^{-1}(|y-y^*|^2 + |u-u^*|_U^2 + \varepsilon^{-1}|v|^2))dt; \right.$$
$$\left. \mathcal{W}y = Bu + v + f \right\}.$$

As in the proof of Theorem 2, we get (3.13) and (see (3.15), (3.16))

$$\mathcal{W}p_\varepsilon = -C^*Cy_\varepsilon + y^* - y_\varepsilon, \tag{6.8}$$

$$B^*p_\varepsilon \in \partial h(u_\varepsilon) + u_\varepsilon - u^* \quad \text{a.e. in } (0,T). \tag{6.9}$$

By (6.7) and (6.8) we have

$$\|B^*p_\varepsilon\|^2_{L^2(0,T;U)} \le C_1 \quad \forall \varepsilon > 0,$$

and so, by virtue of assumption (k), we conclude via the closed range theorem that

$$\{p_\varepsilon^1 + p_\varepsilon^3\} \text{ is bounded in } L^2(0,T;H),$$

where

$$p_\varepsilon = p_\varepsilon^1 + p_\varepsilon^3 + p_\varepsilon^4$$

and $p_\varepsilon^1 \in R(\mathcal{W}), p_\varepsilon^3, p_\varepsilon^4 \in N(\mathcal{W})$, and $B^*p_\varepsilon^4 = 0$ a.e. $t \in (0,T)$. Hence we may pass to the limit in equations (6.8), (6.9) to get (6.5), (6.6), as desired.

The dual problem of (6.1) is (see (3.23), (3.24))

$$\text{Min}\left\{\int_0^T (g_*(v) + h^*(B^*p) + (f,p))dt; \mathcal{W}^*p = -v; v \in L^2(0,T;H)\right\}. \tag{6.10}$$

By using exactly the same argument, it follows that under the assumptions of Theorem 7 or 8, the conclusions of duality Theorem 3 remain valid in the present case.

Now we shall present two examples.

4. *The one-dimensional wave equation.* Consider the control system

$$\begin{aligned} &y_{tt}(x,t) - v^{-1}(x)(v(x)y_x(x,t))_x = Bu(x,t) + f(x,t), \\ &\qquad\qquad (x,t) \in (0,\pi)\times R, \\ &y(0,t) = y(\pi,t) = 0, \qquad\qquad t \in R, \\ &y(x,t+T) = y(x,t), \quad y_t(x,t+T) = y_t(x,t), \quad (x,t) \in (0,\pi)\times R, \end{aligned} \tag{6.11}$$

where $v \in H^2(0,T)$, $v(x) > 0 \ \ \forall x \in [0,\pi]$, $B \in L(L^2(0,\pi), L^2(0,\pi))$, and

$$\text{ess sup}\{(v'(x))^2 - 2v''(x)v(x);\ x \in (0,\pi)\} < 0.$$

In this case $U = L^2(0,\pi)$, $H = L^2(0,\pi)$ is endowed with the scalar product $(y,z) = \int_0^\pi v(x)y(x)z(x)dx$ and

$$Ay = -v^{-1}(vy_x)_x, \qquad D(A) = H_0^1(0,\pi) \cap H^2(0,\pi).$$

Equation (6.11) models the forced vibrations of a nonhomogeneous string as well as the propagation of waves in nonisotropic media. If T is a rational multiple of π then $R(\mathcal{W})$ is closed, $N(\mathcal{W})$ is finite-dimensional [6], and so Theorem 8 is applicable.

5. *The n-dimensional wave equation.* Consider the control system

$$\begin{aligned} &y_{tt} - \Delta y = a(x)u + f, \quad x \in \Omega, \quad t \in R, \\ &y = 0 \text{ in } \partial\Omega \times R; \\ &y(x, t+T) = y(x,t), \qquad y_t(x, t+T) = y_t(x,t), \end{aligned} \tag{6.12}$$

where $\Omega = (0,\pi)^n$, $a \in C(\overline{\Omega})$, $a \not\equiv 0$, $f \in C([0,T]; L^2(\Omega))$, and $f(x, t+T) \equiv f(x,t)$. We may write (6.12) in the form (6.2) where $H = L^2(\Omega)$, $A = -\Delta$, $D(A) = H_0^1(\Omega) \cap H^2(\Omega)$, and $Bu = au \ \forall u \in U = L^2(\Omega)$. If T is a rational multiple of π, then the corresponding operator $\mathcal{W} : L^2(Q) \longrightarrow L^2(Q)$, $Q = \Omega \times (0,T)$, has closed range. Here is the argument (see [14]). If

$$\mathcal{W}y = f; \qquad (y,f) \in L^2(Q) \times L^2(Q),$$

then

$$y = \sum_{m \in Z, k \in N^n} f_{mk}(\mu_m^2 - \lambda_k^2)^{-1} e^{i\mu_m t}\varphi_k, \tag{6.13}$$

where $\mu_m = 2m\pi T^{-1}$, $\lambda_k^2 = k_1^2 + k_2^2 + \cdots + k_n^2$, $k_i \in N$, are the eigenvalues of A and φ_k are the corresponding eigenfunctions; f_{mk} are the Fourier coefficients of f. If T is a rational multiple of π we have

$$\inf\{|\lambda_k^2 - \mu_m^2|; \lambda_k^2 = \mu_m^2\} > 0.$$

Then by (6.13) we see that there is a $y_1 \in L^2(Q)$ such that $\mathcal{W}y_1 = f$ and

$$\|y_1\|_{L^2(Q)} \leq C_1 \|f\|_{L^2(Q)} \quad \forall f \in L^2(Q).$$

This implies that $R(\mathcal{W})$ is closed in $L^2(Q)$.

One might suspect that assumption (k) is true in this case and so that Theorem 8 could be applied to the optimal control problem with dynamics (6.12) and payoff

$$2^{-1} \int_Q |Cy(x,t)|_Z^2 dxdt + \int_0^T h(u(t))dt$$

where $C \in L(L^2(\Omega), Z)$. In the special case $a = 1$ the maximum principle follows for general domains Ω, because if we write equation (6.12) under the form (1.2) on the product space $H_0^1(\Omega) \times L^2(\Omega)$, the corresponding pair (A,B) is stabilizable. For a general $a \in C(\overline{\Omega})$ this happens if the support of a has the optical geometric property [16], [12]. Note also that in the case $n = 1$ and $m \equiv 1$, the interiority condition in $(\text{jjj})_2$ can be weakened to a similar condition in $L^\infty(\Omega)$ [2], but the argument uses the special structure of $N(\mathcal{A})$.

We shall conclude this section with the boundary control version of problem (6.1), i.e.,

$$\text{Min}\left\{\int_0^T (2^{-1}|Cy(t)|_Z^2 + h(u(t)))dt; (y,u) \in L^2(0,T;H) \times L^2(0,T;U)\right\} \tag{6.14}$$

subject to

$$z'' + Az = Du + A^{-1}f; \quad z(0) = z(T), \quad z'(0) = z'(T); \quad y = Az. \tag{6.15}$$

Here A is a self-adjoint, positive definite operator on H, $f \in L^2(0,T;H)$, g, h, C are as above, and $D \in L(U,H)$ satisfies the following condition:

$$\int_0^T |B^*S(t)y|_U^2 dt \le C_3|y|^2 \quad \forall y \in H, \tag{6.16}$$

where $B^* = D^*A \in L(D(A),U)$ and $S(t)$ is the sine operator associated with A. In particular, condition (6.16) is verified by the wave equation with Dirichlet boundary input on a bounded open set Ω of R^n with smooth boundary (or if Ω is a parallelepiped) (see [11]). In this case $A = -\Delta$, $D(A) = H_0^1(\Omega) \cap H^2(\Omega)$, $U = H = L^2(\Omega)$, and $D \in L(U, L^2(\Omega))$ is defined as

$$\Delta Du = 0 \quad \text{in } \Omega; \qquad Du = B_0 u \quad \text{in } \partial\Omega, \tag{6.17}$$

where $B_0 \in L(U,U)$. Note also that if $R(\mathcal{W})$ is closed in $L^2(0,T;H)$, then by (6.16) it follows that

$$\|A\mathcal{W}^{-1}(Du + A^{-1}f)\|_{L^2(0,T;H)} \le C_4(\|u\|_{L^2(0,T;U)} + \|f\|_{L^2(0,T;H)}). \tag{6.18}$$

THEOREM 9. *Assume that condition* (6.16) *is satisfied. Then, under the assumptions of Theorems* 7 *and* 8 *with* B^* *defined as above, the pair* (y^*, u^*) *is optimal in problem* (6.14) *if and only if it satisfies system* (6.5), (6.6).

The proof is essentially the same as that of Theorems 2 and 8. We mention only that in this case one uses condition (6.16) and inequality (6.18) in order to pass to the limit in the corresponding approximating equations

$$\mathcal{W}p_\varepsilon = -C^*Cy_\varepsilon - y_\varepsilon + y^*,$$

$$B^*p_\varepsilon \in \partial h(u_\varepsilon) + u_\varepsilon - u^*.$$

REFERENCES

[1] V. BARBU, *Analysis and Control of Nonlinear Infinite Dimensional Systems*, Academic Press, Boston, 1993.

[2] V. BARBU, *Optimal control of the one dimensional periodic wave equation*, Appl. Math. Optim., 35 (1997), pp. 77–90.

[3] V. BARBU, *Optimal control of periodic linear systems in Hilbert space*, in Proceedings of IFIP Conference on Parameter Distributed Systems, Warsaw, Poland, July 1995.

[4] V. BARBU AND T. PRECUPANU, *Convexity and Optimization in Banach Spaces*, D. Reidel, Dordrecht, the Netherlands, 1986.

[5] V. BARBU AND N. PAVEL, *Optimal control of periodic systems*, Appl. Math. Optim., 33 (1996), pp. 169–188.

[6] V. BARBU AND N. PAVEL, *Periodic solutions to nonlinear one dimensional wave equation with x-dependent coefficients*, Trans. Amer. Math. Soc., to appear.

[7] A. BENSOUSSAN, G. DA PRATO, M. DELFOUR, AND S.K. MITTER, *Representation and Control of Infinite Dimensional Control Systems*, Birkhäuser, Boston, Basel, Berlin, 1993.

[8] G. DA PRATO, *Synthesis of optimal control for an infinite dimensional periodic problem*, SIAM J. Control Optim., 25 (1987), pp. 706–714.

[9] A. HARAUX, *Nonlinear Evolution Equations-Global Behavior of Solutions,* Lecture Notes in Math. 841, Springer-Verlag, Berlin, Heidelberg, New York, 1981.

[10] I. LASIECKA AND R. TRIGGIANI, *Algebraic Riccati Equations with Applications to Boundary Point Control Problems. Continuous Theory and Approximation Theory*, Lecture Notes in Control and Inform. Sci., Springer-Verlag, Berlin, New York, 1991.

[11] I. LASIECKA AND R. TRIGGIANI, *Regularity of hyperbolic equations under* $L_2, T; L_2(\Gamma)$*-Dirichlet boundary terms*, Appl. Math. Optim., 10 (1983), pp. 275–286.

[12] J.L. Lions, *Controlabilité exacte, perturbations et stabilisation de systèmes distribués, Tome* 1, Masson, Paris, 1988.
[13] A. Manitius and R. Triggiani, *Function space controllability of linear retarded systems: A derivation from abstract conditions*, SIAM J. Control Optim., 16 (1978), pp. 599–645.
[14] N. Pavel, *Periodic solutions to nonlinear* 2-D *wave equations*, to appear.
[15] J. Prüss, *On the spectrum of C_0-semigroup*, Trans. Amer. Math. Soc., 284 (1984), pp. 847–857.
[16] J. Rauch and M. Taylor, *Exponential decay of solutions to hyperbolic equations in bounded domains*, Indiana Univ. Math. J., 24 (1974), pp. 7–86.
[17] R.T. Rockafellar, *Existence and duality theorems for convex control problem of Bolza*, Trans. Amer. Math. Soc., 159 (1971), pp. 1–40.
[18] J.C. Saut and B. Scheurer, *Unique continuation for some evolution equations*, J. Differential Equations, 66 (1987), pp. 118–139.
[19] K. Yosida, *Functional Analysis*, Springer-Verlag, Berlin, Heidelberg, New York, 1977.

SIAM J. CONTROL OPTIM.
Vol. 35, No. 6, pp. 2157–2182, November 1997

016

BLACKWELL OPTIMALITY IN BORELIAN CONTINUOUS-IN-ACTION MARKOV DECISION PROCESSES*

ALEXANDER A. YUSHKEVICH†

Abstract. We prove the existence of stationary Blackwell optimal policies in Markov decision processes with a Borel state space, compact action sets, and continuous-in-action and bounded transition densities and rewards, satisfying a simultaneous Doeblin-type condition. The proof is based on a compactification of the randomized stationary policy space in a weak-strong topology, on the continuity of Laurent coefficients of the discounted rewards in this topology, and on a lexicographical policy improvement. Until now similar results were obtained for the models with a denumerable state space or with a Borel state space and finite action sets.

Key words. Markov decision process, Borel space, compact action sets, Blackwell optimality

AMS subject classifications. 93E20, 90C40

PII. S0363012995292469

1. Introduction.

In a Markov decision process (MDP) with finite state and action spaces, a Blackwell optimal policy is optimal for all discount factors $\beta < 1$ close to 1. This concept of optimality was introduced in a pioneering work by Blackwell [3]. In another basic work Veinott [29] has shown that Blackwell optimality is the limiting, most selective concept in an infinite series of sensitive criteria starting from the average optimality, and has developed the Laurent expansions technique for the analysis of sensitive criteria. Initially studied in MDPs with finite state and action spaces, sensitive criteria have since been explored in more general MDPs and in some renewal and continuous-time models. Other early contributions to this field are due to Chitashvili [6], [7] (see also his survey in [34]), Denardo [10], Denardo and Miller [11], Lippman [19], Miller and Veinott [20], Puterman [21], Rothblum [22], Sladký [25], [26], and Veinott [28].

Sensitive optimality in MDPs with a countable state space is treated in more recent works by Cavazos-Cadena and Lasserre [4], [5], Dekker and Hordijk [8], [9], Hordijk and Sladký [17], and Lasserre [18]. In the countable models, the definition of Blackwell optimality required a modification suggested in [8]: to get some kind of existence results, it appeared necessary to allow the interval $\beta_0 < \beta < 1$, where the optimal policy exceeds any other policy, to depend on that policy and on the initial state. In discrete-time models with a Borel state space, sensitive optimality began to be studied in our papers [31], [32]. For a more detailed survey and a discussion of the concept of Blackwell optimality we refer the reader to [8] and [31]. These two references are the closest to our work.

Dekker and Hordijk [8] studied the existence of Blackwell optimal policies in MDPs with a countable state space, compact action sets, and continuous-in-action transition probabilities and rewards; the rewards were allowed to be unbounded (or more precisely, unbounded, but bounded in the so-called μ-norms). To get the existence of Blackwell optimal policies in the class of deterministic stationary policies by lexicographically maximizing coefficients of the related Laurent series, they assumed

*Received by the editors September 27, 1995; accepted for publication (in revised form) September 25, 1996. This research was partially supported by National Science Foundation grant DMS-9404177.
http://www.siam.org/journals/sicon/35-6/29246.html

†Department of Mathematics, University of North Carolina at Charlotte, Charlotte, NC 28223 (fma00aay@unccvm.uncc.edu).

either a uniform geometric ergodicity of Markov chains generated by such policies or some substitutes for it directly in terms of those coefficients (in [9] a different uniform recurrence condition was used). They also introduced lexicographical analogues of the policy improvement and of the Bellman optimality equation, which they called the Blackwell optimality equation, and found an elegant way to prove that a policy Blackwell optimal within the class of stationary policies is Blackwell optimal in comparison with all other policies too. All of these ideas are utilized in the present work, with substantial changes due to a more general state space and a different structure of the transition law. However, our approach is limited to the case of bounded rewards, and we do not use the concept of μ-norms.

Progress toward Blackwell optimality in the case of a Borel state space became possible in models with transition densities (instead of more general transition functions). In contrast to noncontrolled Markov chains, the transition densities are rarely used in the theory of MDPs. A paper by Georgin [14] on the discounted and average optimality is an exception, and our assumptions are essentially the same as some of Georgin's hypotheses.

In [31] we proved the existence of Blackwell optimal policies in the case of a Borel state space, a countable action space, and finite action sets. In the present work we extend this result to the case of compact action sets (in a topological Borel space). As in [31], we suppose that the transition density and the reward function are bounded and assume that the transition density satisfies a simultaneous Doeblin-type condition, which guarantees uniform geometric convergence of the multistep transition densities of the involved Markov chains, and therefore a uniform convergence of the Laurent series for the discounted rewards. This condition is simple but restrictive; it means the existence of a minorant in the case of transition densities. For the place of this condition in the variety of recurrence conditions considered in different works on the Borel state MDPs, see Hernández-Lerma, Montes-de-Oca, and Cavazos-Cadena [15]. Similar to works on MDPs with a countable state space, we require the continuity-in-action of transition densities and rewards. But in contrast to [8], where the maximization of Laurent coefficients is performed directly for every initial state in the class of deterministic stationary policies, we have to work with the wider class of randomized stationary policies and with an absolutely continuous initial distribution. We proceed in this way because the space of deterministic stationary policies does not allow a suitable compactification, and in the compactified space of randomized policies each policy is determined only up to a set of measure 0 in the state space.

An essential novelty with respect to [31] is the utilization of a weak-strong topology for the above compactification. This topology, in a more general context of nonstationary dynamic programming, was studied by Schäl [23], [24] and Balder [1]; a different approach for our specific case is given in [33]. The transition from a randomized stationary policy maximizing the Laurent coefficients for an absolutely continuous initial distribution to a deterministic stationary policy maximizing them for each initial state is made by lexicographical policy improvement. The related measurability matters are resolved by means of the Novikov–Kunigui theorem. The proof that a Blackwell optimal policy within the class of stationary policies is at the same time Blackwell optimal in the class of all policies is a combination of a similar proof in [8] with a compactness result in the weak-strong topology.

The paper is organized in the following way. In section 2 we introduce the terminology and notations and formulate the definitions, assumptions, and necessary measurability results. In section 3 we obtain the Laurent expansions of the expected

discounted rewards for stationary policies, define a space $\mathcal{H}$ of sequences H which contains all sequences of the corresponding Laurent coefficients, and introduce operators L^σ on the space $\mathcal{H}$ associated with stationary policies σ and used afterwards for the policy improvement. In section 4 we define the lexicographical ordering in $\mathcal{H}$ and prove a key comparison lemma for the operators L^σ. As a consequence we get the lexicographical policy improvement, the Blackwell optimality equation $H = TH$ where $T = \operatorname{Lexmax}_\sigma L^\sigma$ is the lexicographical Bellman operator, and the necessary and sufficient conserving condition of optimality. In section 5 we aggregate the space Σ of all stationary policies into a space S of measures which is compact in the weak-strong topology, and prove another key result stating that for absolutely continuous initial distributions the Laurent coefficients are continuous functions on S. In section 6 we prove the existence of a deterministic stationary maximizer φ for the equation $H = TH$, so that φ is Blackwell optimal within the class of stationary policies. It follows that the equation $H = TH$ has a unique solution in $\mathcal{H}$. In section 7 we prove that φ is optimal in the class of all policies. Section 8 contains examples. In section 9 we discuss our assumptions and some open problems, in particular, problems related to the concept of strong Blackwell optimality.

Instead of numerous references to [31] with indications of alterations in the proofs, we give here a self-contained and more readable exposition. However, when closely following some proof in [31], we first focus on its idea and omit secondary technicalities, which can be easily reproduced by the reader.

2. Definitions and assumptions. We start by detailing the notations and terminology. In a space M, let $\mathcal{O}_M$ and $\mathcal{B}_M$ denote, respectively, the systems of all open and all measurable sets (if there are any). It is supposed that in a Polish (i.e., complete separable metric) space M the system $\mathcal{O}_M$ is generated by the metric and the system $\mathcal{B}_M$ is generated by $\mathcal{O}_M$. Any Borel set B in a Polish space M (i.e., a set $B \in \mathcal{B}_M$) is considered as a topological space $(B, \mathcal{O}_B)$ with $\mathcal{O}_B = B \cap \mathcal{O}_M$ and as a measurable space with $\mathcal{B}_B = B \cap \mathcal{B}_M$. By a *Borel space* (i.e., standard Borel space) we understand, as usual, an isomorphic image of $(B, \mathcal{B}_B)$, where B is a Borel set in a Polish space. By a *topological Borel space* we understand a homeomorphic image of $(B, \mathcal{O}_B)$, where B is a Borel set in a Polish space. Every topological Borel space E is also considered a Borel space with $\mathcal{B}_E$ generated by $\mathcal{O}_E$. The term "compactness" means everywhere the sequential compactness (in a topological Borel space, these two concepts are equivalent). In a Borel space E, $B(E)$ is the set of all bounded measurable (with respect to $\mathcal{B}_E$) real-valued functions, and $\|\cdot\|$ denotes the corresponding supremum norm. Given two Borel spaces X and A, $\mathcal{B}_{X\times A} = \mathcal{B}_X \times \mathcal{B}_A$. In this case $\mathcal{B}_K = K \cap \mathcal{B}_{X\times A}$ for every $K \in \mathcal{B}_{X\times A}$, and if in addition A is a topological Borel space, then $\operatorname{Car}(K)$ denotes the set of all *Carathéodory functions* on K, i.e., functions $f \in B(K)$ which are continuous in the second coordinate a at every point $(x, a) \in K$, $x \in X$, $a \in A$.

An *MDP* is defined by a state space X, an action space A, action sets A_x, a transition function $P(x, a, B)$, and a reward function $r(x, a)$ ($x \in X$, $a \in A_x \subset A$, $B \in \mathcal{B}_X$). The components of MDP have the following meaning. At each of the time epochs $t = 0, 1, 2, \ldots$ a controller observes the state $x_t \in X$ of the process and selects an action $a_{t+1} \in A_{x_t}$ on the basis of this and the previous information . The selection of a_{t+1} determines the immediate reward $r(x_t, a_{t+1})$ of the controller and the distribution $P(x_t, a_{t+1}, \cdot)$ of the next state x_{t+1}. The aim of the controller is to maximize the expectation of the total discounted reward $r(x_0, a_1) + \beta r(x_1, a_2) + \beta^2 r(x_2, a_3) + \cdots$ for the values of the discount factor $\beta < 1$ arbitrary close to 1.

The following assumptions are supposed throughout the paper.

Assumption 2.1 (measurability and compactness). X is a Borel space, A is a topological Borel space, all A_x are nonempty compact subsets of A, and

$$K = \{(x,a) : a \in A_x, x \in X\} \in \mathcal{B}_{X\times A}.$$

Assumption 2.2 (existence of a bounded continuous transition density). For every $x \in X$, $a \in A_x$, and $B \in \mathcal{B}_X$

$$P(x,a,B) = \int_B p(x,a,y)m(dy),$$

where m is a given probability measure on X (a reference measure), and the function $p \in B(K \times X)$ is nonnegative, continuous in $a \in A_x$ for every pair $(x,y) \in X \times X$, and such that $P(x,a,X) = 1$ for all $(x,a) \in K$.

Assumption 2.3 (simultaneous Doeblin-type condition). There exist a number δ and a set $D \in \mathcal{B}_X$ with $0 < \delta m(D) < 1$ such that

$$p(x,a,y) \geq \delta \text{ for all } (x,a) \in K,\ y \in D. \tag{2.1}$$

Assumption 2.4 (continuity and boundedness of the reward function). $r \in \mathrm{Car}(K)$.

The policies are defined as usual in dynamic programming; our notations are closer to those in [13]. A *deterministic stationary policy*, or *selector*, is a measurable map φ from X to A with its graph in the set K; under such a policy the actions are selected by a rule $a_{t+1} = \varphi(x_t)$, $t = 0,1,\dots$. A *stationary* (i.e., randomized stationary) *policy* is a stochastic kernel σ from X to A (i.e., a function $\sigma(x,C)$, $x \in X$, $C \in \mathcal{B}_A$ measurable in x and such that $\sigma(x,\cdot)$ is a probability measure on A), satisfying the constraints $\sigma(x,A_x) = 1$, $x \in X$. Under such a policy each a_{t+1} is selected by random with a (conditional) distribution $\sigma(x_t,\cdot)$. A *Markov policy* is a sequence $\mu = \{\sigma_1,\sigma_2,\dots\}$ of stationary policies σ_t; under this policy the action a_{t+1} has a (conditional) distribution $\sigma_t(x_t,\cdot)$. Finally, under an *arbitrary policy* π the distribution of a_{t+1} depends on the whole history $x_0a_1x_1a_2\dots x_t$ (we omit a formal definition of π since it is not used in the paper). The sets of all selectors, stationary policies, Markov policies, and arbitrary policies are denoted, respectively, by $\Phi \subset \Sigma \subset \mathrm{M} \subset \Pi$. These sets are nonempty in accordance with the following general measurability and selection result, also utilized in the following.

PROPOSITION 2.5. *Under Assumption* 2.1

(i) *there exists a measurable map φ from X to A such that $\varphi(x) \in A_x, x \in X$;*

(ii) *for every $f \in \mathrm{Car}(K)$ and every $x \in X$ there exists $\hat{f}(x) = \max_{a\in A_x} f(x,a)$, and $\hat{f} \in B(X)$.*

See, for example, [16, pp. 391–392, Theorems 2 and 3] (where only the upper semicontinuity of $f(x,\cdot)$ is assumed).

To every initial state $x \in X$ and policy $\pi \in \Pi$ there corresponds a probability distribution $\mathbf{P}_x^\pi$ in the space of all infinite-horizon trajectories $x_0a_1x_1a_2\dots$, defined by the formulas

$$\begin{aligned}
&\mathbf{P}_x^\pi\{x_0 = x\} = 1,\\
&\mathbf{P}_x^\pi\{a_{t+1} \in C \mid x_0a_1\dots x_t\} = \pi(C \mid x_0a_1\dots x_t),\\
&\mathbf{P}_x^\pi\{x_{t+1} \in B \mid x_0a_1\dots x_ta_{t+1}\} = P(x_t,a_{t+1},B).
\end{aligned}$$

The corresponding expectation is denoted by $\mathbf{E}_x^\pi$. Since r is bounded (Assumption 2.4), the *expected total discounted reward*

$$v_\beta(x,\pi) = \mathbf{E}_x^\pi \sum_{t=0}^{\infty} \beta^t r(x_t, a_{t+1}) \tag{2.2}$$

is well defined for every initial state $x \in X$, policy $\pi \in \Pi$, and discount factor $0 < \beta < 1$. The corresponding *value function* is defined by

$$v_\beta(x) = \sup_{\pi\in\Pi} v_\beta(x,\pi),\ x \in X.$$

According to the original definition due to Blackwell, a policy π^* is optimal if there exists a number β_0 such that

$$v_\beta(x,\pi^*) = v_\beta(x) \ \text{ for all } \ x \in X,\ \beta_0 < \beta < 1. \tag{2.3}$$

This definition worked well in the case of finite spaces X and A. We do not know any existence results for Blackwell optimal policies in the sense (2.3) in MDPs with infinite X and A. Following Dekker and Hordijk [8, p. 399], we will use a weaker definition, in which π^* is compared separately with each other policy π, and β_0 may depend on x and π.

DEFINITION 2.6. *For a set $\Pi' \subset \Pi$, a policy $\pi^* \in \Pi'$ is said to be Blackwell optimal within the class Π', if for every $x \in X$ and every $\pi \in \Pi'$ there exists a number $\beta_0(x,\pi) < 1$ such that*

$$v_\beta(x,\pi^*) \geq v_\beta(x,\pi) \text{ for all } \ \beta_0(x,\pi) < \beta < 1. \tag{2.4}$$

If $\Pi' = \Pi$ then π^* is said to be Blackwell optimal.

In the case when (2.3) holds, i.e., when in (2.4) one may take $\beta_0 < 1$ independent of x and π, the policy π^* is called *strong Blackwell optimal*. It is an easy exercise to show that if X is finite and if there exists a strong Blackwell optimal policy, then every Blackwell optimal policy is strong Blackwell optimal. The existence of deterministic stationary strong Blackwell optimal policies in MDPs with finite state and action spaces was proved by Blackwell [3]. Therefore, in this case, the two concepts of Blackwell optimality coincide.

3. Laurent series and operators related to stationary policies. As in previous works on sensitive optimality, starting from the original paper by Veinott [29], our analysis is based on the Laurent series for discounted rewards in terms of a small parameter as β approaches 1. Following Chitashvili (see Yushkevich and Chitashvili [34]), we take for such a parameter the variable

$$\alpha = 1 - \beta, \qquad 0 < \alpha < 1.$$

In Veinott [29] and most of the subsequent works, the interest rate $\rho = \frac{1-\beta}{\beta}$ is used for this purpose. The expansion formulas in terms of α and ρ are essentially the same, so that the content of this section is mostly standard. However, in contrast to Dekker and Hordijk [8], we need the Laurent expansions for all stationary policies, not only the deterministic ones.

In the case of a stationary policy σ, the sequence $x_0x_1x_2\ldots$ of the observed states is a Markov chain with a transition function

$$Q^\sigma(x,B) = \int_A P(x,a,B)\sigma(x,da) = \int_B q^\sigma(x,y)m(dy), \quad x \in X, \quad B \in \mathcal{B}_X$$

where

$$q^\sigma(x,y)=\int_{A_x} p(x,a,y)\sigma(x,da),\qquad x,y\in X; \tag{3.1}$$

this follows from the structure of $\mathbf{P}_x^\pi$ in the case of $\pi=\sigma\in\Sigma$ and from Assumption 2.2. Moreover, $0\le q^\sigma\le\|p\|$, and as follows from Assumption 2.3 and (3.1),

$$q^\sigma(x,y)\ge\delta \ \text{ for all } \ x\in X,\ y\in D. \tag{3.2}$$

The one-step and multistep transition densities of a Markov chain in general are not unique. We will use their versions uniquely determined by the recurrence relation

$$q_t^\sigma(x,y)=\int_X q_{t-1}^\sigma(x,z)q^\sigma(z,y)m(dz), \tag{3.3}$$

$x,y\in X$, $t=2,3,\dots$, with $q_1^\sigma=q^\sigma$. In these notations

$$\mathbf{P}_x^\sigma\{x_t\in B\}=\int_B q_t^\sigma(x,y)m(dy),\quad x\in X,\quad B\in\mathcal{B}_X. \tag{3.4}$$

For a stationary policy σ also define $r^\sigma\in B(X)$ by

$$r^\sigma(x)=\int_{A_x} r(x,a)\sigma(x,da),\quad x\in X;\qquad \|\,r^\sigma\|\le\|\,r\|. \tag{3.5}$$

Then by the structure of $\mathbf{P}_x^\sigma$ we have $\mathbf{E}_x^\sigma[r(x_t,a_{t+1})\mid x_0a_1\dots x_t]=r^\sigma(x_t)$, so that in view of (3.4) the definition (2.2) of the discounted reward reduces to

$$v_\beta(x,\sigma)=\mathbf{E}_x^\sigma\sum_{t=0}^\infty\beta^t r^\sigma(x_t)=r^\sigma(x)+\sum_{t=1}^\infty\beta^t\int_X q_t^\sigma(x,y)r^\sigma(y)m(dy). \tag{3.6}$$

The condition (3.2) is a special case of the general Doeblin condition for Markov chains considered in Doob [12] (p. 197, Case (b), and pp. 216–217, Examples 2, 3 (continued)). It is proved there that this condition implies a uniform geometric convergence of the density $q_t^\sigma(x,y)$ to a limit $\bar q^\sigma(y)$ as $t\to\infty$, where $\bar q^\sigma(y)$ is a version of the density of the unique stationary distribution of the corresponding Markov chain. Namely,

$$|\,q_t^\sigma(x,y)-q^\sigma(y)\,|\le 2\,\|\,p\|\rho^{t-2},\quad x,y\in X,\quad t=1,2,\dots,\ \sigma\in\Sigma \tag{3.7}$$

where

$$0<\rho=1-\delta m(D)<1.$$

(Formally, to apply Doob's result to our case, one should multiply m and divide q^σ and δ by $\|p\|$.) It follows that there exists a bounded *deviation density*

$$z^\sigma(x,y)=\sum_{t=1}^\infty[q_t^\sigma(x,y)-\bar q^\sigma(y)],\ x,y\in X;\qquad \|z^\sigma\|\le\frac{2\|p\|}{\rho(1-\rho)}. \tag{3.8}$$

Next consider the limit of the expected one-step reward corresponding to σ (the average expected reward known in MDPs):

$$g^\sigma=\lim_{t\to\infty}\int_X q_t^\sigma(x,y)r^\sigma(y)m(dy)=\int_X\bar q^\sigma(y)r^\sigma(y)m(dy);\qquad |\,g^\sigma\,|\le\|r\|, \tag{3.9}$$

and subtract termwise the identity

$$\frac{g^\sigma}{\alpha} = (1+\beta+\beta^2+\cdots)g^\sigma = g^\sigma + \sum_{t=1}^{\infty}\beta^t \int_X \bar{q}^\sigma(y) r^\sigma(y) m(dy)$$

from (3.6). This leads to a relation

$$v_\beta(x,\sigma) - \frac{g^\sigma}{\alpha} = r^\sigma(x) - g^\sigma + \sum_{t=1}^{\infty}\beta^t \int_X [q_t^\sigma(x,y) - \bar{q}^\sigma(y)] r^\sigma(y) m(dy). \tag{3.10}$$

In view of the bounds in (3.5) and (3.7), the integral factor at β^t in (3.10) is majorized by $c\rho^t$ with $c = 2\|r\| \cdot \|p\|\rho^{-2}$. Therefore the power series in (3.10), as a function of a complex variable β, is analytic in the circle C_1 of radius $R_1 = \frac{1}{\rho}$ centered at 0. Consider another circle C_2: $|\beta - 1| < 1-\rho$ centered at 1 and of radius $R_2 = 1-\rho$. Since $\rho < 1$, we have $R_2 = 1-\rho < \frac{1-\rho}{\rho} = \frac{1}{\rho} - 1 < R_1 - 1$, so that $C_2 \subset C_1$. Therefore the right side of (3.10) is analytic in C_2 too, and can be expanded into a converging series in powers of $\beta - 1 = \alpha$ with a radius of convergence $\geq 1-\rho$. Thus from (3.10) we get a Laurent expansion

$$v_\beta(x,\sigma) = \frac{g^\sigma}{\alpha} + \sum_{n=0}^{\infty} h_n^\sigma(x)\alpha^n = \sum_{n=-1}^{\infty} h_n^\sigma(x)\alpha^n, \quad x \in X, \quad 0 < \alpha < 1-\rho. \tag{3.11}$$

By substituting $\beta = 1-\alpha$ into (3.10) and applying the binomial formula, one may express the coefficients h_n^σ in a form of uniformly convergent series containing integrals of q_t^σ, $\bar{q}^\sigma$, and r^σ, and see that $h_n^\sigma \in B(X)$, $n \geq -1$.

More useful formulas for h_n^σ are obtained by a substitution of the series (3.11) into the equation

$$v_\beta(x,\sigma) = r^\sigma(x) + (1-\alpha)\int_X q^\sigma(x,y) v_\beta(y,\sigma) m(dy), \qquad x \in X, \tag{3.12}$$

which is an immediate consequence of (3.3) and (3.6). The calculations are standard, and we briefly outline the main steps, leaving the details to the reader (they may be found also in [31, pp. 264–266]).

At this point it is convenient to introduce the operator notations. The relations

$$\begin{aligned} Q^\sigma f(x) &= \int_X q^\sigma(x,y) f(y) m(dy), \qquad \bar{Q}^\sigma f(x) = \int_X \bar{q}^\sigma(y) f(y) m(dy), \\ Z^\sigma f(x) &= \int_X z^\sigma(x,y) f(y) m(dy), \qquad x \in X \end{aligned} \tag{3.13}$$

define the (evidently bounded) operators Q^σ, $\bar{Q}^\sigma$, and Z^σ on the space $B(X)$. In view of (3.7), $(Q^\sigma)^t \to \bar{Q}^\sigma$ as $t \to \infty$. Also let I be the identity operator, and let $v_\beta^\sigma = v_\beta(\cdot,\sigma)$. In these notations the equation (3.12) becomes

$$v_\beta^\sigma = r^\sigma + (1-\alpha)Q^\sigma v_\beta^\sigma, \tag{3.14}$$

and (3.6) simplifies to

$$v_\beta^\sigma = \left[I + \sum_{t=1}^{\infty}(\beta Q^\sigma)^t\right] r^\sigma = \sum_{t=0}^{\infty}(\beta Q^\sigma)^t r^\sigma. \tag{3.15}$$

A substitution of (3.11) into (3.14) yields an identity

$$\frac{g^\sigma}{\alpha} + h_0^\sigma + h_1^\sigma\alpha + \cdots = r^\sigma + (1-\alpha)\left(\frac{Q^\sigma g^\sigma}{\alpha} + Q^\sigma h_0^\sigma + Q^\sigma h_1^\sigma\alpha + \cdots\right)$$

valid for $0 < \alpha < 1-\rho$. In view of the uniqueness of the Laurent coefficients, it follows that

$$g^\sigma = Q^\sigma g^\sigma;\ h_0^\sigma = r^\sigma + Q^\sigma(h_0^\sigma - g^\sigma);\ h_n^\sigma = Q^\sigma(h_n^\sigma - h_{n-1}^\sigma), n = 1, 2, \ldots . \tag{3.16}$$

These equations are solved with the help of relations

$$\bar{Q}^\sigma Q^\sigma = \bar{Q}^\sigma, \tag{3.17}$$

$$(I + Z^\sigma)(I - Q^\sigma) = I - \bar{Q}^\sigma, \tag{3.18}$$

$$(I + Z^\sigma)Q^\sigma = Z^\sigma + \bar{Q}^\sigma, \tag{3.19}$$

known in the theory of Markov chains and implied by the geometric convergence $(Q^\sigma)^t \to \bar{Q}^\sigma$ obtained in (3.7) (see, for example, [31, p. 264]). By iterating the first equation in (3.16), we get $g^\sigma = (Q^\sigma)^t g^\sigma$, and in the limit, $g^\sigma = \bar{Q}^\sigma g^\sigma$. Next, multiplying the second equation by $\bar{Q}^\sigma$ and utilizing (3.17), we obtain $\bar{Q}^\sigma h_0^\sigma = \bar{Q}^\sigma r^\sigma + \bar{Q}^\sigma h_0^\sigma - \bar{Q}^\sigma g^\sigma$, so that $g^\sigma = \bar{Q}^\sigma r^\sigma$ (as in (3.9)).

This same multiplication applied to the third equation in (3.16) shows that $\bar{Q}^\sigma h_{n-1}^\sigma = 0$ for $n \geq 1$. In particular, $\bar{Q}^\sigma h_0^\sigma = 0$. Multiplying the second equation by $I + Z^\sigma$ we get $(I + Z^\sigma)(I - Q^\sigma)h_0^\sigma = (I + Z^\sigma)(r^\sigma - Q^\sigma g^\sigma)$ or, in view of (3.18) and since $\bar{Q}^\sigma h_0^\sigma = 0$, $h_0^\sigma = (I + Z^\sigma)(r^\sigma - Q^\sigma g^\sigma)$. Here $Q^\sigma g^\sigma = g^\sigma$ is a constant, and therefore $Z^\sigma Q^\sigma g^\sigma = 0$. Hence $h_0^\sigma = (I + Z^\sigma)r^\sigma - g^\sigma$.

In a similar way, the third equation $h_n^\sigma = Q^\sigma(h_n^\sigma - h_{n-1}^\sigma)$, being multiplied by $I + Z^\sigma$, with the help of (3.19) and the relation $\bar{Q}^\sigma h_{n-1}^\sigma = 0$, is finally transformed into $h_n^\sigma = -Z^\sigma h_{n-1}^\sigma$, $n \geq 1$. Summarizing and taking into account the bounds in (3.5), (3.8), and (3.9), we have the following results.

LEMMA 3.1. *For every policy $\sigma \in \Sigma$, the expected discounted reward v_β^σ expands into the Laurent series* (3.11) *with the coefficients*

$$g^\sigma = h_{-1} = \bar{Q}^\sigma r^\sigma, \qquad \|g^\sigma\| \leq \|r\|;$$

$$h_n^\sigma = (-Z^\sigma)^n(r^\sigma + Z^\sigma r^\sigma - g^\sigma),\ \|h_n^\sigma\| \leq (2+\kappa)\kappa^n,\ \kappa = \frac{2\|p\|}{\rho(1-\rho)},\ n \geq 0. \tag{3.20}$$

These coefficients belong to $B(X)$ and are uniquely determined by the equations (3.16).

Notice that $\|p\| \geq 1$ since $\int_X p(x,a,y)m(dx) = 1$ and $m(X) = 1$. Therefore, for the constant κ in (3.20), we have

$$\frac{1}{\kappa} \leq \frac{\rho(1-\rho)}{2} < 1 - \rho.$$

The above results suggest two definitions. In the first of them we specify a vector space $\mathcal{H}$ such that all possible sequences of the Laurent coefficients of $v_\beta(x,\sigma)$ belong to $\mathcal{H}$, and we introduce some notations.

DEFINITION 3.2.

(i) *The space $\mathcal{H}$ consists of all sequences $H = \{h_n\}$ of functions $h_n \in B(X)$, $n = -1, 0, 1, \ldots$, satisfying the following two conditions:* 1) *h_{-1} is a constant (often denoted by g),* 2) *for every $H \in \mathcal{H}$ there exists a constant $C(H)$ such that $\|h_n\| \leq C(H)\kappa^n$, where κ is defined in* (3.20).

(ii) *With every $H \in \mathcal{H}$ we associate a function $V_\beta(x; H)$ defined by the relation*

$$V_\beta(x; H) = \sum_{n=-1}^{\infty} h_n(x)\alpha^n, \quad x \in X, \quad 0 < \alpha = 1 - \beta < \frac{1}{\kappa}. \tag{3.21}$$

(iii) *If for a policy $\pi \in \Pi$ there exists an (evidently unique) element $H \in \mathcal{H}$ such that $V_\beta(x; H) = v_\beta(x, \pi)$, $0 < 1 - \beta < \frac{1}{\kappa}$, then this element is denoted $H^\pi = \{h_n^\pi\}$.*

The second definition relates with every stationary policy σ an operator L^σ on the space $\mathcal{H}$ such that, similar to (3.14),

$$V_\beta(x; L^\sigma H) = r^\sigma(x) + \beta Q^\sigma V_\beta(x; H), \qquad H \in \mathcal{H}. \tag{3.22}$$

In the same way as we obtained (3.16) from (3.14), we come to the following definition.

DEFINITION 3.3. *The operator $L^\sigma, \sigma \in \Sigma$, transforms every $H = \{h_n\} \in \mathcal{H}$ into $H' = \{h'_n\} \in \mathcal{H}$ according to the formulas*

$$g' = g; \quad h'_0 = r^\sigma + Q^\sigma(h_0 - g); \quad h'_n = Q^\sigma(h_n - h_{n-1}), \quad n \geq 1 \tag{3.23}$$

(since $\|Q^\sigma\| = 1$, $H' \in \mathcal{H}$ together with H).

In these notations we have the following evident consequence of Lemma 3.1.

COROLLARY 3.4. *For every $\sigma \in \Sigma$, H^σ is the unique solution in $\mathcal{H}$ of the equation*

$$L^\sigma H = H.$$

4. Lexicographical policy improvement and Blackwell optimality equation. In this section we obtain lexicographical analogues of the policy improvement and of the Bellman optimality equation, known in dynamic programming. These two concepts are closely related: the optimality equation means that it is impossible to improve.

Similar to Definitions 3.2 and 3.3, we will denote throughout the paper by $\{b_n\}$, $\{f_n\}$, etc., sequences of real numbers or of functions on the same space E indexed by $n = -1, 0, 1, 2, \dots$; the sequence as a whole will be denoted by the same capital letter. In particular, instead of $\{0, 0, \dots\}$ we may write 0. For two numerical sequences B and B', the notation $B \prec B'$ (or $B' \succ B$) means that either $b_{-1} < b'_{-1}$ or, for some index $N \geq 0$, $b_n = b'_n$ for all $n < N$, while $b_N < b'_N$. The notations $B \preceq B'$ and $B' \succeq B$ are self-evident. For a family $\{B^\gamma\}$ of numerical sequences depending on a parameter $\gamma \in \Gamma$, the notation $B = \operatorname{Lexmax}_{\gamma\in\Gamma} B^\gamma$ means that $B \succeq B^\gamma$ for all $\gamma \in \Gamma$, and $B = B^\gamma$ for at least one $\gamma \in \Gamma$.

In the case of functions on a space E, the notation $F \preceq F'$ (or $F' \succeq F$) means that $F(x) \preceq F'(x)$ for all $x \in E$. An integral of $F = \{f_n\}$ is understood termwise. Elementary properties of this partial ordering will be used without stating them explicitly. We formulate only the following less evident one.

PROPOSITION 4.1. *Let E be a Borel space with a finite measure ν. If $F = \{f_n\} \preceq F' = \{f'_n\}$, where $f_n, f'_n \in B(E)$, then*

$$\int_E F(x)\nu(dx) \preceq \int_E F'(x)\nu(dx);$$

if in addition $\nu\{x \in E : F(x) \prec F'(x)\} > 0$, then the same holds with the sign "$\prec$".

Proof. Consider $N = \{\min n : \nu\{x : f_n(x) \neq f'_n(x)\} > 0\}$. □

The following comparison lemma is a key one.

LEMMA 4.2. *If $L^\sigma H \succeq H$ for some $\sigma \in \Sigma$ and $H \in \mathcal{H}$, then $H^\sigma \succeq H$. If in addition $L^\sigma H(x_0) \succ H(x_0)$ at some $x_0 \in X$, then $H^\sigma(x_0) \succ H(x_0)$. Similar results are valid with the opposite inequality signs. If $L^\sigma H = H$ then $H^\sigma = H$.*

Proof. To prove the first statement, it is sufficient to show that for an arbitrary fixed $x \in X$ we have $v_\beta(x,\sigma) \geq V_\beta(x;H)$ for β sufficiently close to 1 (see (3.11) and (3.21)). Since $V_\beta(H)$ is linear in H, from (3.22) we have

$$r^\sigma - V_\beta(H) + \beta Q^\sigma V_\beta(H) = V_\beta(L^\sigma H - H).$$

To each term of this relation we apply the (geometrically convergent) operator $I + \beta Q^\sigma + (\beta Q^\sigma)^2 + \cdots$. Then the term r^σ transforms into v_β^σ in accordance with (3.15). The next two terms transform into $-(I + \beta Q^\sigma + (\beta Q^\sigma)^2 + \cdots)(I - \beta Q^\sigma)V_\beta(H) = -V_\beta(H)$, so that we get

$$v_\beta^\sigma - V_\beta(H) = \sum_{t=0}^{\infty} (\beta Q^\sigma)^t V_\beta(L^\sigma H - H).$$

Since $1 + \beta + \beta^2 + \cdots = \frac{1}{\alpha}$ and $Z^\sigma = \sum_1^\infty [(Q^\sigma)^t - \bar{Q}^\sigma]$ (see (3.8)), the operator on the right side can be transformed into

$$\begin{aligned}\sum_0^\infty (\beta Q^\sigma)^t &= \frac{\bar{Q}^\sigma}{\alpha} + \sum_0^\infty \beta^t [(Q^\sigma)^t - \bar{Q}^\sigma] \\ &= \frac{\bar{Q}^\sigma}{\alpha} + (I - \bar{Q}^\sigma) + Z^\sigma + \sum_1^\infty (\beta^t - 1)[(Q^\sigma)^t - \bar{Q}^\sigma].\end{aligned}$$

In terms of densities, we obtain in that way

$$v_\beta(x,\sigma) - V_\beta(x;H) = V_\beta(x; L^\sigma H - H) + \int_X u(\alpha,x,y) V_\beta(y; L^\sigma H - H) m(dy), \tag{4.1}$$

where

$$u(\alpha,x,y) = \frac{\bar{q}^\sigma(y)}{\alpha} - \bar{q}^\sigma(y) + z^\sigma(x,y) + \sum_{t=1}^\infty (\beta^t - 1)[q_t^\sigma(x,y) - \bar{q}^\sigma(y)]. \tag{4.2}$$

Since $L^\sigma H - H \succeq 0$, the first term at the right side of (4.1) is nonnegative for positive α close enough to 0, so that it remains to prove that the integral in (4.1) becomes nonnegative as $\alpha \downarrow 0$.

For this purpose consider the Laurent series for $V_\beta^\sigma(y; L^\sigma H - H)$:

$$V_\beta(y; L^\sigma H - H) = \sum_{n=-1}^\infty l_n(y)\alpha^n, \quad y \in X, \quad 0 < \alpha < \frac{1}{\kappa}.$$

In accordance with Definitions 3.2 and 3.3, here $\{l_n\} \in \mathcal{H}$, so that $\|l_n\| \leq C\kappa^n$ for some constant C, and $l_{-1} = 0$. Since $L^\sigma H - H \succeq 0$, we may define disjoint sets $Y_n, n = 0, 1, 2, \ldots$ in either of the following two ways:

$$\begin{aligned}Y_n &= \{y \in X : \ l_0(y) = l_1(y) = \cdots = l_{n-1}(y) = 0, l_n(y) \neq 0\} \\ &= \{y \in X : \ l_0(y) = l_1(y) = \cdots = l_{n-1}(y) = 0, l_n(y) > 0\}.\end{aligned}$$

Consider also the sets

$$X_1 = \{y \in X :\ \bar{q}^\sigma(y) > 0\},$$
$$X_2 = X_2(x) = \{y \in X :\ \bar{q}^\sigma(y) = 0,\ z^\sigma(x,y) > 0\}.$$

By (3.1) and (3.3), the transition densities $q_t^\sigma(x,y)$ are nonnegative. Therefore, if $\bar{q}^\sigma(y) = 0$ for some $y \in X$, then $z^\sigma(x,y) = \sum_1^\infty q^\sigma(x,y) \geq 0$ for the same y; and if also $z^\sigma(x,y) = 0$, then $u(\alpha,x,y) = 0$ (see (4.2)). Hence the integral in (4.1) splits into a sum of integrals over X_1 and X_2, where

$$u(\alpha,x,y) = \begin{cases} \frac{\bar{q}^\sigma(y)}{\alpha} + O(1),\ \bar{q}^\sigma(y) > 0 & \text{if } y \in X_1, \\ z^\sigma(x,y) + o(1),\ z^\sigma(x,y) > 0 & \text{if } y \in X_2; \end{cases} \tag{4.3}$$

here $O(1)$ and $o(1)$ are uniform in y as $\alpha \to 0$ according to the bounds obtained in (3.7) and (3.8).

To prove that the integral over X_1 becomes nonnegative as $\alpha \downarrow 0$, consider $N = \min\{n :\ m(X_1 Y_n) > 0\}$. If $N = \infty$, then $l_n = 0$ (a.e. m) on X_1 for all $n = 0,1,2,\ldots,$ and the integral is equal to 0. If $N < \infty$, then $l_n = 0$ (a.e. m) on X_1 for $n = 0,1,\ldots,N-1$, so that

$$\begin{aligned} \int_{X_1} u(\alpha,x,y) V_\beta(y; L^\sigma H - H) m(dy) &= \int_{X_1} \left[\frac{\bar{q}^\sigma}{\alpha} + O(1)\right] \sum_{n=N}^\infty l_n(y)\alpha^n m(dy) \\ &= \alpha^{N-1} \int_{X_1} \bar{q}^\sigma(y) l_N(y) m(dy) + o(\alpha^{N-1}) \end{aligned}$$

as $\alpha \downarrow 0$ (here the bounds $\|l_n\| \leq C\kappa^n$ are used). By the selection of N, we have (1) $l_N \geq 0$ (a.e. m) on X_1 since $L^\sigma H - H \succeq 0$ and since all the preceding l_n vanish a.e. on X_1; (2) $l_N > 0$ on the set $X_1 Y_N$ of a positive measure m. By (4.3) the factor $\bar{q}^\sigma$ is also strictly positive on X_1, so that the same is true for the integral factor at α^{N-1}. Hence, in the case $N < \infty$, the integral over X_1 in (4.1) becomes positive as $\alpha \downarrow 0$. Thus, in any case, it is nonnegative for α close enough to 0.

The integral over X_2 is treated in a similar way, with $N = \min\{n :\ m(X_2 Y_n) > 0\}$. In the case $N = \infty$ this integral is 0, while in the case $N < \infty$, in accordance with (4.3),

$$\int_{X_2} u(\alpha,x,y) V_\beta(y; L^\sigma H - H) m(dy) = \alpha^N \int_{X_2} z^\sigma(x,y) l_N(y) m(dy) + o(\alpha^N) > 0$$

as $\alpha \downarrow 0$. For the case "$\succeq$" the lemma is proved.

The case $H^\sigma(x_0) \succ H(x_0)$ follows from the already proven nonnegativity of the integral in (4.1) and from the strict inequality $V_\beta(x_0; L^\sigma H - H) > 0$ implied by the relation $L^\sigma H(x_0) \succ H(x_0)$. The case of the opposite inequalities is similar, only with negative l_N and $V_\beta(x; L^\sigma H - H)$. The equality case is a consequence of the preceding cases, since in that case both nonstrict inequalities are satisfied. □

COROLLARY 4.3 (lexicographical policy improvement). *If $L^\tau H^\sigma \succeq H^\sigma$ for two stationary policies σ and τ, then $H^\tau \succeq H^\sigma$. If, in addition, $L^\tau H^\sigma(x_0) \succ H^\sigma(x_0)$ at some $x_0 \in X$, then $H^\tau(x_0) \succ H^\sigma(x_0)$.*

The next lemma is an immediate consequence of the Laurent expansion (3.11) and the definition of the sign "$\succeq$".

LEMMA 4.4. *For two policies $\sigma,\tau \in \Sigma$, the relation $H^\tau \succeq H^\sigma$ is equivalent to the existence for every $x \in X$ of a number $\beta_0 = \beta(x,\sigma,\tau)$ such that*

$$v_\beta(x,\tau) \geq v_\beta(x,\sigma) \quad \text{for all} \quad \beta_0 < \beta < 1. \tag{4.4}$$

Consider now the following operator T on the space $\mathcal{H}$:

$$TH(x) = \operatorname*{Lexmax}_{\sigma\in\Sigma} L^\sigma H(x), \quad x \in X, \tag{4.5}$$

defined for those $H \in \mathcal{H}$ for which the Lexmax in (4.5) exists (we avoid a lexicographical supremum, since it is something awkward: $\operatorname{Lexsup}_n\{-\frac{1}{n}, 0, 0, \dots\} = \{0, -\infty, -\infty, \dots\}$).

THEOREM 4.5. *For a policy $\tau \in \Sigma$ the following statements are equivalent:*

(i) *τ is Blackwell optimal within the class Σ;*

(ii) *$H^\tau \succeq H^\sigma$ for every $\sigma \in \Sigma$;*

(iii) *H^τ is a solution of the equation*

$$H = TH, \quad H \in \mathcal{H}; \tag{4.6}$$

(iv) *$L^\tau H = H$ for some solution of* (4.6).

Proof. The equivalence of (i) and (ii) follows from Lemma 4.4 and Definition 2.6. It remains to prove the implications (iv)→(iii)→(ii) and (ii)→(iii)→(iv). If (iv) holds, then $H^\tau = H$ by Lemma 4.2; therefore $H^\tau = TH^\tau$, and we have (iii). If $TH^\tau = H^\tau$, then $L^\sigma H^\tau \preceq H^\tau$ for every $\sigma \in \Sigma$ by the definition of T, hence $H^\sigma \preceq H^\tau$ by Corollary 4.3, and (ii) holds. The implication (ii)→(iii) is proved by a contradiction. If (iii) does not hold, then, in view of Corollary 3.4, for some $\sigma \in \Sigma$ and some $x_0 \in X$ we have $L^\sigma H^\tau(x_0) \succeq H^\tau(x_0)$. In this case consider a policy

$$\rho(x,\cdot) = \begin{cases} \sigma(x,\cdot) & \text{if } x = x_0, \\ \tau(x,\cdot) & \text{if } x \neq x_0, \end{cases}$$

for which L^ρ coincides with L^σ at the point x_0 and coincides with L^τ at all other points $x \in X$ (see (3.1), (3.5), and (3.23)). For this policy $L^\rho H^\tau(x_0) \succ H^\tau(x_0)$, and by Corollary 3.4 $L^\rho H^\tau(x) = H^\tau(x)$ at all other $x \in X$. Hence $H^\rho(x_0) \succ H^\tau(x_0)$ by Corollary 4.3, in contradiction with (ii). Finally, (iii) implies (iv) because $L^\tau H^\tau = H^\tau$ (Corollary 3.4). □

Following Dekker and Hordijk [8, p. 402] we use the name *Blackwell optimality equation* for the equation (4.6) and the name *conserving condition* for the relation (iv) of Theorem 4.5. For the operator T we suggest the name *lexicographical Bellman operator*. Another form of this operator is given in section 6, where the existence of a unique solution of the Blackwell optimality equation is also proved.

5. Continuity of Laurent coefficients. Our next goal is to compactify the space Σ in such a way that the Laurent coefficients of $v_\beta(x,\sigma)$ become continuous functions of σ. We will not achieve this goal literally. First, the needed compactification of Σ requires an aggregation of Σ into a space S of measures on K; the topology in S is introduced by means of Carathéodory functions. Furthermore, to turn H^σ into a function on S, it is necessary to replace in $v_\beta(x,\sigma)$ the initial state x by an absolutely continuous initial distribution. Moreover, to prove the continuity of h_n^σ by an induction in n, we need to consider initial "distributions" with arbitrary bounded, maybe negative densities.

DEFINITION 5.1.

(i) *The space S consists of all probability measures s on K satisfying the condition*

$$\Pr_X s = m, \tag{5.1}$$

where m is the reference measure (see Assumption 2.2) and where the projection of a measure is defined by $\Pr_X s(B) = s(K \cap (B \times X))$, $B \in \mathcal{B}_X$.

(ii) *The weak-strong topology in S is defined by the condition $s_n \overset{ws}{\to} s$ if*

$$\lim_{n\to\infty} \int_K f ds_n = \int_K f ds \qquad \text{for every } f \in \mathrm{Car}(K). \tag{5.2}$$

PROPOSITION 5.2. *The space S is (sequentially) compact in the weak-strong topology.*

This is a known result. With a somewhat different class of functions f, in the context of relaxed controls in the deterministic control theory, it is a part of Theorem IV.3.11 in Warga [30, p. 287]. In the form we need here, it follows from the existence of resolution topologies in the dynamic programming proved in Schäl [23], [24] and Balder [1]. Our Proposition 5.2 is a consequence of Theorem 2.1 and Proposition 3.2 in [1, pp. 144, 148], obtained by a reduction of the dynamic programming model to its first step. A simpler direct proof is given in [33].

LEMMA 5.3. *There is a unique mapping $j : \Sigma \to S$ such that*

$$\int_X \int_{A_x} f(x,a)\sigma(x,da)m(dx) = \int_K f ds \qquad \text{for every } f \in B(K) \tag{5.3}$$

if $s = j(\sigma)$, and j maps Σ onto S.

Proof. Evidently, (5.3) is satisfied if and only if the measure s on K is defined by $s(dxda) = \sigma(x,da)m(dx)$. To verify (5.1) for the so-defined $s = j(\sigma)$, it is sufficient to apply (5.3) to the indicator of the set $(B \times A) \cap K$, $B \in \mathcal{B}_X$. It remains to prove that j is a map onto S. For this purpose, given $s \in S$, we first extend s to the product space $X \times A$ by setting $s((X \times A) \setminus K) = 0$ and then factorize s into $s(dxda) = \sigma(x,da)m(dx)$ where $\sigma(x,\cdot)$ is a regular conditional distribution on A given x. (The marginal distribution on X coincides with m in view of (5.1), and the existence of regular conditional distributions in the case of standard Borel spaces is a well-known fact.) The so-obtained σ satisfies the definition of a stationary policy, except that instead of $\sigma(x, A_x) = 1$ we have $\sigma(x, A) = 1$, so that maybe $\sigma(x, A_x) < 1$ for some $x \in X$. From (5.3) with $f = 1$ we get

$$\int_X \sigma(x, A_x)m(dx) = s(K) = 1 = \int_X 1m(dx),$$

and since $\sigma(x, A_x) \le 1$, it follows that $\sigma(x, A_x) = 1$ (a.e. m). It remains to fix some $\sigma_0 \in \Sigma$ and to change $\sigma(x,\cdot)$ to $\sigma_0(x,\cdot)$ on the set $\{x \in X : \sigma(x, A_x) < 1\}$ of measure 0. □

LEMMA 5.4. *Suppose that $f \in \mathrm{Car}(K)$ and that $g(x,s), x \in X, s \in S$ is a bounded function, measurable in x and continuous in s. Then the function*

$$F(s) = \int_K f(x,a)g(x,s)s(dxda), \qquad s \in S,$$

is continuous on S.

Proof. For $s, s' \in S$ we have $F(s') - F(s) = I_1 + I_2$, where

$$I_1 = \int_K f(x,a)[g(x,s') - g(x,s)]ds', \quad I_2 = \int_K f(x,a)g(x,s)(ds' - ds).$$

Here I_2 vanishes as $s' \overset{ws}{\to} s$ by the definition of the topology in S, since the integrand (with s fixed) is a Carathéodory function on K. For I_1, in accordance with (5.1), we have a bound

$$| I_1 | \le \|f\| \int_K | g(x,s') - g(x,s) | s'(dxda) = \|f\| \int_X | g(x,s') - g(x,s) | m(dx).$$

The last integrand is uniformly bounded by $\|g\|$ and converges to 0 at every x as $s' \overset{ws}{\to} s$. Since $m(X) = 1 < \infty$, the integral also converges to 0 by the majorized convergence theorem. □

Next we have to replace the initial states $x \in X$ by initial densities $l \in B(X)$. Avoiding formalities, we will use the same notations for l as for x, so that

$$q_t^\sigma(l,y) = \int_X l(x) q_t^\sigma(x,y) m(dx), \quad t \geq 1, \quad \bar{q}^\sigma(l,y) = \int_X l(x) \bar{q}^\sigma(y) m(dx), \tag{5.4}$$

and $z^\sigma(l,y)$, $H^\sigma(l) = \{h_n^\sigma(l)\}$, $v_\beta(l,\sigma)$ are similar integrals of $z^\sigma(x,y)$, $H^\sigma(x) = \{h_n^\sigma(x)\}$, $v_\beta(x,\sigma)$; to avoid confusion, x, y, z will always denote the states, while $l, l^s, 1$ will denote the elements of $B(X)$. It is clear that the new functions satisfy relations similar to (3.3), (3.8), and (3.11) and are bounded by the same constants as the previous functions, only multiplied by $\|l\|$ (see (3.8), (3.9), and (3.20)).

Our goal is the continuity of $h_n^\sigma(1), n \geq -1$, as functions of $s = j(\sigma)$, but in order to proceed with an induction in n, we need to replace the density 1 by an arbitrary $l \in B(X)$.

LEMMA 5.5. *For every $l \in B(X)$ and $y \in X$, the functions $q_t^\sigma(l,y)$ $(t \geq 1)$, $\bar{q}^\sigma(l,y)$, $z^\sigma(l,y)$, $h_n^\sigma(l,y)$ $(n \geq -1)$ of σ depend only on $s = j(\sigma)$ and are continuous in $s \in S$.*

As soon as this result is proved for some function, we feel free to use s in place of σ in the notation of that function.

Proof. For $q_t^\sigma(l,y)$, $t \geq 1$, we proceed by an induction in t. For $t = 1$ from (3.1), (5.4), and (5.3) we have

$$\begin{aligned} q^\sigma(l,y) &= \int_X \int_{A_x} l(x) p(x,a,y) \sigma(x,da) m(dx) \\ &= \int_K l(x) p(x,a,y) s(dxda) = q^s(l,y), \quad s = j(\sigma), \end{aligned}$$

and the continuity of $q^s(l,y)$ follows from Lemma 5.4 ($l(\cdot)p(\cdot,\cdot,y) \in \mathrm{Car}(K)$ in accordance with Assumption 2.2). From (3.1), (3.3), and (5.3) we have

$$\begin{aligned} q_{t+1}^\sigma(l,y) &= \int_X q_t^\sigma(l,z) q^\sigma(z,y) m(dz) = \int_X q_t^s(l,z) \int_{A_x} p(z,a,y) \sigma(z,da) m(dz) \\ &= \int_K q_t^s(l,z) p(z,a,y) s(dzda) = q_{t+1}^s(l,y), \quad s = j(\sigma), \end{aligned}$$

and by Lemma 5.4, $q_{t+1}^s(l,y)$ is continuous in s together with $q_t^s(l,y)$.

Since a uniform limit of a sequence of continuous functions is again a continuous function, the continuity of $\bar{q}^\sigma$ and z^σ in $s = j(\sigma)$ follows from the relations

$$| q_t^s(l,y) - q^\sigma(l,y) | \leq 2\|l\| \cdot \|p\| \cdot \rho^{t-2}, \quad z^\sigma(l,y) = \sum_{t=1}^{\infty} [q_t^s(l,y) - \bar{q}^s(l,y)]$$

implied by (3.7) and (3.8).

Now consider $h_n^\sigma(l)$. For $n=-1$, in accordance with (3.20), (3.9), (3.5), and (5.4), we have

$$\begin{aligned} h_{-1}^\sigma(l) &= \int_X l(x) \int_X \bar{q}^\sigma(y) r^\sigma(y) m(dy) m(dx) \\ &= \int_X \bar{q}^\sigma(l,y) \int_{A_y} r(a,y)\sigma(y,da) m(dy) \\ &= \int_K \bar{q}^s(l,y) r(y,a) s(dyda) = h_{-1}^s(l), \qquad s = j(\sigma). \end{aligned}$$

Since $\bar{q}^s(l,y)$ is continuous in s, and $r \in \mathrm{Car}(K)$, the continuity of $h_{-1}^s(l)$ follows from Lemma 5.4. For $n=0$ by (3.20), (3.5), and (3.13) we have

$$\begin{aligned} & h_0^\sigma(l) \\ &= \int_X l(x) \left[\int_{A_x} r(x,a)\sigma(x,da) + \int_X z^\sigma(x,y) \int_{A_y} r(y,a), \sigma(y,da) m(dy) \right] m(dx) - h_{-1}^\sigma(l) \\ &= \int_K l(x) r(x,a) s(dxda) + \int_X z^s(l,y) \int_{A_y} r(y,a)\sigma(y,da) m(dy) - h_{-1}^s(l) \\ &= \int_K [l(x) + z^s(l,x)] r(x,a) s(dxda) - h_{-1}^s(l) = h_0^s(l), \quad s = j(\sigma). \end{aligned}$$

Since $z^s(l,x)$ is continuous in s, by Lemma 5.4 the same is true for $h_0^s(l)$.

Further, we proceed by induction. Suppose that for some $n \geq 0$ the continuity of $h_n^s(l)$ $(= h_n^\sigma(l)$ in $s = j(\sigma))$ for every $l \in B(X)$ is already known. In accordance with (3.13) and (3.20), and since $z^\sigma(l,y) = z^s(l,y)$, we have

$$\begin{aligned} h_{n+1}^\sigma(l) &= -\int_X l(x) \int_X z^\sigma(x,y) h_n^\sigma(y) m(dy) m(dx) \\ &= -\int_X z^s(l,y) h_n^\sigma(y) m(dy) = h_n^\sigma(l^s), \quad s = j(\sigma), \end{aligned} \tag{5.5}$$

where $l^s(\cdot) = -z^s(l,\cdot) \in B(X)$ (cf. (5.4)). If also $s = j(\tau)$ for some $\tau \in \Sigma$, then in a similar way $h_{n+1}^\tau(l) = h_n^\tau(l^s)$. By the supposition of induction applied to l^s in place of l, we have $h_n^\sigma(l^s) = h_n^\tau(l^s)$. It follows from (5.5) that $h_{n+1}^\sigma(l) = h_{n+1}^\tau(l)$, so that $h_{n+1}^\sigma(l)$ is indeed a function of $s = j(\sigma)$ and that

$$h_{n+1}^s(l) = h_n^s(l^s), \quad s \in S.$$

It remains to show that the difference

$$h_{n+1}^t(l) - h_{n+1}^s(l) = [h_n^t(l^t) - h_n^t(l^s)] + [h_n^t(l^s) - h_n^s(l^s)]$$

converges to 0 as $t \overset{ws}{\to} s$ $(s,t \in S)$. The second difference at the right side tends to 0 by the supposition of induction applied to l^s in place of l. The first difference is equal to the integral

$$\int_X [z^s(l,y) - z^t(l,y)] h_n^\tau(y) m(dy), \qquad \text{where } j(\tau) = t.$$

Here the integrand is uniformly bounded by the constant $(2+\kappa)\kappa^n \parallel r\| \cdot 2\|l\|\kappa$ (see (3.8) and (3.20)) and converges to 0 at every $y \in X$ as $t \overset{ws}{\to} s$ since $z^s(l,y)$ is continuous

in $s \in S$. Since $m(X) < \infty$, the integral converges to 0 by the majorized convergence theorem. □

COROLLARY 5.6. *The average reward $g^\sigma, \sigma \in \Sigma$, is a continuous function of $s = j(\sigma)$.*

Proof. The proof follows from the continuity of $h^s_{-1}(l)$ in s for every $l \in B(X)$ and the relation $h^\sigma_{-1}(1) = g^\sigma \int_X 1m(dx) = g^\sigma$. □

6. Blackwell optimality within the class of stationary policies. In this section we prove the existence of a selector φ which is Blackwell optimal within Σ. This is done in two steps. First we obtain a policy $\tau \in \Sigma$ which is the best for a given initial density l. After that, we obtain φ from τ by policy improvement. To be definite, we take $l = 1$. The existence of an optimal selector allows us to complete the results of Theorem 4.5 by proving the existence of a unique solution to the Blackwell optimality equation.

Notice that in accordance with Lemma 3.1 and with notations introduced in Lemma 5.5,

$$v_\beta(1, s) = \frac{g^s}{\alpha} + \sum_{n=0}^{\infty} h^s_n(1)\alpha^n, \quad 0 < \alpha < \frac{1}{\kappa}, \quad s \in S, \tag{6.1}$$

where $\{g^s, h^s_0(1), \dots\} = \{h^s_n(1)\} = H^s(1)$ and $H^\sigma(1) = H^s(1)$ if $s = j(\sigma), \sigma \in \Sigma$.

LEMMA 6.1. *There exists a policy $\tau \in \Sigma$ such that*

$$H^\tau(1) = \operatorname*{Lexmax}_{\sigma\in\Sigma} H^\sigma(1). \tag{6.2}$$

Proof. By Lemma 5.5, all the coefficients of the Laurent series (6.1) are continuous in s. By Proposition 5.2, S is a compact space. Therefore $g^s = h^s_{-1}(1)$ attains its maximum over S on a nonempty compact set $S_0 \subseteq S$. In a similar way, by induction, $h^s_n(1)$ attains its maximum over S_n on a nonempty compact set $S_{n+1} \subseteq S_n$, $n = 0, 1, \dots$. The intersection S_∞ of the sets $S \supseteq S_0 \supseteq S_1 \supseteq \cdots$ is also a nonempty compact set, and evidently $H^t(1) = \operatorname{Lexmax}_{s\in S} H^s(1)$ for any $t \in S_\infty$. By Lemma 5.3 the set $j^{-1}(t)$ is nonempty, and by Lemma 5.5 the relation (6.2) holds for any $\tau \in j^{-1}(t)$. □

To perform the next step, we need the analogues $L^a, a \in A$, of the operators L^σ such that

$$L^\sigma H(x) = \int_{A_x} L^a H(x)\sigma(x, da), \quad x \in X, \quad \sigma \in \Sigma \quad H \in \mathcal{H} \tag{6.3}$$

(see Definition 3.3). In accordance with (3.1) and (3.13), the following definition implies (6.3).

DEFINITION 6.2. *For every $H = \{g, h_0, h_1, \dots\} \in \mathcal{H}$, $x \in X$, and $a \in A_x$,*

$$L^a H(x) = \{g,\ r(x, a) + P^a h_0(x) - g,\ P^a(h_1 - h_0)(x),\ P^a(h_2 - h_1)(x), \dots\} \tag{6.4}$$

where

$$P^a f(x) = \int_X p(x, a, y) f(y) m(dy), \quad f \in B(X). \tag{6.5}$$

(It is easy to see that $L^a H^\sigma(x)$ is indeed the sequence of Laurent coefficients of $v_\beta(x, \pi)$, where π assigns the action a at the state x at the first step of the control and coincides with σ afterwards.)

THEOREM 6.3.

(i) *The lexicographical Bellman operator T introduced in* (4.5) *is defined on the whole space $\mathcal{H}$, and*

$$TH(x) = \operatorname*{Lexmax}_{a\in A_x} L^a H(x), \quad x \in X,\ H \in \mathcal{H}; \tag{6.6}$$

(ii) *for every $H \in \mathcal{H}$ there exists a selector φ such that $L^\varphi H = H$;*

(iii) *$TH \in \mathcal{H}$ for every $H \in \mathcal{H}$.*

Proof. We first prove that the maximum in (6.6) is attained and determine the corresponding subsets of the sets A_x. Given $H = \{h_n\} \in \mathcal{H}$, denote by $h_n(x,a)$, $n \geq -1$, the components of $L^a H(x)$ defined in (6.4), and fix x. Since $h_n \in B(X)$, it follows from (6.4)–(6.5) and Assumptions 2.2 and 2.4 by the majorized convergence theorem that all $h_n(x,a)$ are continuous in a on the compact A_x (so that $h_n(\cdot,\cdot) \in \mathrm{Car}(K)$). The constant term $h_{-1}(x,a) = g$ of $L^a H(x)$ attains its maximum $\hat{h}_{-1}(x)$ $(= \hat{g})$ over A_x on the set $A_{-1}(x) = A_x$. The next term $h_0(x,a)$ attains its maximum $\hat{h}_0(x)$ over $A_{-1}(x)$ on a nonempty compact set $A_0(x) \subset A_{-1}(x)$, etc. By an evident induction, we get nonempty compact sets $A_x = A_{-1}(x) \supset A_0(x) \supset A_1(x) \supset \cdots$ such that $h_n(x,a)$ attains its maximum $\hat{h}_n(x)$ over $A_{n-1}(x)$ on $A_n(x), n \geq 0$. The intersection $A_\infty(x)$ of the sets $A_n(x), n \geq 0$, is a nonempty compact subset of A_x. It is clear that $\hat{H}(x) = \{\hat{h}_n(x)\}$ is the lexicographical maximum on the right side of (6.6) and that this maximum is attained if and only if $a \in A_\infty(x)$.

Since x is arbitrary, we have $L^a H\,(x) \preceq \hat{H}(x)$ for every $x \in X, a \in A_x$. Hence by (6.3) and Proposition 4.1, $L^\sigma H \preceq \hat{H}$ for every $\sigma \in \Sigma$.

In accordance with the preceding paragraph, the set

$$K_\infty = \{(x,a) : a \in A_\infty(x), x \in X\}$$

has nonempty compact x-sections $A_\infty(x)$. To choose a selector φ with its graph in K_∞, we need to prove that this set is measurable (belongs to $\mathcal{B}_{X\times A}$). Since $K_\infty = \bigcap_n K_n$, it is sufficient to show that the sets

$$K_n = \{(x,a) : a \in A_n(x), x \in X\}, \qquad n \geq -1,$$

are measurable. This is done by induction in n. The set $K_{-1} = K$ is measurable by Assumption 2.1. Suppose that K_{n-1} is measurable. Since K_{n-1} has nonempty compact x-sections, by Proposition 2.5(ii) applied to the Carathéodory function $h_n(x,a)$ and the set K_{n-1}, the function $\hat{h}_n(x)$ is measurable on X and therefore is also measurable on K_{n-1} as a function of both x and a. It follows that the set

$$K_n = \{(x,a) \in K_{n-1} : h_n(x,a) = \hat{h}_n(x)\}$$

is measurable too, and this completes the induction.

Since K_∞ is measurable and has nonempty compact x-sections, by Proposition 2.5(i) there exists a selector $\varphi \in \Phi$ such that $\varphi(x) \in A_\infty(x)$, $x \in X$. For this selector we have $L^{\varphi(x)} H(x) = \hat{H}(x)$ for all x, so that $L^\varphi H = \hat{H}$. Therefore, because $\varphi \in \Phi \subset \Sigma$, and since we already know that $L^\sigma H \preceq \hat{H}$ for every $\sigma \in \Sigma$, we have $L^\varphi H = \mathrm{Lexmax}_{\sigma\in\Sigma} L^\sigma H$. Thus both (i) and (ii) are proved. Since $L^\varphi H \in \mathcal{H}$ together with H, we also have (iii). □

THEOREM 6.4.

(i) *There exists a selector $\varphi \in \Phi$ Blackwell optimal within Σ;*

(ii) *the Blackwell optimality equation $H = TH$ has a unique solution H^* in $\mathcal{H}$.*

Proof. By Lemma 6.1 there exists a policy $\tau \in \Sigma$ lexicographically maximizing $H^\sigma(1)$ over Σ, and by Theorem 6.3 there exists a selector φ such that

$$L^\varphi H^\tau = TH^\tau. \tag{6.7}$$

Therefore, utilizing the definition (4.5) of T and Corollary 3.4, we have $L^\varphi H^\tau \succeq L^\tau H^\tau = H^\tau$. Hence, by Corollary 4.3, $H^\varphi \succeq H^\tau$. Thus, by Proposition 4.1, $H^\varphi(1) \succeq H^\tau(1)$, and therefore, since (6.2) holds in particular for $\sigma = \varphi$, $H^\varphi(1) = H^\tau(1)$. The last relation together with $H^\varphi \succeq H^\tau$ means that

$$\int_X H^\varphi(x)m(dx) = \int_X H^\tau(x)m(dx)$$

while $H^\varphi(x) \succeq H^\tau(x)$ everywhere on X. Therefore $H^\varphi(x) = H^\tau(x)$ (a.e. m) on X. Since the nonconstant components of H enter into $L^a H(x)$ only integrated with respect to $m(dx)$ (Definition 6.2), it follows that $L^a H^\varphi(x) = L^a H^\tau(x)$ for all $(x, a) \in K$ (Proposition 4.1). Therefore $TH^\varphi = TH^\tau$ and $L^\varphi H^\tau = L^\varphi H^\varphi$, (see (6.3) and (6.6)). In accordance with (6.7) and Corollary 3.4, it follows that

$$TH^\varphi = TH^\tau = L^\varphi H^\tau = L^\varphi H^\varphi = H^\varphi.$$

Thus $H^\varphi \in \mathcal{H}$ is a solution of the Bellman optimality equation $TH = H$, and also, by Theorem 4.5, φ is Blackwell optimal within Σ.

To prove the uniqueness of the solution, suppose that $H^* \in \mathcal{H}$ is another solution of the equation $TH = H$. Then $L^\varphi H^* \preceq TH^* = H^*$, so that by Lemma 4.2 we have $H^\varphi \preceq H^*$. On the other hand, by Theorem 6.3 there exists a selector ψ such that $L^\psi H^* = TH^* = H^*$. By Lemma 4.2, $H^\psi = H^*$, and by Theorem 4.5(ii) $H^\varphi \succeq H^\psi = H^*$. Since both $H^\varphi \preceq H^*$ and $H^\varphi \succeq H^*$, we have $H^\varphi = H^*$. □

7. Blackwell optimality in the space of all policies. In this section we extend to our model the elegant proof of Theorem 5.4 in Dekker and Hordijk [8, pp. 414–416], leading to the conclusion that a policy Blackwell optimal within the class Σ is Blackwell optimal within the class Π too (in fact, Theorem 5.4 in [8] is stated in slightly different terms). For the case of a Borel space X and finite sets A_x, such an extension is done in [31, pp. 283–287]. We give here a more condensed presentation, with novelties in the definitions of measures corresponding to the components of a Markov policy and of the limiting measure. The following version of Proposition 5.2 is used in the proof.

PROPOSITION 7.1. *For every constant $C > 0$, the set S_C of all measures s on K satisfying the condition*

$$\Pr_X s \le Cm \tag{7.1}$$

is (sequentially) compact in the weak-strong topology.

Proof. For the case $C = 1$ this is proved in [33], and our Proposition 5.2 is obtained as a corollary. It is easy to proceed in the opposite direction too. Namely, one must add a fictitious action a^* to A and each of the sets A_x, and assign the measure $m - \Pr_X s$ to $X \times a^*$; then (7.1) will turn into an equality. To reduce the case of an arbitrary $C > 0$ to the case $C = 1$, one must multiply f and divide s by C in the relation (5.2) defining the topology. □

THEOREM 7.2. *If a stationary policy τ is Blackwell optimal within the class Σ of stationary policies, then τ is also Blackwell optimal within the class Π of all policies.*

Proof. According to a well-known result by Strauch [27], for every $x \in X$ and $\pi \in \Pi$ there exists a Markov policy μ such that for each pair $x_t a_{t+1}$ its $\mathbf{P}_x^\pi$-distribution coincides with its $\mathbf{P}_x^\mu$-distribution, and therefore $v_\beta(x,\pi) = v_\beta(x,\mu)$ for all $0 < \beta < 1$. Hence it is sufficient to prove that τ is Blackwell optimal within the class M, i.e., that for every $x \in X$ and $\mu \in \mathrm{M}$ there exists a number $\beta_0 = \beta_0(x,\tau,\mu) < 1$ such that

$$v_\beta(x,\tau) - v_\beta(x,\mu) \geq 0 \qquad \text{for all} \qquad \beta_0 < \beta < 1. \tag{7.2}$$

Here $\mu = \{\sigma_1, \sigma_2, \dots\}$, where $\sigma_t \in \Sigma$. To simplify the subsequent formulas, we denote the components of $H^\tau = H^*$ by $h_{-1} = g, h_0, h_1, \dots$ and write $q^{(t)}, r^{(t)}, Q^{(t)}, L^{(t)}$ instead of q^{σ_t}, r^{σ_t}, Q^{σ_t}, L^{σ_t} $(t \geq 1)$. Consider also the t-step transition densities q_t corresponding to the policy μ, defined, similar to (3.3), as convolutions $q_{t+1} = q_t * q^{(t+1)}$ with $q_1 = q^{(1)}$, and the corresponding t-step operators $Q_0 = I$, $Q_t = Q^{(1)}Q^{(2)}\dots Q^{(t)}$, $t \geq 1$. It is easy to see that the bounds $0 \leq q^\sigma \leq \|p\|$, $\sigma \in \Sigma$, imply the same bounds for q_t and that Q_t are stochastic operators, so that

$$\|q_t\| \leq \|p\|, \ \|Q_t\| \leq 1, \quad t \geq 1. \tag{7.3}$$

Under the policy μ, the consecutive states $x_0\ x_1\ x_2\ \dots$ form a nonstationary Markov chain with the transition operators $Q^{(t)}$ and one-step rewards $r^{(t)}$, so that (2.2) simplifies to

$$v_\beta(x,\mu) = \sum_{t=0}^{\infty} \beta^t Q_t r^{(t+1)}(x), \quad x \in X, \quad 0 < \beta < 1. \tag{7.4}$$

On the other hand, by a remarkable identity exploited in Sladký [25], Hordijk and Sladký [17], and Dekker and Hordijk [8], we can transform the Laurent expansion (3.11) of $v_\beta(x,\tau)$ into

$$v_\beta(x,\tau) = \sum_{n=-1}^{\infty} h_n(x)\alpha^n = \sum_{n=0}^{\infty} \alpha^n \sum_{t=0}^{\infty} \beta^t Q_t[h_n - Q^{(t+1)}(h_n - h_{n-1})](x), \tag{7.5}$$
$$x \in X, \qquad 0 < \alpha = 1 - \beta < \frac{1}{\kappa}.$$

The identity used in (7.5) is an algebraic one, it is a direct consequence of the relations $Q_t Q^{(t+1)} = Q_{t+1}$, $\alpha \sum \beta^t = 1$, $h_{-1} = \text{const.}$; the absolute convergence of the double series in the indicated interval follows from the bounds (3.20) and (7.3) for h_n and Q_t.

Next we fix $x = x_0 \in X$, subtract (7.4) from (7.5), collect the like terms with β^t in (7.4) and with $\alpha^n\beta^t$ in (7.5), and compare the coefficients of the resulting double series with the components of $L^\sigma H$ (see (3.23)). This gives, for the difference in (7.2), a representation of the form

$$v_\beta(x_0,\tau) - v_\beta(x_0,\mu) = \sum_{n=0}^{\infty}\sum_{t=0}^{\infty} a_{tn}\alpha^n\beta^t, \quad 0 < \alpha < \frac{1}{\kappa}, \tag{7.6}$$

where

$$a_{tn} = Q_t b_{tn}(x_0), \quad \{0, b_{t0}, b_{t1}, \dots\} = H - L^{(t+1)}H. \tag{7.7}$$

We need to express the coefficients a_{tn} through the components $f_n(x,a)$ of the difference

$$H(x) - L^a H(x) = \{0, f_0(x,a), f_1(x,a), \dots\}, \quad (x,a) \in K \tag{7.8}$$

(see Definition 6.2). In accordance with (6.3),

$$b_{tn}(x) = \int_{A_x} f_n(x,a)\sigma_{t+1}(x,da), \quad x \in X. \tag{7.9}$$

A substitution of (7.9) into (7.7) gives a representation

$$a_{tn} = \int_X q_t(x_0,x) \int_{A_x} f_n(x,a)\sigma_{t+1}(x,da)m(dx) = \int_K q_t(x_0,x) f_n(x,a) ds_t, \tag{7.10}$$
$$s_t = j(\sigma_{t+1}) \in S, \quad n \geq 0, \quad t \geq 1,$$

while for the coefficients with $t = 0$ we have

$$a_{0n} = b_{0n}(x_0), \quad n \geq 0, \tag{7.11}$$

since $Q_0 = I$. Finally, we change in (7.10) from ds_t to ds'_t where the measure s'_t on K is defined by

$$s'_t(dxda) = q_t(x_0,x)s_t(dxda), \quad t \geq 1. \tag{7.12}$$

Then (7.10) becomes

$$a_{tn} = \int_K f_n(x,a) ds'_t, \quad n \geq 0, \quad t \geq 1. \tag{7.13}$$

Next, since H $(= H^\tau = H^*)$ is the solution of the Blackwell optimality equation, both $H - L^{(t+1)}H$ and $H(x) - L^a H(x)$ are lexicographically nonnegative (see Theorems 4.5 and 6.3). It follows from (7.8), (7.13), and Proposition 4.1 (or directly from (7.7) and (7.11) in the case $t = 0$) that

$$\{a_{t0}, a_{t1}, a_{t2}, \dots\} \succeq 0, \quad t \geq 0. \tag{7.14}$$

From (7.7), (7.11)–(7.13), and the bounds (3.20) and (7.3) for h_n and q_t, we conclude that

$$| a_{tn} | \leq C\kappa^n, \quad n \geq 0, \quad t \geq 0, \tag{7.15}$$

and that

$$s'_t \in S_C, \quad t \geq 1, \tag{7.16}$$

for a sufficiently large $C > 0$ (see Proposition 7.1). In the proof of Theorem 6.3 we have seen that the components $h_n(x,a)$ of $L^a H(x)$ are continuous in a. By (7.8) $f_n(x,a) = h_n(x) - h_n(x,a)$, and therefore

$$f_n \in \mathrm{Car}(K), \quad n \geq 0. \tag{7.17}$$

To prove the theorem, it remains to show that the relations (7.6), (7.14), (7.15) together with (7.13), (7.16), (7.17) imply (7.2) for $x = x_0$.

Following Dekker and Hordijk [8, pp. 414–416], we consider various possibilities for (7.14) to be satisfied.

Case I. All $a_{tn} = 0$. Then (7.2) obviously holds at $x = x_0$ with the equality sign.

Case II. There is an integer $N \geq 0$ such that all a_{tn} with $n < N$ (if any) are zeros, but not all a_{tN} are equal to 0. Then evidently

$$a_{tN} \geq 0 \text{ for all } t \geq 0, \quad \text{while } a_{0N} + a_{1N} + a_{2N} + \cdots = b > 0, \tag{7.18}$$

and this case splits into two subcases: $b = \infty$ and $b < \infty$.

The subcase $b = \infty$ is also rather trivial. From (7.6) and (7.18) we have, for α converging to 0,

$$v_\beta(x_0, \tau) - v_\beta(x_0, \mu) = \alpha^N \sum_{t=0}^{\infty} a_{tN}\beta^t + R(\alpha), \tag{7.19}$$

where in accordance with (7.15), and since $\alpha \sum \beta^t = 1$,

$$| R(\alpha) | \leq C \sum_{n>N} \sum_{t=0}^{\infty} (\alpha\kappa)^n \beta^t = C\kappa^{N+1}\alpha^N(1 - \alpha\kappa)^{-1} = O(\alpha^N).$$

In view of (7.18), and since $\beta = 1-\alpha$, $b = \infty$, the sum of the series in (7.19) approaches ∞ as $\alpha \downarrow 0$, and this implies (7.2) for $x = x_0$.

The subcase $0 < b < \infty$ is the most interesting one. In this subcase $\sum_{t=0}^{\infty} a_{tN}\beta^t = b + o(1)$, so that similar to the preceding subcase, only taking into account the two lowest powers of α, we have

$$\begin{aligned} v_\beta(x_0, \tau) - v_\beta(x_0, \mu) &= \alpha^N \sum_{t=0}^{\infty} a_{tN}\beta^t + \alpha^{N+1} \sum_{t=0}^{\infty} a_{t,N+1}\beta^t + O(\alpha^{N+1}) \\ &= \alpha^N(b + R(\alpha)) + o(\alpha^N), \end{aligned}$$

where

$$R(\alpha) = \alpha \sum_{t=0}^{\infty} a_{t,N+1}\beta^t. \tag{7.20}$$

Since $b > 0$, it is sufficient to prove that $\liminf_{\alpha\downarrow 0} R(\alpha) \geq 0$ or, as follows from (7.20), that $\liminf_{t\to\infty} a_{t,N+1} \geq 0$.

This is proved by contradiction. If $\liminf_{t\to\infty} a_{t,N+1} < 0$, then for some $\varepsilon > 0$ and some sequence of the values of t we must have $\lim a_{t,N+1} = -\varepsilon$. By Proposition 7.1 and in view of (7.16), there exists a subsequence $\{t_j\}$ of the above sequence and a measure $s' \in S_C$ such that $s'_{t_j} \overset{ws}{\to} s'$. By the definition of the weak-strong topology and in view of (7.13) and (7.17), this implies the existence of the limits

$$a_n = \lim_{j\to\infty} a_{t_j n} = \int_K f_n(x, a) ds', \quad n \geq 0. \tag{7.21}$$

Since all a_{tn} with $n < N$ are zeros in Case II, we have $a_n = 0$ for $n < N$. Because $b < \infty$ in the subcase we consider, the series in (7.18) converges; therefore $\lim_{t\to\infty} a_{tN} = 0$, and we also have $a_N = 0$. By the selection of $\{t_j\}$, $a_{N+1} = -\varepsilon$. It follows that

$$\{0, a_0, a_1, \dots, a_N, a_{N+1}, a_{N+2}, \dots\} = \{0, 0, 0, \dots, 0, -\varepsilon, a_{N+2}, \dots\} \prec 0. \tag{7.22}$$

On the other hand, as we have seen,

$$\{0, f_0(x,a), f_1(x,a), \dots\} = H(x) - L^a H(x) \succeq 0, \quad (x,a) \in K. \tag{7.23}$$

By Proposition 4.1 we may integrate the nonstrict lexicographical inequalities. Therefore, (7.22) contradicts (7.21) and (7.23). □

8. Examples.

Example 8.1 (the water regulation problem). We consider the water regulation problem as described in [13, section 2.8]. The water is stored in a reservoir of a finite capacity M. The amount x_t of water in the reservoir after the tth period of time is

$$x_t = (x_{t-1} - a_t + \xi_t) \wedge M, \quad t = 1, 2, \dots, \tag{8.1}$$

where a_t is the extent of water taken for the consumption and ξ_t is the random influx of water during the tth period of time. Here $0 \le a_t \le x_{t-1}$, and ξ_t are nonnegative i.i.d. (independent identically distributed) random variables. The utility of the consumption is measured by a function $R(a_t)$.

To fit into the definitions of section 2, we suppose that the random variables ξ_t have an absolutely continuous distribution function

$$F(x) = \int_0^x f(z)dz, \qquad 0 \le z < \infty. \tag{8.2}$$

Then the components of the model can be defined in the following way. The state space $X = [0, M]$, the action space $A = [0, M]$, the action sets $A_x = [0, x]$, $x \in X$. The reference measure

$$m(B) = \begin{cases} \frac{\lambda(B)}{2M} & \text{if } B \subset [0, M),\ B \in \mathcal{B}_X, \\ \frac{1}{2} & \text{if } B = \{M\}, \end{cases} \tag{8.3}$$

where λ is the Lebesgue measure. The transition density

$$p(x,a,y) = \begin{cases} 0 & \text{if } 0 \le y < x - a, \\ 2Mf(y - x + a) & \text{if } x - a \le y < M, \\ 2\int_{M-x+a}^{\infty} f(z)dz & \text{if } y = M. \end{cases} \quad a \in A_x, \quad x \in X, \tag{8.4}$$

The reward function $r(x,a) = R(a)$, $a \in A_x, x \in X$. (It is easy to see that with these m and p,

$$\int_B p(x,a,y)m(dy) = \mathbf{P}\{x_t \in B \mid x_{t-1} = x, a_t = a\} = \mathbf{P}\{(x - a + \xi_t) \wedge M \in B\}$$

for every interval B, and hence for every Borel set $B \subset [0, M]$, as should be in accordance with (8.1).)

Assumptions 2.1–2.4 are satisfied if

(i) the density $f(x)$ is continuous (and therefore nonnegative and bounded) on the interval $[0, M]$;

(ii) $f(0) = 0$;

(iii) $1 - F(M)$ $(= \mathbf{P}\{\xi_t \ge M\}) > 0$;

(iv) $R(a)$ is a continuous function on $[0, M]$.

Indeed, Assumption 2.1 trivially holds, Assumption 2.2 follows from (i) and (ii), and Assumption 2.4 is implied by (iv). The simultaneous Doeblin-type condition (Assumption 2.3) follows from (iii) with $D = \{M\}$ and $\delta = 2(1 - F(M))$. Thus, conditions (i)–(iv) imply the existence of a deterministic stationary Blackwell optimal policy.

Example 8.2 (the inventory problem). The inventory model goes back to Bellman [2]. In the case of a single commodity, the evolution of the system can be described by the equation

$$x_t = (x_{t-1} + a_t - \xi_t) \vee 0, \qquad t = 1, 2, \dots,$$

where x_{t-1} is the stock level at the beginning of the tth period of time, a_t is the amount of the additional stock that is reordered, the nonnegative i.i.d. random variables ξ_t represent the demands, and the stock level is bounded above by a constant M (the capacity of the storehouse). The (expected) one-step cost is $c(x_{t-1}, a_t)$ (it consists of the costs of ordering, stockholding, and shortage).

This model is "symmetric" to the water regulation problem. As in that model, we suppose that ξ_t have the distribution function (8.2). As above, we have $X = A = [0, M]$, but now $A_x = [0, M - x]$. The reference measure is similar to (8.3), with the only difference being that the atom $\frac{1}{2}$ is at $x = 0$ instead of $x = M$. The transition density, similar to (8.4), is

$$p(x, a, y) = \begin{cases} 2\int_{x+a}^{\infty} f(z)dz & \text{if } y = 0, \\ 2Mf(x + a - y) & \text{if } 0 < y \le x + a, \quad a \in A_x, x \in X, \\ 0 & \text{if } x + a < y \le M, \end{cases}$$

and the reward function is $r(x, a) = -c(x, a)$.

Assumptions 2.1–2.4 are satisfied under the same conditions (i)–(iii) as above and the condition

(iv′) $c(x, a), \qquad 0 \le a \le M - x \le M$ is continuous in a, bounded and measurable.

Notice that in both examples we have a recurrent state x^* such that $P(x, a, x^*) > \delta > 0$ for all $(x, a) \in K$.

Example 8.3 (the stabilization problem on a circle). This is a circular version of the well-known linear regulator problem with Gaussian disturbances (cf. [13, sections 3.11, 6.12, 7.12]). The state of the system is described by an angle x which should be kept close to 0 but is subject to random fluctuations. The actions a are corrections of x. Larger deviations of x from 0 and larger values of $\mid a \mid$ induce larger costs. To be definite, let $X = A = [-1, 1)$ with the topology of a circle, so that $\lim_{x \uparrow 1} x = -1$. Then both X and A are compact spaces. The evolution of the system is described by the equation

$$x_t = x_{t-1} - a_t + \xi_t \bmod(2), \quad t = 1, 2, \dots,$$

where ξ_t are independent normal random variables with $\mathbf{E}\xi_t = 0$, $\mathbf{E}\xi_t^2 = \sigma^2 > 0$. The costs are of a form $R_1(a_t^2) + R_2(x_t^2)$, where R_i are nondecreasing continuous functions on $[0, 1]$ with $R_i(0) = 0$.

Let $A_x = A$ for every $x \in X$, and define the reference measure by $m(dx) = \frac{1}{2}dx$. Then the transition density is

$$p(x, a, y) = \sum_{n=-\infty}^{\infty} \frac{2}{\sqrt{2\pi}\sigma} e^{-\frac{(y-x+a-2n)^2}{2\sigma^2}}, \quad -1 \le x, a, y < 1.$$

This series converges uniformly in (x, a, y), and it follows that p is bounded, strictly positive, and continuous on the compact $X \times A \times X$. Hence $p(\cdot)$ is bounded from below by a positive constant δ, so that Assumption 2.3 holds with $D = X$. Assumptions 2.1 and 2.3 are trivially satisfied. The reward function

$$r(x, a) = -R_1(a^2) - \int_{-1}^{1} R_2(y^2) p(x, a, y) dy$$

is bounded and continuous (in both variables) together with p, R_1, and R_2. It follows that there exists a deterministic stationary Blackwell optimal policy.

9. Comments and open problems. The measurability and compactness (Assumption 2.1) and the continuity in a of the transition law and of the reward function (parts of Assumptions 2.2 and 2.4) are standard in dynamic programming, when proving the existence of exactly optimal policies.

The absolute continuity of the transition probabilities and the boundedness of the transition densities (Assumption 2.2) are restrictions crucial for our analysis. Only a few works in the theory of MDPs are made in the framework of densities, although models of this kind are encountered in applications. As a separate object of study, controlled Markov chains with transition densities are treated by Georgin [14]. In this paper the discounted and the average rewards are studied. In general, our assumptions are more restrictive than in [14]. Hypothesis 1a in [14] is similar to our Assumption 2.2, except that the density is not uniformly bounded there, but (in our notations) $p(x, a, y) \le q(x, y)$ where $\int_X q(x, y) m(dy) < \infty$ for every $x \in X$.

Another crucial point for our analysis is the boundedness of the reward function r (Assumption 2.4). This is a severe restriction not satisfied in many applications. However, if the state space X is a compact set, this assumption is reasonable, and such models are encountered in practice.

The simultaneous Doeblin-type condition (Assumption 2.3) is an easily checked assumption needed only for the uniform geometric convergence of the t-step transition densities $q_t^\sigma(x, y)$ to their limit $\bar{q}^\sigma(y)$ (see (3.7)). This condition is equivalent to the existence of a nonnegative function $f \in B(X)$ such that $p(x, a, y) \ge f(y)$ for all x, a, y, and $\int_X f(y) m(dy) > 0$ (given (2), one may define $f(x) = \delta \mathbf{1}_D(x)$; given f, one may set $D = \{x : f(x) \ge \delta\}$ where $\delta > 0$ is so small that $m(D) > 0$). In terms of the transition function, the above condition means that $P(x, a, B) \ge \nu(B)$ where ν is a nontrivial measure on X (in our case $\nu(dx) = f(x) m(dx)$). In [13, section 7.1] such a measure ν is called a minorant. Models with a minorant are well known in MDPs, especially in the particular case when ν is concentrated at a single point $x^* \in X$, so that a simpler uniform recurrence condition holds: $P(x, a, \{x^*\}) \ge \delta \ (= \nu(x^*)) > 0$ (cf. Examples 8.1 and 8.2). For a comprehensive survey of recurrence conditions in MDPs with a Borel state space and their relation to the convergence of t-step transition probabilities, see Hernández-Lerma, Montes-de-Oca, and Cavazos-Cadena [15]. Our Assumption 2.3 and its version in terms of a minorant are labeled there as R1(a) and R1(b).

We need the geometric convergence of the transition densities for all randomized stationary policies. A simple way to weaken Assumption 2.3 is to extend the more sophisticated convergence conditions known for Markov chains to all chains generated by $\sigma \in \Sigma$. For example, the geometric convergence (3.7) still holds, and our results remain valid, if Assumption 2.3 is replaced by the following condition considered in Doob [12, section 5.5, case (b)] for a single Markov chain.

Assumption 9.1. There exist an integer $k \geq 1$ and numbers $\delta > 0$ and $\varepsilon > 0$ such that for every policy $\sigma \in \Sigma$ there is a set $D^\sigma \in \mathcal{B}_X$ with $m(D^\sigma) \geq \varepsilon$ and with

$$q_k^\sigma(x, y) \geq \delta \text{ for all } x \in X,\ y \in D^\sigma.$$

However, this assumption is difficult to verify. An open problem is whether Assumption 9.1 follows from a similar condition for the deterministic policies $\varphi \in \Phi$.

The existence of a *strong Blackwell optimal policy* is an interesting open question. A related problem is the asymptotics of the value function v_β as β approaches 1. Indeed, we have the following result.

LEMMA 9.2. *The following three statements are equivalent:*

(i) *there exists a strong Blackwell optimal selector* $\varphi \in \Phi$*;*

(ii) *every Blackwell optimal stationary policy* $\sigma \in \Sigma$ *is strong Blackwell optimal;*

(iii) *there exists a number* $0 < \beta_0 < 1$ *such that*

$$v_\beta(x) = \sum_{n=-1}^{\infty} h_n^*(x)\alpha^n, \quad x \in X, \quad \beta_0 < \beta = 1 - \alpha < 1 \tag{9.1}$$

where H^* *is the unique solution of the Blackwell optimality equation* $H = TH$ *in* $\mathcal{H}$.

Proof. (iii)→(ii). By Theorems 4.5 and 6.4, for a Blackwell optimal $\sigma \in \Sigma$ we have $H^\sigma = H^*$, and therefore, if (9.1) holds, $v_\beta(x, \sigma)$ has the same Laurent expansion as $v_\beta(x)$, only in the interval $\rho < \beta < 1$. Hence $v_\beta(x) = v_\beta(x, \sigma)$ if $\rho \vee \beta_0 < \beta < 1$, so that σ is strong Blackwell optimal.

(ii)→(i). The proof follows from Theorems 6.4(i) and 7.2.

(i)→(iii). Given (i), we have $v_\beta(x) = v_\beta(x, \varphi)$, $x \in X$, $\beta_0 < \beta < 1$ for some $\beta_0 < 1$, and (9.1) follows from the Laurent expansion for $v_\beta(x, \varphi)$ (with $\beta_0 \vee \rho$ in place of β_0). □

The main term of the supposed expansion (9.1) is well known. Utilizing an approach due to Cavazos-Cadena and Lasserre [4], in [32] we have justified the next term in (9.1) and proved that

$$v_\beta(x) = \frac{g^*}{\alpha} + h_0^*(x) + o(1)$$

uniformly in x as $\alpha \downarrow 0$ in the case of Assumptions 2.1–2.4, but we have no idea how to obtain the next terms.

Acknowledgments. The author is thankful to the referees and the associate editor for helpful suggestions, remarks, and additional references.

REFERENCES

[1] E. I. BALDER, *On the compactness of the space of policies in dynamic programming*, Stochastic Process. Appl., 32(1989), pp. 141–150.

[2] R. BELLMAN, *Dynamic Programming,* Princeton University Press, Princeton, NJ, 1957.

[3] D. BLACKWELL, *Discrete dynamic programming*, Ann. Math. Statist., 33(1962), pp. 719–726.

[4] R. CAVAZOS-CADENA AND J. B. LASSERRE, *Strong* 1*-optimal stationary policies in denumerable Markov decision processes*, System Control Lett., 11(1988), pp. 65–71.

[5] R. CAVAZOS-CADENA AND J. B. LASSERRE, *A direct approach to Blackwell optimality*, preprint, 1993. Boletín de la Sociedad Matemática Mexicana, to appear.

[6] R. J. CHITASHVILI, *A controlled finite Markov chain with an arbitrary set of decisions*, Theory Probab. Appl., 20(1975), pp. 839–847.

[7] R. J. CHITASHVILI, *A finite controlled Markov chain with small termination probability*, Theory Probab. Appl., 21(1976), pp. 158–163.

[8] R. Dekker and A. Hordijk, *Average, sensitive and Blackwell optimal policies in denumerable Markov decision chains with unbounded rewards*, Math. Oper. Res., 13(1988), pp. 395–420.
[9] R. Dekker and A. Hordijk, *Recurrence conditions for average and Blackwell optimality in denumerable state Markov decision chains*, Math. Oper. Res., 17(1992), pp. 271–289.
[10] E. V. Denardo, *Markov renewal programs with small interest rates*, Ann. Math. Statist., 42(1971), pp. 477–496.
[11] E. V. Denardo and B. L. Miller, *An optimality condition for discrete dynamic programming with no discounting*, Ann. Mat. Statist., 39(1968), pp. 1220–1227.
[12] J. L. Doob, *Stochastic Processes*, Wiley, New York, 1953.
[13] E. B. Dynkin and A. A. Yushkevich, *Controlled Markov Processes*, Springer-Verlag, New York, 1979.
[14] G.-P. Georgin, *Contrôle de chaînes de Markov sur les espaces arbitraires*, Ann. Inst. Henri Poincaré, Sect. B, 14(1978), pp. 271–289.
[15] O. Hernández-Lerma, R. Montes-de-Oca, and R. Cavazos-Cadena, *Recurrence conditions for Markov decision processes with Borel state space: A survey*, Ann. Oper. Res., 28(1991), pp. 29–46.
[16] C. J. Himmelberg, T. Parthasarathy, and F. S. Van Vleck, *Optimal plans for dynamic programming problems*, Math. Oper. Res., 1(1976), pp. 390–394.
[17] A. Hordijk and K. Sladký, *Sensitive optimality criteria in countable state dynamic programming*, Math. Oper. Res., 2(1977), pp. 11–14.
[18] J. B. Lasserre, *Conditions for existence of average and Blackwell optimal stationary policies in denumerable Markov decision processes*, J. Math. Anal. Appl., 136(1988), pp. 479–489.
[19] S. A. Lippman, *Criterion equivalence in discrete dynamic programming*, Oper. Res., 17(1969), pp. 920–923.
[20] B. L. Miller and A. F. Veinott, Jr., *Discrete dynamic programming with a small interest rate*, Ann. Math. Statist., 40(1969), pp. 366–370.
[21] M. Puterman, *Sensitive discount optimality in controlled one-dimensional diffusions*, Ann. Probab., 2(1974), pp. 408–419.
[22] U. G. Rothblum, *Normalized Markov decision chains* I; *sensitive discount optimality*, Oper. Res., 23(1975), pp. 785–795.
[23] M. Schäl, *On dynamic programming: Compactness of the space of policies*, Stochastic Process. Appl., 3(1975), pp. 345–354.
[24] M. Schäl, *On dynamic programming and statistical decision theory*, Ann. Statist., 7(1979), pp. 432–445.
[25] K. Sladký, *On the set of optimal controls for Markov chains with rewards*, Kybernetika (Prague), 10(1974), pp. 350–367.
[26] K. Sladký, *Sensitive optimality criteria for continuous time Markov decision processes*, Trans. 8th Prague Conference Information Theory, Statistical Decision Functions, Random Processes, Vol. B, Praha, Czechoslovakia, 1978, pp. 211–225.
[27] R. E. Strauch, *Negative dynamic programming*, Ann. Math. Statist., 37(1966), pp. 871–890.
[28] A. F. Veinott, Jr., *On finding optimal policies in discrete dynamic programming with no discounting*, Ann. Math. Statist., 37(1966), pp. 1284–1294.
[29] A. F. Veinott, Jr., *Discrete dynamic programming with sensitive optimality criteria*, Ann. Math. Statist., 40(1969), pp. 1635–1660.
[30] J. Warga, *Optimal Control of Differential and Functional Equations*, Academic Press, New York, 1972.
[31] A. A. Yushkevich, *Blackwell optimal policies in a Markov decision process with a Borel state space*, Z. Oper. Res., 40(1994), pp. 253–288.
[32] A. A. Yushkevich, *A note on asymptotics of discounted value function and strong* 0*-discount optimality*, Math. Methods Oper. Res., 44(1996), pp. 223–231.
[33] A. A. Yushkevich, *The compactness of a policy space in dynamic programming via an extension theorem for Carathéodory functions*, Math. Oper. Res., 22(1997), pp. 458–467.
[34] A. A. Yushkevich and R. J. Chitashvili, *Controlled random sequences and Markov chains*, Russian Math. Surveys, 37(1982), pp. 239–274.

SIAM J. CONTROL OPTIM.
Vol. 35, No. 6, pp. 2183–2197, November 1997

017

A CONTROL METHOD FOR ASSIMILATION OF SURFACE DATA IN A LINEARIZED NAVIER–STOKES-TYPE PROBLEM RELATED TO OCEANOGRAPHY*

AZIZ BELMILOUDI† AND FRANÇOISE BROSSIER†

Abstract. A method of control is developed in order to compute the variability of the velocity and pressure (u, p) in an oceanic doman Ω during a time T. The observation is the variability of p of $\Gamma_0 \times]0, T[$, where Γ_0 is the upper surface of Ω and corresponds to the undisturbed sea surface. The observation is deduced from satellite data. The control is the variability of the wind-stress f, which acts as the forcing of the perturbation. The mean circulation in Ω is supposed to be known. The equations verified by (u, p) are linearized around this mean circulation. The continuity of the application: $f \to (u(f), p(f))$ in convenient functional spaces is proved. We deduce from this result the existence and uniqueness of an optimal control, which is characterized by a set of equations including the direct problem of linearized Navier–Stokes type verified by (u, p) and the adjoint problem. The cost function and consequently the optimal control depend on a real parameter α. We prove the convergence of the sequence (f_α), when α tends to 0, toward a wind-stress f which minimizes the distance between the observed and computed pressures. This result is obtained by means of minimizing sequences.

Key words. optimal control, Navier–Stokes, assimilation of surface data, minimizing sequences, oceanography

AMS subject classifications. 35Q30, 49J20, 49K20, 65J10, 76D05, 76U05, 86A22

PII. S0363012995286137

1. Introduction.

The method developed in this paper allows us to obtain the deep oceanic circulation from satellite measurements which are obviously surface observations. Altimetric measurements give the distance between the satellite and the sea surface. It is now possible to extract from these data the sea level variability with a precision in the order of centimeters. The pressure variability at a given depth can be deduced from the sea level and will act as the observation in our model (cf. Appendix).

The oceanic phenomenon that we want to modelize occurs in the tropical Atlantic and Pacific oceans: the circulation there is characterized by steady zonal currents driven by the trade winds and by long waves propagating westward along the equator, driven by the variability of the wind-stress and superimposed to the mean currents. The mean velocity u_0 is known for each tropical season. The total velocity $\tilde{u}$ can therefore be considered as the sum of two terms: u_0, which is given, and a variability u corresponding to the waves. The same expansion is valid for the pressure: $\tilde{p} = p_0 + p$. The equations of motion in an oceanic domain Ω are of Navier–Stokes type. We assume that, at initial time $t = 0$, the flow Ω is the mean circulation (u_0, p_0), which verifies steady equations. In this paper we are dealing only with the linearized case. The equations verified by the perturbation (u, p) are the following:

$$\frac{\partial u}{\partial t} + (u_0.\nabla)u + (u.\nabla)u_0 + F\Lambda u - \nu_h(\Delta_2 u) - \frac{\partial}{\partial z}\left(\nu_v \frac{\partial u}{\partial z}\right) + \frac{1}{\rho}\nabla p = 0, \tag{1.1a}$$

*Received by the editors May 17, 1995; accepted for publication (in revised form) October 1, 1996.

http://www.siam.org/journals/sicon/35-6/28613.html

†LANS–Institut National des Sciences Appliquées 20, Avenue des Buttes de Coësmes, 35043 Rennes Cédex, France (belmilou@perceval.univ-rennes1.fr, fbrossie@perceval.univ-rennes1.fr).

$$\text{div } u = 0, \tag{1.1b}$$

$$u(0) = 0, \tag{1.1c}$$

$\Delta_2 \frac{\partial^2}{\partial x^2} + \frac{\partial}{\partial y^2}$. (x, y, z) are the Cartesian coordinates: x and y are measured in the horizontal plane of the undisturbed sea surface (x toward the east, y toward the north), and z is vertically ascendant. The density ρ is a constant mean value. Afterward we will take $\rho = 1$ g.cm^{-3}. $F = (0, 0, 2\omega \sin\varphi)$ is the Coriolis force, ω is the rotation rate of earth, and φ is the latitude. The dissipative term corresponds to Reynolds stresses with a coefficient of eddy viscosity ν. Observation of turbulent flows leads to distinguish between horizontal eddy viscosity ν_h and vertical eddy viscosity ν_v. From now on ν_h will be constant and ν_v variable.

We consider a time interval $(0, T)$ and an oceanic domain Ω extending on both sides of the equator ($10°S$–$10°N$) and of constant depth H (for example, $H = 3000$ m). The curvature of the earth is neglected. The vertical extension of Ω ($-H \leq z \leq 0$) corresponds to a part of the physical domain. We assume that for depths greater than H the variability is negligible. The perturbation (u, p) is not computed in the thin surface layer $0 \leq z \leq \xi(x, y)$, where $\xi(x, y)$ is the free sea surface observed by the satellite. The perturbation of the mean flow is made of zonal propagating waves. We can thus choose the zonal extension of Ω as the greatest wavelength and impose periodic boundary conditions on the eastern and western boundaries.

The oceanic domain can be defined as $\Omega =]0, Lx[\times]-Ly, Ly[\times]-H, 0[$. Γ denotes its boundary; $\Gamma = \cup_{i=0,5}\Gamma_i$, where Γ_0 denotes the surface ($z = 0$), Γ_5 denotes the bottom ($z = -H$), Γ_1 and Γ_2 denote the western and eastern boundaries, and Γ_3 and Γ_4 denote the southern and northern boundaries. n is the unit outward vector normal to Γ. On account of the phenomena that we want to describe, we set mixed boundary conditions.

- The flow is periodic in the x-direction:

$$u_{/\Gamma_1} = u_{/\Gamma_2}. \tag{1.2a}$$

- We have the following:

$$u_2 = 0, \quad \frac{\partial u_1}{\partial y} = 0, \quad \text{and} \quad \frac{\partial u_3}{\partial y} = 0 \quad \text{on } \Gamma_3 \text{ and } \Gamma_4. \tag{1.2b}$$

- The perturbation vanishes at $z = -H$:

$$u = 0 \ \text{ on } \ \Gamma_5. \tag{1.2c}$$

- The perturbation is driven by the perturbation of the wind-stress f:

$$u_3 = 0, \quad \frac{\partial u_1}{\partial z} = -\frac{f_1}{\nu_v}, \quad \text{and} \quad \frac{\partial u_2}{\partial z} = -\frac{f_2}{\nu_v} \quad \text{on } \Gamma_0 \tag{1.2d}$$

with : $u = (u_1, u_2, u_3)$ and $f = (f_1, f_2)$.

The variability of the pressure on Γ_0, deduced from altimetric data, constitutes the observation. The variability of the wind-stress f acts as the forcing of the perturbation and is unknown. We take f as the control. $(u(f), p(f))$ are the velocity and pressure corresponding to any control f and satisfying problem (1.1), (1.2). The optimal control is defined as the wind-stress minimizing a given cost function which measures the distance between the observed pressure and the pressure $p(f)$ solution of (1.1),

(1.2). The pressure observed on Γ_0 is the trace of a circulation (u_e, p_e) driven by a wind f_e. The solution $(u(f), p(f))$ of problem (1.1), (1.2) induced by the optimal control f has to approach the real circulation (u_e, p_e). The control method then makes it possible to compute the velocity and pressure in all the domain Ω by means of data on a part of the boundary.

In order to solve the problem of optimal control, we have to consider the adjoint equations associated to (1.1). We then apply the control theory introduced by J. L. Lions [4]. More recently, G. I. Marchuk developed the method of adjoint equations in mathematical physics and performed variational data assimilation in environmental problems [6, 7]. The classical control problem arising in meteorology or oceanography is the adjustment of the initial condition in order to obtain velocity and temperature fields which agree with observations in situ. Our purpose is quite different since we control by the variability of the wind-stress in order to restitute the observed surface pressure deduced from altimetric data.

The paper is organized as follows.

Section 2 is devoted to problem (1.1), (1.2) satisfied by the perturbation of the mean flow. Here are some regularity results proven in [2]: if the forcing f is sufficiently regular, then the weak solution of (1.1), (1.2) is such that $u \in L^2(0, T, H^2(\Omega))$, $p \in L^2(0, T, H^1(\Omega))$. We can then define the trace $\gamma_0 p$ of the pressure on Γ_0. We prove the continuity of the applications $f \to u(f)$, $f \to p(f)$, $f \to \gamma_0 p(f)$ in convenient functional spaces.

Section 3 is devoted to the problem of control. We prove the existence and uniqueness of the solution. The optimal control is given as a function of u^*, the solution of the adjoint problem associated with the direct problem (1.1), (1.2). We thus obtain a set of equations characterizing the optimal control.

The cost function and consequently the optimal control depend on a parameter α. In section 4 we prove a convergence result for the sequence (f_α) of optimal controls by means of minimizing sequences.

2. Regularity and continuity results. We introduce the functional spaces

$$H = \{v \in (L^2(\Omega))^3 / \mathrm{div} v = 0, v.n = 0 \text{ on } \Gamma_0 \cup \Gamma_3 \cup \Gamma_4 \cup \Gamma_5, (v.n)/_{\Gamma_1} = -(v.n)/_{\Gamma_2}\},$$
$$W = \{v \in (H^1(\Omega))^3 / v = 0 \text{ on } \Gamma_5, v.n = 0 \text{ on } \Gamma_0 \cup \Gamma_3 \cup \Gamma_4, v/_{\Gamma_1} = v/_{\Gamma_2}\},$$
$$V = \{v \in W / \mathrm{div}\ v = 0\}.$$

The norm $\|\,.\,\|_{1,\Omega}$ and the seminorm $[\,.\,]_{1,\Omega}$ defined on $H^1(\Omega)$ are equivalent in V and W. We then set $\|v\| = \|v\|_V = \|v\|_W = [v]_{1,\Omega}$. $|\,.\,|$ denotes the norm in $L^2(\Omega)$.

We define

$$a(u, v) = \nu_h(\nabla_2 u, \nabla_2 v) + \left(\nu_v \frac{\partial u}{\partial z}, \frac{\partial v}{\partial z}\right), \quad \text{with } \nabla_2 u = \begin{pmatrix} \frac{\partial u}{\partial x} \\ \frac{\partial u}{\partial y} \end{pmatrix},$$

$d(u, v) = (F\Lambda\ u, v)$,

$b(u, v, w) = ((u.\nabla)v, w)$,

$\ell(u, v) = b(u, u_0, v) + b(u_0, u, v) + d(u, v)$, where u_0 is the mean velocity given in V,

$c(v, q) = -(\mathrm{div}\ v, q)$.

a and d are bilinear continuous forms on W, b is a trilinear continuous form on W, a is coercive on W, and we denote by ν the constant of coercivity ($\nu = \mathrm{Min}\ (\nu_h, \mathrm{Min}\ \nu_v)$).

c is a bilinear continuous form on $W \times L^2(\Omega)$ and verifies the Inf Sup condition

$$\exists\ \beta > 0 \quad \text{such that} \quad \inf_{q \in L^2(\Omega)} \sup_{v \in W} \frac{c(v,q)}{|q|\ \|v\|} \geq \beta.$$

The initial problem (1.1), (1.2) verified by the perturbation (u,p) of the mean flow admits two equivalent weak formulations.

Find $u \in L^2(0,T,V)$ such that

$$\left(\frac{\partial u}{\partial t}, v\right) + a(u,v) + \ell(u,v) = -(f, \gamma_0 v) \quad \forall v \in V, \tag{2.1a}$$

$$u(0) = 0 \tag{2.1b}$$

and

find $(u,p) \in L^2(0,T,W) \times L^2(0,T,L^2(\Omega))$ such that

$$\left(\frac{\partial u}{\partial t}, v\right) + a(u,v) + \ell(u,v) + c(v,p) = -(f, \gamma_0 v) \quad \forall v \in W, \tag{2.2a}$$

$$c(u,q) = 0 \quad \forall q \in L^2(\Omega), \tag{2.2b}$$

$$u(0) = 0. \tag{2.2c}$$

$\gamma_0 v$ denotes the trace of v on Γ_0, and $(f, \gamma_0 v)$ denotes the scalar product in $L^2(\Gamma_0)$.

These results and the following concerning the existence and regularity of the weak solution (u,p) are proven in [2].

PROPOSITION 2.1. *For given u_0 and f, $u_0 \in V$, and $f \in L^2(0,T,L^2(\Gamma_0))$, there exists a unique solution u of problem* (2.1) *satisfying*

$$u \in L^2(0,T,V) \cap C^0([0,T],H), \qquad \frac{\partial u}{\partial t} \in L^2(0,T,V').$$

There is a pressure $p \in L^2(0,T,L^2(\Omega))$, defined regardless of any time-dependent function, such that (u,p) verifies the problem (1.1), (1.2) *in a distribution sense.*

We want to choose the pressure p on Γ_0 as the observation of the method of control. The regularity $p \in L^2(\Omega)$ doesn't make it possible to define $\gamma_0 p$, the trace of p on Γ_0. We have to make stronger assumptions in order to have $p \in H^1(\Omega)$.

From now on we assume the following regularity of the wind-stress:

$$f \in L^2(0,T,H^1(\Gamma_0)), \qquad \frac{\partial f}{\partial t} \in L^2(0,T,L^2(\Gamma_0)), \tag{2.3}$$

Equation (2.3) implies that $f \in C^0([0,T],L^2(\Gamma_0))$ a.e. on $[0,T]$ [5].

f must be consistent with the initial and boundary conditions imposed on the velocity u. We then have to impose the following compatibility conditions:

$$f(0) = 0 \quad \text{on } \Gamma_0, \tag{2.4a}$$

$$f_2 = 0 \quad \text{and} \quad \frac{\partial f_1}{\partial y} = 0 \quad \text{on } \gamma_3 \cup \gamma_4, \tag{2.4b}$$

$$f/_{\gamma_1} = f/_{\gamma_2}. \tag{2.4c}$$

Notation: f is defined on the open set $\Gamma_0 =]0, Lx[\times]-Ly, Ly[$. γ denotes the boundary of Γ_0. $\gamma = \cup_{i=1,4}\gamma_i$, where γ_1 and γ_2 are the western and eastern boundaries and γ_3

and γ_4 the southern and northern boundaries. n is the unit outward vector normal to γ.

PROPOSITION 2.2. *If the wind-stress f verifies the regularity* (2.3) *and the initial and boundary conditions* (2.4)*, then the solution (u,p) of problem* (2.1) *is such that*

$$u \in L^2(0,T,H^2(\Omega)), \quad p \in L^2(0,T,H^1(\Omega)).$$

This result is proven in [2] by an extension method, using even–odd reflection on the boundaries. The presence of corners in the open set Ω makes it impossible to directly apply a standard result of regularity given in [1].

Remarks.

(i) Since $p \in H^1(\Omega)$, we can now define the trace of p on the boundary Γ. p is defined regardless of any time-dependent function. We now determine this function by setting the condition

$$\int_{\Gamma_0} p \, d\Gamma = \int_{\Gamma_0} p_d \, d\Gamma,$$

where p_d is the observation.

(ii) The velocity u is such that

$$u \in C^0([0,T],V), \qquad \frac{\partial u}{\partial t} \in L^2(0,T,V) \cap C^0([0,T],H).$$

PROPOSITION 2.3. *If the wind-stress f verifies the regularity* (2.3) *and the conditions* (2.4)*, then the solution (u,p) of problem* (2.1) *satisfies the following estimates:*

(i) $\|u\|_{L^2(0,T,H^2(\Omega))} \le C\|f\|_{\mathcal{U}}$,

(ii) $\|p\|_{L^2(0,T,H^1(\Omega))} \le C\|f\|_{\mathcal{U}}$,

where

$$\mathcal{U} = \left\{ f / f \in L^2(0,T,H^1(\Gamma_0)), \frac{\partial f}{\partial t} \in L^2(0,T,L^2(\Gamma_0)) \right\},$$

$$\|f\|_{\mathcal{U}} = \left(\|f\|^2_{L^2(0,T,H^1(\Gamma_0))} + \left\| \frac{\partial f}{\partial t} \right\|^2_{L^2(0,T,L^2(\Gamma_0))} \right)^{1/2}.$$

We denote by C any positive constant.

Proof. The solution (u,p) of problem (2.1) verifies (1.1a) in a distribution sense. Equation (1.1a) can also be written

$$-\nu_h \Delta_2 u - \frac{\partial}{\partial z}\left(\nu_v \frac{\partial u}{\partial z}\right) + \nabla p = -\left(\frac{\partial u}{\partial t} + F\Lambda u + (u_0.\nabla)u + (u.\nabla)u_0\right).$$

Set $G = -(\frac{\partial u}{\partial t} + F\Lambda u + (u_0.\nabla)u + (u.\nabla)u_0)$. If $u_0 \in V$, then $G \in C^0([0,T],L^2(\Omega))$ [2].

At each time $t \in [0,T]$, the initial problem (1.1), (1.2) can be written as follows: Given $G \in L^2(\Omega)$, find $(u,p) : \Omega \to \mathbb{R}^3 \times \mathbb{R}$ such that

$$-\nu_h \Delta_2 u - \frac{\partial}{\partial z}\left(\nu_v \frac{\partial u}{\partial z}\right) + \nabla p = G \quad \text{in } \Omega, \tag{2.5a}$$

$$\text{div } u = 0 \quad \text{in } \Omega, \tag{2.5b}$$

and satisfying the boundary conditions (1.2).

After extension of problem (2.5) in an open set $\tilde{\Omega} =]0, Lx[\times]-2Ly, 2Ly[\times]-H, 0[$ in order to cancel the corners of Ω, we obtain the following estimate [2]:

$$\|u\|_{H^2(\Omega)} + |\nabla p|_{L^2(\Omega)} \leq C(|G|_{L^2(\Omega)} + \|f\|_{H^1(\Gamma_0)}) \quad \text{a.e. on } [0,T]. \tag{2.6}$$

C is positive, time-independent constant.

Equation (2.6) implies

$$\|u\|^2_{L^2(0,T,H^2(\Omega))} \leq C\left(\|G\|^2_{L^2(0,T,L^2(\Omega))} + \|f\|^2_{L^2(0,T,H^1(\Gamma_0))}\right), \tag{2.7}$$

$$\|\nabla p\|^2_{L^2(0,T,L^2(\Omega))} \leq C\left(\|G\|^2_{L^2(0,T,L^2(\Omega))} + \|f\|^2_{L^2(0,T,H^1(\Gamma_0))}\right). \tag{2.8}$$

If $u_0 \in V$, we have the inequality

$$\|G\|^2_{L^2(0,T,L^2(\Omega))} \leq C\left(\left\|\frac{\partial u}{\partial t}\right\|^2_{L^2(0,T,L^2(\Omega))} + \|u\|^2_{L^2(0,T,L^2(\Omega))} + \|u\|^2_{L^2(0,T,V)}\right).$$

Setting $v = u$, (2.1a) gives

$$\frac{d}{dt}|u|^2 + 2a(u,u) + 2\ell(u,u) = -2(f, \gamma_0 u).$$

The bilinear form a is coercive on V: $a(u,u) \geq \nu\|u\|^2_V$.

$$\ell(u,u) = b(u, u_0, u);$$

therefore, if $u_0 \in V$: $2|\ell(u,u)| \leq C_1|u|^2$.

γ_0 is the application trace on Γ_0, which is continuous from V into $L^2(\Gamma_0)$. Set C_0, the constant of continuity.

$$2|(f, \gamma_0 u)| \leq 2C_0|f|_{L^2(\Gamma_0)}\|u\| \leq C_2\|f\|^2_{H^1(\Gamma_0)} + \nu\|u\|^2.$$

We then obtain from (2.1a)

$$\frac{d}{dt}|u|^2 + \nu\|u\|^2 \leq C_2\|f\|^2_{H^1(\Gamma_0)} + C_1|u|^2. \tag{2.9}$$

Applying the Gronwall lemma now gives

$$|u(t)| \leq C\|f\|_{L^2(0,T,H^1(\Gamma_0))} \quad \forall t \in [0,T],$$

and therefore

$$\|u\|_{L^2(0,T,L^2(\Omega))} \leq C\|f\|_{L^2(0,T,H^1(\Gamma_0))}. \tag{2.10}$$

We deduce from (2.9), (2.10) that

$$\|u\|_{L^2(0,T,V)} \leq C\|f\|_{L^2(0,T,H^1(\Gamma_0))}. \tag{2.11}$$

Setting $v = \frac{\partial u}{\partial t} = u'$ in (2.1a) gives

$$|u'|^2 + a(u,u') + \ell(u,u') = -(f, \gamma_0 u')$$

and after integration with respect to time

$$-2\int_0^t |u'(s)|^2 ds + a(u(t),u(t)) = -2\int_0^t \ell(u,u')ds + 2\int_0^t (f',\gamma_0 u)\,ds$$
$$-2(f(t),\gamma_0 u(t)) \quad \forall\ t\in[0,T].$$

$\ell(u,u') = b(u,u_0,u') + b(u_0,u,u')$, which yields $|\ell(u,u')| \le C\|u\|\ |u'|$, since u_0 is given in V.

Applying the coercivity yields

$$2\int_0^t |u'|^2 + \nu\|u\|^2 \le C\int_0^t \|u\|\ |u'| + 2C_0\int_0^t |f'|\ \|u\| + 2C_0|f|\ \|u\|.$$

Since $f(t) = \int_0^t f'(s)\,ds$, we have $|f(t)|^2 \le C\|f'\|^2_{L^2(0,T,L^2(\Gamma_0))} \forall\ t\in[0,T]$, and we obtain the inequality

$$\|u'\|^2_{L^2(0,T,L^2(\Omega))} \le C\left(\|u\|^2_{L^2(0,T,V)} + \|f'\|^2_{L^2(0,T,L^2(\Gamma_0))}\right),$$

which implies, according to (2.11),

$$\|u'\|^2_{L^2(0,T,L^2(\Omega))} \le C\left(\|f\|^2_{L^2(0,T,H^1(\Gamma_0))} + \|f'\|^2_{L^2(0,T,L^2(\Gamma_0))}\right),$$

i.e.,

$$\|u'\|_{L^2(0,T,L^2(\Omega))} \le C\|f\|_{\mathcal{U}}. \tag{2.12}$$

We deduce from (2.10)–(2.12) that $\|G\|_{L^2(0,T,L^2(\Omega))} \le C\|f\|_{\mathcal{U}}$, and (2.7) now implies estimate (i).

According to (2.8) we also obtain the majoration

$$\|\nabla p\|_{L^2(0,T,L^2(\Omega))} \le C\|f\|_{\mathcal{U}}. \tag{2.13}$$

To prove estimate (ii) we now use the mixed formulation (2.2). According to the Inf Sup condition verified by c, we have

$$|p| \le \frac{1}{\beta}\operatorname*{Sup}_{v\in W}\frac{c(v,p)}{\|v\|}. \tag{2.14}$$

Equation (2.2a) can also be written

$$c(v,p) = -((u',v) + a(u,v) + \ell(u,v) + (f,\gamma_0 v)). \tag{2.15}$$

Equations (2.14) and (2.15) imply

$$|p| \le C(|u'| + \|u\| + |f|),$$

which gives, according to estimates (2.11), (2.12),

$$\|p\|_{L^2(0,T,L^2(\Omega))} \le C\|f\|_{\mathcal{U}}. \tag{2.16}$$

Equations (2.13) and (2.16) prove estimate (ii).

COROLLARY.

(i) *The application $f \longrightarrow (u(f),p(f))$ is linear and continuous from $\mathcal{U}$ into $L^2(0,T,H^2(\Omega)) \times L^2(0,T,H^1(\Omega))$.*

(ii) *The application $f \longrightarrow \gamma_0 p$ is linear and continuous from $\mathcal{U}$ into $L^2(0,T,L^2(\Gamma_0))$.*

Proof. The application trace is continuous from $H^1(\Omega)$ in $L^2(\Gamma_0)$. We thus have

$$\|\gamma_0 p\|_{L^2(0,T,L^2(\Gamma_0))} \le C\|p\|_{L^2(0,T,H^1(\Omega))} \le C\|f\|_{\mathcal{U}}. \tag{2.17}$$

3. Characterization of the optimal control. The problem is controlled by the variability of the wind-stress f. The observation is the pressure on Γ_0, deduced from altimetric measurements. Controls and observations are thus defined on Γ_0, which is a part of the boundary Γ. The control f has to verify the regularity condition (2.3) in order to obtain $p \in H^1(\Omega)$.

Thereby $\mathcal{U} = \{f/f \in L^2(0,T,H^1(\Gamma_0)), \frac{\partial f}{\partial t} \in L^2(0,T,L^2(\Gamma_0))\}$ will be the control space; $\mathcal{H} = L^2(0,T,L^2(\Gamma_0))$ will be the observation space. For each control f, $(u(f),p(f))$ is the solution of the weak problem (2.1) or (2.2), and the cost function J is defined by

$$J(f) = \frac{1}{2}(\|\mathcal{C}p(f) - p_d\|^2_{\mathcal{H}} + \alpha\|f\|^2_{\mathcal{U}}). \tag{3.1}$$

$p_d \in \mathcal{H}$ is the observation. p_d is the trace of the real pressure p_e, which is supposed to be in $L^2(0,T,H^1(\Omega))$. $\mathcal{C}$ is the application trace on Γ_0. $\mathcal{C} : p \in L^2(0,T,H^1(\Omega)) \longrightarrow \mathcal{C}\ p = \gamma_0 p \in \mathcal{H}$. α is a given positive constant ($\alpha \neq 0$). The optimal control problem then is as follows: find $f \in \mathcal{U}$ such that

$$J(f) = \operatorname*{Inf}_{h\in\mathcal{U}} J(h). \tag{3.2}$$

PROPOSITION 3.1. *Problem* (3.2) *admits one unique solution* $f \in \mathcal{U}$.

Proof.

(i) J is continuous on $\mathcal{U}$. Set

$$r(f,h) = (\mathcal{C}p(f), \mathcal{C}p(h))_{\mathcal{H}} + \alpha(f,h)_{\mathcal{U}}$$

and

$$s(f) = (\mathcal{C}p(f), p_d)_{\mathcal{H}} \quad \forall f \text{ and } h \in \mathcal{U}.$$

We have

$$J(f) = \frac{1}{2}r(f,f) - s(f) + \frac{1}{2}\|p_d\|^2_{\mathcal{H}}.$$

According to (2.17), we can write

$$|r(f,h)| \leq C\|f\|_{\mathcal{U}}\|h\|_{\mathcal{U}} \quad \text{and} \quad |s(f)| \leq C\|f\|_{\mathcal{U}},$$

where C denotes any positive constant. The applications r and s are continuous, and therefore J is continuous on $\mathcal{U}$.

(ii) J is coercive on $\mathcal{U}$ since $J(f) \geq \frac{\alpha}{2}\|f\|^2_{\mathcal{U}} \quad \forall\ f \in \mathcal{U}$.

(iii) J is strictly convex.

J is differentiable, and

$$(J'(f), h) = r(f,h) - s(h) \quad \forall\ f \text{ and } h \in \mathcal{U},$$

which gives

$$(J'(f) - J'(h), f - h) = r(f-h, f-h) > \alpha\|f-h\|^2_{\mathcal{U}}.$$

(i), (ii), and (iii) imply that problem (3.2) admits one unique solution in $\mathcal{U}$ ([4]).

$f \in \mathcal{U}$ is solution of problem (3.2) if and only if $J'(f) = 0$. In order to characterize the optimal control f we introduce the adjoint problem associated with problem (2.2).

The bilinear form a is selfadjoint. Let ℓ^* be the adjoint of ℓ. For a wind-stress $f \in \mathcal{U}, (u(f), p(f))$ is the solution of problem (2.2). We note $(u^*(f), p^*(f))$, the adjoint state which is the solution of the adjoint problem (3.3):

$$-\left(\frac{\partial u^*}{\partial t}(f), v\right) + a(u^*(f), v) + \ell^*(u^*(f), v) + c(v, p^*(f)) = 0 \quad \forall\, v \in W, \tag{3.3a}$$

$$-c(u^*(f), q) = (\mathcal{C}p(f) - p_d, \mathcal{C}q) \quad \forall q \in H^1(\Omega), \tag{3.3b}$$

$$u^*(f)(T) = 0. \tag{3.3c}$$

PROPOSITION 3.2. *The adjoint problem* (3.3) *admits one unique solution such that* $u^*(f) \in L^2(0, T, W), p^*(f) \in L^2(0, T, L^2(\Omega))$.

Proof. The compatibility condition $\int_{\Gamma_0} (\mathcal{C}p(f) - p_d)\, d\Gamma = 0$ is satisfied. So, almost everywhere on $]0, T[$, there exists $u_i \in W$ such that

$$-c(u_1, q) = (\text{div } u_1, q) = (\mathcal{C}p(f) - p_d, \mathcal{C}q) \quad \forall\, q \in H^1(\Omega).$$

Set $\tilde{u} = u^*(f) - u_1$. Equations (3.3) can be written

$$-\left(\frac{\partial \tilde{u}}{\partial t}, v\right) + a(\tilde{u}, v) + \ell^*(\tilde{u}, v) + c(v, p^*(f)) = (G, v) \quad \forall\, v \in W, \tag{3.4a}$$

$$-c(\tilde{u}, q) = 0 \quad \forall\, q \in H^1(\Omega), \tag{3.4b}$$

$$\tilde{u}(T) = 0. \tag{3.4c}$$

G depends on u_1 and $G \in L^2(\Omega)$.

Problem (3.4), being similar to problem (2.2), admits one unique solution: $\tilde{u} \in L^2(0, T, W)$. $u^* = \tilde{u} + u_1 \in L^2(0, T, W)$ and verifies equation (3.3). The existence of a solution being proven, demonstrating the uniqueness is fairly simple. There exists a pressure $p^* \in L^2(0, T, L^2(\Omega))$ such that equations (3.4a) and (3.3a) are verified [8].

PROPOSITION 3.3. $J'(f) = 0$ *if and only if* $\alpha f + \Lambda^{-1}\gamma_0 u^*(f) = 0$ *in* $\mathcal{U}$.

Λ is the canonical isomorphism $\mathcal{U} \longrightarrow \mathcal{U}'$ such that

$$\langle h, v\rangle_{\mathcal{U},\mathcal{U}'} = (\Lambda h, v)_{\mathcal{U}'} = (h, \Lambda^{-1} v)_{\mathcal{U}} \quad \forall\, h \in \mathcal{U}, \quad \forall\, v \in \mathcal{U}'.$$

Proof.

$$(J'(f), h) = r(f, h) - s(h) = (\mathcal{C}p(f) - p_d, \mathcal{C}p(h))_{\mathcal{H}} + \alpha(f, h)_{\mathcal{U}} \quad \forall\, f \text{ and } h \in \mathcal{U}.$$

Equation (3.3b) implies

$$(\mathcal{C}p(f) - p_d, \mathcal{C}p(h))_{L^2(\Gamma_0)} = -c(u^*(f), p(h))$$

and therefore

$$(J'(f), h) = \alpha(f, h)_{\mathcal{U}} - \int_0^T c(u^*(f), p(h))\, dt \quad \forall\, f \text{ and } h \in \mathcal{U}.$$

$(u(h), p(h))$ is the solution of problem (2.2),

$$\left(\frac{\partial u(h)}{\partial t}, v\right) + a(u(h), v) + \ell(u(h), v) + c(v, p(h)) = -(h, \gamma_0 v) \quad \forall\, \mathrm{v} \in W,$$

and so, by setting $v = u^*(f)$,

$$-c(u^*(f), p(h)) = \left(\frac{\partial}{\partial t}u(h), u^*(f)\right) + a(u(h), u^*(f)) + \ell(u(h), u^*(f)) + (h, \gamma_0 u^*(f)),$$

we obtain

$$(J'(f), h) = \alpha(f, h)_{\mathcal{U}} + \int_0^T \left(-\left(\frac{\partial}{\partial t}u^*(f), u(h)\right) + a(u^*(f), u(h)) + \ell^*(u^*(f), u(h)\right) dt + \int_0^T (h, \gamma_0 u^*(f)) dt.$$

$u^*(f)$ is the solution of the adjoint problem (3.3). We thus have

$$(J'(f), h) = \alpha(f, h)_{\mathcal{U}} - \int_0^T c(u(h), p^*(f)) dt + \int_0^T (h, \gamma_0 u^*(f))\, dt.$$

According to equation (2.2b),

$$c(u(h), p^*(f)) = 0$$

and thus

$$\begin{aligned}(J'(f), h) &= \alpha(f, h)_{\mathcal{U}} + \int_0^T (h, \gamma_0 u^*(f)) \\ &= \alpha(f, h)_{\mathcal{U}} + (h, \gamma_0 u^*(f))_{L^2(0,T,L^2(\Gamma_0))} \\ &= \alpha(f, h)_{\mathcal{U}} + \langle h, \gamma_0 u^*(f)\rangle_{\mathcal{U},\mathcal{U}'} \\ &= \alpha(f, h)_{\mathcal{U}} + (\Lambda^{-1}\gamma_0 u^*(f), h)_{\mathcal{U}} \\ &= (\alpha f + \Lambda^{-1}\gamma_0 u^*(f), h)_{\mathcal{U}}.\end{aligned}$$

We can now conclude that $J'(f) = 0$ if and only if $\alpha f + \Lambda^{-1}\gamma_0 u^*(f) = 0$ in $\mathcal{U}$.

Characterization of the optimal control. We have proved that the optimal control f, the solution of problem (3.2), is characterized by the following set of equations:

$$\begin{aligned}\left(\frac{\partial}{\partial t}u(f), v\right) + a(u(f), v) + \ell(u(f), v) + c(v, p(f)) &= -(f, \gamma_0 v) \quad \forall\, v \in W, \\ -c(u(f), q) &= 0 \quad \forall\, q \in H^1(\Omega), \\ -\left(\frac{\partial}{\partial t}u^*(f), v\right) + a(u^*(f), v) + \ell^*(u^*(f), v) + c(v, p^*(f)) &= 0 \quad \forall\, v \in W, \\ -c(u^*(f), q) &= (\mathcal{C}p(f) - p_d, \mathcal{C}q) \\ &\quad \forall\, q \in H^1(\Omega), \\ u(f)(0) &= 0, \\ u^*(f)(T) &= 0, \\ \alpha f + \Lambda^{-1}\gamma_0 u^*(f) &= 0 \text{ in } \mathcal{U}.\end{aligned}$$

4. A convergence result. J_α will now denote the cost function

$$J_\alpha(h) = \frac{1}{2}(\|\mathcal{C}p(h) - p_d\|^2_{\mathcal{H}} + \alpha\|h\|^2_{\mathcal{U}}).$$

Problem (3.2) becomes the following: find $f_\alpha \in \mathcal{U}$ such that

$$J_\alpha(f_\alpha) = \operatorname*{Inf}_{h\in\mathcal{U}} J_\alpha(h). \tag{4.1}$$

We set

$$J_0(h) = \frac{1}{2}\|\mathcal{C}p(h) - p_d\|^2_{\mathcal{H}}, \qquad \Pi(h) = \mathcal{C}p(h).$$

We have already proved that Π is a linear continuous application from $\mathcal{U}$ into $\mathcal{H}$. We consider the functional spaces

$$\mathcal{U} = \left\{h/h \in L^2(0,T,H^1(\Gamma_0)), \frac{\partial h}{\partial t} \in L^2(0,T,L^2(\Gamma_0))\right\}, \qquad \mathcal{U}^* = \mathcal{U}_{\operatorname{Ker}\Pi}.$$

LEMMA.
(i) Ker Π *is a closed subset of* $\mathcal{U}$.
(ii) $\mathcal{U}^*$ *is a Hilbert space.*
Proof.
(i) Let (f_m) be a sequence in Ker Π such that (f_m) converges towards f in $\mathcal{U}$. Since Π is continuous, $\Pi(f_m)$ converges toward $\Pi(f)$. $\Pi(f_m) = 0 \quad \forall\ m$; therefore, $\Pi(f) = 0$ and $f \in$ Ker Π.
(ii) Ker Π is a closed subset of the Hilbert space $\mathcal{U}$. Hence Ker Π and $(\text{Ker }\Pi)^\perp$ are Hilbert spaces.

Let $f^* \in \mathcal{U}^*$: f^* is a coset of $\mathcal{U}$ modulo Ker Π; let f be any element of f^*. $f = f_1 + f_2, f_1 \in$ Ker $\Pi, f_2 \in (\text{Ker }\Pi)^\perp$.

We set $\|f^*\|_{\mathcal{U}^*} = \|f_2\|_{\mathcal{U}} = \text{Inf}_{f\in f^*}\|f\|_{\mathcal{U}}$. We define the linear application ψ:

$$\begin{aligned}\Psi : \mathcal{U}^* &\longrightarrow (\text{Ker }\Pi)^\perp,\\ f^* &\longrightarrow f_2.\end{aligned}$$

Ψ is a bijection:
- if $f_2 \in (\text{Ker }\Pi)^\perp, f = f_2 + f_1 \in \mathcal{U}^*\ \forall\ f_1 \in$ Ker Π.
- $\Psi(f^*) = 0$ implies $f \in$ Ker Π, i.e., $f^* = 0$.

According to the definition of the norm in $\mathcal{U}^*$, ψ and ψ^{-1} are continuous. ψ is therefore an homeomorphism; since $(\text{Ker }\Pi)^\perp$ is a Hilbert space, $\mathcal{U}^*$ is also an Hilbert space.

We now introduce the following problem: find $f^* \in \mathcal{U}^*$ such that

$$J_0(f^*) = \operatorname*{Inf}_{h^*\in\mathcal{U}^*} J_0(h^*). \tag{4.2}$$

If there exists $f \in \mathcal{U}$ such that

$$J_0(f) = \inf_{h\in\mathcal{U}} J_0(h), \tag{4.3}$$

then problem (4.2) admits one unique solution $f^* \in \mathcal{U}^*$. The solution f of (4.3) is an element of the coset f^*. This will be verified if we assume that $f \in \mathcal{U}_{ad}$,

$\mathcal{U}_{ad} = \{h \in \mathcal{U} / \|h\|_{\mathcal{U}} \leq c\}$ [4]. f is the variability of the wind-stress and the assumption that f bounded is sensible.

PROPOSITION 4.1. *Let f_α be the solution of problem* (4.1) *and f^* be the solution of problem* (4.2). *When α tends to* 0, *then*

(i) (f_α) *converges toward f^* strongly in $\mathcal{U}^*$.*

(ii) $J_\alpha(f_\alpha)$ *converges toward $J_0(f^*)$.*

Proof.

(ii) Since f^* verifies (4.2), any element $f \in f^*$ verifies (4.3):

$$J_0(f) \leq J_0(h) \quad \forall\ h \in \mathcal{U}.$$

Definitions of J_0 and J_α imply

$$J_0(h) \leq J_\alpha(h) \quad \forall h \in \mathcal{U}.$$

We thus obtain

$$J_0(f) \leq J_\alpha(f_\alpha) \quad \forall\ \alpha \in \mathbb{R}_+^*. \tag{4.4}$$

Since $J_0(f) = \mathrm{Inf}_{h\in\mathcal{U}}\, J_0(h)\ \forall\ \varepsilon > 0, \exists\ h_\varepsilon \in \mathcal{U}$ such that $J_0(h_\varepsilon) \leq \frac{\varepsilon}{2} + J_0(f)$, and therefore $J_\alpha(h_\varepsilon) \leq \frac{\varepsilon}{2} + J_0(f) + \frac{\alpha}{2}\|h_\varepsilon\|^2_{\mathcal{U}}$. Set $\alpha^* = \frac{\varepsilon}{\|h_\varepsilon\|^2_{\mathcal{U}}}$. We thus have $J_{\alpha^*}(h_\varepsilon) \leq \varepsilon + J_0(f)$.

According to (4.4), this gives the following: $\forall\ \varepsilon > 0, \exists\ \alpha^*$ such that $\forall\ \alpha, 0 < \alpha < \alpha^*$, we have

$$J_0(f) \leq J_\alpha(f_\alpha) \leq \varepsilon + J_0(f). \tag{4.5}$$

We deduce from (4.5) that $\lim_{\alpha\to 0} J_\alpha(f_\alpha) = J_0(f) = J_0(f^*)$.

(i) f_α and f verify (4.1) and (4.3), respectively. We thus have

$$(J_\alpha^{'}(f_\alpha), h - f_\alpha) \geq 0 \quad \forall\ h \in \mathcal{U}$$

and

$$(J_0^{'}(f), h - f) \geq 0 \quad \forall\ h \in \mathcal{U};$$

i.e.,

$$(\Pi(f_\alpha) - p_d, \Pi(h - f_\alpha))_{\mathcal{H}} + \alpha(f_\alpha, h - f_\alpha)_{\mathcal{U}} \geq 0 \quad \forall\ h \in \mathcal{U}, \tag{4.6}$$

$$(\Pi(f) - p_d, \Pi(h - f))_{\mathcal{H}} \geq 0 \quad \forall\ h \in \mathcal{U}. \tag{4.7}$$

Setting $h = f$ in (4.6) and $h = f_\alpha$ in (4.7) gives

$$\alpha(f_\alpha, f - f_\alpha)_{\mathcal{U}} \geq \|\Pi(f - f_\alpha)\|^2_{\mathcal{H}} \geq 0,$$

which yields

$$\|f_\alpha\|^2_{\mathcal{U}} \leq (f, f_\alpha)_{\mathcal{U}} \leq \|f\|_{\mathcal{U}}\|f_\alpha\|_{\mathcal{U}},$$

i.e.,

$$\|f_\alpha\|_{\mathcal{U}} \leq \|f\|_{\mathcal{U}}. \tag{4.8}$$

The sequence (f_α) is bounded in $\mathcal{U}$. We can extract from (f_α) a subsequence also denoted (f_α) which converges toward $\tilde{f}$, weakly in $\mathcal{U}$. Since J_α is lower semicontinuous, we have [3], [4]

$$\liminf J_\alpha(f_\alpha) \geq J_\alpha(\tilde{f}) \geq J_0(\tilde{f}).$$

The inequality (4.5) implies

$$\limsup J_\alpha(f_\alpha) \leq J_0(f)$$

and so

$$J_0(\tilde{f}) \leq \liminf J_\alpha(f_\alpha) \leq \limsup J_\alpha(f_\alpha) \leq J_0(f).$$

f verifies (4.3). Therefore, $J_0(\tilde{f}) = J_0(f) = J_0(f^*)$.

The sequence (f_α) converges towards an element $\tilde{f}$ of f^*, weakly in $\mathcal{U}$. According to (4.8) we have $\|f_\alpha\|_{\mathcal{U}} \leq \|\tilde{f}\|_{\mathcal{U}}$, which implies $\limsup \|f_\alpha\|_{\mathcal{U}} \leq \|\tilde{f}\|_{\mathcal{U}}$.

Therefore, (f_α) converges towards $\tilde{f}$ strongly in $\mathcal{U}$ [3]; i.e., (f_α) converges toward f^* strongly in $\mathcal{U}^*$.

5. Conclusion. The sea level variability is obtained from altimetric measurements. We can deduce from these data the variability of the pressure on Γ_0. Γ_0 corresponds to the horizontal undisturbed sea surface and is the upper boundary of the oceanic domain considered in this paper. We have developed a control method in order to calculate the perturbation (u, p) of the mean velocity and pressure corresponding to the observed variability of p on Γ_0. The control is the variability of the wind-stress f which acts as the forcing of the perturbation. The cost function depends on a real parameter α. For any $\alpha > 0$ we have proved the existence and uniqueness of an optimal control f_α. To characterize f_α we must consider the direct problem verified by (u, p) and its adjoint. When α tends to 0, the sequence (f_α) converges toward a wind-stress f. f is an approximation of the unknown real wind-stress f_e, which induced the observed situation. The perturbation (u_α, p_α) driven by f_α is the unique solution of a linearized Navier–Stokes-type problem. Velocity and pressure are continuous functions of the wind-stress; therefore, when α tends to 0, (u_α, p_α) converges toward the solution (u, p) of the linearized Navier–Stokes problem forced by f. Since f approaches f_e, the perturbation (u, p) is an approximation of the real circulation (u_e, p_e) observed by the satellite. We can thus compute the variability of the velocity and pressure in an oceanic domain Ω, during a time T, from satellite observations of the sea surface. The assumptions of the model correspond to the physical features of equatorial waves. It is then easy to deduce from the variability (u, p) the characteristics of the excited waves.

Appendix. Choice of the observation for the control method. The assumption of hydrostatic pressure is standard in oceanography and justified by the difference between horizontal and vertical scalings in an oceanic domain. The equation of hydrostatic pressure is

$$\frac{\partial \tilde{p}}{\partial z} + \rho g = 0, \tag{A.1}$$

where $\tilde{p}$ is the total pressure, ρ is the density, and g is the constant of gravity.

The sea level is a free surface of equation $z = \tilde{\xi}(x, y, t)$. We suppose that the density ρ is constant in the upper layer $0 \leq z \leq \tilde{\xi}$. Integrating (A.1) with respect to

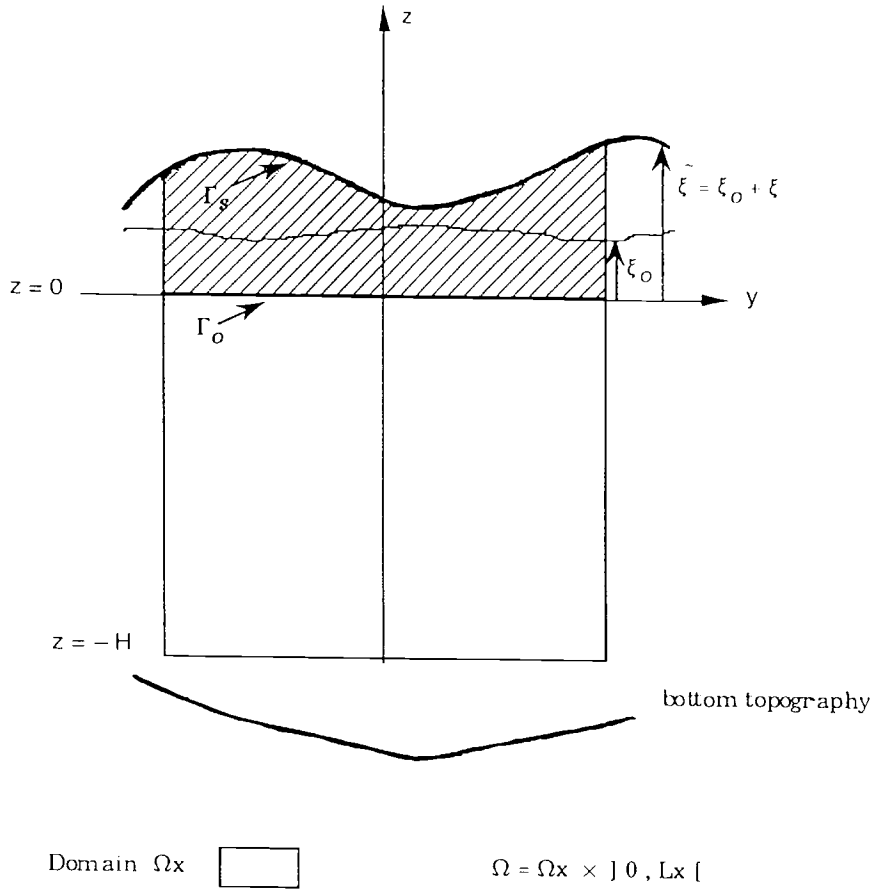


FIG. 1.

z between $z = 0$ and $z = \tilde{\xi}$ gives

$$(\tilde{p})_{z=\tilde{\xi}} = (\tilde{p})_{z=0} - \rho g \tilde{\xi}. \tag{A.2}$$

The surface pressure is equal to the atmospheric pressure: $(\tilde{p})_{z=\tilde{\xi}} = \tilde{p}_{\text{atm}}$.

We assume that the velocity $\tilde{u}$ can be considered as the sum of two terms: a mean given velocity u_0 and a perturbation u. The same expansion is valid for the pressure and the sea level:

$$\tilde{p} = p_0 + p,$$
$$\tilde{\xi} = \xi_0 + \xi.$$

p_0 and ξ_0 are the pressure and the free surface corresponding to the mean flow. p and ξ are the variability of pressure and sea level corresponding to the variability u of the current.

Equation (A.2) is verified by the mean flow and by the complete evolutive situation. This implies

$$p_{\text{atm}} = (p)_{z=0} - \rho g \xi. \tag{A.3}$$

Neglecting the variability of the atmospheric pressure gives

$$(p)_{z=0} = \rho g \xi. \tag{A.4}$$

The variability of the sea level ξ is known from altimetric measurements. Relation (A.4) makes it possible to take the variability of the pressure at a fixed level ($z = 0$) as the observation. Relation (A.4) also induces the choice of the domain Ω. The vertical extension of Ω is $-H \leq z \leq 0$, H being a constant. We denote by Γ_0 the part of the boundary of equation $z = 0$. The observation is therefore given on Γ_0.

It has to be noted that the studied domain Ω differs from the physical domain, which extends from the bottom Γ_B to the free sea surface Γ_S (Fig. 1).

REFERENCES

[1] S. Agmon, A. Douglis, and L. Nirenberg, *Estimates near the boundary for solutions of elliptic, partial differential equations satisfying general boundary conditions* II. Comm. Pure Appl. Math. 17 (1964), pp. 35–92.

[2] A. Belmiloudi and F. Brossier, *Regularity results for a Navier–Stokes type problem related to oceanography*, Acta Appl. Math., 1996, to appear.

[3] H. Brezis, *Analyse fonctionnelle, Theorie et Applications*, Masson, Paris, 1983.

[4] J. L. Lions, *Contrôle optimal de systèmes gouvernés par des équations aux dérivées partielles*, Dunod, Paris, 1968.

[5] J. L. Lions and E. Magenes, *Problèmes aux limites non homogènes et applications*, Tomes 1 et 2, Dunod, Paris, 1968.

[6] G. I. Marchuk, *Mathematical Models in Environmental Problems*, North Holland, Amsterdam, 1986.

[7] G. I. Marchuk, *Adjoint Equation and Analysis of Complex Systems*, Kluwer, Dordrecht, the Netherlands, 1995.

[8] R. Temam, *Navier-Stokes Equations*, North Holland, Amsterdam, 1977.